APPLIED REGRESSION ANALYSIS

SECOND EDITION

N.R.Draper
University of Wisconsin

H.Smith
Mount Sinai School of Medicine

JOHN WILEY & SONS
New York · Chichester · Brisbane · Toronto · Singapo

Н. Дрейпер
Г. Смит

ПРИКЛАДНОЙ РЕГРЕССИОННЫЙ АНАЛИЗ

Издание второе,
переработанное и дополненное

КНИГА 1

Перевод с английского
Ю.П. Адлера и В.Г. Горского

МОСКВА "ФИНАНСЫ И СТАТИСТИКА" 1986

ББК 22.172
Д73

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ ЗА РУБЕЖОМ

ВЫШЛИ ИЗ ПЕЧАТИ

1. Ли Ц., Джадж Д., Зельнер А. Оценивание параметров марковских моделей по агрегированным временным рядам.
2. Райфа Г., Шлейфер Р. Прикладная теория статистических решений.
3. Клейнен Дж. Статистические методы в имитационном моделировании. Вып. 1 и 2.
4. Бард Й. Нелинейное оценивание параметров.
5. Болч Б. У., Хуань К. Д. *Многомерные статистические методы* для экономики.
6. Иберла К. Факторный анализ.
7. Зельнер А. Байесовские методы в эконометрии.
8. Хейс Д. Причинный анализ в статистических исследованиях.
9. Пуарье Д. Эконометрия структурных изменений.
10. Драймз Ф. Распределенные лаги.
11. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. Вып. 1 и 2.
12. Бикел П., Доксам К. Математическая статистика. Вып. 1 и 2.
13. Лимер Э. Статистический анализ неэкспериментальных данных.
14. Песаран М., Слейтер Л. Динамическая регрессия: теория и алгоритмы.
15. Дидэ Э. и др. Методы анализа данных.
16. Бартоломью Д. Стохастические модели социальных процессов.

ГОТОВИТСЯ К ПЕЧАТИ

Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Кн. 2.

Д $\frac{0702000000-118}{010(01)-86}$ 108—86

● ПРЕДИСЛОВИЕ К РУССКОМУ ИЗДАНИЮ

Предлагаемая вниманию читателя монография известных американских статистиков Н. Дрейпера и Г. Смита посвящена регрессионному анализу. Регрессионный анализ по праву может быть назван основным методом современной математической статистики. Идея регрессионного анализа зиждется на мысли о том, что все доступные нам ресурсы важно использовать полно и эффективно, особенно если речь идет о накоплении и переработке информации. А значит, мы говорим не о каком-то частном методе обработки данных, а о предмете, более важном, чем любой конкретный метод.

Регрессионный анализ стал настолько привычным, что мы уже давно не замечаем, как он проявляется в механизмах усреднения, процедурах сглаживания, принципах согласования противоречивых позиций, концепциях оптимальности. Регрессия — это квинтэссенция понятия целесообразности.

С момента выхода перевода первого издания монографии прошло 13 лет. За это время появился целый ряд книг по регрессионному анализу. Среди них следует выделить такие, как: Успенский А. Б., Федоров В. В. Вычислительные аспекты метода наименьших квадратов при анализе и планировании регрессионных экспериментов.— М.: Изд-во МГУ, 1975; Алберт А. Регрессия, псевдоинверсия и рекуррентное оценивание/Пер. с англ. Под ред. Я. З. Цыпкина.— М.: Наука, 1977; Бард Й. Нелинейное оценивание параметров/Пер. с англ. Под ред. В. Г. Горского. — М.: Статистика, 1979; Себер Дж. Линейный регрессионный анализ/Пер. с англ. Под ред. М. Б. Малютова.— М.: Мир, 1980; Демиденко Е. З. Линейная и нелинейная регрессии.— М.: Финансы и статистика, 1981; Петрович М. Л. Регрессионный анализ и его математическое обеспечение на ЕС ЭВМ.— М.: Финансы и статистика, 1982. Однако эти книги не смогли удовлетворить растущие потребности широкого круга специалистов. Они рассчитаны в основном на читателей-математиков. А между тем в силу особой важности регрессионного анализа нужны специализированные руководства для научных работников разного профиля, экономистов, инженеров, врачей, агрономов, т. е. для всех тех, кто связан с математическим описанием разнообразных явлений, процессов и объектов. Этот пробел в известной степени и призвана заполнить монография Н. Дрейпера и Г. Смита.

Второе издание монографии существенно отличается от первого, вышедшего в США в 1966 г. В нее включены новые методы и приемы регрессионного анализа, появившиеся в последние два десятилетия. В результате объем второго издания по сравнению с объемом первого увеличился почти вдвое. В связи с этим перевод монографии Н. Дрейпера и Г. Смита представлен в двух книгах. В первую вошли гл. 1—5, во вторую — гл. 6—10 и приложение.

Работая над переизданием книги, Н. Дрейпер и Г. Смит стремились познакомить читателя с богатым арсеналом методов регрессионного анализа, оконтурить области их успешного применения, разъяснить статистическую (а подчас и геометрическую) природу, суть основных приемов регрессионного анализа, научить правильно пользоваться основными алгоритмами построения регрессий, приобщить читателя к использованию современной вычислительной техники для автоматизации сложных расчетов, связанных с регрессиями.

Авторы в полной мере справились с этими задачами. Книга в целом представляет собой практическое руководство по регрессионному анализу широкого профиля. Она пригодна и для самостоятельного изучения этого важного раздела математической статистики.

Переходя к оценке содержания книги, целесообразно дать краткий очерк становления и развития регрессионного анализа, оказывающего огромное влияние на прогресс во всех областях науки.

Родоначальником регрессионного анализа принято считать К. Гаусса. К. Гаусс (и независимо от него А. Лежандр) на рубеже XVIII и XIX столетий заложили основы метода наименьших квадратов. Поводом для создания метода наименьших квадратов, составляющего математическую основу регрессионного анализа, послужили потребности астрономии, а затем геодезии. Усилиями поколений ученых многих стран была развита и теория, ставшая теперь классической.

Примерно 150 лет, до середины XX в., длился классический период регрессионного анализа. За это время метод «обкатался». К алгебраической процедуре минимизации квадратичной формы, представляющей собой собственно метод наименьших квадратов, прибавилась система статистических постулатов, задающих математическую модель. Были отработаны механизм проверки гипотез об адекватности модели объекта, которая задается известным уравнением, часто полиномом не слишком высокой степени, и процедура проверки гипотез о значимости коэффициентов этого уравнения. Сочетание метода наименьших квадратов с указанными статистическими процедурами и привело к созданию того, что стало называться *регрессионным анализом*. Постепенно расширились и области приложений. Так, например, Д. И. Менделеев начал применять регрессию для описания температурных и иных зависимостей свойств химических веществ. Однако до конца первой мировой войны за пределами астрономии и геодезии метод все же не нашел широкого применения. Появлялись лишь спорадические работы. Любопытно отметить, что после классических работ К. Пирсона в самом начале XX в. теория была хорошо

известна и подробно изложена. Тем не менее существовал отчетливо выраженный временно́й лаг: практические приложения резко отставали.

В 20-е годы сложилось новое направление в экономике — эконометрия. Она взяла на вооружение регрессионные методы и весьма способствовала их распространению. Другой толчок произошел в связи с развитием способов измерения психических свойств личности, имевших большое значение не только для психологии, но и для тесно связанных с нею педагогики, социологии и отчасти медицины. Лишь вторая мировая война и особенно послевоенное время привели к широчайшему внедрению регрессии во все области научных исследований, экономического анализа и промышленного производства. Возник процесс, результаты которого имеют уже экологическое значение. События такого масштаба не могут проходить ни с того ни с сего. В данном случае решающую роль сыграла вычислительная техника. Появление в 50-е годы массового производства ЭВМ привело к регрессионному буму.

Классический регрессионный анализ опирается на некоторую систему постулатов в основном статистического характера. Эти постулаты гласят, что регрессия представляет собой линейную комбинацию некоторых линейно независимых базисных функций от факторов с неизвестными коэффициентами (параметрами). Факторы являются детерминированными. То же справедливо и для параметров. Что же касается откликов (измеряемых зависимых переменных), то считается что это равноточные (с одинаковой дисперсией) некоррелированные случайные величины. Кроме того, предполагается, что это нормально-распределенные случайные величины. И наконец, принимается, что все переменные измеряются в непрерывных шкалах. Такая основа позволяла благополучно довести до числа процесс получения оценок регрессионных коэффициентов и осуществить проверки основных статистических гипотез об уравнении регрессии, его коэффициентах и прогнозируемых значениях отклика.

Заметим кстати, что постулат о равноточности и некоррелированности откликов не является слишком жестким. Если отклики не равноточны и коррелированы, то вычисления коэффициентов регрессии практически не усложняются. Саму процедуру в этом случае называют взвешенным методом наименьших квадратов. По существу, это означает, что указанный постулат можно заменить на более общий, когда предполагается, что априори с точностью до сомножителя известна дисперсионная матрица измеряемых откликов.

Без всяких на то оснований мы до сих пор считали, что набор независимых переменных (факторов) задан однозначно, что все существенные переменные в модели присутствуют и что никаких альтернативных способов выбора факторов нет. Все это, конечно, не так. Выбор переменных, тесно связанный с выбором модели объекта, представляет собой одну из извечных и наисложнейших проблем. Никаких стандартных рецептов здесь нет. Основной аппарат — преобразования. Причем преобразования могут быть содержательными и формальными. Понятно, что содержательные всегда лучше, но не всегда

доступнее. Поэтому чаще встречается такой случай, когда сначала находятся удачные формальные преобразования, а затем выискивается их интерпретация, подбирается физический смысл. Удача на этом пути — всегда событие в науке. Для поиска формальных преобразований разработано множество статистических моделей. Прежде всего — это модели факторного и дискриминантного анализов. Они опираются на линейные преобразования факторного пространства, которые позволяют находить такие новые координаты, что обеспечивается выполнение того или иного условия оптимальности. Разновидностей и модификаций методов подобного рода столь много, что даже перечислить их нет никакой возможности. Постепенно выяснилось, что ЭВМ допускает отказ от жесткой модели объекта исследования и подбор в ходе обработки данных некоторой «наилучшей» модели. После публикаций М. Эфраимсона, известного американского статистика, разработчика первых машинных алгоритмов, предназначенных для решения регрессионных задач, в конце 50-х годов такой новый подход был взят на вооружение, и уже к середине 60-х годов появился целый набор методов, опирающихся на идею последовательного построения подходящих моделей.

Обычная процедура регрессионного анализа исходит из предпосылки, что все нужные данные для построения модели уже собраны. Но ведь на самом деле данные всегда в процессе сбора, их всегда мало. Можно ли обрабатывать данные, которые еще не собраны до конца? Конечно, можно. Для этого разработан целый веер приемов, в основе которых лежит идея последовательного анализа, принадлежащая А. Вальду. Последовательный анализ предполагает, что к некоторому исходному массиву данных добавляется по одной строке и после каждого такого добавления оценки уточняются в свете новой информации. Иногда такой частый пересчет не оправдан и он осуществляется блоками, шагами, после нескольких новых строк. Но можно вообразить и такой вариант, когда одновременно с добавлением новых строк могут вычеркиваться старые, причем не обязательно, чтобы их числа совпадали. Здесь мы оказываемся в рамках моделей стохастической аппроксимации, которые, применительно к нашему случаю, называют еще и текущим регрессионным анализом. В наше время он находит применение в алгоритмах управления некоторыми производственными процессами.

Приемы классического регрессионного анализа в основном обсуждаются в первой книге монографии. Здесь детально рассматриваются исходные предпосылки, процедуры отыскания оценок параметров, свойства этих оценок. Значительное внимание уделяется статистическим аспектам регрессионного анализа, включая проверку гипотез относительно параметров и линейных функций от них. Обосновывается процедура проверки адекватности регрессионной модели.

Процедуры выбора «наилучшей» регрессии из множества возможных сосредоточены в гл. 6, с которой начинается вторая книга.

Исходные предпосылки классического регрессионного анализа выполняются далеко не всегда. Как обнаружить нарушение этих предпосылок? В каких случаях и какие нарушения можно считать допу-

стимыми? Что делать, если нарушения признаются недопустимыми? Эти вопросы давно занимают специалистов по математической статистике.

Мощным средством обнаружения некоторых отклонений от исходных предпосылок регрессионного анализа является анализ остатков, представляющих собой разности между экспериментальными и расчетными значениями откликов. Исследованию остатков посвящена гл. 3 данной книги. Но мало просто обнаружить, что предпосылки нарушены. Нужна конкретная программа действий в указанных условиях. В силу сказанного со всей остротой возникла потребность пересмотра, смягчения основных постулатов регрессионного анализа. Это привело к появлению целого набора новых статистических методов, являющихся продолжением, развитием методов классического регрессионного анализа.

Начнем с пересмотра постулатов относительно базисных функций от факторов и относительно самих факторов. Еще в 20-е годы Р. Фишер разработал дисперсионный анализ. Этот прием, сыгравший огромную роль в развитии планирования эксперимента, породил массу частных моделей и соответствующих методов обработки данных. Понадобилось около 30 лет, чтобы началась консолидация процедур регрессионного и дисперсионного анализа. Стало ясно, что основная особенность задач дисперсионного анализа, если их трактовать в терминах регрессий, состоит не столько в том, что факторы здесь измеряются в дискретных шкалах, сколько в том, что соответствующие базисные функции от факторов оказываются линейно зависимыми. А это в свою очередь приводит к тому, что матрица системы нормальных уравнений вырождена и задача отыскания оценок параметров не имеет единственного решения. Усилиями К. Точера, С. Рао и других исследователей был найден прием, позволяющий свести любую задачу дисперсионного анализа к задаче регрессионного анализа, но с вырожденной матрицей системы нормальных уравнений. Для решения этой системы предлагается использовать так называемые обобщенные обратные матрицы Мура—Пенроуза.

Одновременно шел и «встречный» процесс: дисперсионный анализ начал широко применяться при изучении результатов регрессионного анализа. Это направление отчетливо прослеживается вплоть до наших дней.

От модели дисперсионного анализа оставался один шаг до смешанной модели, в которой представлены как регрессионные, так и дисперсионные переменные. Такая модель стала называться моделью ковариационного анализа. Ее введение тоже связано с именем Р. Фишера. В итоге удалось объединить в рамках одной формальной процедуры регрессионного анализа три типа моделей. Подобное объединение создает удобство при программировании и вычислениях на ЭВМ.

Дисперсионный анализ весьма обстоятельно описан в гл. 9 работы Н. Дрейпера и Г. Смита. Причем авторы уделили большое внимание сопоставлению регрессионного и дисперсионного анализа. В книге рассмотрены разные приемы элиминирования вырожденности исходной системы нормальных уравнений.

В классической регрессии факторы предполагаются детерминированными. Это означает, что в условиях реального эксперимента мы должны знать о них все, знать с бесконечной точностью. Реально ли это? Конечно, нет.

Отказ от детерминированности независимых переменных ведет к новой модели — модели корреляционного анализа. В одном частном случае, для парной корреляции, такая модель играет выдающуюся роль в статистическом анализе. Проявляется это и при исследовании регрессионных моделей. Но все попытки существенного обобщения этой модели на многомерный случай наталкиваются пока на серьезные препятствия. Главный камень преткновения здесь — требования к многомерным функциям распределения, которые не известно ни как обеспечить, ни как проверить.

Трудности многомерного корреляционного анализа привели в 30-е годы к созданию компромиссной модели — модели конфлюэнтного анализа, предложенной Р. Фришем. В этой модели допускается, что при нормально-распределенном отклике факторы тоже могут иметь некоторый разброс значений, тоже нормально-распределенный и усеченный. Причем никаких многомерных условий не налагается. В такой ситуации удается построить процедуру обработки данных, сводящую дело к многократному решению регрессионной задачи.

Теперь коснемся постулата о параметрах моделей. Модели со случайными параметрами рассматриваются в современном дисперсионном анализе, их именуют иногда моделями со случайными компонентами. Отказ от детерминированности параметров регрессионных моделей приводит к более серьезным последствиям, поскольку при этом затрагиваются статистические устои регрессионного анализа. Тем не менее такие модели имеют право на жизнь. Можно себе представить, что иногда существует информация о параметрах регрессионной модели, позволяющая задать некоторое априорное распределение этих величин, рассматриваемых как случайные. Тогда в качестве оценок параметров можно использовать их условные математические ожидания, если только имеют место наблюденные значения откликов. Когда условные распределения параметров используются для получения оценок, говорят о байесовском регрессионном анализе, поскольку условные (апостериорные) распределения и ожидания вычисляются по обобщенной формуле Байеса.

И наконец, обсудим постулаты, относящиеся к отклику регрессионной модели.

Регрессионные модели нередко применяются для описания процессов, развивающихся во времени. Заметим, что отклики при этом могут измеряться дискретно, а в определенные моменты времени непрерывно, на некотором временно́м интервале. В таком случае от рассмотрения случайных величин откликов придется перейти к анализу случайных последовательностей и случайных процессов. А в более общей ситуации, когда процесс развивается и во времени, и в пространстве,— может быть даже и к анализу случайных полей. Это приводит к серьезным осложнениям. Одна из распространенных про-

стейших моделей такого рода — модель авторегрессии. Она предполагает, что отклик зависит не только от ряда изучаемых входных переменных (факторов), но и от времени. Если последнюю зависимость удается выявить, то задача сводится к стандартной, но для преобразованного отклика. Если же нет — требуются специальные, более сложные приемы.

В обычной регрессионной модели предполагается, что неизвестные параметры сосредоточены в зависимости математического ожидания от факторов. Что же касается дисперсий и ковариаций измеряемых откликов, то считается, что они известны с точностью до сомножителя, отождествляемого часто с дисперсией ошибки эксперимента. В реальных задачах информация о дисперсиях и ковариациях откликов отнюдь не столь полна. В этой связи представляет интерес обобщенная регрессионная модель, допускающая зависимость дисперсий измерений от факторов. В эту модель может входить несколько неизвестных параметров. Это обобщение называют F-моделью. Разработана итерационная процедура «ИРДЖИНА» для поочередного оценивания параметров, входящих в выражение для математического ожидания отклика и в выражение для дисперсий измерений. F-модель имеет ряд преимуществ перед классической регрессией. К ней сводится, в частности, модель конфлюэнтного анализа.

Пока нормальный закон считался само собой разумеющимся, особых проблем не возникало. Тем более, что он опирался на авторитет центральной предельной теоремы теории вероятностей. Но когда под воздействием практики от этой догмы пришлось отказаться, стало ясно, что мы существенно зависим от априорной информации о законе распределения отклика. Ее уровень в разных задачах может быть совершенно различным, да и распорядиться ею можно по-разному. Когда мы заранее знаем, каков закон распределения, можно построить процедуру обработки данных, использующую эту информацию. Метод такого рода был разработан также Р. Фишером. Он называется методом максимума правдоподобия. Ясно, что стандартный классический вариант регрессионного анализа — частный случай этого метода. Хотя с вычислительной точки зрения возникающая процедура гораздо менее приятна, чем классическая, ничего страшного она не привносит. Вполне понятно, что учет надежной информации о фактическом законе распределения, скажем логнормальном вместо нормального, улучшит оценки, а в качестве платы за улучшение придется дольше считать по более сложной программе. Но в практике столь высокий уровень априорной информации встречается крайне редко. А что делать, если нам неизвестно истинное распределение?

В математической статистике давно была высказана мысль о том, что возможно получение некоторой полезной информации и в том случае, когда мы не можем или не хотим воспользоваться информацией о законе распределения изучаемой случайной величины. Пока мы верили в нормальность, эта идея не получала признания. Действительно, если нормальность на самом деле выполняется, то такие «непараметрические» процедуры будут существенно менее эффектив-

ными, чем процедуры классической теории. Они годились лишь для каких-то исключительных ситуаций. Когда же нормальность превратилась не более чем в частный случай, пусть распространенный, положение резко изменилось. Выяснилось, что когда отсутствует достаточно обоснованная информация о функции, описывающей регрессию и известной с точностью до параметров, можно построить такую регрессионную процедуру, которая по своей эффективности приближается к классической, а в ряде случаев она оказывается практически единственно возможной. Так появился еще один конкурент классической регрессии — непараметрический регрессионный анализ.

К нашему распределению, какое бы оно ни было, часто примешиваются чужеродные элементы, даже в малых количествах существенно ухудшающие ситуацию. Опыт показывает, что в больших массивах данных появление засорений практически неизбежно. Долгое время разрабатывались методы выявления подозрительных наблюдений, которые называют «дикими» или сорными. Отбрасывание таких наблюдений существенно улучшало положение. Однако, чтобы их выявить, надо снова знать закон распределения. В 1950 г. Дж. Бокс, занимаясь дисперсионным анализом, пришел к мысли о том, что можно не выявлять и не отбрасывать дикие наблюдения, а строить такие процедуры оценки, которые были бы нечувствительны к наличию в выборке засоряющих наблюдений. Он назвал такие процедуры робастными, или устойчивыми. С тех пор теория робастного оценивания вообще и для регрессии в частности быстро развивается. Выведены специальные формулы для робастных оценивателей. Исследованы ранее предложенные методы отыскания параметров регрессии. Выяснилось, что повышенной устойчивостью обладают оценки параметров, полученные по методу минимизации суммы модулей ошибок и максимального модуля ошибки (чебышевский метод оценивания). Новые веяния, относящиеся к робастному оцениванию, кратко описаны в гл. 6 монографии Н. Дрейпера и Г. Смита.

Робастные алгоритмы в известном смысле можно рассматривать как промежуточные компромиссные между параметрическими методами стандартной теории и непараметрическими подходами: они используют некоторую информацию о законах распределения, хотя и «распоряжаются» ею иначе.

По мере того как накапливался опыт работы с регрессиями, все больше и больше обнаруживались их «коварные» свойства. Выяснилось, например, что даже при соблюдении всех исходных постулатов МНК-оценки параметров, несмотря на все их оптимальные свойства, нередко с большой ошибкой оценивают параметры модели. И это вовсе не обязательно связано с плохим выбором условий эксперимента. Часто виновата сама регрессионная модель, ее структура. Если регрессия выражается в виде линейной комбинации экспонент или полиномом высокой степени, то столбцы матрицы $\mathbf{X}$ могут оказаться почти линейно зависимыми. Это явление, называемое мультиколлинеарностью, приводит к плохой обусловленности матрицы системы нормальных уравнений и к неустойчивости оценок параметров.

Плохо обусловленные задачи оценивания регрессии составили целое направление в регрессионном анализе. Они породили специальные, тонкие методы поиска оценок параметров. Практика показала, что повышения устойчивости оценок параметров можно добиться, если отказаться от требования их несмещенности, строго соблюдаемого в обычной регрессии. Так появилась гребневая, или риджрегрессия. Гребневая регрессия достаточно подробно описана в гл. 6 данной книги.

До сих пор речь шла о регрессиях с одномерным откликом. Однако реальные объекты, для описания которых привлекается регрессионный анализ, нередко имеют несколько откликов. В связи с этим представляет интерес многомерная (многооткликовая) регрессия. Появились такие модификации многомерной регрессии, как псевдонезависимая регрессионная модель, модель в виде системы одновременных (синхронных) уравнений.

В первом случае речь идет о ряде стохастически связанных между собой одномерных регрессионных уравнений. Во втором предполагается, что между разными откликами системы существуют линейные связи. Одновременные уравнения находят широкое распространение в эконометрии.

Регрессионные модели, построенные на базе полиномов, носят, как правило, формальный характер. Их используют для описания изучаемых объектов, относительно которых нет достаточно четких количественных представлений. Однако исследователей чаще интересуют содержательные, физические модели, отражающие механизм, сущность явлений. Если разработаны теоретические основы исследуемого явления, то может быть заранее известна структура модели. В этом случае экспериментальные данные служат лишь для определения отдельных параметров. Выбор же типа модели объекта — традиционный удел всякого специалиста. Это вообще один из центральных вопросов науки.

Содержательные, физические модели, как правило, нелинейны по параметрам. Методология их создания составляет один из интенсивно развивающихся разделов математической статистики — нелинейный регрессионный анализ. Нелинейный регрессионный анализ привнес в статистику целый клубок трудно решаемых проблем. Эти проблемы связаны не только с нелинейным характером зависимости откликов от параметров. Как правило, физические модели являются многомерными, отклики нередко связаны между собой. К тому же и сама регрессионная зависимость, связывающая отклики с факторами, выражается неявно. Она обычно представляет собой решение системы алгебраических или дифференциальных уравнений, которое чаще всего не может быть представлено аналитически. В результате появляется проблема точечного оценивания параметров нелинейных моделей. Она намного сложнее, чем в случае линейной параметризации. Оценки — чаще всего смещенные, причем степень смещения оценить нелегко.

Задача оценивания параметров нелинейных моделей, как правило, имеет не одно, а множество решений. А иногда решение вообще от-

сутствует. Неустойчивость оценок резко обостряется. Одним словом, здесь мы сталкиваемся со всеми атрибутами некорректно поставленных задач.

Однако недостаточно просто найти точечные оценки параметров, не менее важно их охарактеризовать статистически, определить их дисперсии и ковариации. В условиях нелинейной параметризации это — нелегкая задача. Линеаризация нелинейных по параметрам зависимостей может приводить к резко искаженным величинам дисперсий и ковариаций оценок параметров. Выходом из этой ситуации может быть использование асимптотических разложений функции отклика, в которых участвуют производные от функции отклика по параметрам более высоких степеней. Заметим, что алгоритм вычисления слагаемых таких разложений является очень трудоемким.

Весьма проблематичными становятся процедуры статистического анализа нелинейной регрессии. Даже если известно, что отклики подчиняются нормальному распределению, то что можно сказать про распределение оценок параметров? Как проверять гипотезы об адекватности модели, о значимости параметров? На эти вопросы пока нет исчерпывающих ответов.

Проблеме оценивания параметров нелинейных моделей посвящена гл. 10 монографии Н. Дрейпера и Г. Смита. Авторам удалось четко проследить сходство и различие между линейным и нелинейным оцениванием.

Регрессионный анализ — методологическая основа теории планирования эксперимента. Многие критерии оптимальности эксперимента заимствованы из соотношений, характеризующих свойства оценок параметров. Планирование эксперимента для линейно параметризованных моделей превратилось в хорошо разработанный, обширный раздел математической статистики. В настоящее время интенсивно развивается планирование эксперимента в случае нелинейной параметризации. В обсуждаемой монографии эти вопросы затронуты вскользь. Интересующийся читатель должен обратиться к специальной литературе (некоторые дополнительные ссылки на литературу приведены в примечаниях переводчиков к гл. 10).

Сейчас наступил новый этап развития вычислительной техники. Появились мини- и микроЭВМ, персональные компьютеры. Повышение быстродействия, увеличение памяти и удешевление ЭВМ, а также значительный прогресс в сервисных устройствах вызвали к жизни новые подходы к анализу данных, основанные на применении вычислительной техники. Это прежде всего относится к имитационному моделированию, предложенному Т. Нейлором. Не менее важное значение имеет концепция анализа данных, вытекающая из работ Дж. Тьюки. Большие надежды возлагают на разработанный и предложенный в 1979 г. Б. Эфроном метод «бутстреп». Все эти методы в совокупности с известными методами многомерной статистической классификации данных обогатили методологию регрессионного анализа. С другой стороны, сама регрессионная модель выступает теперь в качестве цементирующего начала, связывающего эти методы в нечто целостное.

Монография Н. Дрейпера и Г. Смита не охватывает все аспекты регрессионного анализа, что сделать, пожалуй, и невозможно. Важно другое: она дает фундаментальные представления о регрессии — как линейной, так и нелинейной. Опираясь на них, можно при желании углубить и расширить свои знания по регрессионному анализу, обратившись к другим источникам.

Предлагаемая книга может служить путеводителем по соответствующей литературе. К обширной библиографии, завершающей книгу и охватывающей период приблизительно до 1980 г., добавлен список литературы, где в основном приведены работы на русском языке.

Мы надеемся, что книга Н. Дрейпера и Г. Смита представит интерес для советского читателя и поможет статистикам, экономистам, социологам, научным работникам овладеть приемами и методами регрессионного анализа.

Ю. АДЛЕР,
В. ГОРСКИЙ

● ПРЕДИСЛОВИЕ К ПЕРВОМУ ИЗДАНИЮ

В 1962 г. к нам обратились представители химического отдела Американского общества контроля качества (A. S. Q. C.) с предложением подготовить краткий курс регрессионного анализа. В этой связи мы составили ряд конспектов по темам, которые, по нашему мнению, важны для практиков, применяющих регрессионный анализ. Эти конспекты были хорошо приняты, в дальнейшем в них было внесено много дополнений и изменений. Данная книга является результатом этой работы.

Мы попытались объединить в книге ряд методов, развитых для регрессионных задач и распространенных в настоящее время. Так как мы сделали акцент на практическом применении регрессионного анализа, то теоретические результаты во многих случаях приводятся без доказательств. Хотя обучение регрессионному анализу проводится без использования вычислительных машин или со сравнительно примитивной техникой, тем не менее работы по применению регрессии выполняются теперь исключительно с помощью быстродействующих вычислительных машин. Поэтому, хотя данной книгой можно пользоваться вообще без всяких вычислительных средств (или, быть может, только с настольной вычислительной машиной), мы специально употребили в нескольких местах машинные распечатки. Все десятичные знаки в этих данных вряд ли нужны на практике, но числа писались так, будто они получены на обычной вычислительной машине. Мы составили также различные упражнения: некоторые из них можно легко решить «вручную», для решения других, более сложных, была бы полезна, хотя и не абсолютно необходима, вычислительная машина.

Эта книга представляет собой стандартный основной курс множественной линейной регрессии, но она включает также материал, который либо совсем не появлялся в учебниках, либо если и появлялся, то был труднодоступен для понимания. Например, в гл. 3 обсуждается исследование остатков; в гл. 6 рассматриваются процедуры отбора факторов в регрессионных программах разных типов; в гл. 8 обсуждается планирование больших регрессионных исследований; гл. 10 дает основное введение в теорию нелинейного оценивания.

Гл. 1 и 3 вместе представляют собой курс по подбору уравнения прямой линии вообще без использования алгебры матриц. Если же прибавить часть гл. 2, то вдобавок можно получить введение в идеи матричного представления регрессионных задач. Односеместровый

курс регрессионного анализа может быть составлен из материалов гл. 1—7, возможно, с добавлением гл. 8 для статистиков, работающих в промышленности, и студентов, обучающихся управлению промышленностью. Можно изучить всю книгу за один семестр, если предположить, что некоторое предварительное знание части материала у читателя уже имеется. Для более полной проработки необходимо два семестра; это даст возможность преподавателю добавить доказательства утверждений, которые сформулированы, но не доказаны, и позволит более полно обсудить все упражнения, среди них есть довольно трудоемкие.

Предполагается, что читатель обладает основными знаниями элементарной статистики, которые можно получить из типового начального курса. Таблицы F-распределения и t-распределения приводятся.

Мы признательны ряду друзей за помощь в различных аспектах.

Март 1966 г.

Н. Р. ДРЕЙПЕР,
Г. СМИТ

● ПРЕДИСЛОВИЕ КО ВТОРОМУ ИЗДАНИЮ

За 15 лет, что пролетели с момента выхода в свет первого издания этой книги, в регрессионном анализе появилось много новых идей и методов. Так, новые численные алгоритмы и пакеты регрессионных программ сделали совсем простым исследование адекватности проверяемых моделей многими различными методами. Поэтому мы уже давно склонялись к тому, что надо бы дополнить первое издание, причем так, чтобы оно, с одной стороны, отражало ситуацию в прикладном регрессионном анализе вплоть до 1980 г., а с другой — соответствовало нашим мнениям и практическому опыту применения появившихся за это время в регрессии идей и методов.

Было добавлено много нового материала, и книга неизбежно потолстела. А исключен только один параграф 6.8. «Вычислительные аспекты шагового регрессионного метода». Основные добавления мы перечислим ниже. Сделано также множество других вставок и исправлений, слишком многочисленных, чтобы их здесь упоминать. Большая часть дополнительного материала прошла проверку не только в аудиториях университета штата Висконсин [1], но и на краткосрочных курсах в промышленности.

В гл. 1 добавлены параграф 1.7 об обратной регрессии и параграф 1.8 о практических возможностях обычной линейной регрессии.

В гл. 2 мы включили несколько новых параграфов: краткое описание метода наименьших квадратов (МНК) с ограничениями (2.13), несколько замечаний об ошибках в факторах (предикторах) (2.14), и обратную регрессию для случая многих факторов (2.15). А параграф 2.11 о взвешенном методе наименьших квадратов расширен за счет включения численного примера. К этой главе мы дали еще 4 приложения: приложение 2А — несколько полезных свойств матриц, 2Б — математические ожидания сумм квадратов, 2В — некоторые сведения о статистической значимости регрессионной модели, и 2Г — описание неопределенных множителей Лагранжа. В анализе остатков, гл. 3, мы добавили параграфы о сериальной корреляции остатков (3.9), о критерии Дарбина—Уотсона для сериальной корреляции (3.11), о методах определения влияющих наблюдений (3.12) и в при-

[1] Университет штата Висконсин расположен в Мадисоне — столице штата. Это один из крупных центров статистических исследований в США. В нем работает Н. Дрейпер, а руководит департаментом статистики Дж. Бокс (G. E. P. Box).— *Примеч. пер.*

ложении 3А— об использовании нормальных и полунормальных графиков остатков. В гл. 5 появились новые параграфы о семействах преобразований (5.3) и о регрессионном анализе усредненных данных (5.8). Но самым большим дополнениям и изменениям подверглась гл. 6. Здесь мы включили параграфы о статистике C_p Маллоуза и наилучших регрессионных подмножествах (6.1), о гребневой регрессии (ридж-регрессии) (6.7), о критерии предсказанной суммы квадратов (PRESS) (6.8), о регрессии на главных компонентах (6.9), о регрессии на собственных значениях (6.10) и о робастной (устойчивой) регрессии (6.14). В гл. 7 добавлен пример планирования эксперимента для исследования поверхности отклика (7.7). Появились некоторые вставки к методам обоснования моделей из гл. 8. В гл. 9 был расширен регрессионный подход к дисперсионному анализу. Наконец, мы включили новый материал о нелинейных моделях развития (роста) (10.7) и по ряду других вопросов работы с нелинейными моделями (10.8).

Выводы, полученные на основе очень малых наборов данных, часто практически бесполезны. Несмотря на это, в примерах и задачах мы используем несколько искусственно построенных наборов малого числа данных. Сделано это исключительно с целью упростить демонстрацию процесса вычислений, избежав обращения к большим массам чисел. Подобные данные надо и рассматривать в том духе, в каком они представлены. Нет никаких оснований расценивать их иначе, чем средство для иллюстрации, для чего они, собственно, и предназначены.

Учиться по этой книге можно многими способами. В университете штата Висконсин, например, один семестр занимают большая часть гл. 1—4 и куски из гл. 5, 6, 9 и 10. Советуем преподавателям взять за основу материал гл. 1—4, а из остальной части книги черпать по потребности. Никакие параграфы не отмечены звездочками, поскольку то, что можно пропустить «при первом чтении» для одной аудитории, может оказаться существенным для другой.

Мы признательны многим людям. Некоторые из них указаны непосредственно в тексте как авторы определенных сведений. Среди других надо упомянуть многочисленных (за долгие годы) помощников Н. Р. Дрейпера по преподаванию в университете штата Висконсин. Они подготовили решения новых упражнений, часто сами предлагали упражнения, а иногда терпеливо добывали материал, важный для дипломных работ некоторых студентов-выпускников, специализировавшихся в прикладном регрессионном анализе.

Январь 1981 г.

Н. Р. ДРЕЙПЕР,
Г. СМИТ

Глава 1 ● ПОДБОР ПРЯМОЙ МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ

1.0. ВВЕДЕНИЕ. ПОТРЕБНОСТЬ В СТАТИСТИЧЕСКОМ АНАЛИЗЕ

В современной промышленности нет недостатка в «информации» независимо от того, «увешан» ли процесс множеством измерительных приборов или их мало. Показания приборов говорят нам о таких вещах, как начальная температура, концентрация реагента, процент катализатора, температура пара, скорость расходования веществ, давление и т. д., в зависимости от характеристик процесса в данном исследовании. Некоторые из этих показаний получаются через равные интервалы, например каждые пять минут или каждые полчаса, другие измеряются непрерывно. Однако при небольших дополнительных затратах времени и усилий всегда можно иметь непрерывные показания. А анализ конечного продукта можно производить и периодически. В результате такого анализа получаются количественные данные о чистоте продукта, проценте выхода, блеске, сопротивлении разрушению, цвете и о многих других свойствах, имеющих значение для производителя или для потребителя. На многих заводах мы находим гигантские залежи подобных данных. И часто цифры просто коллекционируются без всякого понимания цели или смысла или же во имя целей, ставившихся в прежние годы. Несмотря на то что этих целей больше уже не существует, цифры все же благоговейно собираются час за часом, день за днем, неделя за неделей.

Цель этой книги, однако, не в разъяснении того, какая информация должна или не должна собираться для какого-либо процесса. Цель в другом. Данные только что указанного типа представляются в виде таблиц чисел. В этих числах могут быть завуалированы некоторые соотношения или же эти соотношения могут непосредственно следовать из данных. Мы будем довольно подробно рассматривать некоторые методы выявления основных черт таких соотношений. Сверх того, изучение методов регрессионного анализа может пролить некоторый свет на то, как надо планировать сбор данных, если к тому предоставляется возможность. Это видно, например, из параграфа 1.8.

Для любых задач с изменяющимися количественными переменными представляет интерес исследование влияния (действительного или подозреваемого) некоторых переменных на остальные. Таким влиянием, конечно, может быть простая функциональная связь между переменными; однако во многих физических процессах это скорее исключение, чем правило. Часто, видимо, существует функциональная связь, слишком сложная для понимания или для описания в простых терминах. В таком случае мы можем стремиться подобрать ап-

проксимацию этой функциональной связи с помощью какой-нибудь простой математической функции (скажем, такой, как полином), которая включает подходящие переменные, и сглаживать или аппроксимировать «истинную» функцию в определенной ограниченной области изменения этих переменных. При исследовании такой сглаженной функции мы сможем больше узнать о рассматриваемой «истинной» зависимости и оценить отдельные или совместные эффекты изменения некоторых важных переменных.

Даже тогда, когда по смыслу не существует физической связи между переменными, мы можем стремиться к тому, чтобы отразить ее с помощью математического уравнения данного вида. Если уравнение физически бессмысленно, то оно тем не менее может оказаться весьма ценным для предсказания значений ряда переменных по известным значениям других переменных, быть может, при определенных ограничениях.

В этой книге будем пользоваться одним частным методом получения математической зависимости. Он включает исходное предположение о том, что имеет место определенный тип зависимости, линейной относительно неизвестных параметров (за исключением гл. 10, где рассматриваются нелинейные модели). Неизвестные параметры оцениваются еще при ряде других предположений по имеющимся данным, и получается искомое уравнение. Можно оценить полезность полученного уравнения и проверить, не оказались ли некоторые из предположений ошибочными. Простейшим примером этой процедуры служит подбор прямой по парам наблюдений $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. В данной главе мы рассмотрим его с помощью обычной алгебры. Если же задача включает большое число переменных, то основным становится матричный подход. Он вводится в связи с подбором прямой в гл. 2, которая включает также наиболее важные результаты для более общих регрессионных задач. Часть из этих результатов применяется в гл. 4, где обсуждается задача установления связи переменной Y с двумя переменными X_1 и X_2 с помощью уравнения плоскости. В гл. 5 рассматриваются более сложные модели, а в гл. 6— некоторые методы, используемые при выборе «наилучшего» уравнения. Типичные примеры изучаются в гл. 7, а основные этапы и задачи, связанные с построением моделей, содержатся в гл. 8. В гл. 9 обсуждается регрессионная обработка в задачах дисперсионного анализа, а в гл. 10 дается введение в нелинейное оценивание. В приложении приводятся машинные распечатки, упоминаемые и используемые в различных местах книги.

В гл. 1, 2, 3 и 4 изложен основной курс регрессионного анализа. Читатели, мало знакомые или вовсе не знакомые с матричной алгеброй, будут постепенно знакомиться с ней в гл. 2, остальные же могут пропустить начальные параграфы этой главы.

Читатели с весьма основательными знаниями в области регрессионного анализа могут относиться к последним параграфам гл. 2 как к резюме и обзору и бегло просмотреть конец гл. 2, так же как гл. 1, 3 и 4 (возможно, и 5). Мы надеемся, что последующие главы они найдут интересными и полезными.

Читатели, занимающие промежуточное положение, извлекут пользу из систематической проработки всей книги.

Мы предполагаем, что все, кто воспользуется этой книгой, знакомы с начальным курсом статистики и понимают ее основные идеи. Сюда включаются представления о параметрах, оценках, распределениях (особенно нормальном), среднем и дисперсии случайной величины, ковариации между двумя переменными и проверке простых гипотез, в том числе и с использованием одно- и двусторонних t- и F-критериев. Мы полагаем, однако, что читатели, которые забыли эти понятия или знают их неполно, смогут тем не менее быстро восполнить пробелы [2].

Мы не собирались рассматривать эту книгу как всеобъемлющий учебник по всем аспектам регрессионного анализа. В наши намерения входило дать капитальный основной курс плюс материал, необходимый для решения распространенных практических регрессионных задач.

Теперь мы воспользуемся случаем пораньше ознакомить читателей с приложениями, где содержатся машинные распечатки. Взгляните, например, на начало распечатки в приложении А (см. кн. 2). Здесь вы увидите наблюдения за работой выпарного аппарата на большом промышленном предприятии, разбитые на интервалы. Фиксировались следующие десять переменных:

1. Количество используемого пара в фунтах ежемесячно.
2. Количество активной жирной кислоты в фунтах, накопленное за месяц.
3. Количество готового глицерина-сырца в фунтах.
4. Средняя скорость ветра в милях в час.
5. Число календарных дней в месяце.
6. Число рабочих дней в месяце.
7. Число дней с температурой ниже 32 °F.
8. Средняя температура воздуха (°F).
9. (Средняя скорость ветра)^2.
10. Число пусков.

---

[2] Для «восполнения пробелов» существует обширная литература. Назовем здесь лишь несколько книг, которые могут оказаться полезными: Хальд А. Математическая статистика с техническими приложениями/Пер. с англ. Под ред. Ю. В. Линника.— М.: ИЛ, 1956.— 664 с.; Налимов В. В. Применение математической статистики при анализе вещества.— М.: Физматгиз, 1960.— 430 с.; Пустыльник Е. И. Статистические методы анализа и обработки наблюдений.— М.: Физматгиз, 1968.— 288 с.; Закс Л. Статистическое оценивание/Пер. с нем. Под ред. Ю. П. Адлера и В. Г. Горского.— М.: Статистика, 1976.— 598 с.; Гласс Дж., Стэнли Дж. Статистические методы в педагогике и психологии/Пер. с англ. Под ред. Ю. П. Адлера.— М.: Прогресс, 1976.— 495 с.; Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. Т. 1. Методы обработки данных/Пер. с англ. Под ред. Э. К. Лецкого.— М.: Мир, 1980.— 610 с. Каждая из названных книг содержит всю информацию, необходимую для понимания данной книги.— *Примеч. пер.*

(Способ, которым в действительности были объяснены данные из приложения А, изложен ниже.)

Мы можем различить здесь два основных типа переменных. Назовем их *предсказывающими переменными* (*предикторами*), или *независимыми переменными* (*факторами*), и *зависимыми переменными*, или *переменными-откликами* [3]. Под предикторами, или факторами, мы будем понимать такие переменные, для которых обычно можно устанавливать желаемые значения (например, начальную температуру или скорость подачи катализатора), либо те, которые можно только наблюдать, но не управлять ими (например, влажность воздуха). В результате преднамеренных изменений или изменений, происшедших с независимыми переменными случайно, появляется эффект, который передается на другие переменные, на отклики (например, на окончательный цвет или чистоту химического продукта). В общем, мы будем интересоваться тем, какие изменения предикторов влияют на значения откликов. Если мы сможем обнаружить простое соотношение или зависимость отклика от одного или нескольких факторов, то это, конечно, нам понравится. Разделение на предикторы и отклики не всегда вполне четко и иногда зависит от наших целей. Так, можно рассматривать отклик промежуточной стадии процесса как предиктор для (скажем) конечного цвета продукта. Практически, однако, роли переменных обычно легко различимы. Когда говорят «независимые переменные», не стоит понимать это выражение слишком буквально. В конкретной массе данных две или несколько переменных могут изменяться одновременно некоторым определенным образом, возможно, связанным с методом, лежащим в основе эксперимента. Это обычно нежелательно — прежде всего потому, что ограничивается информация об индивидуальной роли факторов,— но часто неизбежно.

Возвращаясь к приложению А, мы видим, что изучается 25 наборов наблюдений за переменными, по одному набору для каждого из двадцати пяти месяцев. Здесь нас прежде всего интересует количество продукта, произведенное за месяц, а затем его изменения из-за варьирования остальных факторов. Таким образом, мы будем считать переменную 1 откликом (Y), а остальные переменные — предикторами, X_2, X_3, . . . , X_{10}.

Теперь рассмотрим метод анализа, называемый *методом наименьших квадратов* [4]. Его можно применять для обработки данных эксперимента и для получения разумных заключений о свойствах выбранного уравнения. Этот метод часто называют *регрессионным анали-*

---

[3] За годы, прошедшие между первым (1966 г.) и вторым (1981 г.) изданиями этой книги, в англоязычной статистической терминологии произошли любопытные изменения. Мы будем отмечать их по мере появления. Так, например, частный термин «предиктор» (predictor) вытесняет более общий термин «фактор» (или «независимая переменная»). Видимо, это обусловлено высокой частотой задач предсказания в практике. Мы следуем здесь за оригиналом.— *Примеч. пер.*

[4] Далее наряду с полным названием мы будем использовать сокращение МНК. Например, уравнение, полученное методом наименьших квадратов, может именоваться МНК-уравнением.— *Примеч. пер.*

зом [5]. По-видимому, авторство слова «регрессия» принадлежит известному английскому антропологу и метеорологу сэру Фрэнсису Гальтону (1822—1911). Первоначально в неопубликованном докладе «Основные законы наследственности человека», прочитанном в Королевской ассоциации 9 февраля 1877 г., он употребил термин «реверсия» (обращение, движение вспять.— *Примеч. пер.*). Более поздний термин «регрессия» появился в президентском адресе, прочитанном перед секцией Н Британской ассоциации в Абердине в 1885 г. и опубликованном в журнале «Nature» [6] в сентябре 1885 г. (с. 507—510), а также в статье «Регрессия к середине в наследовании роста», опубликованной в «Журнале антропологического института» (Journal of the Anthropological Institute, 1885, **15**, p. 246—263). В этой статье Гальтон рассказывает о своих первоначальных исследованиях (с. 246), в которых «наследники» семян «*не* проявляли тенденции к воспроизведению размеров своих родителей, а, напротив, всегда были ближе к середине, чем они (под серединой имеется в виду среднее арифметическое). А именно: семена были меньше, чем их родители, если родители были велики, и больше, если родители были очень малы . . . Дальнейшие эксперименты показали, что в среднем сыновняя регрессия к середине прямо пропорциональна отклонению родителей от нее». Затем Гальтон переходит к описанию того, как та же самая картина проявилась в данных о «росте 930 взрослых детей и 205 их родителей» [7]. По существу, он показал, что, если Y равен росту ребенка, а X равен росту родителей (на самом деле взвешенному сред-

---

[5] По-видимому, следует различать МНК — вычислительный прием, обеспечивающий минимизацию некоторой заданной квадратичной формы при фиксированном множестве данных, и регрессионный анализ — статистический анализ регрессионной модели, т. е. такой модели, в которой зависимая переменная (отклик) является случайной величиной, а независимые переменные (предикторы) — детерминированные величины. МНК — составная часть регрессионного анализа. На практике эти две вещи часто путают, что иногда приводит к недоразумениям.— *Примеч. пер.*

[6] Королевская ассоциация (Royal Institution или полно — Royal Institution of Great Britain) — это научная организация, проводящая исследовании и распространяющая знания в области физики, астрономии, химии, электроники, физиологии. Находится в Лондоне; основана в 1799 г. Британская ассоциация в Абердине (British Association at Aberdeen или полно —British Association for the Advancement of Learning) проводит ежегодные форумы ученых с докладами о последних достижениях, способствуя тем самым распространению научных знаний. Находится в шотландском портовом городе Абердине; основана в 1831 г. «Nature» («Природа») — научно-популярный журнал, освещающий связи жизни общества с природой; выходит 3 раза в месяц в Лондоне; основан в 1869 г.— *Примеч. пер.*

[7] Здесь речь идет о 930 взрослых детях, однако в краткой биографии Гальтона (см. К а и а е в И. И. Френсис Гальтон.— Л.: Наука, 1972, с. 100) говорится, что детей было 928. Это противоречие вынудило нас обратиться к первоисточнику. Оказалось, что при первом упоминании в описательной части работы Гальтон действительно указал число 930, однако во всех таблицах и вычислениях фигурирует 928. Так что фактически у наших авторов вкралась ошибка. О причинах противоречия в тексте Гальтона остается только гадать. Может быть, он забраковал пару детей по какой-либо причине и забыл исправить текст, а может быть, просто при наборе вводной части выпало слово «около», которое бы все объяснило.— *Примеч. пер.*

нему ростов матерей и отцов; подробности — в исходной статье), то прекрасно подойдет уравнение вида $\hat{Y} = \overline{Y} + (2/3)(X - \overline{X})$, хотя он так и не выражался. (Обозначения объясняются в параграфе 1.1.) Статья Гальтона — увлекательное чтение. Сегодня анализ Гальтона надо было бы назвать «корреляционным анализом», впрочем, этот термин тоже придумал он. Термин «регрессия» вскоре начали применять к зависимостям в совершенно иных ситуациях, чем та, в которой он возник, и даже в таких ситуациях, где предикторные переменные *не* случайны, причем это словоупотребление сохранилось по сей день. В большинстве случаев построения современных моделей нет элемента «регрессии» в первоначальном смысле. Тем не менее слово так прижилось, что мы продолжаем им пользоваться [8]. (Отметим, что метод

---

[8] Сэр Фрэнсис Гальтон внес выдающийся вклад во многие области исследования, но статистическая методология проходит красной нитью практически через все его начинания. Регрессия и корреляция — наиболее важный вклад Гальтона в статистику. В разработке концепции регрессии он, видимо, не имел даже явных предшественников. Что же касается корреляции, то это — термин латинского происхождения (correlatio). Его введение в современную терминологию связано, видимо, с именем Кювье (см.: Канаев И. И. Жорж Кювье (1769—1832).— Л.: Наука, 1976.— 212 с.) и относится к 1806 г., когда Кювье занимался сравнительной анатомией. Затем он положил эту идею, оказавшуюся весьма плодотворной, в основу палеонтологии, родившейся в жарких спорах с Жоффруа Сеит-Илером. Оригинальный взгляд на роль Кювье и проблему корреляции содержится в кн.: Фуко М. Слова и вещи. Пер. с фр.— М.: Прогресс, 1977.— 488 с., особенно в разделе о Кювье, с. 345—366. Фуко ссылается также на книгу о Кювье: Daudin. Les classes zoologiques. Paris, 1930 и на книгу о Жоффруа Сеит-Илере: Cahn Th. La Vie et l'ocuvre d'E. Geoffroy Saint-Hilaire.— Paris: 1962. Вся проблема Кювье явно требует дальнейших исследований. Способ вычисления корреляции был известен многим исследователям, занимавшимся многомерным нормальным распределением и основанной на нем теорией ошибок измерений. Среди них были К. Ф. Гаусс (1777—1855), Огюст Браве (1811—1863) и Фрэнсис Эджворт (1845—1926). Так, Браве получил формулу, по которой мы сегодня вычисляем коэффициент корреляции, в 1846 г. Но только Гальтон понял, что корреляция — это мера связи между переменными. Карл Пирсон (1857—1936) превратил эту концепцию в статистическую теорию. Интересно, что Гальтон, видимо, не ссылался на работы Кювье (во всяком случае мы таких ссылок не нашли). Это скорее всего означает, что он не воспринимал работы Кювье как предшествующие своим. Вместе с тем известно, что сам Кювье был хорошо известен Гальтону и интересовал его. В приложении к своей книге о наследственном гении Гальтон привел краткую справку о Кювье и его жизни.

Кроме биографии Гальтона, упомянутой в предыдущем примечании, отметим еще: Филипчеико Ю. А. Фрэнсис Гальтон и Грегор Мендель.— М.: ГИЗ, 1925, с. 3—56; Pearson K. The Life, letters and labours of Francis Galton. Cambridge: University Press, 1914—1930 (in 3 vol.).

Современное состояние проблемы, с которой Гальтон начинал работы по корреляции и регрессии, отражено в кн.: Гинзбург Э. Х. Описание наследования количественных признаков.— Новосибирск: Наука, 1984.— 249 с. Исторические подробности читатель найдет в работах: Математика XIX века. Математическая логика, алгебра, теория чисел, теория вероятностей. Под ред. А. Н. Колмогорова, А. П. Юшкевича.— М.: Наука, 1978, с. 229—235, особо с. 233; Докторов Б. З. «Принцип корреляции» и развитие математической теории корреляции.— В кн.: Успехи биометрии.— Л.: Изд-во ЛГУ, т. 72, вып. 5, с. 8—23, библ. 60 назв.; Rodriguez R. N. Correlation.— In: Encyclopedia of statistical sciences.— New York: Wiley, 1982, 2, p. 193—204.

наименьших квадратов был известен задолго до появления Гальтона; см. с. 32.)

Мы начнем изучать метод наименьших квадратов в связи с простейшим приложением — подбором «наилучшей» прямой по данным для двух переменных X и Y, а затем обсудим возможность распространения результатов на случаи, когда рассматривается большее число факторов.

1.1. ПРЯМОЛИНЕЙНАЯ ЗАВИСИМОСТЬ МЕЖДУ ДВУМЯ ПЕРЕМЕННЫМИ

Во многих экспериментальных работах мы хотим исследовать, как изменения одной переменной влияют на другую. Иногда две переменные связаны точным уравнением прямой линии. Например, если сопротивление R простой цепи поддерживается постоянным, то протекающий ток I меняется линейно при линейном изменении напряжения V в соответствии с законом Ома $I = V/R$. Если бы мы не знали закона Ома, то могли бы найти зависимость эмпирически, изменяя V и измеряя I, поддерживая тем временем R фиксированным. Тогда мы бы увидели, что график зависимости I от V дает более или менее прямую линию, проходящую через начало координат. Мы сказали «более или менее», так как, хотя зависимость фактически точная, наши измерения могут содержать малые ошибки, и поэтому точки на графике, возможно, не попадут строго на линию, а будут разбросаны вокруг нее случайным образом. Однако для предсказания I по частным значениям V (при фиксированном R) мы будем использовать прямую, проходящую через начало координат. Иногда линейная зависимость не точна (даже без учета ошибки). Но тем не менее она может иметь смысл. Пусть, например, рассматриваются рост и вес взрослых мужчин из некоторой данной популяции. Если мы нанесем на график пары чисел (Y_1, Y_2) = (рост, вес), то результат будет примерно соответствовать рис. 1.1. (Такое изображение обычно называют диаграммой рассеяния, или точечной диаграммой [9].)

Важный вклад в развитие учения о корреляции внесла отечественная статистическая школа в лице таких ее представителей, как А. А. Чупров, Е. Е. Слуцкий. Н. С. Четвериков и др. См., например: Слуцкий Е. Е. Теория корреляции и элементы учения о кривых распределения. Киев, б. указ. изд., 1912.— 211 с.; Чупров А. А. Очерки по теории статистики.— М.: Госстатиздат, 1959, 319 с. и др.-

Заметим еще, что, хотя термин «регрессия» критиковали едва ли не все, кто писал о нем, он, несмотря ни на что, «жив и здоров». Можно предположить, что его удивительная устойчивость связана с переосмыслением значения. Постепенно исходная антропометрическая задача, занимавшая Гальтона, была забыта, а интерпретация вытеснилась благодаря ассоциативной связи с понятием «регресс», т. е. движение назад. Сначала берутся данные, а уж потом, задним числом, проводится их обработка. Такое понимание пришло на смену традиционной, еще средневековой, априорной модели, для которой данные были лишь инструментом подтверждения. Негативный оттенок, присущий понятию «регресс», думается, и вызывает психологический дискомфорт, поскольку воспринимается одновременно с понятиями, описывающими такой прогрессивный метод, как регрессионный анализ.— *Примеч. пер.*

[9] В советской литературе используют еще термины «поле корреляции», «поле рассеяния».— *Примеч. пер.*

Заметим, что для любого заданного роста встречаются различные веса и наоборот. Такая вариация может, в частности, получиться из-за ошибки измерений, но прежде всего это, конечно, следствие разброса между индивидами. Поэтому не приходится ожидать никакого единственного однозначного уравнения связи между ростом и весом. Однако мы можем обнаружить, что средний наблюденный вес при заданном росте растет с увеличением роста. Геометрическое место точек средних наблюденных весов при данных ростах (при изменении роста) назовем *регрессионной кривой* веса от роста. Обозна-

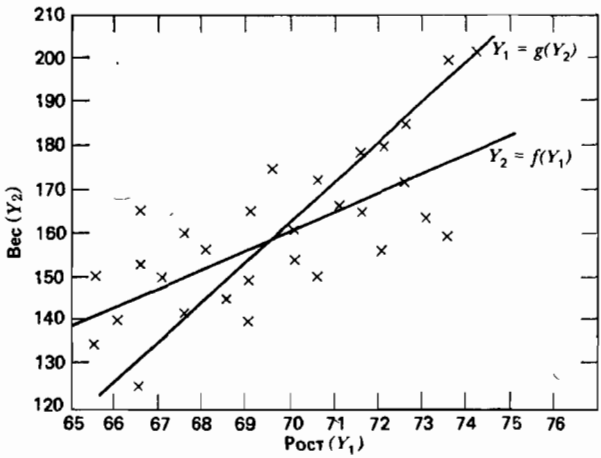


Рис. 1.1. Рост и вес 30 мужчин-американцев

чим это следующим образом: $Y_2 = f(Y_1)$. Существует также и регрессионная кривая роста от веса, подобная рассмотренной, которую мы можем записать так: $Y_1 = g(Y_2)$. Предположим, что обе эти «кривые» есть прямые (какими они могут и не быть). Вообще эти две кривые не есть *одно и то же*, что хорошо видно на рисунке.

Пусть мы теперь располагаем надежными данными по индивидуальным ростам, но не знаем соответствующих весов, которые хотим оценить. Что мы должны сделать? Мы должны найти из регрессионной линии веса от роста средние наблюденные веса индивидов данных ростов и использовать эти средние как оценки весов, которыми мы не располагали.

Пары случайных переменных, таких, как пара (рост—вес), имеют двумерное распределение вероятностей некоторого типа. Если мы установим связь между зависимой случайной величиной Y и величиной X, которая является переменной, но *не является случайной* переменной, то уравнение Y относительно X будет называться *уравнением регрессии*. Хотя это название, строго говоря, некорректно, оно, как мы уже говорили, укоренилось и широко распространено.

Почти всюду в этой книге мы будем предполагать, что переменные-предикторы не подвержены случайной вариации, тогда как отклики, напротив, подвержены. С практической точки зрения весьма редко такое предположение оказывается безупречным, но если это не так, то требуются гораздо более сложные методы построения зависимостей. Чтобы обойти возникшую трудность, мы используем метод наименьших квадратов только в таких ситуациях, где можно предположить, что вся возможная случайная вариация в любом предикторе столь мала по сравнению с наблюдаемым *диапазоном* его из-

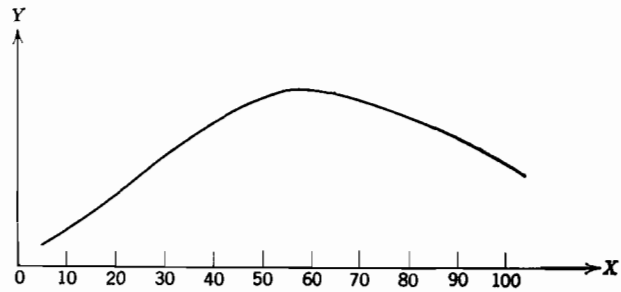


Рис. 1.2. Зависимость отклика от фактора

менения, что ею вполне можно пренебречь. И это предположение выполняется не часто, но оно подразумевается в каждой работе, посвященной методу наименьших квадратов, где предикторы считаются «фиксированными». (В таком контексте слово «фиксированный» означает «неслучайный», а вовсе не то, что предикторы вообще не могут иметь различных значений или уровней.) За дальнейшими подробностями обратитесь к параграфу 2.14.

Можно убедиться, что знать вид зависимости весьма полезно и когда зависимость строго линейная, и когда она линейна лишь для средних значений. (Зависимость может быть, конечно, более сложной, чем линейная, но мы будем пока рассматривать именно этот простой случай.)

Прямолинейная зависимость может быть полезна, даже если *известно*, что такое соотношение вообще не может быть верным. Рассмотрим зависимость отклика от фактора, показанную на рис. 1.2. Она, очевидно, нелинейна в диапазоне $0 \leqslant X \leqslant 100$. Однако если бы мы заинтересовались главным образом интервалом $0 \leqslant X \leqslant 45$, то линейное уравнение для наблюдений в *этих пределах* могло бы обеспечить вполне адекватное представление. Конечно, построенное уравнение неприменимо для значений X, выходящих за эти границы, так как оно не может обеспечить разумного предсказания.

(Подобные замечания можно сделать и в тех случаях, когда рассматривается более чем одна независимая переменная (предиктор). Пусть мы хотим исследовать, каким образом отклик Y зависит от факторов $X_1, X_2, \ldots, X_k$. Мы получаем уравнение регрессии для дан-

ных, которые «покрывают» некоторую область в «пространстве X»[10]. Пусть точка $X_0 = (X_{10}, X_{20}, \ldots, X_{k0})$ лежит *вне* области, покрываемой исходными данными. Хотя математически можно получить предсказанное значение $Y(X_0)$ для отклика в точке X_0, мы должны ясно понимать, что доверять такому предсказанию крайне опасно и опасность возрастает при удалении X_0 от исходной области, если, конечно, не привлекается некоторая имеющаяся дополнительная информация, делающая уравнение регрессии пригодным в широкой области пространства X. Заметим, что иногда трудно понять сразу, что интересующая нас точка лежит за пределами данной области многомерного пространства. Возьмем в качестве простого примера область, ограниченную эллипсом на рис. 1.3, внутри которой лежат все точки (X_1, X_2), а соответствующие им значения Y, лежащие на перпендикуляре к странице, здесь не показаны. Мы видим, что в область попадают точки, для которых $1 \leqslant X_1 \leqslant 9$ и $2{,}4 \leqslant X_2 \leqslant 6{,}3$. Тем не менее, хотя обе координаты точки P лежат в этих пределах, сама точка находится вне области. Если рассматривается больше переменных, то легко могут возникнуть недоразумения такого типа.)

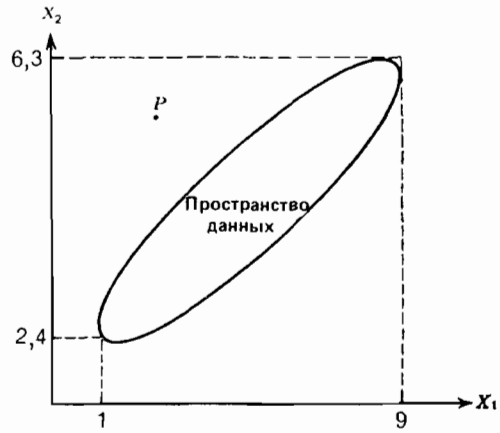


Рис. 1.3. Точка вне области, покрываемой данными

1.2. ЛИНЕЙНАЯ РЕГРЕССИЯ: ПОДБОР ПРЯМОЙ

Мы упоминали, что уравнение прямой может быть полезно во многих ситуациях для обобщения наблюдаемой зависимости одной переменной от другой. Теперь покажем, как такое уравнение можно получить методом наименьших квадратов, когда имеются экспериментальные данные. Выделим в машинной распечатке на с. 30 двадцать пять наблюдений переменной 1 (количество пара (в фунтах) [11], израсходованного за месяц) и переменной 8 (средняя температура воздуха в градусах Фаренгейта). Соответствующие пáры наблюдений приведены в табл. 1.1 и нанесены на график рис. 1.4.

---

[10] Наряду с таким термином будут употребляться его синонимы: «факторное пространство» и «пространство предикторов».— *Примеч. пер.*

[11] Так как все примеры в книге носят иллюстративный характер, а при обработке данных числовой материал часто подвергается перекодировке, мы решили сохранить систему единиц, которой пользуются авторы, и не переводить данные в СИ. — *Примеч. пер.*

Предположим, что линия регрессии переменной, которую мы обозначим Y, от переменной (X) имеет вид $\beta_0 + \beta_1 X$. Тогда можно записать линейную модель [12]:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \tag{1.2.1}$$

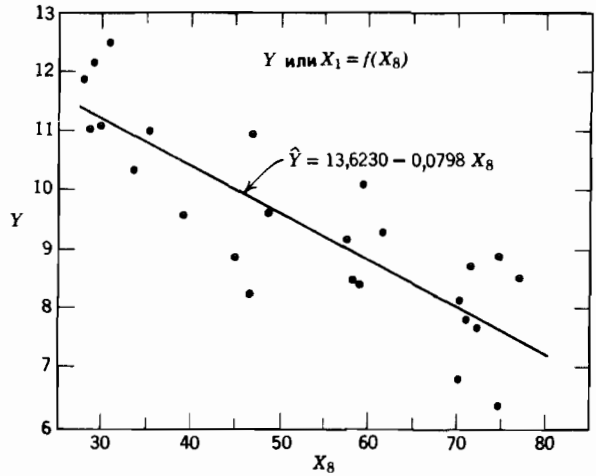


Рис. 1.4. Данные и подобранная прямая

так что для данного X соответствующее значение Y состоит из величины $\beta_0 + \beta_1 X$ плюс добавка ε, при учете которой любой индивидуальный Y получает возможность не попасть на линию регрессии.

Таблица 1.1. **Двадцать пять наблюдений переменных 1 и 8**

Номер опыта	Номер переменной		Номер опыта	Номер переменной	
	1 (Y)	8 (X)		1 (Y)	8 (X)
1	10,98	35,3	14	9,57	39,1
2	11,13	29,7	15	10,94	46,8
3	12,51	30,8	16	9,58	48,5
4	8,40	58,8	17	10,09	59,3
5	9,27	61,4	18	8,11	70,0
6	8,73	71,3	19	6,83	70,0
7	6,36	74,4	20	8,88	74,5
8	8,50	76,7	21	7,68	72,1
9	7,82	70,7	22	8,47	58,1
10	9,14	57,5	23	8,86	44,6
11	8,24	46,4	24	10,36	33,4
12	12,19	28,9	25	11,08	28,6
13	11,88	28,1			

[12] Наряду с термином «линейная модель» используется выражение «модель первого порядка». Хотя в данном случае они оба верны, их все же надо различать, так как они относятся к разным вещам (см. ниже).— *Примеч. пер.*

Уравнение (1.2.1) — это *модель*, в которую мы верим [13]. Начнем с предположения, что эта модель установлена, но на последующих стадиях будем проверять, так ли это на самом деле. Предположение о математической модели процесса необходимо с многих статистических точек зрения. Следует подчеркнуть, что то, что мы обычно делаем, есть *постулирование* модели либо *предварительное допущение о ее правильности*. Модель надо всесторонне критически исследовать в разных аспектах. Это наше «мнение» о ситуации на первой стадии исследования и это «мнение» может измениться, если мы найдем на более поздней стадии, что факты против него. Величины β_0 и β_1 называют *параметрами* модели.

(Примечание. Когда мы говорим, что модель линейна или нелинейна, мы имеем в виду линейность или нелинейность *по параметрам*. Величина наивысшей степени предиктора в модели называется *порядком* [14] модели. Например,

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$$

есть регрессионная модель второго порядка (по X) и линейная (по β). Если только специально не оговаривается, что модель нелинейна, а это может быть сделано, то имеется в виду линейная по параметрам модель, а слово «линейная» обычно опускается. Порядок модели может быть любым. Обозначение вида β_{11} часто используется в полиномиальных моделях, где параметр β_1 соотносится с X, в то время как β_{11} соотносится с $X^2 = XX$. Естественное обобщение обозначений такого рода встречается, например, в параграфах 5.1 и 7.7.)

Итак, в уравнении (1.2.1) величины β_0, β_1 и ε неизвестны, причем величину ε на самом деле будет трудно исследовать, поскольку она меняется от наблюдения к наблюдению. Однако β_0 и β_1 остаются постоянными, и, хотя мы не умеем находить их точно без изучения всех возможных сочетаний Y и X, мы можем использовать информацию, содержащуюся в двадцати пяти наблюдениях табл. 1.1, для получения *оценок* b_0 и b_1 параметров β_0 и β_1. Запишем это в таком виде:

$$\hat{Y} = b_0 + b_1 X, \tag{1.2.2}$$

где Y (читается «Y с крышечкой») обозначает *предсказанное* значение Y для данного X, когда b_0 и b_1 определены [15]. Уравнение (1.2.2) можно использовать как предсказывающее уравнение; подстановка в него значения X позволяет предсказать «истинное» среднее значение Y для этого X.

Общепринято обозначение оценок параметров маленькими латинскими буквами, а самих параметров — греческими: b_0 и b_1 и β_0 и β_1 соответственно. Правда, довольно часто встречаются и такие обозна-

[13] Иногда предпочитают «менее субъективные» выражения. Например, говорят: «Это модель, которой мы задаемся» или «которую мы постулируем». По существу, однако, различие здесь чисто терминологическое.— *Примеч. пер.*

[14] Здесь неявно предполагается, что в качестве модели выступает алгебраический полином.— *Примеч. пер.*

[15] Следовательно, совокупность предсказанных значений образует линию (поверхность) регрессии, показанную, например, на рис. 1.4. Наряду с данным термином применяется и термин «расчетное значение».— *Примеч. пер.*

чения для оценок: $\hat{\beta}_0$ и $\hat{\beta}_1$. Да мы и сами воспользуемся ими в гл. 10.

Нашей процедурой оценивания будет *метод наименьших квадратов.* Возник спор, насчет того, кто же первый предложил этот метод. По-видимому, он был разработан независимо Карлом Фридрихом Гауссом (1777—1855) и Адриеном Мари Лежандром (1752—1833), ибо Гаусс начал им пользоваться до 1803 г. (он настаивал на дате около 1795 г., но доказательств для этой более ранней даты нет), а Лежандр опубликовал первое сообщение в 1805 г. Когда Гаусс в 1809 г. написал, что он пользовался методом наименьших квадратов раньше, чем были опубликованы результаты Лежандра, началась ссора из-за приоритета. Эти данные тщательно изучены и обсуждены в работе Плэкетта из цикла «Исследования по истории теории вероятностей и статистики» (см.: Plackett R. L. Studies in the history of probability and statistics. XXIX. The discovery of the method of least squares.— Biometrika, 1972, 59, p. 239—251), которую мы настоятельно рекомендуем читателю. Еще рекомендуем публикации: Eisenhart C. The meaning of «least» in least squares.— Journal of the Washington Academy of Sciences. 1964, 54, p. 24—33 (перепечатано в Precision Measurement and Calibration, ed. H. H. Ku. National Bureau of Standards Special Publication 300, 1969, 1) и статью «Карл Фридрих Гаусс» из Международной энциклопедии социальных наук (Gauss, Carl Friedrich. International Encyclopedia of the Social Sciences.— New York: Macmillan Co., Free Press Div., 1968, 6, p. 74—81), а также связанную с этой проблемой работу: Stigler S. M. Gergonnes's 1815 paper on the design and analysis of polynomial regression experiments. Historia Mathematica, 1974, 1, p. 431—447 (см. с. 433) [16].

---

[16] Спор между Лежандром и Гауссом носил драматический характер. Вот две краткие цитаты из писем. В 1809 г. как только вышла в свет работа Гаусса, Лежандр отправил ему большое эмоциональное письмо (от 30 мая), где он, в частности, написал: «Не существует открытия, которое нельзя было бы приписать себе . . . , но если не дать тому доказательство, состоящее в указании места, где оно опубликовано, то это утверждение становится беспредметным и представляет собой только обиду для истинного автора открытия». Отвечая Лапласу, который взял на себя роль посредника, Гаусс писал (30 января 1812 г.): «Мне не пришло в голову, что г. Лежандр может придавать такое значение идее столь простой, что нужно скорее удивляться тому, что ее не было сто лет назад, чем сердиться по поводу того, что я пользовался ею раньше него». (Цит. по статье: Гнеденко Б. В. О работах Гаусса по теории вероятностей.— В кн.: Карл Фридрих Гаусс/Под ред. И. М. Виноградова.— М.: Изд-во АН СССР, 1956, с. 217—240.) О жизни Гаусса еще см., например: Белл Э. Т. Творцы математики/Пер. с англ. Под ред. С. Н. Киро.— М.: Просвещение, 1979, с. 178—217; Wussing H. Gauss.— Leipzig: 1974.— 100 S; Dunnington G. W. Carl Fridrich Gauss: Titan of science.— New York: Hafner, 1955. (Есть болгарский перевод, которым мы пользовались за неимением оригинала: Дънингтън Дж. Карл Фридрих Гаус. Титан на науката/Прев. Г. Гаргов.— София: Наука и изкуство. 1983.— 348 с.)

Драма, однако, на этом не кончилась. Позже выяснилось, что у Гаусса был еще один конкурент, опередивший его с публикацией. Это был малоизвестный в Европе американский математик Роберт Адриан Эдрейн (Adrain) (1775—1843). В 1808 г. он опубликовал статью: Adrain R. Research concerning the probabilites of the errors which happen in making observations.— The analyst or mathematical museum.— Philadelphia, 1808, V. 1, N 4, p. 93—109, содер-

При некоторых предположениях, которые обсуждаются в гл. 2, этот метод обладает определенными свойствами. Пусть мы имеем множество из n наблюдений $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. (В нашем примере $n = 25$.) Тогда уравнение (1.2.1) можно записать в виде

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \tag{1.2.3}$$

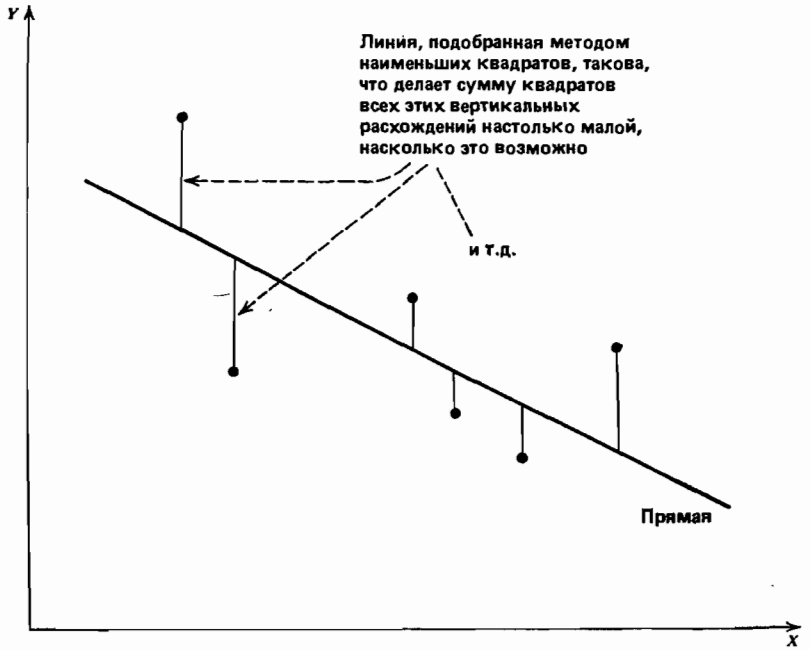


Рис. 1.5. Вертикальные отклонения, минимизирующие сумму квадратов в методе наименьших квадратов

жащую, помимо прочего, и описание метода наименьших квадратов. Трудно было бы найти более неподходящее место для публикации: журнал просуществовал всего 1 год. Анализ творчества этого математика см. в содержательной книге: Майстров Л. Е. Теория вероятностей. Исторический очерк.— М.: Наука, 1967, с. 176—178 и в статье: Шейнин О. Б. О работах Роберта Эдрейна по теории ошибок и ее приложениям.— В кн.: Историко-математические исследования, вып. XVI, 1965, с. 325—336. Конечно, по содержанию Эдрейну трудно состязаться с самим Гауссом, но это вовсе не означает, что его результаты не заслуживают серьезного внимания. Во всяком случае он первым применил в 1818 г. МНК для оценки сжатия земного эллипсоида вращения по результатам градусных измерений и получил исключительно точные для своего времени значения полуосей этого эллипсоида. См.: Математика XIX века. Математическая логика, алгебра, теория чисел, теория вероятностей/Под. ред. А. Н. Колмогорова, А. П. Юшкевича.— М.: Наука, 1978, с. 199.

Наконец, все аспекты этой истории кратко изложены в работе: Harter H. L. Least squares.— In: Encyclopedia of statistical sciences.— New York: Wiley, 1983, 4, p. 593—598.— *Примеч. пер.*

где $i = 1, 2, \ldots, n$. Следовательно, сумма квадратов отклонений от «истинной» линии есть

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2. \tag{1.2.4}$$

Будем подбирать значения оценок b_0 и b_1 так, чтобы их подстановка вместо β_0 и β_1 в уравнение (1.2.4) давала наименьшее возможное (минимальное) значение S, см. рис. 1.5. (Заметим, что X_i, Y_i — это фиксированные числа, которые нам известны.) Мы можем определить b_0 и b_1, дифференцируя уравнение (1.2.4) сначала по β_0, затем по β_1 и приравнивая результаты к нулю. Тогда

$$\begin{aligned} \frac{\partial S}{\partial \beta_0} &= -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i), \\ \frac{\partial S}{\partial \beta_1} &= -2 \sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i), \end{aligned} \tag{1.2.5}$$

так что для оценок b_0 и b_1 имеем

$$\begin{aligned} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) &= 0, \\ \sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i) &= 0, \end{aligned} \tag{1.2.6}$$

где при приравнивании выражений (1.2.5) к нулю мы подставили (b_0, b_1) вместо (β_0, β_1). Из (1.2.6) имеем:

$$\begin{aligned} \sum_{i=1}^{n} Y_i - n b_0 - b_1 \sum_{i=1}^{n} X_i &= 0, \\ \sum_{i=1}^{n} X_i Y_i - b_0 \sum_{i=1}^{n} X_i - b_1 \sum_{i=1}^{n} X_i^2 &= 0 \end{aligned} \tag{1.2.7}$$

или

$$\begin{aligned} b_0 n + b_1 \sum_{i=1}^{n} X_i &= \sum_{i=1}^{n} Y_i, \\ b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n} X_i Y_i. \end{aligned} \tag{1.2.8}$$

Эти уравнения называют нормальными.

Решение уравнений (1.2.8) относительно угла наклона прямой — b_1 дает

$$b_1 = \frac{\Sigma X_i Y_i - [(\Sigma X_i)(\Sigma Y_i)]/n}{\Sigma X_i^2 - (\Sigma X_i)^2/n} = \frac{\Sigma (X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma (X_i - \overline{X})^2}, \tag{1.2.9}$$

где суммирование всегда ведется от $i = 1$ до n, а два выражения для b_1 — это обе правильные, но несколько различные формы одной и

той же величины. Так как по определению

$$\overline{X}=(X_1+X_2+\ \ldots\ +X_n)/n=\Sigma X_i/n,$$
$$\overline{Y}=(Y_1+Y_2+\ \ldots\ +Y_n)/n=\Sigma Y_i/n,$$

имеем:

$$\Sigma\,(X_i-\overline{X})\,(Y_i-\overline{Y})=\Sigma X_iY_i-\overline{X}\Sigma Y_i-\overline{Y}\Sigma X_i+n\overline{X}\,\overline{Y}=$$
$$=\Sigma X_iY_i-n\overline{X}\,\overline{Y}=\Sigma X_iY_i-(\Sigma X_i)(\Sigma Y_i)/n.$$

Отсюда следует эквивалентность числителей в (1.2.9), а заодно, при замене Y на X, эквивалентность знаменателей. Величина ΣX_i^2 называется [17] *нескорректированной суммой квадратов X-ов*, а $(\Sigma X_i)^2/n$ — *коррекцией на среднее значение X-ов*. Разность между ними называется *скорректированной суммой квадратов X-ов*. Аналогично ΣX_iY_i называется *нескорректированной суммой смешанных (парных) произведений*, а $(\Sigma X_i)(\Sigma Y_i)/n$ — *коррекцией на среднее*. Разность между ними называется *скорректированной суммой произведений X и Y*.

Первая форма уравнения (1.2.9) обычно используется для вычисления b_1 на микрокалькуляторе, поскольку с ней гораздо легче работать и нет нужды в громоздких подсчетах для каждого X_i и Y_i выражений $(X_i-\overline{X})$ и $(Y_i-\overline{Y})$ соответственно. Полезно иметь в виду, что для уменьшения ошибок округления лучше всего сохранять в процессе счета столько знаков после запятой, сколько возможно. (Такая стратегия хороша и вообще. Округлять лучше всего на «стадии выдачи результатов», а не на промежуточных этапах.) Многие из цифровых компьютеров дадут более точные ответы, если воспользоваться второй формой уравнения (1.2.9). Это обусловлено машинной системой округления.

Здесь и далее возьмем удобные обозначения и запишем:

$$S_{XY}=\Sigma\,(X_i-\overline{X})\,(Y_i-\overline{Y})=\Sigma\,(X_i-\overline{X})\,Y_i=\Sigma X_i\,(Y_i-\overline{Y})=$$
$$=\Sigma X_iY_i-(\Sigma X_i)(\Sigma Y_i)/n=\Sigma X_iY_i-n\overline{X}\,\overline{Y}.$$

Заметим, что все эти выражения эквивалентны. Аналогично можно записать:

$$S_{XX}=\Sigma\,(X_i-\overline{X})^2=\Sigma\,(X_i-\overline{X})\,X_i=\Sigma X_i^2-(\Sigma X_i)^2/n=$$
$$=\Sigma X_i^2-n\overline{X}^2;$$
$$S_{YY}=\Sigma\,(Y_i-\overline{Y})^2=\Sigma\,(Y_i-\overline{Y})\,Y_i=\Sigma Y_i^2-(\Sigma Y_i)^2/n=\Sigma Y_i^2-n\overline{Y}^2.$$

Вот легко запоминающаяся формула для b_1:

$$b_1=S_{XY}/S_{XX}. \qquad (1.2.9а)$$

Решение уравнения (1.2.8) относительно свободного члена (отрезка на оси ординат при $X=0$) b_0 дает

$$b_0=\overline{Y}-b_1\overline{X}. \qquad (1.2.10)$$

[17] Ниже вводится терминология, восходящая к дисперсионному анализу. Ее употребление иногда очень удобно.— *Примеч. пер.*

С помощью подстановки уравнения (1.2.10) в уравнение (1.2.2) можно получить оцениваемое уравнение регрессии:

$$\hat{Y}_i = \overline{Y} + b_1 (X_i - \overline{X}), \qquad (1.2.11)$$

где b_1 определяется уравнением (1.2.9).

Отметим, что если в (1.2.11) положить $X_i = \overline{X}$, то окажется, что $\hat{Y} = \overline{Y}$. А это означает, что точка $(\overline{X}, \overline{Y})$ лежит на подобранной прямой. Выполним теперь эти вычисления, пользуясь данными табл. 1.1. Мы найдем, что:

$$n = 25$$

$$\Sigma Y_i = 10{,}98 + 11{,}13 + \ldots + 11{,}08 = 235{,}60,$$

$$\overline{Y} = 235{,}60/25 = 9{,}424,$$

$$\Sigma X_i = 35{,}3 + 29{,}7 + \ldots + 28{,}6 = 1315,$$

$$\overline{X} = 1315/25 = 52{,}60,$$

$$\Sigma X_i Y_i = (10{,}98)(35{,}3) + (11{,}13)(29{,}7) + \ldots + (11{,}08)(28{,}6) = 11821{,}4320,$$

$$\Sigma X_i^2 = (35{,}3)^2 + (29{,}7)^2 + \ldots + (28{,}6)^2 = 76323{,}42,$$

$$b_1 = \frac{\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)/n}{\Sigma X_i^2 - (\Sigma X_i)^2/n},$$

$$b_1 = \frac{11821{,}4320 - (1315)(235{,}60)/25}{76323{,}42 - (1315)^2/25} = \frac{-571{,}1280}{7154{,}42},$$

$$b_1 = -0{,}079829.$$

Поэтому подобранное уравнение есть

$$\hat{Y}_i = \overline{Y} + b_1 (X_i - \overline{X}),$$

$$\hat{Y}_i = 9{,}4240 - 0{,}079829\,(X_i - 52{,}60),$$

$$\hat{Y}_i = 13{,}623005 - 0{,}079829 X_i.$$

Построенная линия регрессии нанесена на рис. 1.4. Мы можем составить таблицу предсказанных значений Y_i для каждого из 25 значений X_i, для которого известно наблюденное значение Y_i, и найти *остатки* $Y_i - \hat{Y}_i$, как это сделано в табл. 1.2. Остатков получается столько же, сколько исходных данных.

Отметим, что так как

$$\hat{Y}_i = \overline{Y} + b_1 (X_i - \overline{X}),$$

то

$$Y_i - \hat{Y}_i = (Y_i - \overline{Y}) - b_1 (X_i - \overline{X}),$$

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_i) = \sum_{i=1}^{n} (Y_i - \overline{Y}) - b_1 \sum_{i=1}^{n} (X_i - \overline{X}) = 0.$$

Значит и сумма остатков будет равна нулю. На практике из-за ошибок округления она может оказаться не точно равной нулю.

Таблица 1.2. **Результаты наблюдений, расчетные значения и остатки**

Номер опыта	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	Номер опыта	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
1	10,98	10,81	0,17	13	11,88	11,38	0,50
2	11,13	11,25	—0,12	14	9,57	10,50	—0,93
3	12,51	11,17	1,34	15	10,94	9,89	1,05
4	8,40	8,93	—0,53	16	9,58	9,75	—0,17
5	9,27	8,72	0,55	17	10,09	8,89	1,20
6	8,73	7,93	0,80	18	8,11	8,03	0,08
7	6,36	7,68	—1,32	19	6,83	8,03	—1,20
8	8,50	7,50	1,00	20	8,88	7,68	1,20
9	7,82	7,98	—0,16	21	7,68	7,87	—0,19
10	9,14	9,03	0,11	22	8,47	8,98	—0,51
11	8,24	9,92	—1,68	23	8,86	10,06	—1,20
12	12,19	11,32	0,87	24	10,36	10,96	—0,60
				25	11,08	11,34	—0,26

В любой регрессионной задаче сумма остатков всегда равна нулю, если член β_0 входит в модель. Это следствие первого из нормальных уравнений. Исключение β_0 из модели приводит к тому, что отклик обращается в нуль, когда все предикторы равны нулю. Такое предположение слишком сильно и потому обычно не справедливо. В линейной модели $Y = \beta_0 + \beta_1 X + \varepsilon$ исключение β_0 означает, что линия проходит через точку $X = 0$, $Y = 0$, т. е. что она *отсекает нулевой отрезок* $\beta_0 = 0$ при $X = 0$. Заметим здесь, до более подробного обсуждения в параграфе 5.4, что исключение β_0 из модели всегда возможно с помощью «центрирования» данных, но это совершенно не то же самое, что приравнивание $\beta_0 = 0$. Если, например, мы запишем уравнение (1.2.1) в виде

$$Y - \overline{Y} = (\beta_0 + \beta_1 \overline{X} - \overline{Y}) + \beta_1 (X_i - \overline{X}) + \varepsilon$$

или

$$y = \beta_0' + \beta_1 x + \varepsilon,$$

где $y = Y - \overline{Y}$, $\beta_0' = \beta_0 - \beta_1 \overline{X} - \overline{Y}$, $x = X - \overline{X}$, то оценки для β_0' и β_1 будут такими:

$$b_1 = \frac{\Sigma (x_i - \overline{x})(y_i - \overline{y})}{\Sigma (x_i - \overline{x})^2} = \frac{\Sigma (X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma (X_i - \overline{X})^2}$$

в соответствии с уравнением (1.2.9) и

$$b_0' = \overline{y} - b_1 \overline{x} = 0,$$

так как $\overline{x} = \overline{y} = 0$ при любом значении b_1. Поэтому с полным успехом можно записать центрированную модель, совсем опуская свободный член β_0' (отрезок):

$$Y - \overline{Y} = \beta_1 (X - \overline{X}) + \varepsilon.$$

Мы потеряли один параметр, но это соответствует потере в данных, так как величины $Y_i - \overline{Y}$ ($i = 1, 2, \ldots, n$) представляют собой только (n—1) различных элементов информации (в связи с тем, что их сумма равна нулю), тогда как $Y_1, Y_2, \ldots, Y_n$ содержат n различных элементов информации. «Потерянная» часть информации была эффективно использована для надлежащей корректировки модели, позволяющей исключить свободный член.

1.3. ТОЧНОСТЬ ОЦЕНКИ РЕГРЕССИИ

Теперь мы изучим вопрос о том, какая точность может быть приписана нашей оценке линии регрессии. Рассмотрим следующее тождество:

$$Y_i - \hat{Y}_i = Y_i - \overline{Y} - (\hat{Y}_i - \overline{Y}). \tag{1.3.1}$$

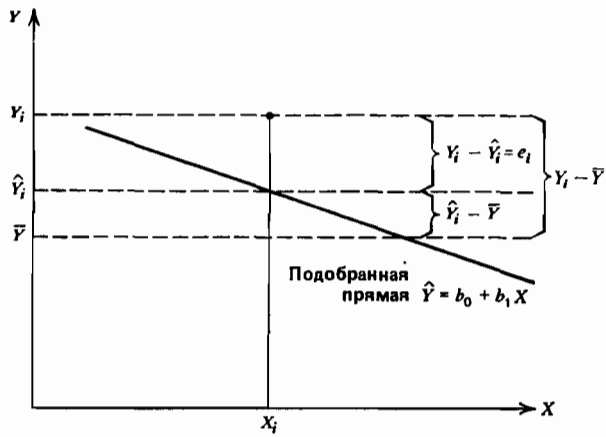


Рис. 1.6. Геометрический смысл тождества (1.3.1)

Что это означает геометрически для подбора прямой, показано на рис. 1.6. Остаток $e_i = Y_i - \hat{Y}_i$ представляет собой разность между двумя величинами: 1) отклонением наблюдаемого значения отклика Y_i от общего среднего откликов $\overline{Y}$ и 2) отклонением предсказанного значения отклика $\hat{Y}_i$ от того же общего среднего $\overline{Y}$. Заметим, что среднее арифметическое предсказанных значений $\hat{Y}_i$ равно

$$\Sigma \hat{Y}_i / n = \Sigma (b_0 + b_1 X_i)/n = (nb_0 + b_1 n\overline{X})/n = b_0 + b_1 \overline{X} = \overline{Y}.$$

Иными словами, среднее арифметическое предсказанных значений $\hat{Y}_i$ то же, что и среднее арифметическое наблюдаемых откликов Y_i. Из этого факта еще раз вытекает, что $\Sigma e_i = \Sigma (Y_i - \hat{Y}_i) = n\overline{Y} - n\overline{Y} = 0$, как было установлено ранее.

Уравнение (1.3.1) можно переписать еще и так:

$$(Y_i - \overline{Y}) = (\hat{Y}_i - \overline{Y}) + (Y_i - \hat{Y}_i).$$

Если мы возведем обе части этого выражения в квадрат и просуммируем от $i = 1$ до n, то получим [18]:

$$\Sigma (Y_i - \bar{Y})^2 = \Sigma (\hat{Y}_i - \bar{Y})^2 + \Sigma (Y_i - \hat{Y}_i)^2. \qquad (1.3.2)$$

Воспользовавшись уравнением (1.2.11) с подстрочным индексом i, можно показать, что член, содержащий парное произведение (СРТ)[19], а именно $\text{СРТ} = 2\Sigma (\hat{Y}_i - \bar{Y}) \cdot (Y_i - \hat{Y}_i)$, исчезает, поскольку

$$\hat{Y}_i - \bar{Y} = b_1 (X_i - \bar{X}),$$

$$Y_i - \hat{Y}_i = Y_i - \bar{Y} - b_1 (X_i - \bar{X}).$$

Отсюда следует, что член, содержащий парное произведение, равен:

$$\text{СРТ} = 2\Sigma b_1 (X_i - \bar{X}) \{(Y_i - \bar{Y}) - b_1 (X_i - \bar{X})\} =$$
$$= 2b_1 \{S_{XY} - b_1 S_{XX}\} = 0$$

по уравнению (1.2.9а). Отсюда также ясно, что

$$\Sigma (\hat{Y}_i - \bar{Y})^2 = \Sigma b_1^2 (X_i - \bar{X})^2 = b_1^2 S_{XX} = b_1 S_{XY}. \qquad (1.3.3)$$

Теперь мы можем вернуться к обсуждению уравнения (1.3.2). Величина $(Y_i - \bar{Y})$ — это отклонение i-го наблюдения от общего среднего, следовательно, левая часть уравнения (1.3.2) — это сумма квадратов отклонений относительно среднего наблюдений, сокращенно — SS *относительно среднего*, а также *скорректированная сумма квадратов Y-ов*. Так как $\hat{Y}_i - \bar{Y}$ есть отклонение i-го наблюдения от его предсказанного или вычисленного значения (i-й *остаток*), а $\hat{Y}_i - \bar{Y}_i$ — отклонение предсказанного значения i-го наблюдения от среднего, то мы можем выразить уравнение (1.3.2) словесно следующим образом [20]:

$$\begin{pmatrix}\text{Сумма квадратов}\\ \text{относительно}\\ \text{среднего}\end{pmatrix} = \begin{pmatrix}\text{Сумма квадратов}\\ \text{относительно}\\ \text{регрессии}\end{pmatrix} + \begin{pmatrix}\text{Сумма квадратов,}\\ \text{обусловленная}\\ \text{регрессией}\end{pmatrix}.$$

Отсюда следует, что разброс Y-ов относительно их среднего можно приписать в некоторой степени (поскольку есть член $\Sigma (Y_i - \hat{Y}_i)^2$) тому факту, что не все действительные наблюдения лежат на линии регрессии. А если бы это было не так, то сумма квадратов относительно регрессии была бы равна нулю! Из этих рассуждений ясно, что пригодность линии регрессии для целей предсказания зависит от того,

---

[18] Уравнение (1.3.2) в дисперсионном анализе играет фундаментальную роль. Можно даже сказать, что в нем в зародыше содержится весь дисперсионный анализ.— *Примеч. пер.*

[19] Аббревиатура СРТ соответствует английскому cross-product term, т. е. член, содержащий парное произведение.— *Примеч. пер.*

[20] Название первого слагаемого в правой части надо дополнить указанием на то, что сумма скорректированная. Тогда более точное название: «сумма квадратов, обусловленная регрессией, скорректированная на среднее».— *Примеч. пер.*

какая часть SS относительно среднего приходится на SS, обусловленную регрессией [21], и какая — соответствует SS относительно регрессии. Мы будем удовлетворены, если SS, обусловленная регрессией, будет много больше, чем SS относительно регрессии, или, что то же самое, если отношение R^2 = (SS, обусловленная регрессией)/(SS, относительно среднего) будет не слишком сильно отличаться от единицы.

Всякая сумма квадратов связана с числом, называемым ее *степенями свободы* [22]. Это число показывает, как много независимых элементов информации, получающихся из n независимых чисел Y_1, Y_2, . . . , Y_n, требуется для образования данной суммы квадратов. Например, для SS относительно среднего требуется (n—1) независимый элемент (из чисел $Y_1-\overline{Y}$, $Y_2-\overline{Y}$, . . . , $Y_n-\overline{Y}$ независимы только (n—1), так как сумма всех n чисел при определении среднего приравнивалась к нулю). Мы можем вычислить SS, обусловленную регрессией, используя единственную функцию от Y_1, Y_2, . . . , Y_n, а именно b_1 (так как $\Sigma(\hat{Y}_i-\overline{Y})^2=b_1^2\Sigma(X_i-\overline{X})^2$), и поэтому данная сумма квадратов имеет одну степень свободы. По разности SS относительно регрессии имеет (n—2) степени свободы. Это отражает тот факт, что рассматриваемые остатки получены для модели прямой линии, которая требует оценивания *двух* параметров. Вообще, остаточная сумма квадратов основывается на числе степеней свободы, равном числу наблюдений минус число оцениваемых параметров. Следовательно, в соответствии с уравнением (1.3.2), мы можем разложить степени свободы таким образом:

$$n - 1 = 1 + (n - 2). \qquad (1.3.4)$$

Пользуясь уравнениями (1.3.2) и (1.3.4), мы можем построить таблицу *дисперсионного анализа*, представленную в табл. 1.3. «Средний квадрат» получается при делении каждой суммы квадратов на соответствующее ей число степеней свободы.

Более общая форма таблицы дисперсионного анализа, которая здесь нам не понадобится, но будет полезна позднее (см. параграф 2.2), получается при добавлении в таблицу корректирующего фактора для среднего Y-ов, который по причинам, указанным в параграфе 2.2, называется SS (b_0). Такая таблица имеет вид табл. 1.4. (Обратите внимание, что в шапке используются сокращения.) (Альтернативный способ представления табл. 1.4 состоит в том, чтобы опустить строку, обозначенную «Общий, скорректированный», и не воспользоваться упомянутым выше правилом. А строка «Общий» станет тогда суммой оставшихся трех строк.)

Когда вычисления для табл. 1.3 и 1.4 идут на микрокалькуляторе, остаточная сумма SS редко подсчитывается так, как показано в таб-

---

[21] Которая скорректирована на среднее.— *Примеч. пер.*

[22] В статистике числом степеней свободы некоторой величины часто называют разность между числом различных опытов и числом констант, найденных по этим опытам независимо друг от друга. В тексте далее говорится об одном частном применении этого понятия к суммам квадратов.— *Примеч. пер.*

Таблица 1.3. **Таблица дисперсионного анализа** (ANOVA). **Основное разложение**

Источник вариации	Число степеней свободы	Суммы квадратов SS	Средние квадраты MS
Обусловленный регрессией	1	$\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2$	$MS_{рег}$
Относительно регрессии (остаток)	$n-2$	$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	$s^2=\frac{SS}{(n-2)}$ *
Общий, скорректированный на среднее Y	$n-1$	$\sum_{i=1}^{n}(Y_i-\overline{Y})^2$	

* В некоторых регрессионных программах распечатка организована так, что величина $\Sigma(Y_i-\overline{Y})^2/(n-1)=S_{YY}/(n-1)$ обозначается через s^2. Для нас это было бы верно только в том случае, если бы подбираемая модель имела вид $Y=\beta+\varepsilon$. Тогда регрессионная сумма квадратов, обусловленная коэффициентом b_0, была бы равна $n\overline{Y}^2=(\Sigma Y_i)^2/n$, а величина S_{YY} как раз служила бы остаточной суммой квадратов для соответствующей построенной модели $\hat{Y}=\overline{Y}$ (как это будет видно в общем случае, например, из табл. 1.4).

Таблица 1.4. **Таблица дисперсионного анализа** (ANOVA), **включающая** SS (b_0)

Источник	Число степеней свободы	SS	MS
Обусловленный $b_1 \mid b_0$	1	$SS(b_1 \mid b_0)=\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2$	$MS_{рег}$
Остаток	$n-2$	$\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	s^2
Общий, скорректированный	$n-1$	$\sum_{i=1}^{n}(Y_i-\overline{Y})^2$	
Корректирующий фактор (обусловленный b_0)	1	$SS(b_0)=\left(\sum_{i=1}^{n}Y_i\right)^2/n=n\overline{Y}^2$	
Общий	n	$\sum_{i=1}^{n}Y_i^2$	

лице, а обычно получается делением SS (b_1/b_0) на «общую, скорректированную SS». Сумму квадратов, обусловленную регрессией, можно, как показано ниже, вычислять множеством способов. (Суммирование везде идет по $i = 1, 2, \ldots, n$.)

$$SS(b_1 \mid b_0) = \Sigma (\hat{Y}_i - \overline{Y})^2 = b_1 \{\Sigma (X_i - \overline{X})(Y_i - \overline{Y})\} = b_1 S_{XY} = \tag{1.3.5}$$

$$= \frac{\{\Sigma (X_i - \overline{X})(Y_i - \overline{Y})\}^2}{\Sigma (X_i - \overline{X})^2} = \frac{S_{XY}^2}{S_{XX}} = \tag{1.3.6}$$

$$= \frac{\{\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)/n\}^2}{\Sigma X_i^2 - (\Sigma X_i)^2/n} = \frac{S_{XY}^2}{S_{XX}} = \tag{1.3.7}$$

$$= \frac{\{\Sigma (X_i - \overline{X}) Y_i\}^2}{\Sigma (X_i - \overline{X})^2}. \tag{1.3.8}$$

Мы оставляем читателю возможность самостоятельно убедиться в том, что эти формулы алгебраически эквивалентны тем, что фигурировали ранее на с. 35 и 39. В таком виде уравнение (1.3.5) проще всего использовать на микрокалькуляторе, поскольку оба сомножителя уже получены при подборе уравнения прямой. Правда, округление при вычислении b_1 может послужить причиной неточности, поэтому мы советуем при вычислениях применять формулу (1.3.7), где деление производится в последний момент.

Отметим, что общую скорректированную сумму квадратов можно записать и вычислять следующим образом:

$$S_{YY} = \Sigma (Y_i - \overline{Y})^2 = \Sigma Y_i^2 - (\Sigma Y_i)^2/n = \tag{1.3.9}$$

$$= \Sigma Y_i^2 - n\overline{Y}^2. \tag{1.3.10}$$

Обозначение SS $(b_1 \mid b_0)$ читается так: «сумма квадратов для b_1 с учетом поправки на b_0». Причины такого обозначения объясняются в параграфах 2.2 и 2.7.

Средний квадрат относительно регрессии s^2 дает оценку *дисперсии относительно регрессии, основанную на n—2 степенях свободы* [23]. Мы будем обозначать эту величину $\sigma^2_{Y \cdot X}$. Если уравнение регрессии будет оцениваться из неопределенно большого числа наблюдений, то дисперсия относительно регрессии будет представлять ошибку измерения, с которой любое измеренное значение Y предсказывается для данного значения X по известному уравнению (см. примечание в параграфе 1.4, с. 45).

Теперь мы выполним вычисления для нашего примера, а затем обсудим ряд подходов, с помощью которых можно исследовать уравнение регрессии. Сумма квадратов SS, обусловленная регрессией,

[23] Эту дисперсию часто называют остаточной.— *Примеч. пер.*

с учетом (1.3.7) есть

$$\frac{\{\Sigma X_t Y_i - (\Sigma X_t)(\Sigma Y_i)/n\}^2}{\{\Sigma X_t^2 - (\Sigma X_t)^2/n\}} = (-571{,}1280)^2/7154{,}42 = 45{,}5924.$$

Полная (скорректированная) сумма квадратов есть

$$\Sigma Y_t^2 - (\Sigma Y_t)^2/n = 2284{,}1102 - (235{,}60)^2/25 = 63{,}8158.$$

Наша оценка величины $\sigma^2_{Y.X}$ — это $s^2 = 0{,}7923$. Она основана на 23 степенях свободы. Что такое величина F, будет объяснено позднее.

Таблица 1.5. **Таблица дисперсионного анализа для примера**

Источник	Чнсло степеней свободы	SS	MS	Вычисленное значение F
Регрессия	1	45,5924	45,5924	57,54
Остаток	23	18,2234	$s^2 = 0{,}7923$	
Общий, скорректированный	24	63,8158		

Упрощенная таблица дисперсионного анализа

Упрощенная таблица дисперсионного анализа содержит только столбцы «Источник» и «Число степеней свободы». Во многих случаях, как, например, в параграфе 1.8, где сравнивается несколько возможных расположений опытов (планов экспериментов) еще до их реализации, полезно для выяснения того, какой из них окажется более предпочтительным, сравнить соответствующие упрощенные таблицы дисперсионного анализа.

1.4. ИССЛЕДОВАНИЕ УРАВНЕНИЯ РЕГРЕССИИ

До настоящего момента мы не использовали предположений о распределении вероятностей. Надо было лишь выполнить некоторое число конкретных алгебраических операций. Теперь же введем основные предположения (постулаты) о том, что в модели

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n,$$

1) остаток ε_i есть [24] случайная величина со средним, равным нулю, и дисперсией (неизвестной) σ^2, т. е.

$$E(\varepsilon_i) = 0, \qquad V(\varepsilon_i) = \sigma^2;$$

[24] Для обозначения математического ожидания (среднего) и дисперсии в литературе применяется много способов. Так, наряду с E ( ) распространено M ( ), вместо V ( ) применяют D ( ) и т. д. Ввиду разнобоя в обозначениях мы следуем оригиналу.— *Примеч. пер.*

2) остатки ε_i и ε_j некоррелированы при $i \neq j$, так что $\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$. Поэтому

$$E(Y_i) = \beta_0 + \beta_1 X_i, \qquad V(Y_i) = \sigma^2.$$

Значения Y_i и Y_j некоррелированы при $i \neq j$. Следующее, не столь необходимое предположение — о чем мы будем помнить при использовании:

3) остаток ε_i есть нормально-распределенная случайная величина со средним 0 и дисперсией σ^2 по (1), т. е.

$$\varepsilon_i \sim N(0,\ \sigma^2).$$

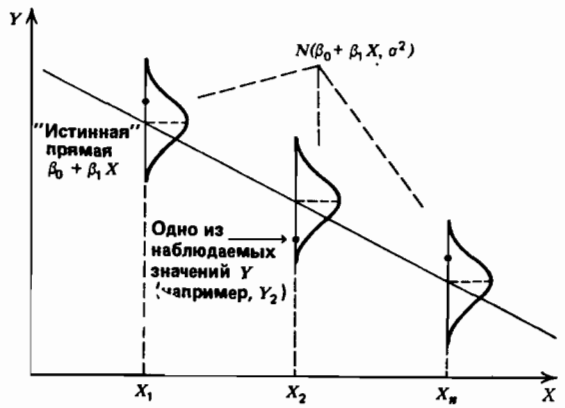


Рис. 1.7. Предполагается, что каждое наблюдение отклнка имеет нормальное распределение относнтельно вертикали со среднм, получаемым нз постулированной модели. Дисперсии же всех нормально распределенных величин предполагаются одннаковыми н равными σ^2

При добавлении этого предположения остатки ε_i и ε_j становятся не только некоррелированными, но и обязательно независимыми.

Эта ситуация показана на рис. 1.7.

(Примечания: 1. Дисперсия σ^2 может быть или не быть равной $\sigma^2_{Y \cdot \vec{X}}$, дисперсии относительно регрессии, которая упоминалась ранее. Если постулированная модель не соответствует «истинной», то $\sigma^2 < \sigma^2_{Y \cdot X}$. Из этого следует, что s^2 — остаточный средний квадрат, который в любом случае оценивает $\sigma^2_{Y \cdot X}$, — служит оценкой σ^2, если только модель корректна. Если $\sigma^2_{Y \cdot X} > \sigma^2$, то мы будем говорить, что постулируемая модель некорректна, или *страдает неадекватностью*. Пути преодоления этой трудности обсуждаются ниже.

2. Во многих реальных ситуациях ошибки, в соответствии с центральной предельной теоремой, подчиняются нормальному распреде-

лению [25]. Если член, содержащий ошибку, таков, что ε оказывается суммой ошибок от нескольких причин, то независимо от того, как могут быть распределены отдельные ошибки, их сумма ε будет распределена с тенденцией по мере увеличения числа слагаемых, в соответствии с центральной предельной теоремой, все больше и больше приближаться к нормальному распределению. Практически экспериментальная ошибка может слагаться из ошибки прибора, ошибки, обусловленной небольшими утечками в системе, ошибки измерения количества используемого катализатора и т. д. Поэтому предположение о нормальности часто правдоподобно. Во всяком случае мы будем позже проверять это предположение при исследовании остатков (см. гл. 3).

Воспользуемся теперь нашими предположениями для исследования уравнения регрессии.

Стандартное отклонение [26] углового коэффициента b_1, доверительный интервал для β_1

Мы знаем, что

$$b_1 = \Sigma (X_i - \bar{X})(Y_i - \bar{Y}) / \Sigma (X_i - \bar{X})^2 = \Sigma (X_i - \bar{X}) Y_i / \Sigma (X_i - \bar{X})^2 =$$

(так как второй член числителя * сокращается

$$\Sigma (X_i - \bar{X}) \bar{Y} = \bar{Y} \Sigma (X_i - \bar{X}) = 0) = \{(X_1 - \bar{X}) Y_1 + \ldots + (X_n - \bar{X}) Y_n\} / \Sigma (X_i - \bar{X})^2.$$

---

[25] Содержание центральной предельной теоремы сводится к тому, что суммарное действие большого числа независимых случайных величин с произвольными законами распределения приводит к случайной величине с нормальным законом. Это и обусловливает особый статус нормального распределения среди всех прочих. Огромную роль в формировании концепции нормального распределения сыграл К. Ф. Гаусс. Иногда нормальный закон даже называют распределением Гаусса. Мысль о том, что реально наблюдаемые случайные величины слагаются из большого числа других случайных величин, и сегодня кажется вполне разумной. Но представление об их *независимом* воздействии кажется гораздо более жестким, чем во времена Гаусса. Это и приводит к постепенной утрате нормальным законом доминирующего положения. Тем не менее основные результаты данной книги относятся к случаю, когда он выполняется. Заметим, что формулировка и развитие центральной предельной теоремы связаны с именами Бернулли и Муавра, Лапласа и Гаусса, Чебышева и Ляпунова. В настоящее время советская школа теории вероятностей вносит важный вклад в исследование этой концепции. См., например: Heyde C. C. Central limit theorem.— In: Encyclopedia of statistical sciences.— New York: Wiley, 1983, 4, p. 651—655.— *Примеч. пер.*

[26] Во втором издании книги термин «стандартная ошибка» (standard error) систематически заменялся на термин «стандартное отклонение» (standard deviation). Причины такой замены не вполне ясны. Возможно, «отклонение» звучит менее обязывающе, чем «ошибка». Мы следуем за авторами и отмечаем те случаи, где они сами не исправили старые термины. — *Примеч. пер.*

* Этот член вообще можно было бы выбросить, но принято писать числитель для b_1 в симметричной форме. См. определение S_{XY} выше в уравнении (1.2.9а).

Далее, дисперсия некоторой функции

$$F = a_1 Y_1 + a_2 Y_2 + \ldots + a_n Y_n$$

равна:

$$V(F) = a_1^2 V(Y_1) + a_2^2 V(Y_2) + \ldots + a_n^2 V(Y_n),$$

если Y_i попарно некоррелированы и a_i — константы. Кроме того, если $V(Y_i) = \sigma^2$, то

$$V(F) = (a_1^2 + a_2^2 + \ldots + a_n^2)\,\sigma^2 = (\Sigma a_i^2)\,\sigma^2.$$

В выражении для b_1 $a_i = (X_i - \overline{X})/\Sigma (X_i - \overline{X})^2$, так как X_i можно рассматривать как константы. Отсюда после преобразований

$$V(b_1) = \frac{\sigma^2}{\Sigma (X_i - \overline{X})^2} = \frac{\sigma^2}{S_{XX}}. \qquad (1.4.1)$$

(Примечание. Интересно следствие этого результата. Допустим, что перед сбором некоторых данных мы хотим подобрать значения X_i, при которых проводятся наблюдения Y_i, причем таким образом, чтобы $V(b_1)$ минимизировалось. Тогда X_i надо выбирать так, чтобы максимизировалось $\Sigma (X_i - \overline{X})^2$. Теоретический ответ на этот вопрос таков, что значения X_i должны будут стремиться к $\pm \infty$. В практической интерпретации это значит, что X_i должны локализоваться на границах области, где реализуется эксперимент. Если, например, мы хотим реализовать четыре опыта, то их следует провести по два на каждой границе. Этот результат имеет смысл и верен, если предварительные надежды на существование модели первого порядка *абсолютно справедливы, причем во всей возможной области значений X-ов*. Если это не так, а в практике такие модели никогда не бывают верными, то результат может оказаться совсем ошибочным. Дж. Боксом и Н. Дрейпером (Journal of the American Statistical Association, 1959, **54**, p. 622—654) было практически показано, что если «область интереса» X-ов есть (с учетом масштаба) интервал $(-R, R)$ и если $\overline{X} = 0$, а мы хотим подобрать прямую, но допускаем, что «истинная» модель может иметь второй порядок, то подходящее значение для $\Sigma (X_i - \overline{X})^2$ — не бесконечность, а число, несколько большее, чем $NR/3$, где N — число факторов, включенных в модель, если только модель близка к правильной или ошибка не очень велика. Общая мораль: выводы, полученные при минимизации дисперсии ошибки, правильны только тогда, когда постулируемая модель корректна, и они могут быть ошибочными во многих практических задачах планирования эксперимента.)

Стандартное отклонение b_1 есть корень квадратный из дисперсии, т. е.

$$\text{ст. откл. } (b_1) = \frac{\sigma}{\{\Sigma (X_i - \overline{X})^2\}^{1/2}}.$$

Если σ неизвестна и мы применяем вместо нее оценку s, предполагая,

что модель корректна, то оценка стандартного отклонения b_1 есть

$$\text{оц. ст. откл.}\,(b_1) = \frac{s}{\{\Sigma\,(X_i - \overline{X})^2\}^{1/2}}. \qquad (1.4.2)$$

В иной терминологии вместо *оцениваемое стандартное отклонение* говорят *стандартная ошибка* или что-нибудь еще в этом роде.

Если мы предполагаем, что разброс наблюдений относительно линии нормален, т. е. что ошибки ε_j все принадлежат некоторому нормальному распределению, $N\,(0,\ \sigma^2)$, то можно показать, что $100\,(1-\alpha)$ %-ные доверительные интервалы для β_1 получаются, если вычислить

$$b_1 \pm \frac{t\left(n-2,\ 1-\frac{1}{2}\alpha\right)s}{\{\Sigma\,(X_i - \overline{X})^2\}^{1/2}}, \qquad (1.4.3)$$

где $t\left(n-2,\ 1-\frac{1}{2}\alpha\right)$ — это $100\left(1-\frac{1}{2}\alpha\right)$%-ная точка t-распределения с $(n-2)$ степенями свободы (они основаны на числе степеней свободы, с которым найдена оценка s^2).

С другой стороны, если это целесообразно, мы можем проверить нуль-гипотезу о том, что β_1 равно β_{10}, где β_{10} — частное значение, которое может быть нулем, против альтернативы, что β_1 отлично от β_{10} (обычно пишут: «$H_0 : \beta_1 = \beta_{10}$ против $H_1 : \beta_1 \neq \beta_{10}$»). Для этого надо вычислить

$$t = \frac{(b_1 - \beta_{10})}{\{\text{оц. ст. откл.}\,(b_1)\}} = \frac{(b_1 - \beta_{10})\,\{\Sigma\,(X_i - \overline{X})^2\}^{1/2}}{s} \qquad (1.4.4)$$

и сравнить $|t|$ с $t\,(n-2,\ 1-\frac{1}{2}\alpha)$ из таблицы t-критерия с $(n-2)$ степенями свободы — числом, на котором основана оценка s^2. В таком виде критерий будет двусторонним со $100\,\alpha$ %-ным уровнем значимости. Продолжим вычисления для нашего примера.

Пример (*продолжение, см. табл. 1.1*)

$$V\,(b_1) = \sigma^2/\Sigma\,(X_i - \overline{X})^2 = \sigma^2/7154{,}42,$$

$$\text{оц. } V\,(b_1) = s^2/7154{,}42 = 0{,}7923/7154{,}42 = 0{,}00011074,$$

$$\text{оц. ст. откл.}\,(b_1) = \sqrt{\text{оц. } V(b_1)} = 0{,}0105.$$

Положим $\alpha = 0{,}05$, так что $t\,(23;\ 0{,}975) = 2{,}069$. Тогда 95 %-ные доверительные границы для β_1 будут $b_1 \pm t\,(23;\ 0{,}975)\cdot s/\{\Sigma\,(X_i - \overline{X})^2\}^{1/2}$, или $-0{,}0798 \pm (2{,}069)\,(0{,}0105)$, что дает интервал, $-0{,}1015 \leqslant \beta_1 \leqslant -0{,}0581$. Словом, истинное значение β_1 лежит в интервале (от $-0{,}1015$ до $-0{,}0581$) и это установлено с 95 %-ной доверительной вероятностью.

Теперь можно проверить нуль-гипотезу о том, что «истинное» значение β_1 — нуль или, иными словами, что между температурой воз-

духа и количеством использованного пара нет линейной зависимости. Запишем (используя $\beta_{10} = 0$):

$$H_0: \beta_1 = 0, \qquad H_1: \beta_1 \neq 0,$$

и далее

$$t = b_1/\text{оц. ст. откл.}\,(b_1) = -0{,}0798/0{,}0105 = -7{,}60.$$

Так как $|t| = 7{,}60$ превосходит соответствующее критическое значение t (23; 0,975) = 2,069, то $H_0: \beta_1 = 0$ отвергается. (Фактически 7,60 превышает даже t (23: 0,9995). Мы выбрали здесь двусторонний 95 %-ный критерий именно так, чтобы доверительный интервал и t-критерий имели один и тот же уровень вероятности. В этом случае можно фактически проверить гипотезу, просто выясняя, как описано выше, включает ли доверительный интервал нуль.) В нашем случае надо отбросить мысль о том, что между Y и X не может быть линейной связи.

Если бы оказалось, что наблюденное значение $|t|$ стало меньше критического значения, мы *не смогли бы отвергнуть* гипотезу. Заметим, что мы избежали слова «принять», так как обычно нельзя принять гипотезу. Более того, мы можем только сказать, что на основе определенных данных наблюдений ее не удается отвергнуть. Однако может случиться так, что, располагая другим массивом данных, мы обнаружим факты, противоречащие нашей гипотезе и тем самым отвергающие ее.

Если, например, мы видим человека, который плохо одет, то мы можем выдвинуть гипотезу, H_0: «Этот человек беден». Если этот человек ходит пешком, чтобы сэкономить на автобусе, или не завтракает ради экономии, мы не имеем оснований для отбрасывания этой гипотезы. Дальнейшие наблюдения такого рода убедят нас, что H_0 верна, но мы не можем, однако, ее принять, если не знаем всего об этом человеке. Причем даже одного-единственного наблюдения, говорящего против H_0, скажем, что этот человек владеет банковским счетом на сумму 500 тыс. дол., будет достаточно, чтобы ее отвергнуть.

После того как мы получили доверительный интервал для β_1, нет необходимости находить величину $|t|$ для проверки гипотезы с помощью t-критерия. Достаточно исследовать доверительный интервал для β_1 и посмотреть, содержит ли он значение β_{10}. Если это так, то гипотезу $\beta_1 = \beta_{10}$ нельзя отвергнуть, а если *не* так, то она отвергается. Это можно увидеть из уравнений (1.4.4), $H_0: \beta_1 = \beta_{10}$ отвергается при α-уровне, если $|t| > t\,(n-2,\ 1-1/2\alpha)$, откуда следует, что

$$|b_1 - \beta_{10}| > t\left(n-2,\ 1-\frac{1}{2}\alpha\right)\cdot s/\{\Sigma\,(X_i-\bar{X})^2\}^{1/2},$$

т. е., что β_{10} лежит за пределами, соответствующими уравнению (1.4.3).

Стандартное отклонение свободного члена; доверительный интервал для β_0

Доверительный интервал для β_0 и проверку гипотезы о том, что β_0 равно или не равно некоторому заданному числу, удается построить, в общем, аналогично тому, как было описано выше для β_1. Мы можем

показать (детали в параграфе 2.3), что

$$\text{ст. откл. } (b_0) = \left\{\frac{\Sigma X_i^2}{n\Sigma (X_i - \overline{X})^2}\right\}^{1/2} \sigma. \qquad (1.4.5)$$

Замена σ на s дает оц. ст. откл. (b_0). Отсюда получаем 100 (1—α) %-ные доверительные пределы для β_0

$$b_0 \pm t\left(n-2,\ 1-\frac{1}{2}\alpha\right)\left\{\frac{\Sigma X_i^2}{n\Sigma (X_i - \overline{X})^2}\right\}^{1/2} s. \qquad (1.4.6)$$

Критерий t для нуль-гипотезы $H_0 : \beta_0 = \beta_{00}$ против альтернативы $H_1 : \beta_0 \neq \beta_{00}$, где β_{00} — заданное значение, будет отвергать ее с 100 α %-ным уровнем значимости, если β_{00} попадет за доверительные границы, или не будет ее отвергать, если β_{00} попадет внутрь интервала.

Проверку гипотезы H_0 можно выполнить и иначе, находя величину

$$t = (b_0 - \beta_{00}) / \left\{\frac{\Sigma X_i^2}{n\Sigma (X_i - \overline{X})^2}\right\}^{1/2} s \qquad (1.4.7)$$

и сравнивая ее с процентной точкой t (n—2, 1—0,5 α), ибо (n—2) — это число степеней свободы, на котором основана s^2-оценка для σ^2.

Примечание. Возможно также построение *совместной доверительной области* для β_0 и β_1 одновременно, если применить формулу (2.6.15).

Стандартное отклонение $\hat{Y}$

Мы показали, что подобранное уравнение регрессии имеет вид

$$\hat{Y} = \overline{Y} + b_1 (X - \overline{X}),$$

где как $\overline{Y}$, так и b_1 подвержены ошибкам, которые будут влиять на $\hat{Y}$. Далее, если a_i и c_i — константы и

$$a = a_1 Y_1 + a_2 Y_2 + \ldots + a_n Y_n,$$
$$c = c_1 Y_1 + c_2 Y_2 + \ldots + c_n Y_n,$$

то в случае некоррелированности Y_i и Y_j при $i \neq j$ и при условии $V(Y_i) = \sigma^2$ для всех i имеем

$$\text{cov}(a, c) = (a_1 c_1 + a_2 c_2 + \ldots + a_n c_n)\sigma^2. \qquad (1.4.8)$$

Из этого следует, что замена $a = \overline{Y}$ влечет $a_i = 1/n$, а замена $c = b_1$ влечет $c_i = (X_i - \overline{X}) / \Sigma (X_i - X)^2$, так что

$$\text{cov}(\overline{Y}, b_1) = 0,$$

т. е. $\overline{Y}$ и b_1 — некоррелированные случайные величины. Поэтому дисперсия предсказываемого среднего значения Y (или $\hat{Y}_0$ при заданном X_0) в зависимости от X есть

$$V(\hat{Y}_0) = V(\overline{Y}) + (X_0 - \overline{X})^2 V(b_1) = \frac{\sigma^2}{n} + \frac{(X_0 - \overline{X})^2 \sigma^2}{\Sigma (X_i - \overline{X})^2}. \qquad (1.4.9)$$

Отсюда

$$\text{оц. ст. откл.}\,(\hat{Y}_0) = s\left\{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\Sigma\,(X_i - \overline{X})^2}\right\}^{1/2}. \qquad (1.4.10)$$

Следовательно, эта величина достигает минимума, когда $X_0 = \overline{X}$, и возрастает по мере того, как мы «удаляем» X_0 от $\overline{X}$ в любом направлении. Другими словами, чем больше разность между X_0 и средним значением $\overline{X}$, тем больше ошибка, с которой мы будем предсказывать среднее значение Y для данного X_0. Это интуитивно хорошо понятно. Говоря несколько вольно, мы можем ожидать «наилучшее» предсказание в «центре» области наблюдений нашего X и не должны ожидать хорошего предсказания при удалении от «центра». Для значений X за пределами наших опытов, т. е. за областью наблюдений, мы должны ожидать тем худших предсказаний, чем дальше мы уходим от области наблюденных значений.

Пример (*продолжение*)

$$n = 25, \quad \Sigma\,(X_i - \overline{X})^2 = 7154{,}42,$$

$$s^2 = 0{,}7923, \quad \overline{X} = 52{,}60.$$

$$\text{оц. } V\,(\hat{Y}_0) = s^2\left\{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\Sigma\,(X_i - \overline{X})^2}\right\} = 0{,}7923\left\{\frac{1}{25} + \frac{(X_0 - 52{,}60)^2}{7154{,}42}\right\}.$$

Если $X_0 = \overline{X}$, то $\hat{Y}_0 = \overline{Y}$ и

$$\text{оц. } V\,(\hat{Y}_0) = 0{,}7923\left\{\frac{1}{25}\right\} = 0{,}031692,$$

т. е.

$$\text{оц. ст. откл.}\,(\hat{Y}_0) = \sqrt{0{,}031692} = 0{,}1780.$$

А если $X_0 = 28{,}6$, то

$$\text{оц. } V\,(\hat{Y}_0) = 0{,}7923\left\{\frac{1}{25} + \frac{(28{,}60 - 52{,}60)^2}{7154{,}42}\right\} = 0{,}095480.$$

Значит, оц. ст. откл. $(\hat{Y}_0) = \sqrt{0{,}095480} = 0{,}3090$.

Соответственно, когда $X_0 = 76{,}60$, оц. ст. откл. $(\hat{Y}_0)$ равна тоже 0,3090.

95 %-ные доверительные пределы для «истинного» среднего значения Y при данном X_0 определяются выражением $\hat{Y}_0 \pm (2{,}069)$ оц. ст. откл. $(\hat{Y}_0)$. Ситуация иллюстрируется на рис. 1.8; две кривые по обе стороны от линии регрессии определяют 95 %-ные доверительные пределы и показывают, как меняются данные пределы при изменении X_0. Эти кривые — гиперболы.

Пределы можно интерпретировать следующим образом. Предположим, что повторные выборки величин Y_i имеют тот же самый объем и взяты при тех же фиксированных значениях X, которые использовались при построении приведенной выше линии. Тогда из всех

95 %-ных доверительных интервалов, построенных для среднего значения Y и отвечающих данному значению X, скажем, X_0, 95 % будут содержать «истинное» значение среднего Y при X_0. Если сделано только одно предсказание $\hat{Y}_0$, скажем, при $X = X_0$, то вероятность того, что найденный для этой точки ($X = X_0$) интервал будет содержать «истинное» среднее, равна 0,95.

Дисперсия и стандартное отклонение, показанные выше, относятся к предсказываемому *среднему значению* Y при данном X_0. Так

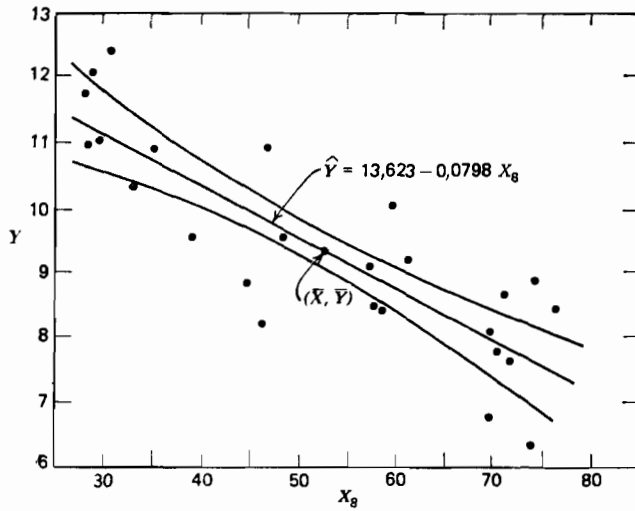


Рис. 1.8. 95 %-ные доверительные интервалы для «истинного» среднего значения Y

как фактические значения Y варьируют около «истинного» среднего значения с дисперсией σ^2 (не зависимой от V (Y)), предсказанное значение индивидуального наблюдения будет по-прежнему определяться величиной $\hat{Y}$, но с дисперсией

$$\sigma^2 \left\{ 1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right\} \qquad (1.4.11)$$

и с соответствующим значением оценки при подстановке s^2 вместо σ^2. Доверительные пределы можно найти уже указанным способом, т. е. мы вычисляем 95 %-ный доверительный интервал для нового наблюдения, который будет симметричен относительно $\hat{Y}_0$ и длина которого будет зависеть от оценки этой новой дисперсии:

$$\hat{Y}_0 \pm t(v,\ 0{,}975) \left\{ 1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right\}^{1/2} s, \qquad (1.4.12)$$

где v — число степеней свободы, на котором основана оценка s^2 (равное здесь n—2). Доверительный интервал для среднего из q новых наблюдений $\hat{Y}_0$ находится аналогично исходя из следующего.

Пусть $\overline{Y}_0$ есть среднее из q новых наблюдений при X_0 (где может быть равно 1, как в предыдущем случае). Тогда

$$\overline{Y}_0 \sim N(\beta_0+\beta_1 X_0,\ \sigma_0^2),$$
$$\hat{Y}_0 \sim N(\beta_0+\beta_1 X_0,\ V(\hat{Y}_0)),$$

так что

$$\overline{Y}_0-\hat{Y}_0 \sim N(0,\ \sigma_0^2+V(\hat{Y}_0))$$

и $[(\overline{Y}_0-\hat{Y}_0)/\text{оц. ст. откл. } (\overline{Y}_0-\hat{Y}_0)]$ распределено как $t(\nu)$, где ν — число степеней свободы, на котором основана s^2, оценка σ^2. Поэтому

$$\text{вер.}\left\{|\overline{Y}_0-\hat{Y}_0| \leqslant t(\nu,\ 0{,}975)\left[s^2\left(\frac{1}{q}+\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\Sigma(X_i-\overline{X})^2}\right)\right]^{1/2}\right\}=0{,}95,$$

так что мы можем построить доверительный интервал для $\overline{Y}_0$ относительно $\hat{Y}_0$:

$$\hat{Y}_0 \pm t(\nu,\ 0{,}975)\left[\frac{1}{q}+\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\Sigma(X_i-\overline{X})^2}\right]^{1/2} s. \qquad (1.4.13)$$

Эти пределы, конечно, шире, чем для среднего значения Y при данном X_0, так как ожидается, что 95 % будущих наблюдений при X_0 (для $q=1$) или будущих средних из q наблюдений (для $q>1$) лежат внутри них.

(Примечание. Для получения совместных доверительных кривых, пригодных для всей регрессионной функции, на всем ее протяжении, надо было бы заменить

$$t(v;\ 1-1/2\alpha) \text{ на } \{2F(2,\ n-2,1-\alpha)\}^{1/2}$$

(см., например: Miller R. G. Simultaneous Statistical Inference.— New York: McGraw-Hill, 1966, p. 110—116 [27]).

Доверительные пределы на практике строят редко. Однако сама идея важна, а подходящий доверительный интервал на основе любого значения $\hat{Y}$ всегда можно найти численно с помощью общей алгебраической формулы при каком угодно числе значений X.)

F-критерий значимости регрессии

Так как Y_i — случайные величины, любая функция от них тоже будет случайной величиной; в частности, две функции: MS_R — средний квадрат, обусловленный регрессией, и s^2 — средний квадрат, обусловленный остаточной вариацией, которые введены в таблице дисперсионного анализа в параграфе 1.3, тоже будут случайными. Эти функции имеют свои собственные распределения, средние, диспер-

[27] Удобную сводку всех относящихся к делу формул (с примером) можно найти в кн.: Закс Л. Статистическое оценивание/Пер. с нем. Под ред. Ю. П. Адлера, В. Г. Горского.— М.: Статистика, 1976, с. 406—408.— *Примеч. пер.*

сии и моменты. Можно показать, что их средние значения будут:

$$E(MS_R) = \sigma^2 + \beta_1^2 \Sigma (X_i - \overline{X})^2, \qquad (1.4.14)$$

$$E(s^2) = \sigma^2,$$

где если Z — случайная величина, то $E(Z)$ обозначает ее среднее или математическое ожидание. Положим, что ошибки ε_i — независимые случайные величины с распределением $N(0, \sigma^2)$. Тогда можно показать, что *если* $\beta_1 = 0$, то величина MS_R, умноженная на свое число степеней свободы (в данном случае на единицу), следует χ^2-распределению с тем же самым числом степеней свободы. Более того, $(n-2)\, s^2/\sigma^2$ тоже следует χ^2-распределению с $(n-2)$ степенями свободы. Так как эти две случайные величины независимы, то из статистической теории вытекает, что отношение

$$F = \frac{MS_R}{s^2} \qquad (1.4.15)$$

подчиняется F-распределению с (здесь) 1 и $(n-2)$ степенями свободы при условии, что $\beta_1 = 0$. Этот факт можно теперь использовать как критерий выполнимости равенства $\beta_1 = 0$. Мы сравним отношение $F = MS_R/s^2$ со 100 $(1-\alpha)$ %-ной табличной точкой $F(1, n-2)$-распределения, чтобы посмотреть, можно ли на основе имеющихся данных рассматривать β_1 как число, отличное от нуля.

Пример (*продолжение*). Из табл. 1.5 мы видели, что искомое отношение $F = 45{,}5924/0{,}7923 = 57{,}54$. Если мы посмотрим процентные точки $F(1; 23)$-распределения, то увидим, что 95 %-ная точка $F(1; 23; 0{,}95) = 4{,}28$. Так как расчетное значение F превышает критическое значение F из таблицы (т. е. $F = 57{,}54 > 4{,}28$), то мы отбрасываем гипотезу $H_0 : \beta_1 = 0$ с риском ошибиться не более чем в 5 % случаев.

(Примечание. В частном случае при построении прямой этот F-критерий для «регрессии» точно такой же, как t-критерий для $\beta_1 = 0$, приведенный выше. По этой причине отношение

$$F = \frac{MS_R}{s^2} = \frac{b_1 \{\Sigma (X_i - \overline{X})(Y_i - \overline{Y})\}}{s^2} = \frac{b_1^2 \Sigma (X_i - \overline{X})^2}{s^2} = \qquad (1.4.16)$$

(по определению b_1)

$$= \left[\frac{b_1 \{\Sigma (X_i - \overline{X})^2\}^{1/2}}{s} \right]^2 = t^2$$

в силу уравнения (1.4.4). Так как величина $F(1, n-2)$ есть квадрат величины $t(n-2)$, то и результат проверки будет тем же. Если рассматривается больше коэффициентов регрессии, то общий F-критерий для регрессии, являющийся обобщением рассматриваемого здесь, не соответствует t-критерию для коэффициента. Однако критерии для индивидуальных коэффициентов можно построить или в форме t, или $t^2 = F$, опираясь на аналогичные рассуждения. F-критерий часто встречается в машинных программах.)

В нашем примере наблюдаемое значение F было 57,54, а $t = -7{,}60$. Заметим, что $(-7{,}6)^2 = 57{,}76$. Это, с учетом ошибки округления, равно значению F.

Объясняемая доля разброса

Мы определили, что

R^2 = (SS, обусловленная регрессией)/(полная SS, скорректированная на среднее $\overline{Y}$)[28] $= \dfrac{\Sigma(\hat{Y}_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2}$, (1.4.17)

где оба суммирования ведутся по i от 1 до n. Тогда R^2 измеряет «долю общего разброса относительно среднего $\overline{Y}$, объясняемую регрессией». Ее часто выражают в процентах, умножая на 100. Фактически R — это корреляция (см. (1.6.5)) между Y и $\hat{Y}$ и его обычно называют *множественным коэффициентом корреляции*.

Пример *(продолжение)*. Из табл. 1.5 имеем:

$$R^2 = \frac{45{,}5924}{63{,}8158} = 0{,}7144.$$

Таким образом, полученное уравнение регрессии, $\hat{Y} = 13{,}623 - 0{,}0798X$, на 71,44 % объясняет общий разброс данных относительно среднего $\overline{Y}$.

Коэффициент R^2 самое большее может достигнуть величины 1 (или 100 %), когда все значения X различны. Ну а если в данных есть повторяющиеся опыты, то, как показано в параграфе 1.5, величина R^2 не может достигнуть 1, как бы хороша ни была модель. Это обусловлено отнюдь не качеством модели, а объясняется вариацией в данных из-за «чистой» ошибки опыта (ошибки воспроизводимости). Алгебраическое доказательство приведенного факта дается в решении упражнения 13 из гл. 1.

1.5. НЕАДЕКВАТНОСТЬ И «ЧИСТАЯ» ОШИБКА

Мы уже отмечали, что построенная линия регрессии — это расчетная линия, основанная на некоторой модели или предположениях. Но предположения мы не можем принимать слепо, а должны *рассматривать их как предварительные*. При некоторых обстоятельствах (условиях) можно проверить, корректна ли наша модель. Прежде всего мы можем изучить проявления предполагаемой некорректности модели. Вспомним, что $e_i = Y_i - \hat{Y}_i$ — *остатки* при $X = X_i$. Это величины, на которые действительные наблюдаемые значения Y_i отличаются от значений $\hat{Y}_i$, вычисленных по уравнению. Как показано в параграфе 1.2, $\Sigma e_i = 0$. Остатки содержат всю мыслимую информа-

[28] Точнее R^2 = (*SS*, обусловленная регрессией, скорректированная на среднее)/(полная *SS*, скорректированная на среднее). Использование R^2 как меры адекватности модели наталкивается на ряд трудностей. Их обсуждение можно найти в кн.: Демиденко Е. З. Линейная и нелинейная регрессия.— М.: Финансы и статистика, 1981, с. 34—41.— *Примеч. пер.*

цию относительно того, почему построенная модель недостаточно правильно объясняет наблюдаемый разброс значений зависимой переменной Y. (Об исследовании остатков см. гл. 3.) Пусть $\eta_i = E(Y_i)$ обозначает величину среднего для «истинной» модели при $X = X_i$. Тогда мы можем записать:

$$Y_i - \hat{Y}_i = (Y_i - \hat{Y}_i) - E(Y_i - \hat{Y}_i) + E(Y_i - \hat{Y}_i) =$$
$$= \{(Y_i - \hat{Y}_i) - (\eta_i - E(\hat{Y}_i))\} + (\eta_i - E(\hat{Y}_i)) = q_i + B_i,$$

где

$$q_i = \{(Y_i - \hat{Y}_i) - (\eta_i - E(\hat{Y}_i))\}, \qquad B_i = \eta_i - E(\hat{Y}_i).$$

Величина B_i — это ошибка смещения при $X = X_i$. Если модель верна, то $E(\hat{Y}_i) = \eta_i$ и $B_i = 0$. Если же модель не верна, то $E(\hat{Y}_i) \neq \eta_i$ и $B_i \neq 0$ и его значение зависит от «истинной» модели и значения X_i. Переменная q_i — это случайная величина, имеющая нулевое среднее, так как

$$E(q_i) = E(Y_i - \hat{Y}_i) - (\eta_i - E(\hat{Y}_i)) = \eta_i - E(\hat{Y}_i) - (\eta_i - E(\hat{Y}_i)) = 0,$$

и это верно независимо от того, будет ли модель правильна (т. е. будет ли $E(\hat{Y}_i)$ равно η_i).

Можно показать, что q_i коррелированны и величина $q_1^2 + q_2^2 + \ldots + q_n^2$ имеет математическое ожидание, или среднее значение $(n-2)\,\sigma^2$, где $V(Y_i = V(\varepsilon_i) = \sigma^2$ — дисперсия ошибки. Исходя из этого можно далее показать, что остаточный средний квадрат, т. е. величина

$$\frac{1}{n-2}\left\{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right\}, \tag{1.5.1}$$

имеет математическое ожидание, или среднее значение σ^2, если постулированная модель корректна, и $\sigma^2 + \Sigma B_i^2/(n-2)$, если модель не корректна. Если модель корректна, т. е. $B_i = 0$, то остатки будут (коррелированными) случайными отклонениями q_i и остаточный средний квадрат можно использовать как оценку дисперсии ошибки σ^2.

Однако если модель не корректна, т. е. $B_i \neq 0$, то остатки содержат оба компонента: случайный (q_i) и систематический (B_i). Мы можем рассматривать их соответственно как случайную ошибку разброса и систематическую ошибку смещения. Таким образом, остаточная сумма квадратов будет иметь тенденцию к разбуханию и перестанет служить удовлетворительной мерой случайных вариаций, имеющихся в наблюдениях. Однако так как средний квадрат *есть* случайная величина, то может оказаться, что он не будет иметь большого значения, даже если смещение существует. С некоторыми аналогичными задачами в общей проблеме регрессии можно познакомиться в параграфе 2.12.

В простом случае подбора прямой обычно можно определить ошибку смещения, непосредственно исследуя график с данными (см.,

например, рис. 1.10). Если модель более сложна и (или) включает больше переменных, то это невозможно. Если существует априорная оценка σ^2 (под «априорной оценкой» мы понимаем оценку, полученную на основе ранее выполненных опытов, в которых варьировались изучаемые условия), то можно увидеть (или проверить по F-критерию), значимо ли остаточная сумма квадратов превышает нашу априорную оценку. Если она значимо больше, то мы говорим, что имеет место неадекватность и следует пересмотреть модель, поскольку в данной форме она непригодна. Если априорной оценки σ^2 нет, но измерения Y повторялись (два раза или более) при одинаковых значениях X, то мы можем использовать эти повторения для получения оценки σ^2. Про такую оценку говорят, что она представляет «чистую» ошибку, потому что если сделать X одинаковыми для двух наблюдений, то только случайные вариации могут влиять на результаты и создавать разброс между ними. Такие различия обычно будут обеспечивать получение оценки σ^2, которая более надежна, чем оценки, получаемые из любых других источников. По этой причине имеет смысл при планировании экспериментов ставить опыты с повторениями.

(Примечание. Важно понимать, что повторение опытов может быть в некотором смысле верным и неверным. Пусть, например, мы будем пытаться применять регрессионный метод к зависимости Y (тест на коэффициент интеллекта (КИ)) от X (рост человека) [29]. Можно получить верные повторные точки, если измерять отдельно КИ у двух людей абсолютно одинакового роста. Если, однако, мы измеряем *дважды* КИ одного человека, то сможем получить вовсе не правильные повторные точки в нашем смысле, а только «переподтвержденную» единственную точку. Она будет содержать информацию о разбросе метода испытаний, являющемся составной частью разброса σ^2, но *не сможет* обеспечить информацию относительно разброса в КИ между людьми с одинаковым ростом, определяющим σ^2 в нашей задаче. В химических экспериментах последовательные наблюдения, выполненные при установившемся состоянии, тоже не дают верных повторных точек. Если же, однако, некоторое множество условий проведения опыта устанавливать заново после промежуточных опытов при других уровнях X и в отсутствии дрейфа уровня отклика [30], то удается получить верные повторные опыты. Имея это в виду, к повторяющимся опытам, обнаруживающим вопреки ожиданиям заметное согласие, следует всегда относиться с осторожностью и подвергать их дополнительному исследованию.)

Когда в данных содержатся повторные опыты, нам нужны дополнительные обозначения для множества наблюдений Y при одном и том же значении X. Пусть мы имеем m различных значений X и

---

[29] Желающих подробнее ознакомиться с КИ мы отсылаем к кн.: Айзенк Г. Ю. Проверьте свои способности. Пер. с англ.— М.: Мир, 1972.— *Примеч. пер.*

[30] Это значит, что все существенные факторы либо подконтрольны, либо остались неизменными.— *Примеч. пер.*

к j-му из этих значений X_i, где $j = 1, 2, \ldots, m$, относятся n_j наблюдений. Тогда мы говорим, что

$Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ — n_1 повторных наблюдений при X_1,

$Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ — n_2 повторных наблюдений при X_2,

Y_{ju} — u-е наблюдение ($u = 1, 2, \ldots, n_j$) при X_j,

$Y_{m1}, Y_{m2}, \ldots, Y_{mn_m}$ — n_m повторных наблюдений при X_m.

Всего получается

$$n = \sum_{j=1}^{m} \sum_{u=1}^{n_j} 1 = \sum_{j=1}^{m} n_j$$

наблюдений. Вклад суммы квадратов, связанной с «чистой» ошибкой для n_1 наблюдений при X_1, будет равен внутренней сумме квадратов Y_{1u} относительно их среднего $\bar{Y}_1$, т. е.:

$$\sum_{u=1}^{n_1} (Y_{1u} - \bar{Y}_1)^2 = \sum_{u=1}^{n_1} Y_{1u}^2 - n_1 \bar{Y}_1^2 = \sum_{u=1}^{n_1} Y_{1u}^2 - \left(\sum_{u=1}^{n_1} Y_{1u} \right)^2 / n_1. \quad (1.5.2)$$

Объединяя внутренние суммы квадратов для всех серий повторных опытов, мы получим общую сумму квадратов «чистых» ошибок в виде

$$\sum_{j=1}^{m} \sum_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)^2 \quad (1.5.3)$$

со степенями свободы

$$n_e = \sum_{j=1}^{m} (n_j - 1) = \sum_{j=1}^{m} n_j - m. \quad (1.5.4)$$

Отсюда средний квадрат «чистых» ошибок равен:

$$s_e^2 = \frac{\Sigma_{j=1}^{m} \Sigma_{u=1}^{n_j} (Y_{ju} - \bar{Y}_j)^2}{\Sigma_{j=1}^{m} n_j - m} \quad (1.5.5)$$

и он служит оценкой σ^2 безотносительно к тому, корректна ли подобранная модель. Словом, эта величина — полная сумма квадратов «между повторениями (параллельными опытами)», деленная на общее число степеней свободы.

(Примечание. Если имеются только два наблюдения Y_{j1} и Y_{j2} в точке X_j, то

$$\sum_{u=1}^{2} (Y_{ju} - \bar{Y}_j)^2 = \frac{1}{2} (Y_{j1} - Y_{j2})^2. \quad (1.5.6)$$

Это удобная форма для вычислений. Такая SS имеет одну степень свободы.)

Таким образом, сумма квадратов «чистых» ошибок фактически оказывается частью остаточной суммы квадратов, что мы теперь и покажем. Остаток для u-го наблюдения при X_j можно записать в виде

$$Y_{ju} - \hat{Y}_j = (Y_{ju} - \bar{Y}_j) - (\hat{Y}_j - \bar{Y}_j), \quad (1.5.7)$$

воспользовавшись тем обстоятельством, что все повторные точки при любом X_j имеют одно и то же предсказанное значение $\hat{Y}_j$. Если мы возведем в квадрат обе части этого выражения, а затем просуммируем их по u и по j, то получим

$$\sum_{j=1}^{m}\sum_{u=1}^{n_j}(Y_{ju}-\hat{Y}_j)^2=\sum_{j=1}^{m}\sum_{u=1}^{n_j}(Y_{ju}-\bar{Y}_j)^2+\sum_{j=1}^{m}n_j(\hat{Y}_j-\bar{Y}_j)^2, \quad (1.5.8)$$

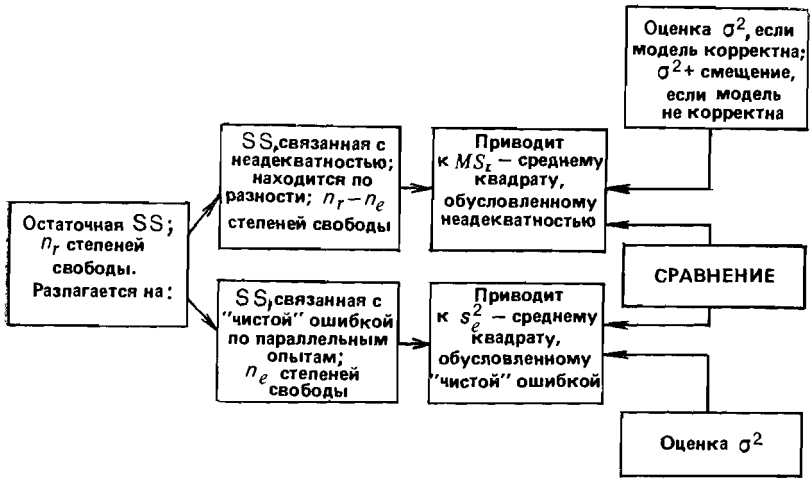


Рис. 1.9. Разложение остаточной суммы квадратов на суммы квадратов, обусловленные неадекватностью и «чистой» ошибкой

причем парные произведения исчезают при суммировании по u для каждого j. Слева в уравнении (1.5.8) стоит остаточная сумма квадратов. Первый член в правой части — это сумма квадратов чистых ошибок. Последний член мы называем суммой квадратов неадекватности. Отсюда следует, что сумму квадратов, обусловленную «чистой» ошибкой, можно ввести в таблицу дисперсионного анализа, как показано на рис. 1.9. Обычный прием — это сравнение отношений $F = MS_L/s_e^2$ со 100 $(1-\alpha)$ %-ной точкой F-распределения при (n_r-n_e) и n_e степенях свободы. Если это отношение является:

1) значимым, то это показывает, что модель, по-видимому, неадекватна. Можно попытаться изучить, когда и как встречается неадекватность. (См. комментарии к различным графикам остатков в гл. 3. Заметим, однако, что графики остатков — стандартная процедура, которая должна применяться в любом регрессионном анализе, а не только в тех случаях, когда неадекватность может быть продемонстрирована с помощью этого критерия.);

2) незначимым, то это показывает, что, по-видимому, нет оснований сомневаться в адекватности модели и что как средний квадрат, связанный с «чистой» ошибкой, так и средний квадрат, обусловленный

неадекватностью, могут использоваться как оценки σ^2. Объединенная оценка σ^2 может быть получена из суммы квадратов, связанной с «чистой» ошибкой, и суммы квадратов, связанной с неадекватностью, путем объединения их в остаточную сумму квадратов и деления ее на остаточное число степеней свободы n_r, что дает $s^2 =$ («остаточная» SS)/n_r. (Обратите внимание, что остатки *все же* должны исследоваться — см. замечания после нижеследующего примера, с. 61.)

Мы уже отмечали выше, что повторные опыты должны быть действительно повторными. Если же это не так, то s^2 будет проявлять склонность к переоценке σ^2, а F-критерий для проверки неадекватности в свою очередь будет иметь тенденцию к ошибочному «определению» отсутствия неадекватности.

Пример. Так как предыдущий пример, который включал данные из приложения А, не содержал параллельных опытов, мы рассмотрим специально построенный пример (табл. 1.6), иллюстрирующий материал этого параграфа о неадекватности и «чистой» ошибке. По следующим данным была оценена линия регрессии $\hat{Y} = 1{,}436 + 0{,}338X$. Таблица дисперсионного анализа представлена табл. 1.7. Заметим, что *на этом этапе* значение F для регрессии не проверяется, поскольку мы еще не знаем, адекватна ли модель.

Таблица 1.6. Двадцать четыре наблюдения с частичными повторами

Номер наблюдения	Y	X	Номер наблюдения	Y	X	Номер наблюдения	Y	X
1	2,3	1,3	9	1,7	3,7	17	3,5	5,3
2	1,8	1,3	10	2,8	4,0	18	2,8	5,3
3	2,8	2,0	11	2,8	4,0	19	2,1	5,3
4	1,5	2,0	12	2,2	4,0	20	3,4	5,7
5	2,2	2,7	13	5,4	4,7	21	3,2	6,0
6	3,8	3,3	14	3,2	4,7	22	3,0	6,0
7	1,8	3,3	15	1,9	4,7	23	3,0	6,3
8	3,7	3,7	16	1,8	5,0	24	5,9	6,7

Таблица 1.7. Таблица дисперсионного анализа для данных из табл. 1.6

Источник	Число степеней свободы	SS	MS	F-отношение
Регрессия	1	6,326	6,326	6,569 значимо при уровне $\alpha = 0{,}05$, если нет неадекватности
Остаток	22	21,192	$0{,}963 = s^2$	
Общий, скорректированный	23	27,518		

1. SS, связанная с «чистой» ошибкой, из повторений при $X = 1{,}3$ есть $1/2\,(2{,}3-1{,}8)^2 = 0{,}125$ с 1 степенью свободы.

2. SS, связанная с «чистой» ошибкой, из повторений при $X = 4{,}7$ есть $(5{,}4)^2 + (3{,}2)^2 + (1{,}9)^2 - 3\,\{(5{,}4 + 3{,}2 + 1{,}9)/3\}^2 = 43{,}01 - (10{,}5)^2/3 = 43{,}01 - 36{,}75 = 6{,}26$ с 2 степенями свободы. Аналогичные вычисления дают следующие величины:

Уровень X	$\sum_{u=1}^{n_j}(Y_{ju}-\overline{Y}_j)^2$	Число степеней свободы	Уровень X	$\sum_{u=1}^{n_j}(Y_{ju}-\overline{Y}_j)^2$	Число степеней свободы
1,3	0,125	1	4,0	0,240	2
2,0	0,845	1	4,7	6,260	2
3,3	2,000	1	5,3	0,980	2
3,7	2,000	1	6,0	0,020	1
			Итого	12,470	11

Теперь можно переписать данные дисперсионного анализа, как показано в табл. 1.8. Отношение $F = MS_L/s_e^2 = 0{,}699$ не значимо, так как оно меньше единицы *. Поэтому на основе такого критерия по крайней мере нет оснований сомневаться в адекватности нашей модели и можно использовать $s^2 = 0{,}963$ как оценку для σ^2, чтобы иметь возможность воспользоваться F-критерием для проверки значимости всей регрессии.

Таблица 1.8. **Дисперсионный анализ (демонстрация неадекватности)**

Источник	Число степеней свободы	SS	MS	F-отношение
Регрессия	1	6,326	6,326	6,569 значимо при $\alpha = 0{,}05$
Остаток	22	21,192	$0{,}963 = s^2$	
Неадекватность	11	8,722	$0{,}793 = MS_L$	0,699 (не значимо)
«Чистая» ошибка	11	12,470	$1{,}134 = s_e^2$	
Общий, скорректированный	23	27,518		

Этот последний F-критерий состоятелен, только если нет неадекватности представления результатов нашей моделью. Чтобы подчеркнуть этот момент, мы подытожим все необходимые действия, когда наши данные содержат повторные наблюдения:

* При взгляде на таблицы нижних процентных точек, приведенные в конце кн. 2, мы должны заметить, что при любых степенях свободы все процентные точки либо больше единицы, либо равны ей. Значит, если вы наблюдаете значение F, которое меньше единицы, то сразу ясно, что оно не может быть значимо.

1. Подобрать модель, составить простую таблицу дисперсионного анализа с двумя входами: регрессией и остатком. Но для общей регрессии пока не использовать F-критерий.

2. Вычислить сумму квадратов, связанную с «чистой» ошибкой и разложить остаточную сумму квадратов, как на рис. 1.9. (Ну а если «чистой» ошибки нет, то остается проверять неадекватность посредством анализа графиков остатков (см. гл. 3).)

3. Применить F-критерий для неадекватности. Если критерий неадекватности не значим, т. е. нет смысла сомневаться в адекватности модели, то перейти к пункту 4б.

4а. Значимая неадекватность. Прекратить анализ подобранной модели и искать пути улучшения модели методами анализа остатков (см. гл. 3). *Не* применять F-критерий для общей регрессии (см. с. 157) и не пытаться строить доверительные интервалы. Если нет адекватности подобранной модели, то не верны предпосылки, которые лежат в основе этих операций.

4б. Неадекватность *не* значима. Снова объединить суммы квадратов для «чистых» ошибок и неадекватности в остаточную сумму квадратов. Использовать остаточный средний квадрат s^2 в качестве оценки для $V(Y) = \sigma^2$, применить F-критерий для общей регрессии, получить доверительные пределы для «истинного» среднего значения Y, вычислить R^2 и т. д. А графики для остатков все-таки надо строить и надо исследовать их особенности (см. гл. 3).

Заметим, что если модель «проходит через все барьеры», это еще не означает, что *она* правильна; просто нет оснований считать ее неадекватной имеющимся данным. Если неадекватность обнаружена, то может понадобиться другая модель, возможно, квадратичная вида $Y = \alpha + \beta X + \gamma X^2 + \varepsilon$. На рис. 1.10 показаны некоторые ситуации, которые могут возникнуть, когда прямая строится по данным шаг за шагом

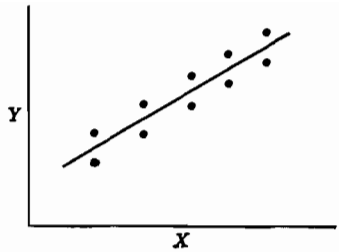
Влияние повторных опытов на R^2

Как мы отмечали в параграфе 1.4, невозможно, чтобы величина R^2 достигла 1, если есть повторные опыты, сколько бы членов ни использовалось в модели. (Тривиальное исключение появляется, когда $s_e^2 = 0$, что случается крайне редко при повторении опытов.) Никакая модель не может изменить вариацию, обусловленную «чистой» ошибкой (см. решение упражнения 13 из гл. 1).

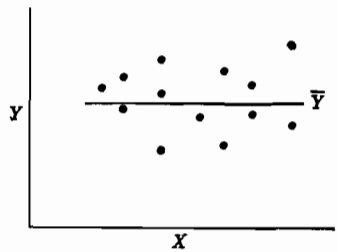
Для демонстрации этого в нашем последнем примере напомним, что сумма квадратов, обусловленная «чистой» ошибкой, равна 12,470 при 11 степенях свободы. То, что модель подогнана к этим данным, не имеет значения, все равно величина 12,470 остается неизменяемой и необъясняемой. Следовательно, максимум R^2, достижимый при этих данных, есть

$$\text{Max } R^2 = \frac{\text{Общая SS, скоррект.} - \text{SS, обусл. «чистой» ошибкой}}{\text{Общая SS, скоррект.}} =$$

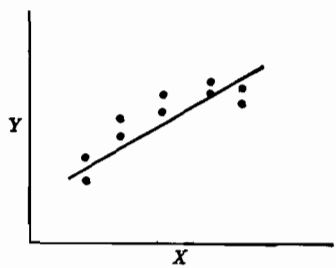
$$= (27{,}518 - 12{,}470)/27{,}518 = 0{,}5468,$$



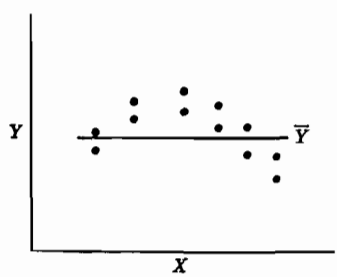
Случай 1:
(1) Проверяется модель $Y=\beta_0+\beta_1 X+\varepsilon$.
(2) Нет неадекватности.
(3) Линейная регрессия значима.
(4) Используется модель $\hat{Y}=b_0+b_1 X$.



Случай 2:
(1) Проверяется модель $Y=\beta_0+\beta_1 X+\varepsilon$.
(2) Нет неадекватности.
(3) Линейная регрессия незначима.
(4) Используется модель $\hat{Y}=\overline{Y}$.



Случай 3:
(1) Проверяется модель $Y=\beta_0+\beta_1 X+\varepsilon$.
(2) Неадекватность значима.
(3) Проверяется модель $Y=\beta_0+\beta_1 X+\beta_{11}X^2+\varepsilon$.



Случай 4:
(1) Проверяется модель $Y=\beta_0+\beta_1 X+\varepsilon$.
(2) Неадекватность значима.
(3) Проверяется модель $Y=\beta_0+\beta_1 X+\beta_{11}X^2+\varepsilon$.

(Примечание. Член β_1 может оказаться значимо отличным от нуля, когда остаточная ошибка определяется в предположении об отсутствии $\beta_{11}X^2$ (см. гл. 6).)

Рис. 1.10. Типичные ситуации при линейной регрессии

или 54,68 %. Однако то значение R^2, что фактически достигнуто для подобранной модели, равно:

$$R^2 = 6{,}326/27{,}518 = 0{,}2299, \quad \text{или} \quad 22{,}99\%.$$

Иными словами, мы можем объяснить 0,2299/0,5468 = 0,4202, или около 42 %, того, что вообще может быть объяснено. Этот результат, хоть он и не слишком впечатляющ, выглядит привлекательнее. Такие расчеты часто позволяют глубже понять, чего модель действительно стоит по сравнению с тем, что она могла бы стоить в лучшем случае.

«Чистая» ошибка в многофакторном случае

Приведенные выше для случая одной переменной формулы применимы и в общем, сколько бы предикторов X_1, X_2, . . . ни оказалось в данных. Единственный момент, который надо иметь в виду, состоит в том, что у повторных опытов должны совпадать все координаты, т. е. они должны иметь одни и те же значения для X_1, совпадающие значения для X_2 и т. д. Например, следующие 4 отклика для 4 точек

$$(X_1, X_2, X_3, X_4) = (4, 2, 17, 1), (4, 2, 17, 1), (4, 2, 17, 1), (4, 2, 17, 1)$$

дают повторные опыты. Однако 4 точки

$$(X_1, X_2, X_3, X_4) = (4, 2, 17, 1), (4, 2, 16, 1), (4, 2, 18, 1), (4, 2, 19, 1)$$

уже не дают повторных опытов, поскольку координаты X_3 во всех этих случаях различны.

Приблизительные повторы

Некоторые наборы данных не имеют или имеют очень мало повторных опытов, зато в них есть *приблизительные повторы*, т. е. множества опытов, которые очень близки друг к другу в пространстве X по сравнению с общим разбросом точек в этом пространстве. В таких случаях мы можем воспользоваться этими псевдоповторами так, как будто они обычные повторы и вычислить по ним приближенную сумму квадратов, связанную с «чистой» ошибкой. Тогда ее можно использовать в анализе стандартным способом. Пример такого использования приведен в упражнении 12 из гл. 1.

1.6. КОРРЕЛЯЦИЯ МЕЖДУ X И Y

Когда мы выдвигали постулат о линейности модели $Y = \beta_0 + \beta_1 X + \varepsilon$, мы предварительно полагали, что Y можно без учета ошибок выразить как функцию первого порядка от X. В такой зависимости X обычно предполагается «фиксированным» (неслучайным), т. е. не имеющим вероятностного распределения, в то время как Y обычно предполагается случайной величиной, имеющей распределе-

ние вероятностей со средним $\beta_0 + \beta_1 X$ и дисперсией $V(\varepsilon)$. Если даже это для X и не совсем так, во многих практических ситуациях можно действовать так, как будто это верно. (Дальнейшее обсуждение см. в параграфе 2.14.)

Теперь, для большей общности, рассмотрим две случайные величины, скажем, U и W, с некоторым непрерывным совместным двумерным распределением вероятностей $f(U, W)$. Тогда мы определим коэффициент корреляции между U и W как

$$\rho_{UW} = \frac{\text{Covar}(U, W)}{\{V(U)\,V(W)\}^{1/2}}, \tag{1.6.1}$$

где

$$\text{Covar}(U, W) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{U - E(U)\}\{W - E(W)\} f(U, W)\, dU dW, \tag{1.6.2}$$

и

$$V(U) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{U - E(U)\}^2 f(U, W)\, dU dW, \tag{1.6.3}$$

а

$$E(U) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} U f(U, W)\, dU dW. \tag{1.6.4}$$

Значения $V(W)$ и $E(W)$ определяются аналогично в терминах W. (Если распределения дискретны, то, как обычно, интегрирование заменяется суммированием.)

Можно показать, что $-1 \leqslant \rho_{UW} \leqslant 1$. Величина ρ_{UW} служит мерой линейной зависимости между случайными величинами U и W. Если, например, $\rho_{UW} = 1$, то U и W идеально положительно коррелированы и все возможные значения U и W лежат на прямой с положительным наклоном в плоскости U—W. Если же $\rho_{UW} = 0$, то говорят, что величины не коррелированы, т. е. не связаны друг с другом линейно. Это *не* означает, что U и W статистически независимы, как можно узнать из любого элементарного учебника. Ну а если $\rho_{UW} = -1$, то U и W идеально отрицательно коррелированны и все возможные значения U и W снова лежат на прямой, на этот раз с отрицательным наклоном в плоскости U—W.

Если имеется выборка объема n из величин $(U_1, W_1), (U_2, W_2), \ldots, (U_n, W_n)$ с совместным распределением, то величина

$$r_{UW} = \frac{\sum_{i=1}^{n}(U_i - \overline{U})(W_i - \overline{W})}{\{\sum_{i=1}^{n}(U_i - \overline{U})^2\}^{1/2}\{\sum_{i=1}^{n}(W_i - \overline{W})^2\}^{1/2}}, \tag{1.6.5}$$

называемая *выборочным коэффициентом корреляции* между U и W, оценивает ρ_{UW} и представляет собой эмпирическую меру *линейной зависимости* между U и W. Причем $n\overline{U} = \Sigma U_i$, $n\overline{W} = \Sigma W_i$. (Если

перед всеми суммами поставить множители $1/(n-1)$, то r_{UW} примет вид ρ_{UW} с дисперсиями и ковариацией, замененными их выборочными оценками.) Подобно ρ_{UW} r_{UW} лежит между -1 и 1.

Если величины U_i и W_i ($i = 1, 2, \ldots, n$) представляют собой скорее постоянные, чем выборочные значения из некоторого распределения, то r_{UW} можно все же использовать как меру линейной зависимости. Поскольку множество значений (U_i, W_i), $i = 1, 2, \ldots, n$, может рассматриваться как полное конечное распределение, r_{UW} будет действительно скорее генеральным, чем выборочным, значением, т. е. в этом случае $r_{UW} = \rho_{UW}$. (Если перед всеми суммами в уравнение (1.6.5) добавить множители $1/n$, то как раз получится уравнение (1.6.1) для дискретного случая.)

Если мы сталкиваемся с ситуацией, где $X_1, X_2, \ldots, X_n$ представляют собой значения из конечного X-распределения, а соответствующие им наблюдения $Y_1, Y_2, \ldots, Y_n$ — фактические значения случайных величин, средние значения которых зависят от соответствующих X-ов (как в этой главе), то коэффициент корреляции ρ_{XY} можно все-таки определить по уравнению (1.6.1) при условии, что все интегралы по X в выражениях вроде уравнений (1.6.2) — (1.6.4) заменяются суммами по дискретным значениям $X_1, X_2, \ldots, X_n$. Выражение (1.6.5), с заменой U и W на X и Y, можно, конечно, применить для оценки ρ_{XY} по r_{XY}, если имеется выборка наблюдений $Y_1, Y_2, \ldots, Y_n$ для n значений X: $X_1, X_2, \ldots, X_n$ соответственно.

В этой книге мы будем пользоваться выражением для r_{UW} из уравнения (1.6.5). Его фактические названия и роли будут зависеть от того, можно ли рассматривать величины U и W как выборочные или как генеральные. Мы будем называть все такие величины r_{UW}, *корреляциями (коэффициентами корреляции) между* U и W, рассматривая их как подходящие меры линейной связи, между различными величинами, представляющими интерес. Указанные выше различия зависят от того, являются ли действительные значения выборочными или же они генеральные. Это, однако, не обязательно учитывать для наших целей, и мы будем игнорировать такие различия.

Если корреляция r_{XY} не равна нулю, это значит, что в нашем множестве данных существует некоторая линейная зависимость между конкретными значениями X_i и Y_i при $i = 1, 2, \ldots, n$. В рассматриваемой регрессионной ситуации мы предполагаем, что значения X_i не подвержены воздействию случайных ошибок (или по крайней мере такое приближение можно считать удовлетворительным, поскольку подобные постулаты редко выполняются строго, что обсуждается в параграфе 2.14), а значения Y_i имеют случайный разброс относительно среднего, зависящего от модели. Позже, когда мы начнем рассматривать больше чем одну предикторную переменную, мы еще будем пользоваться коэффициентом корреляции (например, уравнение (1.6.5) с X_1 и X_2 вместо U и W. Этот коэффициент мы можем тогда назвать r_{12})для измерения линейной зависимости между конкретными значениями (X_{1i}, X_{2i}), встречающимися в наборе данных. Ни в одном из этих случаев у нас нет выборки из некоторого двумерного распределения.

Наконец, последний чрезвычайно важный момент. Значение коэффициента корреляции r_{XY} указывает только на силу линейной зависимости между X и Y. Из него не вытекает *никакого* заключения о типе причинной связи между X и Y. Такое ложное заключение во многих случаях приводит к ошибочным выводам. (Несколько примеров таких выводов, вроде: «Блохи делают человека здоровым», см. в гл. 8 книги Даррелла Хаффа «Как лгать с помощью статистики» (H u f f D. How to lie with statistics.— New York: W. W. Norton, 1954). За пределами Северной Америки эта книга известна благодаря массовому изданию в мягкой обложке в серии «Пеликан» [31].)

Корреляция и регрессия

Допустим, что имеются данные $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. Применяя уравнение (1.6.5), мы можем получить $r_{XY} = r_{YX}$, а если постулировать модель $Y = \beta_0 + \beta_1 X + \varepsilon$, то можно получить и оценку коэффициента регрессии b_1 по уравнению (1.2.9). Раньше мы подчеркивали тот факт, что r_{XY} представляет собой меру линейной зависимости между X и Y. Теперь же перейдем к вопросу о том, как связаны между собой r_{XY} и b_1. Прежде всего отметим, что форма уравнения (1.6.5) не изменится, если перенести начало координат или изменить масштабы для U и W. Сравнивая уравнение (1.6.5) при замене U и W на X и Y с уравнением (1.2.9), мы видим, что

$$b_1 = \left\{ \frac{\Sigma (Y_i - \overline{Y})^2}{\Sigma (X_i - \overline{X})^2} \right\}^{1/2} r_{XY}, \tag{1.6.6}$$

где суммирование ведется по $i = 1, 2, \ldots, n$. Иными словами, b_1 — это «взвешенный» вариант величины r_{XY}, причем «взвешивание» происходит с помощью отношения «разброса» Y_i к «разбросу» X_i. Если мы запишем

$$(n-1)\, s_Y^2 = \Sigma (Y_i - \overline{Y})^2,$$
$$(n-1)\, s_X^2 = \Sigma (X_i - \overline{X})^2,$$

то

$$b_1 = \frac{s_Y}{s_X} r_{XY}. \tag{1.6.7}$$

Таким образом, b_1 и r_{XY} весьма близки, но интерпретируются поразному. Коэффициент r_{XY} измеряет связь между X и Y, в то время как b_1 измеряет величину изменения переменной Y, которую можно предсказать, если изменение переменной X равно единице. В более общих задачах коэффициенты регрессии тоже связаны с корреляциями типа (1.6.5), но более сложным образом (см. параграф 5.4).

Приведем еще два соотношения:

$$r_{XY} = R = (\text{знак } b_1)\,(R^2)^{1/2} \tag{1.6.8}$$

[31] Эта серия научно-популярных и научных книг с эмблемой пеликана выпускается издательством «Пингвин» — одним из крупнейших издательств в Англии (Penguin Books), основанным в 1936 г.— *Примеч. пер.*

только для случая подбора прямой, где R — множественный коэффициент корреляции, квадрат которого равен:

$$R^2 = \frac{\Sigma (\hat{Y}_i - \overline{Y})^2}{\Sigma (Y_i - \overline{Y})^2} \qquad (1.6.9)$$

по определению из параграфа 1.4. Кроме того,

$$r_{Y\hat{Y}} = R, \qquad (1.6.10)$$

т. е. R равно корреляции между имеющимися наблюдениями Y_i и предсказанными значениями $\hat{Y}_i$. Уравнение (1.6.10) справедливо для любой линейной регрессии с любым числом предикторов (тогда как уравнение (1.6.8) верно только для уравнения прямой). Читатель сможет лучше понять эти соотношения, выполнив соответствующие алгебраические упражнения (см. решение упражнения 16 из гл. 1).

Проверка значимости коэффициента парной корреляции

Допустим, что мы нашли коэффициент парной корреляции r_{XY} (он будет далее обозначаться r без подстрочных индексов), который служит оценкой для какого-то истинного (но неизвестного) параметра ρ. Мы можем получить доверительный интервал для ρ или проверить нуль-гипотезу $H_0 : \rho = \rho_0$, где ρ_0 — определенное значение (быть может, и нуль), против любой из альтернативных гипотез H_1: $\rho \neq \rho_0$ или $\rho > \rho_0$, воспользовавшись приближением, известным как z-преобразование Фишера *[32]. Вот это примерное соотношение [33]:

$$z' = \frac{1}{2} \ln\left(\frac{1+r}{1-r}\right) = \tanh^{-1}(r) \sim N\left(\tanh^{-1}\rho,\ \frac{1}{n-3}\right). \qquad (1.6.11)$$

---

* Подробнее о Р. А. Фишере см. в кн.: Fisher-Box Joan. R. A. Fisher: The life of a Scientist.— New York: Wiley, 1978.

[32] Рональд Айлмер Фишер (1890—1962) — выдающийся английский ученый, внесший огромный вклад в развитие современной статистики. Он, в частности, разработал метод максимума правдоподобия и современную концепцию планирования эксперимента. С его именем связано множество понятий математической статистики, в частности тот критерий, который обсуждается ниже в тексте. Биография Фишера на русском языке нам неизвестна. О его роли см., например: Неймам Дж. Текущие задачи математической статистики.— В кн.: Международный математический конгресс в Амстердаме 1954 г. (Обзорные доклады)/Пер. с англ. Под. ред. С. В. Фомина.— М.: Физматгиз, 1961, с. 229—258. Есть русский перевод одной из его книг: Фишер Р. Статистические методы для исследователей. Пер. с англ.— М.: Госстатиздат, 1958.— 268 с.— *Примеч. пер.*

[33] Подробности о преобразовании Фишера и его применениях см., например, в работе: Закс Л. Статистическое оценивание/Пер. с нем.; Под ред. Ю. П. Адлера, В. Г. Горского.— М.: Статистика, 1976, с. 393—398. Обозначение tanh принятое в тексте, соответствует нашему th — гиперболический тангенс.— *Примеч. пер.*

Отсюда приближенный $100(1-\alpha)$ %-ный доверительный интервал для ρ получается из решения уравнения

$$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)\pm z\left(1-\frac{\alpha}{2}\right)\left\{\frac{1}{n-3}\right\}^{1/2}=\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \tag{1.6.12}$$

где $z(1-\alpha/2)$ — верхняя $\alpha/2$ %-ная точка распределения $N(0,1)$ для двух значений ρ, соответствующих двум альтернативам со знаками плюс и минус в первой части уравнения (1.6.12). Статистика, лежащая в основе критерия для проверки гипотезы H_0, такова:

$$z=(n-3)^{1/2}\left\{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)-\frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)\right\}. \tag{1.6 13}$$

Она сравнивается с наперед выбранными процентными точками распределения $N(0,1)$. Три альтернативные гипотезы требуют: $H_1:\rho\neq\rho_0$ — двустороннего критерия, $H_1:\rho>\rho_0$ — одностороннего критерия для верхнего «хвоста» распределения и, наконец, $H_1:\rho<\rho_0$ — одностороннего критерия для нижнего «хвоста» распределения. Поскольку t-распределение с бесконечным числом степеней свободы совпадает с единичным (нормированным) нормальным распределением, поиск процентных точек для этого случая можно вести в последней строке таблицы нормального распределения в конце кн. 2. А для других процентных точек пользуйтесь таблицами t-распределения там же

Пример. Пусть $n=103$, $r=0{,}5$. Выберем $\alpha=0{,}05$. Тогда уравнение (1.6.12) сводится к

$$\frac{1}{2}\ln 3\pm 0{,}196=\frac{1}{2}\ln\{(1+\rho)/(1-\rho)\},$$

и 95 %-ный доверительный интервал для ρ получается от 0,339 до 0,632. Любое значение ρ_0 за пределами этого интервала означало бы, что нулевая гипотеза $H_0:\rho=\rho_0$ отвергается благодаря параллельности арифметических процедур на 5 %-ном уровне двусторонним критерием против альтернативы $H_1:\rho\neq\rho_0$.

Пусть мы хотим проверить $H_0:\rho=0{,}6$ против $H_1:\rho<0{,}6$ на 1 %-ном уровне. Требуемая процентная точка равна $-2{,}326$ из столбца 0,02 (поскольку требуется критерий только для нижнего «хвоста» распределения) таблицы t-распределения с бесконечным числом степеней свободы. Из уравнения (1.6.13) находим статистику, лежащую в основе этого критерия:

$$z=10\left\{\frac{1}{2}\ln 3-\frac{1}{2}\ln(1{,}6/0{,}4)\right\}=-1{,}438.$$

Она оказывается выше, чем процентная точка $-2{,}326$, следовательно, мы не отвергаем H_0 на 1 %-ном уровне значимости.

Хороший обзор этого материала содержится в кн: Pearson E. S., Hartley H. O. Biometrika tables for statisticians. Cambridge University Press, 1958, 1, p. 28—32, 139.

1.7. ОБРАТНАЯ РЕГРЕССИЯ (СЛУЧАЙ ПРЯМОЙ ЛИНИИ)

Допустим, что мы подобрали уравнение прямой $\hat{Y} = b_0 + b_1 X$ по множеству данных (X_i, Y_i), $i = 1, 2, \ldots, n$, а теперь хотим для определенного значения Y, скажем Y_0, получить предсказанное значение $\hat{X}_0$, соответствующее значению X, да еще хотя бы какой-нибудь доверительный интервал, устанавливаемый для X вокруг $\hat{X}_0$. Вот практический пример такой задачи. Величина X_i представляет собой

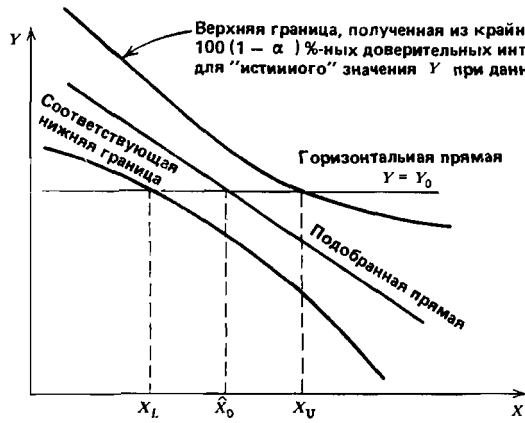


Рис. 1.11. Обратная регрессия: оценивание X по $\hat{X}_0$ для данного значения Y_0 и получения 100(1—α)%-ных «фидуциальных интервалов» для X

оценку возраста дерева, полученную по подсчету годовых колец, тогда как Y_i — соответствующая оценка, полученная методом радиоуглеродной датировки [34]. Подобранная прямая — это «калибровочная кривая» для радиоуглеродной датировки относительно более точных данных по подсчету годовых колец. Теперь применение метода радиоуглеродной датировки к некоторому объекту дает значение Y_0. Какие утверждения мы можем сделать насчет истинного возраста нашего объекта? Эта задача называется задачей обратной регрессии. (В других примерах Y_0 может оказаться истинным средним значением или средним арифметическим q наблюдений.)

Есть несколько альтернативных способов получения (одного и того же) решения задач этого типа. Для начала допустим, что Y_0 — истинное среднее значение, а не единичное наблюдение или среднее арифметическое q наблюдений. Интуитивно кажется, что разумно поступить так. Нарисовать полученную прямую и кривые, соединяющие конечные точки 100 (1—α) %-ных доверительных интервалов

[34] Метод датировки по изотопу углерода ^{14}C был предложен в 1947 г. американским физиком Либби в связи с задачами археологии. Теперь применяется и в других областях, особенно в геологии. Подробности можно найти, например, в кн: Ваганов П. А. Физикн дописывают историю.— Л.: Изд-во ЛГУ, 1984.— 215 с. (особо с. 61—111).— *Примеч. пер.*

для истинного среднего значения Y при данном X (рис. 1.11). На высоте Y_0 провести горизонтальную линию, параллельную оси X. Там, где эта линия пересечет кривые доверительных интервалов, опустить перпендикуляры на ось X, что и даст нижний и верхний $100(1-\alpha)$ %-ные «фидуциальные пределы», обозначенные на рис. 1.11 X_L и X_U соответственно [35]. Перпендикуляр, опущенный на ось X из точки пересечения двух прямых, дает обратную оценку X, определяемую как решение уравнения $Y_0 = \beta_0 + \beta_1 \hat{X}_0$ относительно $\hat{X}_0$, а именно:

$$\hat{X}_0 = (Y_0 - b_0)/b_1,$$

Для получения значений X_L и X_U можно поступить и так. На рисунке X_L — это X-координата точки пересечения прямой

$$Y = Y_0 \quad (\text{т. е. } Y = b_0 + b_1 \hat{X}_0) \tag{1.7.1}$$

и кривой

$$Y = Y_{X_L} - ts\left\{\frac{1}{n} + \frac{(X_L - \overline{X})^2}{S_{xx}}\right\}^{1/2}, \tag{1.7.2}$$

где

$$S_{xx} = \Sigma (X_i - \overline{X})^2, \quad Y_{X_L} = b_0 + b_1 X_L, \quad t = t(\nu, \ 1-\alpha/2)$$

— обычная процентная точка для t-критерия, а ν — число степеней свободы для s^2. Приравнивание уравнений (1.7.1) и (1.7.2), сокращение b_0, перенесение квадратного корня из левой части уравнения в правую, возведение обеих частей в квадрат для избавления от корня приводит к следующему квадратному уравнению относительно X_L:

$$PX_L^2 + 2QX_L + R = 0, \tag{1.7.3}$$

где

$$\begin{aligned} P &= b_1^2 - t^2 s^2 / S_{xx}, \\ Q &= t^2 s^2 \overline{X} / S_{xx} - b_1^2 \hat{X}_0, \\ R &= b_1^2 \hat{X}_0^2 - t^2 s^2 / n - t^2 s^2 \overline{X}^2 / S_{xx}. \end{aligned} \tag{1.7.4}$$

---

[35] Концепция «фидуциальных пределов» — одно из детищ Р. А. Фишера — возникла в 30-е годы. Она представляет собой попытку построения доверительных пределов для искомой оценки, не зависящих от априорного распределения той случайной величины, оценка которой как раз ищется. Фидуциальный подход вызвал дискуссию, вскрывшую и некоторые связанные с ним трудности. В настоящее время употребляется мало в ожидании дальнейших исследований. Точную постановку вопроса и ссылки см. в кн.: Рао С. Р. Линейные статистические методы и их применение/Пер. с англ. Под ред. Ю. В. Линника — М.: Наука, 1968, с. 291—292; Кендалл М., Стьюарт А. Статистические выводы и связи/Пер. с англ. Под ред. А. Н. Колмогорова.— М.: Наука, 1973, гл. 21, с. 183—218; Климов Г. П. Инвариантные выводы в статистике.— М.: Изд-во МГУ, 1973.— 186 с., библ. 87 назв. В тексте термин взят в кавычки, чтобы показать, что он употребляется «фигурально» (см. ниже).— *Примеч. пер.*

Мы получим то же самое уравнение для X_U, так что X_L и X_U оказываются корнями уравнения (1.7.3). Таким образом, после некоторых преобразований, должно получиться:

$$\left.\begin{matrix} X_U \\ X_L \end{matrix}\right\} = \overline{X} + \frac{b_1(Y_0-\overline{Y}) \pm ts\left\{\left[(Y_0-\overline{Y})^2/S_{xx}\right] + \left(b_1^2/n\right) - \left(t^2s^2/nS_{xx}\right)\right\}^{1/2}}{b_1^2 - \left(t^2s^2/S_{xx}\right)} = \tag{1.7.5}$$

или, в ином виде,

$$= \hat{X}_0 + \frac{(\hat{X}_0-\overline{X})\,g \pm (ts/b_1)\left\{\left[(\hat{X}_0-\overline{X})^2/S_{xx}\right] + (1-g)/n\right\}^{1/2}}{1-g}, \tag{1.7.6}$$

где $g = t^2s^2/(b_1^2S_{xx})$. Когда g «мало», скажем 0,05 или еще меньше, допустимо, принять, что оно равно нулю, и получить приближенный ответ. Заметим, что можно записать

$$g = t^2/\{b_1/(s^2/S_{xx})^{1/2}\}^2 =$$

$$= \left\{\frac{t(\nu,\ 1-\alpha/2)\%\text{-ная точка}}{t\text{-статистика, полученная из } b_1\text{, деленного на его ст. ош.}^{36}}\right\}^2. \tag{1.7.7}$$

Таким образом, чем «более значим» коэффициент b_1, тем больше будет знаменатель g и тем меньше будет сам g. Понятно, что g будет возрастать, если модуль b_1 мал, и будет плохо определен, если начнет возрастать s^2 или убывать S_{xx}, или и то и другое одновременно. Обратное оценивание, как правило, не имеет большого практического значения, если регрессия не достаточно хорошо определена, т. е. если b_1 не значим, откуда следует, что величина g должна была бы быть меньше, чем (скажем) приблизительно 0,20. (Значение t-критерия, равное 2,236, например, должно бы это обеспечить.)

Когда же линия регрессии определена недостаточно хорошо, могут возникнуть странности. Так, например, корни X_L и X_U могут оказаться комплексными или же они будут действительными, но оба лягут по одну сторону от линии регрессии. Картинки обычно сразу делают очевидным, почему возникли эти странности, когда линия оценена недостаточно хорошо, гиперболы, определяемые конечными точками доверительных интервалов, обычно плохо выгибаются или резко загибаются вниз (или вверх, как может быть в данном случае). На рис. 1.12 показаны два примера.

Другой способ записи квадратного уравнения (1.7.3), который допускает обобщение на случай многих предикторов, можно упростить. Вот он:

$$\{-Y_0 + b_0 + b_1X\}^2 = t^2s^2\left\{\frac{1}{n} + \frac{(X-\overline{X})^2}{S_{xx}}\right\}, \tag{1.7.8}$$

где X представляет X_L в уравнении (1.7.3) либо X_U.

[36] Здесь авторы сохранили термин «стандартная ошибка». Возможно, это просто недосмотр.— *Примеч. пер.*

(Примечание. Представленные выше вычисления — это вычисления истинного среднего значения. Для получения более общей формулы, в которой Y_0 фигурирует не как истинное среднее значение, а как среднее из q наблюдений, надо везде в уравнениях (1.7.2), (1.7.4), (1.7.5), (1.7.6) и (1.7.8) заменить $1/n$ на $1/q + 1/n$, как описано выше. Тогда $q = 1$ даст формулу для отдельного наблюдения, применимую для единственного нового результата, как в примере с радиоуглеродной датировкой из этого параграфа. А когда положим $q = \infty$, то получим формулы для истинного среднего значения, приведенные выше.

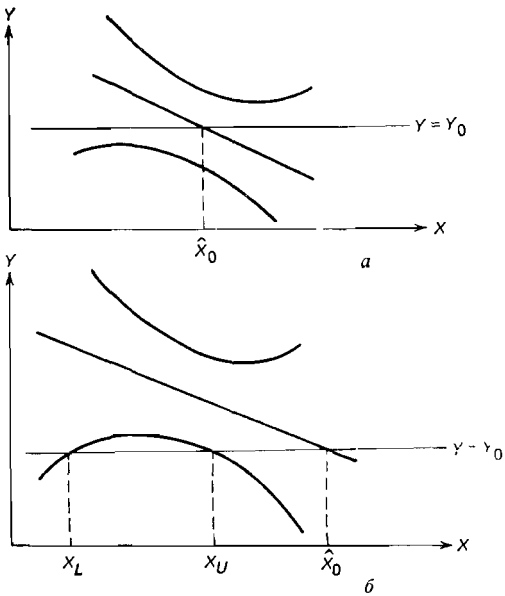


Рис. 1.12. Странности обратной регрессии: (а) комплексные корни; (б) действительные корни, но лежащие по одну сторону от линии регрессии. В обстоятельствах такого рода обратная регрессия не могла бы иметь большого практического значения

Мы заимствовали термин «фидуциальные пределы» для (X_L, X_U) у И. Вильямса, который превосходно изложил эту тему в гл. 6 своей книги «Регрессионный анализ» (Williams E. J. Regression analysis.— New York: Wiley, 1959). Вместо того чтобы ввязываться в теоретическую дискуссию об использовании этого термина, мы просим читателя рассматривать такие интервалы просто как обратные доверительные интервалы для X при данном Y_0.)

1.8. НЕКОТОРЫЕ СЛЕДСТВИЯ ИЗ ГЛ. 1, ИМЕЮЩИЕ ПРАКТИЧЕСКОЕ ЗНАЧЕНИЕ

В этой главе мы занимались подбором прямой для модели $Y = \beta_0 + \beta_1 X + \varepsilon$ по множеству данных (X_i, Y_i), $i = 1, 2, \ldots, n$. Тщательному и всестороннему анализу был подвергнут вопрос о том, как лучше подбирать нашу прямую и могут ли повторные наблюдения или любые особенности данных указывать на то, что следовало бы предпочесть иную модель. Когда рассматривается только один предиктор, а постулируемая модель — прямая, в качестве альтернативных моделей чаще всего выступают полиномы более высокого порядка по X, например квадратичный вида $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$, кубический и т. д. Теперь мы воспользуемся всей этой информацией и рассмотрим задачу выбора стратегии эксперимента для случая одного предиктора с практической точки зрения.

Решения о стратегии эксперимента

Пусть некий экспериментатор хочет собрать данные об отклике Y при n выбранных значениях управляемого предиктора для определения эмпирической зависимости между Y и этим предиктором. Положим, что предиктор не подвержен действию случайной ошибки (по меньшей мере, что отсутствие такого воздействия — удовлетворительное приближение), но что Y подвержен случаю. Еще будем считать, что все n значений предиктора не обязательно различны, т. е. что повторные опыты допускаются. По началу экспериментатор спрашивает и хочет получить ответы на массу вопросов.

1. Какой диапазон значений предиктора представляет для него интерес в настоящий момент? Часто это трудно решить. Диапазон должен быть достаточно широк, чтобы стали возможны полезные выводы, вместе с тем он должен быть достаточно узок, чтобы стало возможным представление результатов наипростейшей моделью. Когда же решение уже принято, интервал может быть кодирован (— 1,1) без нарушения общности. Если, например, для температуры выбран диапазон $140\ ^\circ\mathrm{F} \leqslant T \leqslant 200\ ^\circ\mathrm{F}$, то кодирование

$$X = (T - 170)/30$$

даст интервал $-1 \leqslant X \leqslant 1$. В общем, преобразование имеет вид

$$X = \frac{\text{Натуральная переменная} - \text{Середина натурального интервала}}{\text{Половина натурального диапазона}}.$$

2. Какого рода зависимость, как предчувствует экспериментатор, окажется правильной в выбранном диапазоне? Что это — модель первого порядка (т. е. прямая), второго порядка (т. е. квадратичная) или нечто иное? Для принятия решения ему понадобятся не только все его собственные знания, но скорее всего он станет еще искать способа воспользоваться опытом других. Для определенности давайте положим, что экспериментатор верит в возможность зависимости первого порядка, однако он неабсолютно в этом уверен.

3. А что если зависимость, предварительное решение относительно которой принято выше, в пункте (2), ошибочна? Какую альтернативу экспериментатор считает наиболее перспективной? Если, например, он верит, что истинная модель — это прямая, то он, по-видимому, должен ожидать, что при ее неверности надо рассматривать какую-нибудь криволинейную зависимость квадратичного типа. Менее вероятная возможность заключается в том, что действительная модель окажется кубической. Как правило, он будет на всякий случай решать, что, *может быть*, одного порядка ему слишком мало. Иначе он должен был бы, вероятно, сначала постулировать более высокий порядок модели.

4. Каков разброс, присущий отклику? Иначе говоря, чему равна $V(Y) = \sigma^2$? У экспериментатора может быть богатый опыт работы с аналогичными данными, тогда он может «знать», чему же равна σ^2. Более характерно, что он хочет присоединить к своему эксперименту повторные опыты, чтобы можно было оценить σ^2 одновременно с получением зависимости между Y и X, а заодно и проверить обычное

предположение о постоянстве σ^2 во всем диапазоне значений предиктора.

5. Сколько опытов может понадобиться? Экспериментатор знает только ограничения на средства, персонал, оборудование и время. Сколько опытов достаточно с учетом важности задачи и расходов?

6. Сколько мест (т. е. различных значений X) стоит выбрать? Сколько повторных опытов имеет смысл проводить в каждом месте?

Давайте теперь продолжим наше обсуждение на конкретном примере.

Пример. Допустим, наш экспериментатор решил, что во всем диапазоне $-1 \leqslant X \leqslant 1$ кодированного предиктора наиболее правдоподобна линейная зависимость, что в крайнем случае возможна квадратичная альтернатива, что дисперсии σ^2 он не знает и что возможны 14 опытов. Так при каких же значениях X (т. е. в каких местах) стоит проводить опыты, по скольку в каждом из этих мест и на каком основании?

На рис. 1.13 представлен ряд возможностей, которые он может рассматривать. (Каждой точке соответствует опыт; столбики точек соответствуют повторным опытам.) Давайте взглянем, что требует каждый из этих вариантов.

Каждый план с самого начала имеет 14 степеней свободы. Две из них идут на оценки параметров b_0 и b_1, после чего получается 12 остаточных степеней свободы, которые надо разделить между неадекватностью и «чистой» ошибкой. Строки (1) и (2) в табл. 1.9 показывают, как эти остаточные степени свободы разбиваются в различных планах.

В строке (3) приведены значения

$$\{\Sigma (X_i - \bar{X})^2\}^{-1/2},$$

которые, по уравнению (1.4.1), пропорциональны стандартному от-

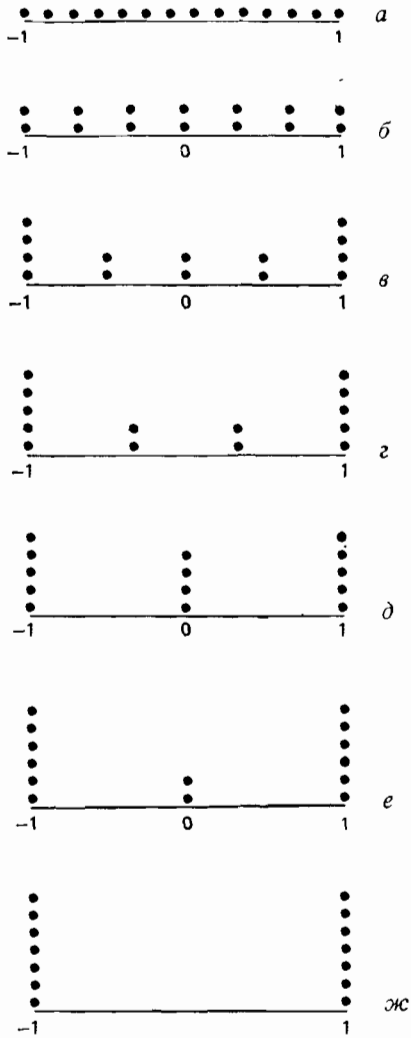


Рис. 1.13. Некоторые возможные расположения опытов для получения данных при подборе прямой: (а) 14 мест; (б) 7 мест; (в) 5 мест; (г) 4 места; (д) 3 места; (е) 3 места; (ж) 2 места. Что лучше, а что хуже при условиях, описанных в тексте? Места расположены равномерно в случаях (а) — (е)

Таблица 1.9. **Характеристики различных стратегий, представленных на рис. 1.13**

	(а)	(б)	(в)	(г)	(д)	(е)	(ж)
(1) Число степеней свободы для неадекватности	12	5	3	2	1	1	0
(2) Число степеней свободы «чистой» ошибки	0	7	9	10	11	11	12
(3) Стандартное отклонение $(b_1)/\sigma$	0,43	0,40	0,33	0,31	0,32	0,29	0,27
(4) Число мест p	14	7	5	4	3	3	2

клонению коэффициента b_1 подобранной прямой. В строке (4) показано число параметров, которое *можно* найти по данным соответствующего плана. По плану с p местами можно подобрать полином порядка p—1 (с p параметрами, считая и β_0). Вторая причина того, что эти данные приведены, заключается в том, что места пропорциональны (когда n и σ^2 фиксированы) величине $p\sigma^2/n$, а она есть среднее арифметическое для фиксированного места $V(\hat{Y}(X))$, усредненное по всем точкам плана, по которому оценивается полином порядка p—1. Иными словами,

$$\sum_{i=1}^{n} V\{\hat{Y}(X_i)\}/n = p\sigma^2/n.$$

Этот результат верен и в общем случае для любой линейной модели. А в случае прямой, когда $p = 2$, его можно вывести из уравнения (1.4.11), заменив подстрочный индекс i на 0 и просуммировав по $i = 1, 2, \ldots, n$. Общее доказательство приводится в упражнении 13 из гл. 2.

Заметим, что число степеней свободы для неадекватности равно числу различных мест для X в данных минус число параметров в постулированной модели. Действительно, так как в нашем примере есть два параметра, подлежащих оценке, β_0 и β_1, то разность между числами, стоящими в строках (4) и (1) табл. 1.9, всюду равна 2.

Комментарии к табл. 1.9.

Поскольку в нашем примере требуется, чтобы σ^2 оценивалась через «чистую» ошибку, стратегия (а) оказывается в данном случае плохой. А поскольку мы не в состоянии проверить адекватность, то и вариант (ж) тоже автоматически исключается.

Возьмем теперь случай (б). Действительно ли существенно использовать семь разных уровней, когда главной альтернативой нам служит квадратичная модель? Конечно нет, поскольку нам вовсе не нужно так много уровней для проверки этой альтернативы. Более того, из оставшихся планов этот имеет наибольшее стандартное от-

клонение $(b_1)/\sigma$. Следовательно, мы исключаем случай (б) из рассмотрения.

Ясно, что наилучший выбор заключается для нас в вариантах (в), (г), (д) или (е). А какой именно среди них выбрать, зависит от предпочтений экспериментатора. Всего трех уровней строго необходимо, чтобы можно было проверить неадекватность при квадратичной альтернативе, но при этом остается лишь одна степень свободы для неадекватности, как в случаях (д) и (е). Причем с точки зрения стандартного отклонения $(b_1)/\sigma$ лучше взять последний из них. План (г) оставляет для неадекватности две степени свободы, а план (в), возможно, идет слишком далеко, имея целых пять уровней. Таким образом, окончательный выбор осуществляется между вариантами (е) и (г), причем вариант (е), по-видимому, чуть более предпочтителен, если квадратичная альтернатива — это все, что можно себе представить.

Быть может, самое важное в этом обсуждении не то, какой конкретный план лучше всего выбрать, а решительное отбрасывание планов, которые в каких-то иных обстоятельствах вполне могли показаться разумными. План (а) был бы очень плохим выбором — кто же требует 14 уровней для оценки уравнения прямой? А план (ж) дает наименьшую дисперсию углового коэффициента b_1, но им нельзя пользоваться во всех случаях, если мы хотим иметь возможность проверять неадекватность против квадратичной (или действительно любой) альтернативы. Когда нужно выбрать план при наличии списка альтернатив, мы советуем проводить подробное рассмотрение данных, аналогичных тем, что представлены в табл. 1.9. Такое представление может быть и полезно, и поучительно.

Упражнения

1. В одной работе исследовалось влияние температуры на выход химического продукта. Были собраны следующие данные (в кодированной форме):

X	Y	X	Y
—5	1	1	9
—4	5	2	13
—3	4	3	14
—2	7	4	13
—1	10	5	18
0	8		

1) Какими будут оценки наименьших квадратов коэффициентов β_0 и β_1 в предположении, что модель имеет вид: $Y = \beta_0 + \beta_1 X + \varepsilon$? Каково уравнение для предсказания?

2) Постройте таблицу дисперсионного анализа и проверьте гипотезу $H_0 : \beta_1 = 0$ с α-риском 0,05.

3) Каковы доверительные пределы ($\alpha = 0{,}05$) для β_1?

4) Каковы доверительные пределы ($\alpha = 0{,}05$) для «истинного» среднего значения величины Y, когда $X = 3$?

5) Каковы доверительные пределы ($\alpha = 0{,}05$) для разности между «истинными» средними значениями величины Y, когда $X_1 = 3$ и когда $X_2 = -2$?

6) Можно ли здесь установить, что испытывается лучшая модель?

7) Прокомментируйте исследованное число уровней температуры в связи с оценкой β_1 в предполагаемой модели.

2. С целью определения эффекта фактора X (такого, как температура) на некоторую характеристику качества выходного продукта Y (такую, как плот-

ность) рассматривается ряд опытов. На каждое из пяти значений X приходится по четыре наблюдения.

1) В каком порядке вы будете проводить двенадцать наблюдений, требуемых для данной задачи?

2) Когда исследование было фактически закончено, получились следующие результаты:

$$\overline{X} = 5{,}0; \quad \Sigma (X_i - \overline{X})^2 = 160{,}0; \quad \Sigma (X_i - \overline{X})(Y_i - \overline{Y}) = 80{,}0;$$

$$\overline{Y} = 3{,}0; \quad \Sigma (Y_i - \overline{Y})^2 = 83{,}2.$$

Пусть модель имеет вид $Y = \beta_0 + \beta_1 X + \varepsilon$.

а) Постройте уравнение регрессии.

б) Составьте таблицу дисперсионного анализа.

в) Определите 95 %-ные доверительные пределы для «истинного» среднего значения величины Y, когда (1) $X = 5{,}0$ и (2) $X = 9{,}0$.

3) Положим, что фактические данные расположены так, как показано на рис. 1.14, и что сумма квадратов, обусловленная повторениями («чистая» ошибка), равна 42. Основываясь на этой дополнительной информации, ответьте на следующие вопросы:

а) Является ли построенное вами в пункте 2а) уравнение регрессии адекватным? Дайте обоснование и приведите результаты проверки на неадекватность.

б) Применимы ли доверительные пределы, вычисленные в пункте 2в)? Если нет, то приведите ваши соображения.

в) Если используемая в пункте 2) модель представляется вам непригодной, предложите возможную альтернативу.

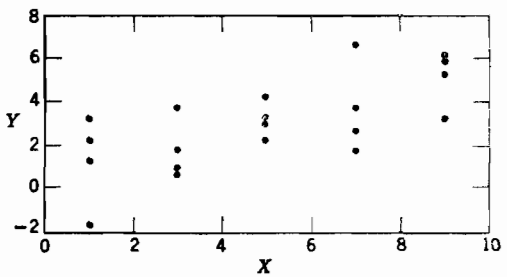


Рис. 1.14.

4) Положим, что фактические данные расположены так, как показано на рис. 1.15, и что сумма квадратов, обусловленная повторениями, равна 42,0. Ответьте на вопросы 3а), 3б) и 3в), пользуясь этой информацией.

5) Положим, что фактические данные расположены так, как показано на рис. 1.16, и что сумма квадратов, обусловленная повторениями, равна 23.2. Ответьте на вопросы 3а), 3б) и 3в), пользуясь этой информацией.

3. Тринадцать образцов медно-никелевых сплавов (Cu/Ni = 90/10), каждый с определенным содержанием железа, были испытаны на коррозию в установке с колесом [37]. Колесо вращалось в соленой морской воде со скоростью 30 футов/с в течение 60 дней. Коррозия определялась по потере в весе в мнлли-

[37] Сплав меди с никелем, в котором концентрация никеля колеблется от 5 до 30 %, иазывается мельхиором. Это сплав, известный своими хорошими механическими свойствами в сочетании с высокой коррозионной стойкостью, находит широкое применение в судостроении, производстве медицинских инструментов, монет и столовой посуды. Сплавы системы медь—никель образуют непрерывный ряд твердых растворов, поэтому трудно рассчитывать на улучшение их коррозионных свойств введением третьего компонента. Тем не менее авторы данного примера считали, что действие добавки железа (Fe) до 2 % может улучшить стойкость этих сплавов. Двухмесячное испытание в установке с колесом, имитирующей активный контакт с морской водой, должно дать сравнительные результаты, по которым и надо сделать заключение о роли железа. Правда, в эксперименте такого объема и с таким немотнвированиым планом не приходится надеяться на однозначиый ответ.— *Примеч. пер.*

граммах на квадратный дециметр в день (МДД). Были собраны следующие данные:

X (Fe)	Y (потери в МДД)	X (Fe)	Y (потери в МДД)
0,01	127,6	1,44	92,3
0,48	124,0	0,71	113,1
0,71	110,8	1,96	83,7
0,95	103,9	0,01	128,0
1,19	101,5	1,44	91,4
0,01	130,1	1,96	86,2
0,48	122,0		

Определите, можно ли оправдать влияние содержания железа на сопротивление коррозии сплавов Cu—Ni 90/10 в морской воде в рамках линейной модели. Примите $\alpha = 0{,}05$.

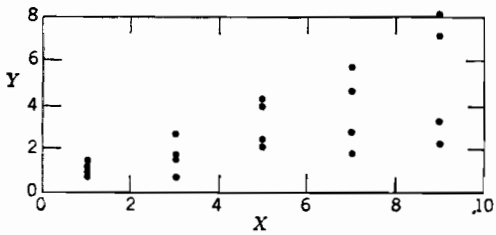


Рис. 1.15.

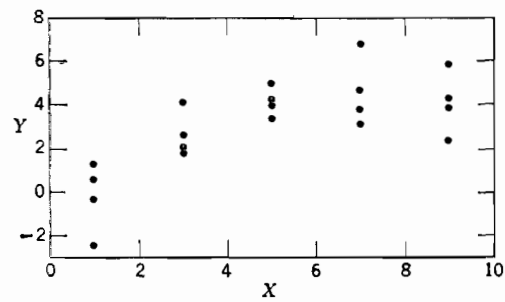


Рис. 1.16.

4. (Источник. Ergun S. Application of the principle of least squares to families of straight lines. Industrial and engineering chemistry, 1956, 48, November, p. 2063—2068.)

1) Покажите, что МНК-оценки $a_1, a_2, \ldots, a_m, b$ параметров $\alpha_1, \alpha_2, \ldots, \alpha_m, \beta$ для семейства прямых линий $E(Y_i) = \alpha_i + \beta Y_i$, $i = 1, 2, \ldots, m$, даются выражениями:

$$b = \frac{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu}-\overline{X}_i)(Y_{iu}-\overline{Y}_i)}{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu}-\overline{X}_i)^2},$$

$$a_i = \overline{Y}_i - b\overline{X}_i,$$

где

$X_{i1}, X_{i2}, \ldots, X_{iu}, \ldots, X_{in_i}$,

$Y_{i1}, Y_{i2}, \ldots, Y_{iu}, \ldots, Y_{in_i}$

обозначают наблюденные значения X_i и Y_i, которые относится к i-й линии, $i = 1, 2, \ldots, m$. Докажите также, что остаточная сумма квадратов — это

$$S^2 = \sum_{i=1}^{m}\sum_{u=1}^{n_i}(Y_{iu}-\overline{Y}_i)^2 - b^2\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu}-\overline{X}_i)^2$$

с $\left(\sum_{i=1}^{m} n_i - m - 1\right)$ степенями свободы, что

$$\sigma_b^2 = \frac{\sigma^2}{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu}-\overline{X}_i)^2}$$

и что

$$\sigma_{a_i}^2 = \frac{\sigma^2}{n_i}\left\{1 + \frac{n_i \bar{X}_i^2}{\sum_{i=1}^{m}\sum_{u=1}^{n_i}(X_{iu} - \bar{X}_i)^2}\right\}.$$

2) Покажите, что МНК-оценки a, b_1, b_2, . . . , b_m параметров α, β_1, β_2, . . . β_m для семейства прямых линий $E(Y_i) = \alpha + \beta_i X_i$, $i = 1, 2, \ldots, m$, даются выражениями:

$$a = \frac{\sum_{i=1}^{m} n_i \left(\bar{Y}_i - \bar{X}_i \left\{\sum_{u=1}^{n_i} X_{iu} Y_{iu} / \sum_{u=1}^{n_i} X_{iu}^2\right\}\right)}{\sum_{i=1}^{m} n_i \left(1 - n_i \bar{X}_i^2 / \sum_{u=1}^{n_i} X_{iu}^2\right)},$$

$$b_i = \frac{\left\{\sum_{u=1}^{n_i} X_{iu} Y_{iu} - a \sum_{u=1}^{n_i} X_{iu}\right\}}{\sum_{u=1}^{n_i} X_{iu}^2},$$

где X_{iu}, Y_{iu} те же, что и выше. Докажите также, что остаточная сумма квадратов — это

$$S^2 = \sum_{i=1}^{m}\sum_{u=1}^{n_i} Y_{iu}^2 - \sum_{i=1}^{m} b_i^2 \sum_{u=1}^{n_i} X_{iu}^2 + a^2 \sum_{i=1}^{m} n_i - 2a \sum_{i=1}^{m} n_i \bar{Y}_i$$

с $\left(\sum_{i=1}^{m} n_i - m - 1\right)$ степенями свободы, что

$$\sigma_a^2 = \sigma^2 / \sum_{i=1}^{m} n_i \left(1 - n_i \bar{X}_i^2 / \sum_{u=1}^{n_i} X_{iu}^2\right)$$

и что

$$\sigma_{b_i}^2 = \left\{\frac{1}{\sum_{u=1}^{n_i} X_{iu}^2} + \frac{\left(\sum_{u=1}^{n_i} X_{iu}\right)^2 \Big/ \left(\sum_{u=1}^{n_i} X_{iu}^2\right)^2}{\sum_{i=1}^{m} n_i \left(1 - n_i \bar{X}_i^2 / \sum_{u=1}^{n_i} X_{iu}^2\right)}\right\} \sigma^2.$$

5. Критерий равенства угловых коэффициентов β_i для m линий, представленных моделями первого порядка

$$Y_{iu} - \bar{Y}_i = \beta_i (X_{iu} - \bar{X}_i) + \varepsilon_{iu}, \qquad i = 1, 2, \ldots, m,$$

можно построить следующим образом. Пусть имеются данные

$$X_{i1}, X_{i2}, \ldots, X_{iu}, \ldots, X_{in_i} \text{ (фиксированные)}$$

и

$$Y_{i1}, Y_{i2}, \ldots, Y_{iu}, \ldots, Y_{in_i} \quad (\varepsilon_{iu} \sim N(0, \sigma^2) \text{ (независимые)}$$

для оценивания параметров i-й линии. МНК-оценка для β_i есть

$$b_i = \left\{ \frac{\sum_{u=1}^{n_i} (X_{iu} - \bar{X}_i)(Y_{iu} - \bar{Y}_i)}{\sum_{u=1}^{n_i} (X_{iu} - \bar{X}_i)^2} \right\}$$

с суммой квадратов (1 степень свободы)

$$\mathrm{SS}(b_i) = b_i^2 \left\{ \sum_{u=1}^{n_i} (X_{iu} - \bar{X}_i)^2 \right\}$$

и остаточной суммой квадратов (n_i—2 степени свободы)

$$S_i = \sum_{u=1}^{n_i} (Y_{iu} - \bar{Y}_i)^2 - \mathrm{SS}(b_i).$$

Если мы положим $\beta_i = \beta$ для всех i, то получим МНК-оценку для β:

$$b = \left\{ \frac{\sum_{i=1}^{m} \sum_{u=1}^{n_i} (X_{iu} - \bar{X}_i)(Y_{iu} - \bar{Y}_i)}{\sum_{i=1}^{m} \sum_{u=1}^{n_i} (X_{iu} - \bar{X}_i)^2} \right\}$$

с суммой квадратов (1 степень свободы)

$$\mathrm{SS}(b) = b^2 \left\{ \sum_{i=1}^{m} \sum_{u=1}^{n_i} (X_{iu} - \bar{X}_i)^2 \right\}$$

и остаточной суммой квадратов (Σn_i—$2m$ степени свободы)

$$S = \sum_{i=1}^{m} \sum_{u=1}^{n_i} (Y_{iu} - \bar{Y}_i)^2 - \mathrm{SS}(b).$$

Можно построить таблицу дисперсионного анализа:

ANOVA

Источник	Число степеней свободы	SS	MS	F
b	1	$\mathrm{SS}(b)$	M_1	$F_1 = M_1/s^2$
Все $b_i \mid b$	$m-1$	$\sum_{i=1}^{m} \mathrm{SS}(b_i) - \mathrm{SS}(b)$	M_2	$F_2 = M_2/s^2$
Остаток	$\sum_{i=1}^{m} n_i - 2m$	по разности	s^2 (оценка σ^2, если модели первого порядка корректны)	
Общий	$\sum_{i=1}^{m} n_i - m$	$\sum_{i=1}^{m} \sum_{u=1}^{n_i} (Y_{iu} - \bar{Y}_i)^2$		

Гипотеза $H_0 : \beta_i = \beta$ проверяется сравнением F_2 с соответствующей процентной точкой распределения $F\{(m-1), (\sum_{i=1}^{m} n_i - 2m)\}$. Если H_0 не отвергается, то b используется как общая оценка угла наклона всех линий. (Это частный случай проверки линейной гипотезы. Можно также построить критерий и для равенства свободных членов двух линий.) Критерий F_1 используется для проверки $H_0 : \beta_0 = 0$.

Примените рассмотренную выше процедуру для следующих данных:

u	X_1	Y_1	X_2	Y_2	X_3	Y_3
1	3,5	24	3,2	22	3,0	32
2	4,1	32	3,9	33	4,0	36
3	4,4	37	4,9	39	5,0	47
4	5,0	40	6,1	44	6,0	49
5	5,5	43	7,0	53	6,5	55
6	6,1	51	8,1	57	7,0	59
7	6,6	62			7,3	64
8					7,4	64

6. Предполагается, что влажность исходной смеси продуктов влияет на плотность конечного продукта. Влажность смеси выбиралась, а плотность конечного продукта измерялась. Были получены следующие данные:

Влажность смеси (кодированная) X	Плотность (кодированная) Y	Влажность смеси (кодированная) X	Плотность (кодированная) Y
4,7	3	5,9	6
5,0	3	5,6	6
5,2	4	5,0	4
5,2	5	$\Sigma X = 63{,}6$	$\Sigma Y = 62$
5,9	10	$\Sigma X^2 = 339{,}18$	$\Sigma Y^2 = 390$
4,7	2	$\Sigma x^2 = 2{,}10$	$\Sigma y^2 = 69{,}67$
5,9	9		
5,2	3	$\overline{X} = 5{,}3$	$\overline{Y} = 5{,}17$
5,3	7	$\Sigma XY = 339{,}1$	
		$\Sigma xy = 10{,}5$	

1) Постройте по данным модель $Y = \beta_0 + \beta_1 X + \varepsilon$.
2) Установите 95 %-ные доверительные интервалы для β_1.
3) Содержатся ли в данных какие-нибудь основания для испытания более сложной модели? (Возьмите $\alpha = 0{,}05$).

7. Стоимость эксплуатации транспортных винтовых самолетов, видимо, растет с возрастом самолета. Собраны следующие данные:

Возраст, лет X	6-месячная стоимость, дол. Y	Возраст, лет X	6-месячная стоимость, дол. Y	Возраст, лет X	6-месячная стоимость, дол. Y
4,5	619	5,0	890	6,0	764
4,5	1049	5,0	1522	6,0	1373
4,5	1033	5,5	987	1,0	978
4,0	495	5,0	1194	1,0	466
4,0	723	0,5	163	1,0	549
4,0	681	0,5	182		

1) Определите, имеет ли смысл линейная зависимость (возьмите $\alpha = 0{,}10$)?
2) Можно ли выбрать лучшую модель?

8. При организации производства некоторого продукта было предложено заменить анализ потерь в бутылках анализом потерь в чашках, который дороже, чем применяемая лабораторная методика. Зато анализ потерь в чашках дал бы возможность лучше управлять процессом, так как он экономит время. Если бы удалось доказать, что потери в чашках — функция потерь в бутылках, то стоило бы принять решение о переходе к анализу потерь в чашках. Имеются следующие данные:

Потери в бутылках, % X	Потери в чашках, % Y	Потери в бутылках, % X	Потери в чашках, % Y	Потери в бутылках, % X	Потери в чашках, % Y
3,0	3,1	2,7	3,6	3,2	4,1
3,1	3,9	3,1	3,1	2,1	2,6
3,0	3,4	2,7	3,6	3,0	3,1
3,6	4,0	2,7	2,9	2,6	2,8
3,8	3,6	3,3	3,6		

Каковы ваши выводы? (Возьмите $\alpha = 0{,}05$).

9. Считается, что число консервных банок, поврежденных при перевозках в товарных вагонах,— это функция скорости вагонов при толчках. Методом случайного отбора были выбраны тринадцать вагонов для проверки того, насколько это соответствует действительности. Собраны следующие данные

Скорость вагона при толчках X	Число поврежденных банок Y	Скорость вагона при толчках X	Число поврежденных банок Y	Скорость вагона при толчках X	Число поврежденных банок Y
4	27	3	109	7	168
3	54	3	28	3	47
5	86	4	75	8	52
8	136	3	53		
4	65	5	33		

Каковы ваши выводы? (Возьмите $\alpha = 0{,}05$).

10. Экспериментально было установлено влияние температуры процесса дезодорации на цвет конечного продукта. Получены следующие данные:

Температура X	Цвет Y	Температура X	Цвет Y	Температура X	Цвет Y
460	0,3	410	0,5	410	0,7
450	0,3	450	0,5	400	0,6
440	0,4	440	0,6	420	0,6
430	0,4	430	0,6	410	0,6
420	0,6	420	0,6	400	0,6

1) Постройте модель $Y = \beta_0 + \beta_1 X + \varepsilon$.
2) Имеет ли эта модель смысл? (Используйте $\alpha = 0{,}05$).

3) Постройте 95 %-ный доверительный интервал для «истинного» среднего значения Y при любом заданном значении X, скажем при X_0.

11. Приведенные ниже данные (представленные Томом Уитекером) содержат 34 пары значений переменных X и Y, где

X — среднее арифметическое значение уровня афлатоксина в единичной минимальной пробе земляного ореха [38], равной 120 фунтам (единиц на миллиард),

Y — процент незараженных орехов в партии-99.

Y	X	Y	X	Y	X
0,971	3,0	0,942	18,8	0,863	46,8
0,979	4,7	0,932	18,9	0,811	46,8
0,982	8,3	0,908	21,7	0,877	58,1
0,971	9,3	0,970	21,9	0,798	62,3
0,957	9,9	0,985	22,8	0,855	70,6
0,961	11,0	0,933	24,2	0,788	71,1
0,956	12,3	0,858	25,8	0,821	71,3
0,972	12,5	0,987	30,6	0,830	83,2
0,889	12,6	0,958	36,2	0,718	83,6
0,961	15,9	0,909	39,8	0,642	99,5
0,982	16,7	0,859	14,3	0,658	111,2
0,975	18,8				

1) Нанесите данные на график (где Y — ордината, а X — абсцисса) и без всяких вычислений «на глазок» проведите прямую, которая вам кажется «самой лучшей прямой, какую только можно провести через даные точки». Сохраните этот рисунок, он вам понадобится позже для сравнения. Считаете ли вы, что ваша прямая — это «хорошее приближение»?

2) Теперь вычислите ΣX_i, ΣY_i, ΣX_i^2, ΣY_i^2 и $\Sigma X_i Y_i$, где все суммы берутся от $i = 1$ до 34.

3) Постройте методом наименьших квадратов модель $Y = \beta_0 + \beta_1 X + \varepsilon$. Нанесите эту прямую на ваш график и проверьте, хороша ли была «подгонка на глазок»?

4) Найдите остатки, сохраняя три знака после запятой. Убедитесь, что $\Sigma e_i = 0$ с точностью до ошибки округления.

5) Постройте таблицу дисперсионного анализа в форме табл. 1.4.

6) Найдите стандартные ошибки [39] коэффициентов b_0 и b_1.

7) Выведите формулу для стандартной ошибки значения $\hat{Y}$ и постройте 95 %-ные доверительные границы для «истинного» среднего значения Y. (На-

[38] Афлатоксин — один из видов токсинов (ядов), продукт метаболизма (обмена веществ) некоторых микроорганизмов, развивающихся на отдельных видах плесени. Есть основания подозревать, что он оказывает на живые организмы канцерогенное действие (т. е. способствует развитию раковых заболеваний). Земляные орехи, особенно при хранении, склонны к гнилостным процессам, приводящим к образованию такой плесени, которая продуцирует (вырабатывает) афлатоксин. Поэтому анализ земляных орехов на это вещество имеет большое значение при определении качества и годности партий орехов. Поскольку концентрация афлатоксина ничтожна, приходится находить минимальную единичную пробу, объем которой достаточен для надежной идентификации вещества в миллиардных долях (1 на 10^9). В данной работе такая навеска равна 120 фунтам (несколько более чем 54,43 кг). Число 99, видимо, просто обозначает номер партии орехов.— *Примеч. пер.*

[39] Здесь и строкой ниже авторы снова забыли заменить «ошибку» на «отклонение».— *Примеч. пер.*

иесите на график полдюжины точек так, чтобы они охватывали весь диапазон значений X, и соедините их гладкими кривыми.)

8) Проверьте всю регрессию по F-критерию и установите, какую долю вариации относительно среднего $\overline{Y}$ объясняет подобранная прямая.

12. В данных из упражнения 11 фактически есть только две пары действительно повторных опытов при $X = 18{,}8$ и при $X = 46{,}8$. Правда, на практике часто применяют один трюк, состоящий в том, чтобы рассматривать «достаточно близкие друг к другу» точки как приблизительно повторные, и использовать их в качестве основы для вычисления приближенной суммы квадратов «чистой» ошибки. Ну а после этого можно действовать как ни в чем не бывало. Главная трудность здесь состоит в том, чтобы придать точный смысл словам «достаточно близкие друг к другу».

В рассматриваемых данных было бы интересно выяснить, к чему приводит представление о следующих семи множествах опытов как о множествах приблизительно повторных опытов:

$X = 9{,}3;\ 9{,}9.$

$X = 12{,}3;\ 12{,}5;\ 12{,}6.$

$X = 18{,}8;\ 18{,}8;\ 18{,}9.$

$X = 21{,}7;\ 21{,}9.$

$X = 46{,}8;\ 46{,}8$ (понятно, что это как раз точное повторение!).

$X = 70{,}6;\ 71{,}1;\ 71{,}3.$

$X = 83{,}2;\ 83{,}6.$

Вычислите приближенную сумму квадратов «чистой» ошибки, по этим наборам данных и перейдите к соответствующему приближенному анализу, тщательно следуя за текстом параграфа 1.5. Сделайте выводы.

13. Для подобранной прямой мы можем найти величину

$$R^2 = \mathrm{SS}(b_1|b_0)/\Sigma\,(Y_i - \overline{Y})^2.$$

Каково максимально возможное значение R^2, если:

1) В данных нет повторных опытов?

2) В данных *есть* настоящие повторные опыты?

Как вы думаете, можно ли распространить ваши выводы на общие регрессионные ситуации?

14. По приведенным ниже данным подобрана линейная модель $\hat{Y} = 1{,}692 - 0{,}0546X$ и получена таблица дисперсионного анализа. Продолжите исследование.

X	Y	$\hat{Y}$	e	X	Y	$\hat{Y}$	e
10	—2	1,1	—3,1	40	0	—0,5	0,5
10	—4	1,1	—5,1	40	1	—0,5	1,5
20	1	0,6	0,4	40	2	—0,5	2,5
20	3	0,6	2,4	50	—2	—1,0	—1,0
30	2	0,1	1,9	50	—3	—1,0	—2,0
30	5	0,1	4,9	50	—4	—1,0	—3,0

ANOVA

Источник	Число степеней свободы	SS	MS	F
b_0	1	0,083		
$b_1 \mid b_0$	1	7,240	7,240	0,845
Остаток	10	85,677	8,568	
Общий	12	93,000		

15. На одном из слушаний Федеральной комиссии по энергетике [40] (США) несколько лет назад были представлены следующие данные (которые здесь округлены):

X	Y	X	Y
13,3	3,5	26,3	6,0
16,9	5,1	30,1	9,5
19,9	4,8	42,6	8,1
23,2	6,7		

Переменные таковы:

предиктор X — процент жидкости в добываемом из скважин газе;

отклик Y — единичная стоимость в центах процесса добычи газа.

Подберите по этим данным прямую и, пользуясь таблицей дисперсионного анализа, найдите остаточную вариацию.

Свидетели, приведшие эти данные, экстраполировали прямую до точек $X = 0$ и $X = 100$ для подсчета единичных стоимостей в таких ситуациях, когда добыча вообще не содержит жидкостей, и, наоборот, когда она сплошь жидкая. Найдите значения $\hat{Y}(0)$ и $\hat{Y}(100)$ и определите 95 %-ные доверительные пределы для «истинного» среднего значения Y при $X = 0$ и $X = 100$. Каковы ваши заключения?

16. Покажите, что для подобранной прямой линии $r^2_{XY} = R^2 = r^2_{Y\hat{Y}}$. (Последнее равенство верно и в общем случае.)

17. Рассмотрите гипотезу, согласно которой названия статей в журналах обычно бывают длиннее для коротких статей и короче для длинных. Проверьте эту гипотезу следующим образом. Возьмите несколько номеров любого журнала из области ваших интересов и составьте список для некоторого числа статей (чем больше, тем лучше, но берите, скажем, за последние двадцать пять лет) таких данных:

X — число страниц в статье или заметке,

Y — число слов в названии статьи или заметки (слова, которые пишутся через черточку, считайте как одно слово).

Нанесите эти данные на график и подберите прямую $Y = \beta_0 + \beta_1 X + \varepsilon$ методом наименьших квадратов. Если есть повторные точки (или хотя бы близкие к повторным), то проверьте неадекватность. Затем, если нет адекватности или если нет информации о «чистой» ошибке, да еще и в остатках не просматривается никакой особой структуры, которая могла бы служить указанием на

[40] Федеральная комиссия по энергетике — орган правительства США. *Примеч. пер.*

неадекватность, то проверьте гипотезу $H_0 : \beta_1 = 0$ против альтернативы $H_1 : \beta_1 < 0$ с помощью одностороннего t-критерия. Каковы ваши выводы?

18. (Источник. Данные о росте кристаллов льда[41] из работы Района, Уишарта и Шоу, сотрудников Государственной организации научных и промышленных исследований Австралии, см.: Ryan B. F., Wishart E. R. and Shaw D. E. The growth rates and densities of ice crystals between — 3 °C and — 21 °C.— Journal of the Atmospheric of Sciences, 1976, 33, p. 842—850).

Кристаллы льда помещаются в камеру, внутри которой поддерживаются постоянная температура (—5 °C) и постоянный уровень насыщения воздуха водяными парами. Наблюдается рост кристаллов во времени. Здесь приведены 43 набора значений аксиальной длины кристаллов (длины, измеренной вдоль оси роста) (A) в микронах для времени (T) от 50 до 180 с от момента введения кристаллов. Каждое измерение представляет отдельный полный эксперимент. Опыты велись много дней и были рандомизированы относительно наблюдений времени. (Порядок, в котором на самом деле проводились опыты, нам не известен.) Было бы хорошо узнать, может ли линейная модель $A = \beta_0 + \beta_1 T + \varepsilon$ служить адекватным представлением роста во времени аксиальной длины кристаллов льда. Проведите полный анализ и сделайте выводы.

T	A	T	A	T	A	T	A
50	19	100	30, 29, 33	125	28	155	41,33
60	20, 21	105	35, 32	130	31,32	160	40, 30, 37
70	17, 22	110	30, 28, 30	135	34,25	165	32
80	25, 28	115	31, 36, 30	140	26,33	170	35
90	21, 25, 31	120	36, 25, 28	145	31	180	38
95	25			150	36,33		

19. Обратитесь к данным из табл. 1.1 на с. 30 и к их последующему анализу.

1) Допустим, что мы точно знаем, что «истинное» среднее значение $Y_0 = 10$. Воспользуйтесь методами обратного оценивания для получения оценки $\hat{X}_0$, соответствующей значению X, и постройте 95 %-ные «фидуциальные» пределы (X_L, X_U).

2) Зафиксируйте то значение g, которое вы получили в 1). Теперь получите значения границ (X_L, X_U) для приближения «$g = 0$» и сравните его с тем, что получено в 1).

3) Если значение Y_0 было бы не «истинным» средним значением, а просто результатом единственного нового наблюдения, то что стало бы с $\hat{X}_0$ и (X_L, X_U)?

4) Зафиксируйте то значение g, которое вы получили в 3). Теперь получите значения границ (X_L, X_U) для приближения «$g = 0$» и сравните его с тем, что получено в 3).

Используйте точные цифры и сохраняйте достаточное число знаков после запятой для минимизации ошибок округления.

20. Вспомните, что

$$V(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right].$$

[41] Данное исследование — звено в цепи работ, направленных на создание методов управления окружающей средой, в частности управления осадкообразованием. Рост кристаллов льда в некоторых типах облаков чреват градом, который ежегодно наносит народному хозяйству огромный ущерб. Анализ кинетики роста кристаллов льда может помочь разработке эффективных методов прекращения или даже предотвращения этого нежелательного процесса. Работы в этом направлении ведутся во многих странах, в том числе и у нас.— *Примеч. пер.*

Представьте себе, что некий экспериментатор говорит вам, что он нарисовал геометрические места крайних точек 95 %-ных доверительных границ для «истинного» среднего значения Y при заданном X. «Для всех практических целей,— сказал он,— во всем диапазоне значений X от 150 до 170, который представляет для меня интерес, эти геометрические места оказались прямыми, параллельными моей прямой». Могли ли вы ожидать, что диапазон значений X в его данных:

1) гораздо меньше, чем 20 единиц X?
2) примерно равен 20 единицам X?
3) гораздо больше, чем 20 единиц X?

Почему?

21. 1) Вы получили некоторые данные (X_i, Y_i), $i=1, 2, \ldots, n$. Вас просят построить по этим данным уравнение прямой $\hat{Y} = b_0 + b_1 X$ методом наименьших квадратов и выполнить «полный анализ». Как только вы закончили, приходит экспериментатор и говорит: «Я только что обнаружил, что все мои X-ы смещены. Каждое значение X, которое я вам дал, составляет только 90 % истинного значения. Пожалуйста, если можно, переделайте анализ». Используя $1-q$ вместо 90 % (так вы сможете трактовать эту задачу более общо), объясните, какая часть «полного анализа» изменится в связи с такой новой информацией и каким образом.

(Указание. Новое значение

$$S_{XX} = S^*_{XX} = S_{XX}/(1-q)^2,$$

$$\text{откуда} \quad b_1^* = S^*_{XY}/S^*_{XX} = b_1(1-q),$$

и т. д.)

2) Что бы вы делали, если бы оказалось, что смещения всех наблюдений различны; например, если бы было $(1-q_i)$ при $i = 1, 2, \ldots, n$?

3) Если бы в 1) или 2) значения q или q_i были бы так малы, чтобы величинами q^2 или q_i^2 можно было бы пренебречь, то как изменился бы обычный полный анализ?

22. (Источник. Wilson M. E., Mather L. E. Life expectancy. Letter to the Editor.— Journal of the American Medical Association, 1974, 229, № 11, p. 1421—1422.)

Приведенные ниже 50 пар наблюдений взяты из исследования докторов Л. Матера и М. Уилсона. Рассматривались такие переменные:

X — возраст человека к моменту смерти (округленный до ближайшего целого года),

Y — длина «линии жизни» на левой руке в сантиметрах (округленная до ближайших 0,15 см).

Многие люди верят, что продолжительность их жизни зависит линейно от длины их «линии жизни» [42]. Какой свет проливают приведенные здесь данные на такую веру? Вы можете считать, что:

$$\Sigma X = 3333, \qquad \Sigma X^2 = 231933, \qquad \Sigma XY = 30549{,}75,$$

$$\Sigma Y = 459{,}9, \qquad \Sigma Y^2 = 4308{,}57.$$

---

[42] Хиромантия — предсказание судьбы человека по линиям его руки — одно из распространенных средневековых учений, доживших и до наших дней. Представляет собой смесь суеверий, мистики и отдельных научных положений. По этому вопросу имеется огромная литература. Вот одна из наиболее обстоятельных работ: Дебароль А. Тайны руки. Искусство узнавать жизнь, характер и будущность каждого посредством простого исследования руки. Пер. с фр.— М.: В Университетской типографии (Катков и К°), 1868,— 608 с. (Мы привели эту ссылку в современной орфографии.) Один из самых эффективных путей борьбы с подобными пережитками — прямой научный эксперимент. Такой эксперимент и проделали авторы данного примера. Его результат достаточно красноречив.— *Примеч. пер.*

X — возраст, лет	Y — длина, см	X — возраст, лет	Y — длина, см	X — возраст, лет	Y — длина, см	X — возраст, лет	Y — длина, см
19	9,75	61	7,20	68	9,00	75	10,20
40	9,00	62	7,95	69	7,80	76	6,00
42	9,60	62	8,85	69	10,05	77	8,85
42	9,75	65	8,25	70	10,50	80	9,00
47	11,25	65	8,85	71	9,15	82	9,75
49	9,45	65	9,75	71	9,45	82	10,65
50	11,25	66	8,85	71	9,45	82	13,20
54	9,00	66	9,15	72	9,45	83	7,95
56	7,95	66	10,20	73	8,10	86	7,95
56	12,00	67	9,15	74	8,85	88	9,15
57	8,10	68	7,95	74	9,60	88	9,75
57	10,20	68	8,85	75	6,45	94	9,00
58	8,55			75	9,75		

23. В газете The Chicago Maroon за пятницу 10 ноября 1972 г. сообщалось, что на оптовом рынке ожидаются следующие цены на марочные портвейны в расчете на бутылку:

Год	Цена, дол.	Год	Цена, дол.	Год	Цена, дол	Год	Цена, дол.
1890	50,00	1934	15,00	1941	10,00	1952	4,99
1900	35,00	1935	13,00	1944	5,99	1955	5,98
1920	25,00	1940	6,98	1948	8,98	1960	4,98
1931	11,98			1950	6,98		

Источник. Данные из работы Стива Стиглера, 1976.

1) Нанесите данные на график и исследуйте их. Имеет ли смысл строить регрессию для отклика «цена» от предиктора «год»? Какие неудобства вам удалось здесь обнаружить?

2) Какое преобразование переменной «год» имело бы смысл? (Указание. Предположите, например, что сейчас 1972 г. Как вы обычно говорите о своем «годе рождения»?) Постройте график зависимости цены от вашего нового предиктора и исследуйте его. Какого рода преобразование цены кажется вам здесь разумным, чтобы сделать данные «лучше ложащимися на прямую»?

3) Постройте график для данных $Y = \ln$ (цена) при $Z =$ возраст бутылки. Методом наименьших квадратов подберите по этим данным прямую, вычислите остатки и постройте таблицу дисперсионного анализа.

4) Что вы можете сказать про цены на марочный портвейн на основе такого представления данных и вашего анализа? Какого (с точностью до ближайшего цента) прироста цены за год вы ожидаете, если подобная структура цен сохранится и на будущее?

5) В следующем объявлении в той же газете три года спустя, 25 ноября 1975 г., во вторник, говорилось, что предлагается марочный портвейн 1937 г. по цене 20,00 дол. за бутылку. Если бы можно было считать, что прямолинейная зависимость сохранилась и что она приложима также и к этой новой точке, то каким представляется прирост цены в год на бутылку за истекшие три года? Сохраняет ли здесь силу ваш ответ на вопрос из пункта 4) или похоже, что ежегодное нарастание цены ускорилось?

24. (Источник. Anscombe F. J. Graphs in statistical analysis. American statistician, 27, 1973, p. 17—21.)

Подберите уравнения прямых по модели $Y = \beta_0 + \beta_1 X + \varepsilon$ для всех четырех приведенных ниже наборов данных и убедитесь в том, что *для каждого из этих наборов* справедливо: $n = 11$, $\overline{X} = 9$, $\overline{Y} = 7{,}5$, $\widehat{Y} = 3 + 0{,}5X$, $S_{XX} = 110$, регрессионная $SS = S^2_{XY}/S_{XX} = 27{,}5$ (1 ст. св.), остаточная $SS = S_{YY} - S^2_{XY}/S_{XX} = 13{,}75$ (9 ст. св), оц. ст. откл. $(b_1) = 0{,}118$, $R^2 = 0{,}667$.

Постройте графики для всех четырех наборов данных и объясните, в чем их различия и что их главным образом характеризует.

(Обратите внимание, что для данных из наборов 1—3 все значения X-ов одни и те же, а значения Y-ов различны.)

Наборы данных		1—3	1	2	3	4	4
Переменные		X	Y	Y	Y	X	Y
Номера опытов	1	10	8,04	9,14	7,46	8	6,58
	2	8	6,95	8,14	6,77	8	5,76
	3	13	7,58	8,74	12,74	8	7,71
	4	9	8,81	8,77	7,11	8	8,84
	5	11	8,33	9,26	7,81	8	8,47
	6	14	9,96	8,10	8,84	8	7,04
	7	6	7,24	6,13	6,08	8	5,25
	8	4	4,26	3,10	5,39	8	5,56
	9	12	10,84	9,13	8,15	8	7,91
	10	7	4,82	7,26	6,42	8	6,89
	11	5	5,68	4,74	5,73	19	12,50

25. Представьте себе, что вас попросили с помощью имитационного моделирования выполнить следующее:

1) выбрать «истинную» прямую линию $\eta = \beta_0 + \beta_1 X$;

2) отобрать n значений X : $X_1, X_2, \ldots, X_n$;

3) генерировать n случайных ошибок ε_i, $i = 1, 2, \ldots, n$, из распределения $N(0, \sigma^2)$ и отсюда получить n «наблюдений» по $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$;

4) по этим «наблюдениям» построить прямую $\widehat{Y} = b_0 + b_1 X$;

5) найти набор остатков; $e_i = Y_i - \widehat{Y}_i$;

6) повторить шаги 3)—5) всего N раз, используя (например) такие значения параметров: $n = 11$, $X_i = -1 + (i-1)\,0{,}2$, $\sigma^2 = 1$, $N = 1000$.

Объясните, почему не были указаны конкретные значения для параметров β_0 и β_1?

(Указание. Докажите, что остатки e_i не зависят от β_0 и β_1. Значит, могут использоваться любые значения. Проще всего взять $\beta_0 = \beta_1 = 0$.)

Ответы к упражнениям

1. 1) $b_1 = 158/110 = 1{,}44$; $b_0 = 102/11 = 9{,}27$; $\widehat{Y} = 9{,}27 + 1{,}44\,X$.

2) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий разброс (скорректированный)	10	248,18		

Источник рассеяния	Число степеней свободы	SS	MS	F
Регрессия	1	$\frac{(158)^2}{110}$	226,95	96,17*
Остаток	9	21,23	2,36	

Гипотеза $H_0 : \beta_1 = 0$ проверяется при $\alpha = 0{,}05$ путем сравнения вычисленной F (1,9)-статистики с критическим значением F (1,9) для $\alpha = 0{,}05$. Из таблицы F-критерия при $\alpha = 0{,}05$ находим F (1, 9, 0,95) = 5,12. Так как 96,17 больше, чем 5,12, гипотеза о том, что $\beta_1 = 0$, отвергается.

3) Доверительные пределы для β_1 (95 %) : $1{,}11 \leqslant \beta_1 \leqslant 1{,}77$.

4) 95 %-ные доверительные пределы для истинного среднего значения Y при $X_0 = 3$: 12,15 $\leqslant$ истинное среднее Y при $X_0 = 3 \leqslant 15{,}03$.

5) 95 %-ные доверительные пределы для разности между истинным средним значением Y при $X_1 = 3$ и истинным средним значением Y при $X_2 = -2$. Сначала определим алгебраическую разность между $\hat{Y}_1$ и $\hat{Y}_2$.

$$\hat{Y}_1 = b_0 + b_1 (3), \quad \hat{Y}_2 = b_0 + b_1 (-2).$$

Следовательно,

$$\hat{Y}_1 - \hat{Y}_2 = b_1 (3 + 2) = 5b_1 = 5 (1{,}44) = 7{,}20;$$

$$s^2_{(\hat{Y}_1 - \hat{Y}_2)} = 25 s^2_{b_1} = 25 \left(\frac{2{,}36}{110} \right) = 0{,}53635;$$

$$s_{(\hat{Y}_1 - \hat{Y}_2)} = \sqrt{0{,}53635} = 0{,}732;$$

$$ts_{(\hat{Y}_1 - \hat{Y}_2)} = (2{,}262)(0{,}732) = 1{,}656.$$

Таким образом, 95 %-ные доверительные границы для истинной разности равны: 7,20—1,66 $\leqslant$ истинная разность $\leqslant$ 7,20 + 1,66, или 5,54 $\leqslant$ истинная разность $\leqslant$ 8,86.

6) Рассчитываем остатки и рассматриваем «картинки».

X	Y	$\hat{Y}$	$Y - \hat{Y}$	X	Y	$\hat{Y}$	$Y - \hat{Y}$
—5	1	2,07	—1,07	1	9	10,71	—1,71
—4	5	3,51	1,49	2	13	12,15	0,85
—3	4	4,95	—0,95	3	14	13,59	0,41
—2	7	6,39	0,61	4	13	15,03	—2,03
—1	10	7,83	2,17	5	18	16,47	1,53
0	8	9,27	—1,27				

Никакой очевидной альтернативы для модели нет.

7) Если пробное предположение о модели первого порядка приемлемо, то нет смысла использовать одиннадцать экспериментальных уровней. Конечно, необходимо по крайней мере два уровня, чтобы оценить параметры модели, и хотя бы еще один, чтобы определить кривизну истинной модели, если кривизна имеется. С помощью проведения повторных опытов при некоторых или всех уровнях мы можем получить оценку дисперсии, отвечающей чистой ошибке σ^2, для проверки адекватности. Таким образом, для эксперимента примерно одного и того же объема можно будет выбрать три как можно более удаленных друг от друга уровня, например границы области X и центр, и провести по 4 опыта на каждом из этих уровней. Это приведет к таблице дисперсионного анализа в форме, показанной далее.

Дисперсионный анализ ($n = 12$)

Источник рассеяния	Число степеней свободы
Общий (скорректированный)	11
Регрессия	1
Остаток	10
неадекватность	1
«чистая» ошибка	9

Так как у нас только одна степень свободы для проверки адекватности, это не совсем хорошо. Несколько лучше было бы выбрать по три опыта при каждом из четырех уровней:

Дисперсионный анализ ($n = 12$)

Источник рассеяния	Число степеней свободы
Общий (скорректированный)	11
Регрессия	1
Остаток	10
неадекватность	2
«чистая» ошибка	8

Возможны и другие варианты, см. параграф 1.8.

2. 1) В рандомизированном порядке.

2а) $\hat{Y} = 0{,}5 + 0{,}5X$.

2б)

Источник рассеяния	Число степеней свободы	SS	MS
Скорректированный общий $\Sigma (Y_i - \bar{Y})^2$	19	83,2	
Обусловленный регрессией $\dfrac{[\Sigma (X_i - \bar{X})(Y_i - \bar{Y})]^2}{\Sigma (X_i - \bar{X})^2}$	1	40,0	40,0 *
Остаток	18	43,2	2,4

2в) (1) $\hat{Y} = 3{,}0 \pm 0{,}73 = 2{,}27;\ 3{,}73.$

(2) $\hat{Y} = 5{,}0 \pm 1{,}26 = 3{,}74;\ 6{,}26.$

3а) Остаток	18	43,2	2,4	
Неадекватность	3	1,2	0,4	незначимо
«Чистая» ошибка	15	42,0	2,8	

Неадекватность незначима (т. е. модель признается адекватной).

3б) Да.

4а) Такой же дисперсионный анализ и вывод, как и в 3а).

4б) Доверительные пределы неприменимы. Дисперсия ошибки зависит от уровня Y.

4в) Пригодна модель первого порядка.

5а) Остаток	18	43,2	2,4
Неадекватность	3	20,0	6,67 *
«Чистая» ошибка	15	23,2	1,55

Значимость члена, связанного с неадекватностью, указывает на неадекватность модели.

5б) Так как модель неверна, доверительные пределы будут необоснованными.

5в) Предлагается модель второго порядка.

3. Наилучшим уравнением прямой будет следующее:

$$\hat{Y} = b_0 + b_1 X = 129{,}7872 - 24{,}0199X.$$

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
Общий	12	3396,62		
Регрессия	1	3293,77	3293,77	
Остаток	11	102,85	9,35	
неадекватность	5	91,08	18,22	9,30*
«чистая» ошибка	6	11,77	1,96	

Модель неадекватна.

4. Решение не приводится.

5. Модели: $Y_{iu} - \overline{Y}_i = \beta_i (X_{iu} - \overline{X}_i)$,
$i = 1, 2, \ldots, m$;
$u = 1, 2, \ldots, n$.

1)

X_1	Y_1	$(X_{1u} - \overline{X}_1)$	$(Y_{1u} - \overline{Y}_1)$
3,5	24	—1,529	—17,286
4,1	32	—0,929	—9,286
4,4	37	—0,629	—4,286
5,0	40	—0,029	—1,286
5,5	43	0,471	1,714
6,1	51	1,071	9,714
6,6	62	1,571	20,714

$$\overline{X}_1 = 5{,}029, \quad \overline{Y}_1 = 41{,}286, \quad n_1 = 7,$$

$$b_1 = \left\{\sum_{u=1}^{7} (X_{iu} - \overline{X}_i)(Y_{iu} - \overline{Y}_i)\right\} \Big/ \left\{\sum_{u=1}^{7} (X_{iu} - \overline{X}_i)^2\right\} = 81{,}542858/7{,}434 = 10{,}969.$$

$$\mathrm{SS}(b_1) = b_1^2 \left\{\sum_{u=1}^{7} (\overline{X}_{iu} - \overline{X}_i)^2\right\} = (120318961)(7{,}434) = 894{,}451.$$

2)

X_2	Y_2		$(X_{iu}-\overline{X}_i)$	$(Y_{iu}-\overline{Y}_i)$
3,2	22	$\overline{X}_2 = 5{,}533$	—2,333	—19,333
3,9	33	$\overline{Y}_2 = 41{,}333$	—1,633	—8,333
4,9	39		—0,633	—2,333
6,1	44	$n_2 = 6$	0,567	2,667
7,0	53		1,467	11,667
8,1	57		2,567	15,667

$$b_2 = \left\{\sum_{u=1}^{6}(X_{iu}-\overline{X}_i)(Y_{iu}-\overline{Y}_i)\right\} \Big/ \left\{\sum_{u=1}^{6}(X_{iu}-\overline{X}_i)^2\right\} = 119{,}033334/17{,}573334 =$$

$$= 6{,}773520.$$

$$SS(b_2) = b_2^2\left\{\sum_{u=1}^{6}(X_{iu}-\overline{X}_i)^2\right\} = (45{,}880573)\,(17{,}573334) = 806{,}274633.$$

3)

X_3	Y_3	$(X_{iu}-\overline{X}_i)$	$(Y_{iu}-\overline{Y}_i)$	X_3	Y_3	$(X_{iu}-\overline{X}_i)$	$(Y_{iu}-\overline{Y}_i)$
3,0	32	—2,775	—18,750	6,5	55	0,725	4,250
4,0	36	—1,775	—14,750	7,0	59	1,225	8,250
5,0	47	—0,775	—3,750	7,3	64	1,525	13,250
5,0	49	0,225	—1,750	7,4	64	1,625	13,250

$$\overline{X}_3 = 5{,}775, \quad \overline{Y}_3 = 50{,}750,$$

$$n_3 = 8,$$

$$b_3 = \left\{\sum_{u=1}^{8}(X_{iu}-\overline{X}_i)(Y_{iu}-\overline{Y}_i)\right\} \Big/ \left\{\sum_{u=1}^{8}(X_{iu}-\overline{X}_i)^2\right\} =$$

$$= (135{,}650000)/(18{,}495) = 7{,}334415,$$

$$SS\,(b_3) = (53{,}793643)\,(18{,}495) = 994{,}913427,$$

$$b = \left\{\sum_{i=1}^{m=3}\sum_{u=1}^{n_i}(X_{iu}-\overline{X}_i)(Y_{iu}-\overline{Y}_i)\right\} \Big/ \left\{\sum_{i=1}^{m=3}\sum_{u=1}^{n_i}(X_{iu}-\overline{X}_i)^2\right\} =$$

$$= (81{,}542858 + 119{,}033334 + 135{,}650)/(7{,}434 + 17{,}573334 + 18{,}495) =$$

$$= 336{,}226192/43{,}502334 = 7{,}728923.$$

$$SS\,(b) = b^2\left\{\sum_{i=1}^{3}\sum_{u=1}^{n_i}(X_{iu}-\overline{X}_i)^2\right\} = (59{,}736251)\;\;(43{,}502334) = 2598{,}666299,$$

SS, обусловленная всеми $b_i|b = 894{,}451000 + 806{,}274633 + 994{,}913427 -$

$- 2598{,}666343 = 96{,}972717$, остаток=полная SS — SS (b) — SS,

обусловленная всеми $b_i\,|\,b = 2792{,}261906 - 2598{,}666343 - 96{,}972717 =$

$= 96{,}622846.$

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
b	1	2598,666343	2598,666343	403,42
Все $b_i \mid b$	2	96,972717	48,486359	7,53
Остаток	15	96,622846	6,441523	
Общий	18	2792,261906		

$$H_0: \beta_i = \beta, \qquad F_2 = 7{,}53 > F\,(2,\ 15,\ 0{,}95) = 3{,}68,$$

$\therefore$ H_0 отвергается.

6. 1) $\hat{Y} = -21{,}33 + 5X$.

2) $2{,}984 \leqslant \beta_1 \leqslant 7{,}016$.

3) **Дисперсионный анализ**

Источник	Число степеней свободы	SS	MS	F
Скорректированный общий	11	69,67		
Обусловленный регрессией	1	52,50	52,50	
Остаток	10	17,17	1,72	
неадекватность	4	5,50	1,375	0.706 (не значимо при $\alpha = 0{,}05$)
«чистая» ошибка	6	11,67	1,945	

Оказывается, что модель адекватна.

7. 1) $\hat{Y} = 323{,}628 - 131{,}717X$.

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
Скорректированный общий	16	2 305 042		
Обусловленный регрессией	1	1 099 641,1	1 099 641,10	13,68 значимо при [43] $\alpha = 0{,}05$
Остаток	15	1 205 400,9	80 360,06	

[43] Ответ противоречит заданию. Требуется проверить гипотезу при $\alpha = 0{,}10$, а в ответе при $\alpha = 0{,}05$. Очевидно, это недосмотр авторов. Отметим, однако, что смысл ответа сохранится и при правильном выборе уровня значимости. Все сказанное в этом примечании следует повторить и для следующей таблицы дисперсионного анализа.—*Примеч. пер.*

2) **Дисперсионный анализ**

Источник	Число степеней свободы	SS	MS	F
Скорректированный общий	16	2 305 042		
Обусловленный регрессией	1	1 099 641,1		
Остаток	15	1 205 400,9		
неадекватность	5	520 648,6	104 129,72	1,52 не значимо при $\alpha = 0{,}05$
«чистая» ошибка	10	684 752,3	68 475,23	

Линейная зависимость оказывается приемлемой.

8. Предсказывающее уравнение: $Y = 1{,}222 + 0{,}723X$.

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F	$F_{0,95}$
Скорректированный общий	13	2,777			
Обусловленный регрессией	1	1,251	1,251	9,850	4,75
Остаток	12	1,526	0,127		

$9{,}850 > F(1,12,\ 0{,}95) = 4{,}75$. · . отвергается $H_0 : \beta_1 = 0$.

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F	$F_{0,95}$
Скорректированный общий	13	2,777			
Обусловленный регрессией	1	1,251			
Остаток	12	1,526			
неадекватность	7	0,819	0,117	0,830	4,88
«чистая» ошибка	5	0,707	0,141		

$0{,}830 < F\ (7,\ 5,\ 0{,}95) = 4{,}88$, . · . неадекватность незначима.

Вывод: используйте для предсказания уравнение: потери в чашках (%) = = 1,222 + (0,723) [потери в бутылках (%)].

9. Предсказывающее уравнение: $\hat{Y} = 17{,}146 + 11{,}836X$.

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F	$F_{0,95}$
Скорректированный общий	12	22 126,308			
Обусловленный регрессией	1	6 034,379	6034,379	4,125	4,840
Остаток	11	16 091,929	1462,903		

$4{,}125 < F\,(1,\ 11,\ 0{,}95) = 4{,}840$, $\therefore$ не отвергается $H_0 : \beta_1 = 0$. Регрессия незначима.

$$R^2 = \frac{\text{SS, обусловленная регрессией}}{\text{SS, скорректированная общая}} = \frac{6034{,}379}{22126{,}308} = 27{,}27\%.$$

Выводы. (1) Модель неадекватна. (2) Необходимо дальнейшее исследование альтернативных переменных. (3) «Чистая» ошибка нуждается в проверке.

10. 1) $\hat{Y} = 2{,}5372000 - 0{,}004718X$.

2) **Дисперсионный анализ**

Источник	Число степеней свободы	SS	MS	F	$F_{0,95}$
Скорректированный общий	14	0,209333			
Обусловленный регрессией	1	0,110395	0,110395	14,50	4,67
Остаток	13	0,098938	0,007611		

$14{,}50 > F\,(1,\ 13,\ 0{,}95) = 4{,}67$, $\therefore$ отвергается $H_0 : \beta_1 = 0$

Регрессия значима, если нет неадекватности.

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F	$F_{0\ 95}$
Остаток	13	0,098938			
Неадекватность	5	0,018938	0,003788	0,38	3,69
«Чистая» ошибка	8	0,080000	0,010000		

$0{,}38 < F\,(5,\ 8,\ 0{,}95) = 3{,}69$; $\therefore$ неадекватность незначима.

3) 95 %-ные доверительные интервалы для истинного среднего значения Y, вычисленные в четырех точках: $X = 0$, $X = \overline{X}$, $X = 400$, $X = 460$:

при $X = 0$ $\hat{Y} \pm (2{,}160)\,(0{,}527) = \hat{Y} = \pm\ 1{,}138$,

при $X = \overline{X}$ $\hat{Y} \pm (2{,}160)\,(0{,}022) = \hat{Y} \pm 0{,}048$,

при $X = 400$ $\hat{Y} \pm (2{,}160)\,(0{,}039) = \hat{Y} \pm 0{,}084$,

при $X = 460$ $\hat{Y} \pm (2{,}160)\,(0{,}048) = \hat{Y} \pm 0{,}104$.

11. 1) Нанесите данные на график и проведите линию «на глаз». Здесь такая прямая может оказаться несколько отличной от того, что получится позже методом наименьших квадратов.

2) $\Sigma X_u = 1244{,}5$, $\Sigma Y_u = 30{,}458$,

$\Sigma X_u^2 = 73920{,}05$, $\Sigma Y_u^2 = 27{,}573638$,

$\Sigma X_u Y_u = 1032{,}4865$.

Для дальнейших вычислений разумно сохранить все знаки в этих суммах квадратов.

3) $b_1 = -0{,}00290351$, $b_0 = \overline{Y} - b_1\overline{X} = 1{,}00210$.

Эта прямая содержит (например) точки (0, 1,0021) и (100, 0,7117).

4) В этом ответе могут появиться любые разности, но их сумма должна быть достаточно малой и лишь слегка отличаться от нуля в силу ошибок округления.

Дисперсионный анализ

5)

Источник	Число степеней свободы	SS	MS
Регрессия (b_0)	1	27,28500	
Регрессия ($b_1 \mid b_0$)	1	0,23914	0,23915
Остаток	32	0,04950	$s^2 = 0{,}001547$
Общий	34	27,57364	

6) ст. ош.[44] $(b_1) = s/\{\Sigma X_u^2 - (\Sigma X_u)^2/n\}^{1/2} = 0{,}00023$,

ст. ош. $(k_2) = s\left[\Sigma X_u^2/\{n\Sigma X_u^2 - (\Sigma X_u)^2\}\right]^{1/2} = 0{,}01089$.

7) ст. ош. $(\hat{Y}_0) = s\,\{1/n + (X_0 - \overline{X})^2/(\Sigma X_u^2 - (\Sigma X_u)^2/n)\}^{1/2}$.

Формула для любого частного значения X_0 получается после подстановки этого значения (вместо X_0). Тогда 95 %-ные доверительные границы для «истинного» среднего значения Y при X_0 даются выражением

$$\hat{Y}_0 \pm t\,(32,\ 0{,}975)\ \text{ст. ош. } (\hat{Y}_0).$$

Вместо t (32, 0,975) мы можем взять близкое t (30, 0,975) = 2,042 или интерполировать по таблицам. График для доверительных границ должен выглядеть примерно так, как на рис. 1.8.

8) Статистика F-критерия для всей регрессии равна 0,23915/0,001547 = = 154,6 по сравнению с табличным F (1, 30, 0,95) = 4,17. Мы должны были бы интерполировать на 32 степени свободы, но совершенно ясно, что в данном слу-

[44] Здесь, как и в задании к этому упражнению, вместо «отклонений» сохранились «ошибки».— *Примеч. пер.*

чае в этом нет нужды. Таким образом, мы отбрасываем нуль-гипотезу о том, что $\beta_1 = 0$. $R^2 = 0{,}83$, т. е. 83 % вариации относительно среднего $\bar{Y}$ объясняется с помощью нашей линейной регрессии.

12. Мы нашли приближенную сумму квадратов для «чистой» ошибки, равной 0,01678 с 10 степенями свободы. Поэтому получается такая таблица дисперсионного анализа:

ANOVA

Источник	Число степеней свободы	SS	MS
Общий (скорректированный)	33	0,2886	
Регрессия $(b_1 \mid b_0)$	1	0,23915	0,23915
Остаток	32	0,04950	$s^2 = 0{,}001547$
неадекватность	22	0,03272	$MS_L = 0{,}001487$
«чистая» ошибка	10	0,01678	$s_e^2 = 0{,}001678$

Для проверки адекватности вычислим F-статистнку, равную $MS_L/s_e^2 = 0{,}8862$, что явно указывает на отсутствие неадекватности, поскольку табличное F (22, 10, 0,95) = 2,75. Поэтому мы объединим суммы квадратов, обусловленные неадекватностью н «чистой» ошибкой, для вычисления s^2. З а к л ю ч е н и е. Представляется, что данные адекватно описываются линейной регрессией Y на X. Найденной зависимостью можно воспользоваться для предсказания истинного среднего значения Y при любом заданном X_0, а также для построения доверительных границ, которые заранее указывают, с какой точностью такое предсказание может быть сделано, полагая, что модель корректна.

13. Выводы таковы: 1) R^2 может быть равен 1, если в данных нет повторных опытов, но 2) R^2 не может быть равен 1, если только имеют место повторные опыты, поскольку модель не в состоянии объяснить сумму квадратов, обусловленную «чистой» ошибкой. Эти утверждения справедливы и в общей регрессионной ситуации, как показывают следующие выкладки.

Пусть наблюдения будут таковы:

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \text{ в первой области пространства } X,$$

$$Y_{21}, Y_{22}, \ldots, Y_{2n_2} \text{ во второй области пространства } X,$$

$$\vdots$$

$$Y_{k1}, Y_{k2}, \ldots, Y_{kn_k} \text{ в } k\text{-й области пространства } X.$$

$$R^2 = 1 - \frac{\sum_{r=1}^{k}\sum_{u=1}^{n_r}(Y_{ru} - \hat{Y}_{ru})^2}{\sum_{r=1}^{k}\sum_{u=1}^{n_r}(Y_{ru} - \bar{Y})^2} = 1 - \frac{\text{Остаточная SS}}{\text{Общая скорректированная SS}}.$$

Теперь заметим, что $\hat{Y}_{ru} = \hat{Y}_r$ одинаковы для всех u. Пусть $\bar{Y}_r = \sum_{u=1}^{n_r} Y_{ru}/n_r$ будет средним откликом в r-й области. Тогда

$$\sum_{u=1}^{n_r}(Y_{ru} - \hat{Y}_{ru})^2 = \sum_{u=1}^{n_r}\{(Y_{ru} - \bar{Y}_r) + (\bar{Y}_r - \hat{Y}_r)\}^2 = \sum_{u=1}^{n_r}(Y_{ru} - \bar{Y}_r)^2 +$$

$$+\sum_{u=1}^{n_r}(\overline{Y}_r-\widehat{Y}_r)^2+2(\overline{Y}_r-\widehat{Y}_r)\sum_{u=1}^{n_r}(Y_{ru}-\overline{Y}_r)=\sum_{u=1}^{n_r}(Y_{ru}-\overline{Y}_r)^2+$$

$$+n_r(\overline{Y}_r-\widehat{Y}_r)^2,$$

последнее суммирование дает нуль. Следовательно,

$$R^2=1-\frac{\sum_{r=1}^{k}\sum_{u=1}^{n_r}(Y_{ru}-\overline{Y}_r)^2+\sum_{r=1}^{k}n_r(\overline{Y}_r-\widehat{Y}_r)^2}{\sum_{r=1}^{k}\sum_{u=1}^{n_r}(Y_{ru}-\overline{Y})^2}.$$

Отсюда вытекает, что R^2 может достичь 1, если, во-первых, все $Y_{ru}=\overline{Y}_r$ и, во-вторых, $\overline{Y}_r=\widehat{Y}_r$.

Однако первое верно, только если нет повторных опытов, т. е. все $n_r=1$, или все повторы в некоторой области идентичны, и это верно для всех областей.

Из второго следует, что подобранная модель проходит точно по всем точкам средних, что иногда может случиться.

Значит и в общем, если только существует «чистая» ошибка, то $R^2<1$.

14. Прежде чем проверять регрессию, испытаем адекватность. Если модель окажется неадекватной, то F-критерий для регрессии и вычисление доверительных границ, как и все последующее, не имеют оснований. Мы так вычисляем «чистую» ошибку:

X	Вклад в «чистую» ошибку		Число степеней свободы
10	$\frac{1}{2}[-2-(-4)]^2$	$=2$	1
20	$\frac{1}{2}[1-3]^2$	$=2$	1
30	$\frac{1}{2}[2-5]^2$	$=4{,}5$	1
40	$0^2+1^2+2^2-3^2/_3$	$=2$	2
50	$(-2)^2+(-3)^2+(-4)^2-(-9)^2/_3$	$=2$	2
SS «чистой» ошибки		$=12{,}5$	7

Разложение остаточной суммы квадратов

Источник	Число степеней свободы	SS	MS	F
Неадекватность	3	73,177	24,392	$F(3,\ 7)=13{,}66$, указывает на значимую неадекватность
«Чистая» ошибка	7	12,5	1,786	
Остаток	10	85,677		

Модель страдает неадекватностью. Следующий шаг — построение графиков для остатков и их внимательное изучение в надежде увидеть, нельзя ли как-нибудь улучшить модель.

15. Предсказывающее уравнение: $\hat{Y} = 2{,}0464 - 0{,}1705X$.

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
Регрессия $(b_1 \mid b_0)$	1	16,514	16,514	9,47
Остаток	5	8,723	$s^2 = 1{,}745$	значимо на уровне 2,76 %
Общий, скорректированный	6	25,237		
SS (b_0)	1	272,813		
Общий	7	298,050		

$$R^2 = \frac{16{,}514}{24{,}237} = 0{,}6544\,, \qquad \hat{Y}\,(0) = 2{,}05,$$

$$\hat{Y}\,(100) = 19{,}10,$$

$$\text{Оц. } V\{\hat{Y}(X_0)\} = s^2\left\{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\Sigma (X_i - \overline{X})^2}\right\} =$$

$$= 1{,}745\left\{\frac{1}{7} + \frac{(X_0 - 24{,}614286)^2}{568{,}168571}\right\},$$

$$\text{Оц. } V(\hat{Y}(0)) = 1{,}745\,\{0{,}142857 + 1{,}066344\} = 2{,}110056 = (1{,}452603)^2,$$

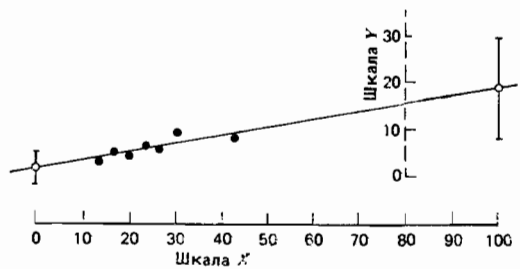
$$\text{Оц. } V(\hat{Y}(100)) = 1{,}745\,\{0{,}142857 - 10{,}002324\} = 17{,}70334 = (4{,}207534)^2.$$

95 % ные доверительные границы для истинного среднего значения Y следующие:

$$\hat{Y} \pm t_v\left(1 - \frac{1}{2}\,\alpha\right)\sqrt{\text{Оц. } V(\hat{Y})} = \hat{Y} \pm 2{,}571\,\sqrt{\text{Оц. } V(\hat{Y})}\,,$$

$X = 0 : 2{,}05 \pm 3{,}73 =$ от $-1{,}68$ до $5{,}78$,

$X = 100 : 19{,}10 \pm 10{,}82 =$ от $8{,}28$ до $29{,}92$ (см. рис. ниже).



К решению упражнения 15

Это широковато (по сравнению, скажем, с днапазоном значений Y, равным 6 единицам), даже если модель верна во всем интервале от 0 до 100, что создает неопределенность по двум причинам:

1. Оба значения 0 и 100 лежат «*за тридевять земель*» от области эксперимента.

2. Опыт показывает, что нет никакого смысла верить в линейную зависимость в столь широком днапазоне, как от 0 до 100.

Вывод: Предсказания на $X = 0$ и $X = 100$ должны вызывать большие опасения.

16. 1) $$r_{XY}^2 = \frac{\{\Sigma (X_i - \overline{X})(Y_i - \overline{Y})\}^2}{\{\Sigma (X_i - \overline{X})^2\}\{\Sigma (Y_i - \overline{Y})^2\}} = \frac{\mathrm{SS}(R \mid b_0)}{\Sigma (Y_i - \overline{Y})^2} = R^2.$$

2) $$r_{Y\hat{Y}} = \frac{\Sigma (\hat{Y}_i - \overline{\hat{Y}}_i)(Y_i - \overline{Y})}{\{\Sigma (Y_i - \overline{\hat{Y}}_i)^2\}^{1/2}\{\Sigma (Y_i - \overline{Y})^2\}^{1/2}}.$$

Кроме того, $\hat{Y}_i = b_0 + b_1 X_i$, $\overline{\hat{Y}}_i = b_0 + b_1 \overline{X} = \overline{Y}$. Следовательно, мы можем подставить $\hat{Y}_i - \overline{\hat{Y}}_i = b_i (X_i - \overline{X})$, сократить b_1 вверху и внизу и получить r_{XY}.

17. Решение зависит от собранных данных.

18. (Частное решение). **Подобранное уравнение таково:**

$$\hat{A} = 14{,}410649 + 0{,}130768T.$$

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
Общий	43	39 554,000		
b_0	1	38 221,488		
$b_1 \mid b_0$	1	764,646	764,646	55,2 значимо на уровне 1 %
Остаток	41	567,866	$s^2 = 13{,}851$	
Неадекватность	20	243,866	12,19	<1 не значимо
«Чистая» ошибка	21	324,000	15,43	

$$R^2 = 0{,}574.$$

Здесь в остатках вроде бы нет никаких примечательных особенностей. Модель значима. Из вариации относительно среднего 57,4 % удается объяснить с помощью модели, а максимально возможная доля в данном случае составляет 100 (1332,512—324)/1332,512 = 75,7 %. (Остальная вариация обусловлена «чистой» ошибкой). Таким образом, получено вполне удовлетворительное приближение, хотя могло бы быть и лучше.

Вот остатки, приведенные в том же порядке и округленные до ближайшего целого:

—2	—3	3, 2, 6	6, —2
—2, —1	0, 1	7, 4	5, —5, 2
—7, —2	2, —7	1, —1, 1	—4
0, 3	—7, 0	2, 7, 1	—2
—5, —1 5	—2	6, —5, —2	0
—2	2, —1		

19. 1) $\hat{X}_0 = 45{,}38$, $(X_L, X_U) = (39{,}56,\ 50{,}05)$.

2) $g = 0{,}074418$. Для $g = 0$ $(X_L, X_U) = (40{,}37,\ 50{,}40)$ весьма близко к значениям из пункта 1), но не тождественно.

3) $\widehat{X}_0 = 45{,}38$ (так же, как и в пункте 1). В уравнение (1.7.8) в выражение в прямых скобках вводится дополнительная единица, что дает конечные точки $(X_L, X_u) = (19{,}33,\ 70{,}28)$. Это гораздо шире, чем раньше.

4) $g = 0{,}074418$ (так же, как и в пункте 2). Для $g = 0$ $(X_L, X_u) = (21{,}78,\ 68{,}99)$. Большой диапазон, охватываемый этими двумя интервалами, делает их довольно сильно похожими на интервалы из пункта 3), несмотря на большое численное различие.

20. Единственная возможность, чтобы доверительные границы выглядели как параллельные прямые, состоит в том, чтобы член второго порядка в выражении для $V(\widehat{Y}_0)$ оказался малым, а значит, при $150 \leqslant X_0 \leqslant 170$, какой бы ни был $\widehat{X}$, числитель члена второго порядка должен оставаться большим для заданного значения X_0. Таким образом, $S_{XX} = \Sigma (X_i - \overline{X})^2$ должно быть очень большим, и правильный ответ 3).

21. 1) $b_1^* = b_1 (1-q)$. Поскольку в таблице дисперсионного анализа $SS(b_1^* \mid b_0) = SS_{XY}^{*2}/S_{XX}^* = S_{XY}^2/S_{XX}$, ни в дисперсионном анализе, ни в F-критерии никаких изменений не произойдет. Однако подобранные значения коэффициентов и вычисления доверительных границ изменятся. Подставьте во все формулы $X_i/(1-q)$ вместо X_i.

2) В этом случае произойдут сложные изменения, вызванные тем фактом, что $(1-q_i)$ нельзя объединить при суммировании. Подставьте во все формулы $X_i/(1-q_i)$ вместо X_i.

3) Если это случилось, то подставьте $1-2q_i$ вместо $(1-q_i)^2$. Когда $q_i = q$, все q можно объединять при суммировании.

22. Мы можем построить прямую $\widehat{Y} = 9{,}930 - 0{,}010987X$. Вот таблица дисперсионного анализа:

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
b_0	1	4230,1602		
$b_1 \mid b_0$	1	1,1777	1,1777	0,73
Остаток	48	77,2321	1,6090	
неадекватность	29	53,9783	1,7993	1,39
«чистая» ошибка	19	23,2538	1,2919	
Общий	50	4308,5700		

$F(30,\ 18,\ 0{,}95) = 2{,}11$; $F(1,\ 48,\ 0{,}95) = 4{,}05$. Неадекватность не проявилась, и угловой коэффициент линии регрессии не значим.

Выводы. Эти данные *не* подтверждают идею о том, что продолжительность жизни зависит от длины «линии жизни».

Примечания. 1) Вклады в «чистую» ошибку при $X = 56$, 75 и 82 (относительно) крайне велики, поскольку относятся к двум крайним значениям отклика 6,45 и 13,20 соответственно. 2) Для улучшения анализа стоило бы принять во внимание сопутствующую переменную «рост» или что-нибудь еще в этом роде и скорректировать наблюдения с учетом этого возможного источника вариации.

23. Пункты 1) и 2) приводят вас к мысли о том, что в качестве предиктора стоит использовать *возраст*, и к рассмотрению возможных преобразований цены и/или возраста, вроде того, что в пункте 3).

3) $\mathbf{Y} = (3{,}91,\ 3{,}56,\ \ldots,\ 1{,}61)'$,

$\mathbf{Z} = (82,\ 72,\ \ldots,\ 12)'$,

$$\hat{Y} = (1{,}143181 + 0{,}0346564Z).$$

$$100e = (-8,\ -8,\ 27,\ -8,\ 25,\ 13,\ -31,\ 8,\ -32,\ 22,\ 3,\ -23,\ 6,\ 5)'.$$

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
b_0	1	80,6880	—	
$b_1 \mid b_0$	1	6,4075	6,4075	157,4
Остаток	12	0,4884	$s^2 = 0{,}0407$	
Общий	14	87,5859		

Регрессия высоко значима, а $R^2 = 0{,}9292$.

4) Заметим, что после того, как были сделаны преобразования, трудно ответить непосредственно на вопрос о прямолинейности. Ежегодный прирост цен равен b_1 (цена) и зависит не только от b_1, но и от цены. Поэтому проще всего сказать, что $Y = \ln$ (цены) возрастает, как b_1, на год.

5) Снова нужны некоторые предосторожности. Если мы предскажем Y (на основе нашего уравнения) для $Z = 38$, то получим $\hat{Y}_{38} = 2{,}46$. Вместе с тем в 1975 г. фактически $Y = ln\ (20) = 3{,}00$, т. е. наше предсказание оказалось на 0,56 меньше, а это довольно большое число. Если линейная зависимость и сохранится к 1975 г., то это вовсе не означает, что она останется той же, но на основе одной-единственной новой точки трудно сказать, то ли вся линия сдвинулась вверх, т. е. вырос только b_0 (что возможно), то ли увеличился только b_1 (что неправдоподобно, поскольку более молодой портвейн должен быть и более дешевым!), то ли, наконец, возросли и b_0, и b_1 (что возможно). Для дальнейшего исследования нужно больше данных.

24. Исчерпывающее обсуждение этого упражнения содержится в цитированной работе. О способе обобщения таких данных см.: Searle S. R., Firey P. A. Computer generation of data sets for homework exercises in simple regression.— American Statistician, 34, February 1980, p. 51—54.

25. Решение не приводится.

Глава 2 ● МАТРИЧНЫЙ ПОДХОД К ЛИНЕЙНОЙ РЕГРЕССИИ

2.0. ВВЕДЕНИЕ

Теперь мы изложим пример, приведенный в гл. 1, в терминах матричной алгебры. Применение матриц дает много преимуществ. Не последнее из них — общность. Как только задача записывается и разрешается в матричной форме, ее решение приложимо к любой регрессионной задаче такого рода независимо от того, сколько членов содержится в уравнении регрессии.

Матрица — это прямоугольная таблица символов или чисел, обычно обозначаемая с помощью одной полужирной буквы, например, **Q** или **q**. Существует несколько правил действий с такими таблицами. Довольно сложные выражения или уравнения могут быть зачастую представлены очень просто с помощью нескольких букв, надлежащим образом определенных и сгруппированных.

Мы не будем вводить матрицы формально, а используем их при изложении примера. Читатель, достаточно хорошо знакомый с матрицами, может сразу перейти к параграфу 2.5 или 2.6. Читатель, не имеющий таких знаний, должен тщательно проработать параграфы 2.1—2.5. В последующих параграфах он будет встречать некоторые места, возможно, трудные при первом чтении. Поэтому мы рекомендуем снова вернуться к ним после прочтения гл. 4. Свободное владение материалом, изложенным в параграфах 2.6—2.12, необходимо для того, чтобы справляться со сложными регрессионными задачами, но не обязательно для полного понимания последующих глав.

2.1. ПОДБОР УРАВНЕНИЯ ПРЯМОЙ В МАТРИЧНЫХ ОБОЗНАЧЕНИЯХ; ОЦЕНКИ ПАРАМЕТРОВ β_0 И β_1

Введем следующие определения:

$\mathbf{Y}$ — *вектор наблюдений* Y,

$\mathbf{X}$ — *матрица независимых переменных*,

$\boldsymbol{\beta}$ — *вектор параметров, подлежащих оцениванию*,

$\boldsymbol{\varepsilon}$ — *вектор ошибок*,

$\mathbf{1}$ — *вектор, образованный из единиц*.

Исходя из данных табл. 1.1 и уравнения (1.2.3) из гл. 1, мы можем записать для нашего основного примера:

$$\mathbf{Y} = \begin{bmatrix} 10{,}98 \\ 11{,}13 \\ 12{,}51 \\ 8{,}40 \\ \vdots \\ 10{,}36 \\ 11{,}08 \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & 35{,}3 \\ 1 & 29{,}7 \\ 1 & 30{,}8 \\ 1 & 58{,}8 \\ \vdots & \vdots \\ 1 & 33{,}4 \\ 1 & 28{,}6 \end{bmatrix},$$

$$\beta = \begin{vmatrix} \beta_0 \\ \beta_1 \end{vmatrix}, \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \vdots \\ \varepsilon_{24} \\ \varepsilon_{25} \end{bmatrix}. \qquad (2.1.1)$$

Заметим, что

$\mathbf{Y}$ есть 25×1-вектор,
$\mathbf{X}$ — 25×2-матрица,
β — 2×1-вектор,
ε — 25×1-вектор.

(Любая матрица, содержащая один столбец, называется вектор-столбцом; матрицу с одной строкой именуют вектор-строкой; 1×1-матрица есть просто обычное число или скаляр.)

С целью сокращения принято не указывать в матрицах или векторах все элементы, вместо отсутствующих элементов приводятся точки. Вектор-столбец $\mathbf{1}$, строго говоря, нам не требуется на этой стадии, но его удобно определить здесь. Он очень полезен при проведении матричных операций. Заметим, что матрица $\mathbf{X}$ состоит из двух вектор-столбцов. Первый из них есть просто $\mathbf{1}$, второй не имеет специального названия. Это вектор с элементами в виде величин X. В анализе данных его обычно называют «X-столбцом». Некоторые авторы называют также первый столбец в матрице $\mathbf{X}$ «X_0-столбцом», имея в виду, что он соответствует предикторной переменной X_0, тождественно равной 1. Переменные такого типа, значения которых выбираются в некотором смысле произвольно, обычно называют фиктивными. Они широко используются у нас в гл. 5 и 9 [1].

[1] Матричный подход является исключительно плодотворным для описания как линейных, так и нелинейных регрессий. Поэтому читатель, заинтересованный в использовании регрессионного анализа, должен не пожалеть усилий для ознакомления с теорией матриц. Для начальноиго знакомства с матрицами мы рекомендуем следующие книги: Сирл С., Госман У. Матричная ал-

Правила операций с матрицами

Правила перемножения матриц и векторов требуют, чтобы размеры перемножаемых матриц были *согласованными* между собой [2]. Так, например, если $\mathbf{A}$ есть $n \times p$-матрица, то она может быть:

1) первым сомножителем в произведении данной матрицы на $p \times q$-матрицу, результатом такого перемножения будет $n \times p \times p \times q = n \times q$-матрица;

2) вторым сомножителем в произведении $m \times n$-матрицы на данную матрицу, результатом такой операции будет $m \times n \times n \times p =$ = матрица.

Следовательно, произведения $\boldsymbol{\beta}\mathbf{X}$, например, не существует, так как $\boldsymbol{\beta}$ есть 2×1-матрица, а $\mathbf{X}$ — 25×2-матрица. Но $\mathbf{X}\boldsymbol{\beta}$ существует и выражается в следующей форме:

$$\mathbf{X}\boldsymbol{\beta} = \underset{25\times 2}{\begin{bmatrix} 1 & 35{,}3 \\ 1 & 29{,}7 \\ \cdot & \cdot \\ 1 & 28{,}6 \end{bmatrix}} \underset{2\times 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}} = \underset{25\times 1}{\begin{bmatrix} \beta_0 + 35{,}3\beta_1 \\ \beta_0 + 29{,}7\beta_1 \\ \cdot \\ \beta_0 + 28{,}6\beta_1 \end{bmatrix}}. \qquad (2.1.2)$$

В качестве более общего примера рассмотрим произведение

$$\underset{3\times 3}{\mathbf{A}} \quad \underset{3\times 2}{\mathbf{B}} \quad \underset{3\times 2}{\mathbf{C}}$$

$$\begin{bmatrix} 1 & 2 & 4 \\ -1 & 0 & 1 \\ 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & 1 \\ 3 & 5 \end{bmatrix} = \begin{bmatrix} 17 & 21 \\ 2 & 6 \\ 11 & 6 \end{bmatrix}.$$

гебра в экономике. Пер. с англ.— М.: Статистика, 1974.— 376 с.; Гильберт А. Как работать с матрицами. Пер. с нем.— М.: Статистика, 1981.— 160 с. Более подготовленный читатель может воспользоваться книгами: Гантмахер Ф. Р. Теория матриц.— М.: Наука, 1967.— 576 с.; Беллман Р. Введение в теорию матриц. Пер. с англ.— М.: Наука, 1969.— 368 с.; Ланкастер П. Теория матриц. Пер. с англ.— М.: Наука, 1978.— 280 с.; Ланцош К. Практические методы прикладного анализа. Пер. с англ.— М.: Физматгиз, 1961.— 524 с.; Кемени Дж., Снелл Дж., Томпсон Дж. Введение в конечную математику. Пер. с англ. Под ред. И. М. Яглома.— М.: Мир, 1965.— 487 с. (см. гл. 5, с. 259—356); Деруссо П., Рой Р., Клоуз Ч. Пространство состояний в теории управления/Пер. с англ. Под ред. М. В. Меерова.— М.: Наука, 1970.— 620 с. (см. гл. 4, с. 206—327); Graybill F. A. Introduction to Matrices with Applications in Statistics. —Belmont, California: Wadsworth Publ. Co, 1969.— 372 p.; Современная математика для инженеров/Под ред. Э. Ф. Беккенбаха; Пер. с англ. Под ред. И. Н. Векуа.— М.: ИЛ, 1959.— 500 с. (см. 305—345); Мишина А. П., Проскуряков И. В. Высшая алгебра.— М.: Физматгиз, 1962.— 300 с. Последние две книги могут служить справочными пособиями.— *Примеч. пер.*

[2] Согласованность размеров двух перемножаемых матриц состоит в том, что число столбцов первой матрицы (стоящей в произведении слева) должно быть равно числу строк второй (правой) матрицы.— *Примеч. пер.*

Чтобы найти элемент i-й строки и j-го столбца матрицы **C**, необходимо взять i-ю строку матрицы **A** и j-й столбец матрицы **B**, вычислить попарные произведения соответствующих элементов и сложить их. Например,

вторая строка матрицы **A** есть — 1 0 1,

первый столбец матрицы **B** есть 1 2 3.

В таком случае элемент второй строки и первого столбца матрицы **C** равен:

$$-1\ (1) + 0\ (2) + 1\ (3) = 2.$$

Определение. Если сумма парных произведений соответствующих элементов i-й строки первой матрицы и j-го столбца второй матрицы равна нулю, то говорят, что строка i первой матрицы *ортогональна* столбцу j второй матрицы. Аналогичное определение справедливо при перемножении строк или столбцов одной матрицы.

Сумма двух матриц или векторов есть просто матрица, элементы которой представляют собой суммы соответствующих элементов складываемых матриц или векторов. Например,

$$\mathbf{X\beta}+\boldsymbol{\varepsilon}=\begin{bmatrix}\beta_0+35{,}3\beta_1\\ \beta_0+29{,}7\beta_1\\ \cdots\\ \beta_0+28{,}6\beta_1\end{bmatrix}+\begin{bmatrix}\varepsilon_1\\ \varepsilon_2\\ \cdots\\ \varepsilon_{25}\end{bmatrix}=\begin{bmatrix}\beta_0+35{,}3\beta_1+\varepsilon_1\\ \beta_0+29{,}7\beta_1+\varepsilon_2\\ \cdots\\ \beta_0+28{,}6\beta_1+\varepsilon_{25}\end{bmatrix}. \quad (2.1.3)$$

Две матрицы или два вектора, которые суммируются, должны иметь одинаковые размеры. Разность между двумя матрицами определяется аналогично сказанному выше с заменой слова «сумма» на слово «разность». Если две матрицы или два вектора равны, то их соответствующие элементы также равны. Таким образом, из матричного уравнения

$$\mathbf{Y}=\mathbf{X\beta}+\boldsymbol{\varepsilon} \quad (2.1.4)$$

следует, что

$$\begin{aligned}10{,}98&=\beta_0+35{,}3\beta_1+\varepsilon\\ &\cdots\cdots\cdots\cdots\\ 11{,}08&=\beta_0+28{,}6\beta_1+\varepsilon_{25}\end{aligned} \quad (2.1.5)$$

или

$$Y_i=\beta_0+\beta_1X_i+\varepsilon_i\ \ (i=1,\ \ldots,\ 25) \quad (2.1.6)$$

для каждого из 25 наблюдений. Следовательно, матричное уравнение (2.1.4) и уравнение (2.1.6) выражают одну и ту же модель. Уравнение (2.1.6) идентично уравнению (1.2.3).

При представлении модели в матричной форме некоторые трудности для начинающих обычно связаны только с выбором матрицы **X**. Наиболее простой способ составления этой матрицы состоит в следующем. Надо выписать сначала все параметры модели в виде вектор-столбца $\boldsymbol{\beta}$, и тогда становится ясным, что соответствующие **X**-столбцы должны быть выбраны так, чтобы получить модель в данной алгебраи-

ческой форме из произведения $\mathbf{X}\boldsymbol{\beta}$. Например, если модель имеет вид $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$, то вектор $\boldsymbol{\beta}$ представляет собой столбец, содержащий упорядоченную последовательность элементов β_0, β_1 и β_{11}, а соответствующие **X**-столбцы должны формироваться из 1 (или значений X_0, если пользоваться таким обозначением) значений X и значений X^2. Следовательно i-я строка матрицы **X** будет иметь вид $(1, X_i, X_i^2)$, где X_i есть i-е из n наблюдений. Заметим, что при изменении порядка расположения элементов вектора $\boldsymbol{\beta}$ необходимо провести соответствующее переупорядочение вектор-столбцов матрицы **X**.

Транспонирование и обращение

Теперь мы определим операцию транспонирования матрицы. Транспонированная матрица — это матрица, которая получается из исходной, если ее строки записать в виде столбцов, сохранив порядок их расположения. Таким образом, строки и столбцы исходной матрицы совпадают с соответствующими столбцами и строками транспонированной матрицы. Результат транспонирования матрицы **M** записывается с помощью символа **M′**, например

$$\mathbf{M} = \underset{3\times 2}{\begin{bmatrix} 3 & 2 \\ 1 & 4 \\ 7 & 0 \end{bmatrix}}, \qquad \mathbf{M}' = \underset{2\times 3}{\begin{bmatrix} 3 & 1 & 7 \\ 2 & 4 & 0 \end{bmatrix}}$$

Поскольку

$$\boldsymbol{\varepsilon}' = (\varepsilon_1 \ \varepsilon_2, \ldots, \varepsilon_n),$$

можно записать

$$\varepsilon_1^2 + \varepsilon_2^2 + \ldots + \varepsilon_n^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}.$$

Аналогично

$$Y_1^2 + Y_2^2 + \ldots + Y_n^2 = \mathbf{Y}'\mathbf{Y},$$
$$n\overline{Y} = Y_1 + Y_2 + \ldots + Y_n = \mathbf{1}'\mathbf{Y},$$
$$n\overline{Y}^2 = (\Sigma Y_i)^2/n = \mathbf{Y}'\mathbf{1}\mathbf{1}'\mathbf{Y}/n.$$

Далее

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \ldots & 1 \\ 35{,}3 & 29{,}7 & \ldots & 28{,}6 \end{bmatrix} \begin{bmatrix} 1 & 35{,}3 \\ 1 & 29{,}7 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 28{,}6 \end{bmatrix} =$$

$$= \begin{bmatrix} 25 & 1315 \\ 1315 & 76323{,}42 \end{bmatrix}.$$

Вообще для модели в виде прямой линии справедливо

$$\mathbf{X'X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & X_n \end{bmatrix} = \begin{bmatrix} n & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2 \end{bmatrix}. \tag{2.1.7}$$

Кроме того,

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ 35{,}3 & 29{,}7 & \dots & 28{,}6 \end{bmatrix} \begin{bmatrix} 10{,}98 \\ 11{,}13 \\ \cdot \\ \cdot \\ \cdot \\ 11{,}08 \end{bmatrix} = \begin{bmatrix} 235{,}60 \\ 11821{,}4320 \end{bmatrix}.$$

Таким образом, в общем случае, при подгонке уравнения прямой линии

$$\mathbf{X'Y} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ X_1 & X_2 & \dots & X_n \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix} = \begin{bmatrix} \Sigma Y_i \\ \Sigma X_i Y_i \end{bmatrix}. \tag{2.1.8}$$

Это означает, что нормальные уравнения (1.2.8) могут быть записаны так:

$$\mathbf{X'Xb} = \mathbf{X'Y}, \tag{2.1.9}$$

где $\mathbf{b}' = (b_0, b_1)$. Решение этой системы уравнений дает МНК-оценки (b_0, b_1) параметров β_0, β_1. Но как теперь решить эти уравнения в матричной форме? Для этого надо ввести определение *обратной* матрицы. Такая матрица существует только тогда, когда исходная матрица является квадратной и ее определитель (величина, которую мы здесь определять не будем, но будем использовать в некоторых примерах) не равен нулю. Это последнее условие обычно выражается словами: *матрица является неособенной*. Мы будем полагать, что в наших задачах матрицы неособенные, если, конечно, не оговорено обратное. В регрессионных задачах приходится обращать матрицу $\mathbf{X'X}$. Если она особенная и, следовательно, для нее не существует обратной матрицы, то это обусловлено тем, что некоторые из нормальных уравнений являются линейными комбинациями остальных, см., например, уравнения (9.4.3). В этом случае мы будем иметь меньше уравнений, чем неизвестных, подлежащих определению. И здесь единственные оценки не могут быть получены, если на оцениваемые параметры не

наложены какие-либо дополнительные условия (дополнительные комментарии по этому поводу см. в гл. 9).

Допустим, теперь, что $\mathbf{M}$ есть неособенная $p \times p$-матрица. Матрица $\mathbf{M}^{-1}$, обратная к $\mathbf{M}$, есть $p \times p$ и такая, что

$$\mathbf{M}^{-1}\mathbf{M} = \mathbf{M}\mathbf{M}^{-1} = \mathbf{I}_p,$$

где $\mathbf{I}_p$ — *единичная матрица порядка* p, которая содержит единицы на всех позициях главной диагонали (т. е. диагонали, идущей из левого верхнего угла в правый нижний) и нули на остальных позициях. Например,

$$\mathbf{I}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Если порядок единичный матрицы очевиден, то индекс часто опускается. Единичная матрица играет такую же роль в перемножении матриц, как и единица в перемножении чисел: она оставляет сомножители неизменными. Обратная матрица единственна.

Формулы для обращения матриц порядка 2 и 3 имеют вид:

$$\mathbf{M}^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \begin{bmatrix} d/D & -b/D \\ -c/D & a/D \end{bmatrix}, \tag{2.1.10}$$

где $D = ad - bc$ есть детерминант (определитель) 2×2-матрицы $\mathbf{M}$.

$$\mathbf{Q}^{-1} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & k \end{bmatrix}^{-1} = \begin{bmatrix} A & B & C \\ D & E & F \\ G & H & K \end{bmatrix}, \tag{2.1.11}$$

где

$$A = (ek - fh)/Z, \quad B = -(bk - ch)/Z, \quad C = (bf - ce)/Z,$$
$$D = -(dk - fg)/Z, \quad E = (ak - cg)/Z, \quad F = -(af - cd)/Z,$$
$$G = (dh - eg)/Z, \quad H = -(ah - bg)/Z, \quad K = (ae - bd)/Z$$

и где

$$Z = a(ek - fh) - b(dk - fg) + c(dh - eg) =$$
$$= aek + bfg + cdh - ahf - dbk - gec$$

есть детерминант матрицы $\mathbf{Q}$[3].

Матрицы вида $\mathbf{X}'\mathbf{X}$, встречающиеся в регрессионных задачах, всегда симметричны. У этой матрицы элемент i-й строки и j-го столбца

[3] Различные методы вычисления определителей описаны в книгах, посвященных матрицам (см. примечание к с. 105).— *Примеч. пер.*

равен элементу j-строки и i-го столбца [4]. Следовательно, транспонирование симметричной матрицы не меняет ее. Это легко видеть, если применить общее правило $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$ для транспонирования произведения матриц. Так как $(\mathbf{A}')' = \mathbf{A}$, мы можем записать $(\mathbf{X}'\mathbf{X})' = \mathbf{X}'\mathbf{X}$. (Разбирая некоторые простые численные примеры, мы разъясним еще этот момент.) Если матрица $\mathbf{M}$ порядка 2 симметрична, то $b = c$ и обратная матрица будет также симметричной. Если матрица $\mathbf{Q}$, упомянутая выше, симметрична, то $b = d$, $c = g$, $f = h$. Тогда, переобозначая матрицу $\mathbf{Q}$ в $\mathbf{S}$, мы получим симметричную обратную матрицу

$$\mathbf{S}^{-1} = \begin{bmatrix} a & b & c \\ b & e & f \\ c & f & k \end{bmatrix}^{-1} = \begin{bmatrix} A & B & C \\ B & E & F \\ C & F & K \end{bmatrix}, \qquad (2.1.12)$$

где

$$A = (ek - f^2)/Y, \qquad B = -(bk - cf)/Y, \qquad C = (bf - ce)/Y,$$
$$E = (ak - c^2)/Y, \qquad F = -(af - bc)/Y, \qquad K = (ae - b^2)/Y$$

и где

$$Y = a(ek - f^2) - b(bk - cf) + c(bf - ce) = aek + 2bcf - af^2 - b^2k - c^2e$$

есть детерминант матрицы $\mathbf{S}$. Обратная матрица от любой симметричной матрицы есть, следовательно, симметричная матрица.

Матрицы, имеющие порядок больше трех, обычно трудно обращать, если они не имеют специальной формы. Матрица, которая легко обращается независимо от порядка,— это *диагональная* матрица, содержащая ненулевые элементы только на главной диагонали при условии, что остальные элементы — нули. Обратная матрица от нее получается путем обращения всех ненулевых элементов и сохранения их на тех же позициях, что и в исходной матрице. Например,

$$\begin{bmatrix} a_1 & & & 0 \\ & a_2 & & \\ & & \ddots & \\ 0 & & & a_n \end{bmatrix}^{-1} = \begin{bmatrix} 1/a_1 & & & 0 \\ & 1/a_2 & & \\ & & \ddots & \\ 0 & & & 1/a_n \end{bmatrix}. \qquad (2.1.13)$$

(Заметим, что в данном случае цифра $\mathbf{0}$ используется для обозначения больших треугольных блоков, состоящих из нулей. Что, впрочем, часто очевидно.) Этот результат мы используем в гл. 7.

Другой случай упрощения имеет место, когда некоторые столбцы матрицы $\mathbf{X}$ ортогональны ко *всем* остальным столбцам. Матрица $\mathbf{X}'\mathbf{X}$ приобретает тогда блочную форму

$$\begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix},$$

[4] Иными словами, здесь имеет место симметрия элементов квадратной матрицы относительно ее главной диагонали, соединяющей левый верхний элемент с правым нижним.— *Примеч. пер.*

где, например, **P** может быть $p \times p$-матрицей, **R** — $r \times r$-матрицей, а символ **0** в данном случае используется для обозначения блоков разного размера, состоящих из нулей. При этом блок, стоящий в правом верхнем углу, имеет размер $p \times r$, а блок, находящийся в левом нижнем углу,— $r \times p$-матрица. Обратная матрица имеет вид

$$\begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{P}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}^{-1} \end{bmatrix}. \tag{2.1.14}$$

Например, если

$$\mathbf{P} = \begin{bmatrix} 1 & 3 \\ 2 & 8 \end{bmatrix}, \qquad \mathbf{P}^{-1} = \begin{bmatrix} 4 & -3/2 \\ -1 & 1/2 \end{bmatrix},$$

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 3 & 2 \\ 4 & 1 & 1 \end{bmatrix}, \qquad \mathbf{R}^{-1} = \begin{bmatrix} -1/9 & -1/9 & 1/3 \\ -2/3 & 1/3 & 0 \\ 10/9 & 1/9 & -1/3 \end{bmatrix},$$

то

$$\begin{bmatrix} 1 & 3 & 0 & 0 & 0 \\ 2 & 8 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 2 & 3 & 2 \\ 0 & 0 & 4 & 1 & 1 \end{bmatrix}^{-1} = \begin{bmatrix} 4 & -\frac{3}{2} & 0 & 0 & 0 \\ -1 & \frac{1}{2} & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{9} & -\frac{1}{9} & \frac{1}{3} \\ 0 & 0 & -\frac{2}{3} & \frac{1}{3} & 0 \\ 0 & 0 & \frac{10}{9} & \frac{1}{9} & -\frac{1}{3} \end{bmatrix}.$$

Если имеется больше двух ненулевых блоков, то возможно очевидное обобщение. Важно при этом заметить, что ненулевые блоки должны стоять только на главной диагонали, а внедиагональные блоки должны быть нулевыми [5]. Лишь в этом случае обобщение приведенного выражения возможно.

Формула обращения (2.1.14) применима даже в том случае, когда строки и столбцы, включающие ненулевые элементы, перемешаны, при условии, что матрица может быть разбита на части (подобные **P** и **R**, указанным выше), полностью отделяющиеся друг от друга с по-

[5] Надо отметить, однако, что этот прием носит довольно частный характер и применяется редко. Существует множество других, более общих приемов обращения матриц, описанных в литературе по матрицам (см. примечание к с. 105), а также в литературе по вычислительной математике: Фадеев Д. К., Фадеева В. Н. Вычислительные методы линейной алгебры.— М.: Физматгиз, 1963.— 763 с. (см. гл. 2, с. 179—183 и др.); Райс Дж. Матричные вычисления и математическое обеспечение/Пер. с англ. Под ред. В. В. Воеводина.— М.: Мир, 1984.— 264 с.,; Maindonald J. H. Statistical Computation.— New York: John Wiley and Sons, 1984.— 370 p. (см. p. 40—43).— *Примеч. пер.*

мощью нулей. Например, используя те же числа, запишем матрицу

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 3 & 0 \\ 2 & 0 & 3 & 0 & 2 \\ 0 & 2 & 0 & 8 & 0 \\ 4 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

Ее можно представить в блочном виде и обращать отдельные блоки независимо. Заметим, что вторая и четвертая строки *и* столбцы полностью изолированы или отделены от первого, третьего и пятого столбцов нулями. Следовательно, ненулевые элементы во второй и четвертой строках и столбцах образуют 2×2-матрицу, которая может быть обращена независимо, в то время как остальные ненулевые элементы образуют совершенно изолированную 3×3-матрицу, которую тоже можно обратить отдельно. Таким образом, обратная матрица имеет вид

$$\begin{bmatrix} -\frac{1}{9} & 0 & -\frac{1}{9} & 0 & \frac{1}{3} \\ 0 & 4 & 0 & -\frac{3}{2} & 0 \\ -\frac{2}{3} & 0 & \frac{1}{3} & 0 & 0 \\ 0 & -1 & 0 & \frac{1}{2} & 0 \\ \frac{10}{9} & 0 & \frac{1}{9} & 0 & -\frac{1}{3} \end{bmatrix}.$$

Результаты такого типа будут приведены в гл. 7. Корректность всех этих операций обращения можно подтвердить с помощью умножения исходной матрицы на обратную как слева, так и справа. В итоге должна получиться единичная матрица соответствующего порядка. На практике, когда размер матрицы превышает 3×3 и матрица не обладает формой, допускающей упрощение, нахождение обратной матрицы может стать громоздкой процедурой. Эту работу обычно проводят с помощью вычислительной машины. Некоторые авторы (см. библиографию) предлагают «ручные» методы обращения матриц, но мы с ними здесь дела иметь не будем [6].

Теперь мы должны обратить матрицу $\mathbf{X}'\mathbf{X}$ в нашем примере. Это матрица размера 2×2 и общей формы, соответствующей выражению

---

[6] С «ручными» приемами обращения матриц можно познакомиться в кн.: Гильберт А. Как работать с матрицами. Пер. с нем.— М.: Статистика, 1981.— 160 с.— *Примеч. пер.*

(2.1.7). Используя уравнение (2.1.10), найдем обратную матрицу в виде

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} \dfrac{\Sigma X_i^2}{n\Sigma (X_i - \overline{X})^2} & \dfrac{-\overline{X}}{\Sigma (X_i - \overline{X})^2} \\ \dfrac{-\overline{X}}{\Sigma (X_i - \overline{X})^2} & \dfrac{1}{\Sigma (X_i - \overline{X})^2} \end{bmatrix}. \qquad (2.1.15)$$

Если *каждый* элемент матрицы содержит общий сомножитель, то его можно вынести перед матрицей как сомножитель. (И, наоборот, если матрица умножается на константу C, то это значит, что для получения итоговой матрицы надо каждый элемент исходной матрицы умножить на C.) Следовательно, альтернативная форма имеет вид

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{n\Sigma (X_i - \overline{X})^2} \begin{bmatrix} \Sigma X_i^2 & -\Sigma X_i \\ -\Sigma X_i & n \end{bmatrix}. \qquad (2.1.16)$$

Так как матрица $\mathbf{X}'\mathbf{X}$ симметрична, обратная матрица $(\mathbf{X}'\mathbf{X})^{-1}$, как указывалось выше, также симметрична. Величина, стоящая перед скобкой в (2.1.16), есть обратная величина от детерминанта матрицы $\mathbf{X}'\mathbf{X}$, который обозначается в виде $\det(\mathbf{X}'\mathbf{X})$ или $|\mathbf{X}'\mathbf{X}|$. Принимая за основу (2.1.15) и пользуясь данными нашего примера, мы найдем, что

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0{,}4267941 & -0{,}0073535 \\ -0{,}0073535 & 0{,}0001398 \end{bmatrix}.$$

Решение нормальных уравнений

Если умножить обе части уравнения (2.1.9) на $(\mathbf{X}'\mathbf{X})^{-1}$ слева, то мы получим

$$(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\,\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

т. е.

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \qquad (2.1.17)$$

поскольку $(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{X} = \mathbf{I}$. Этот важный результат следует запомнить, так как решение нормальных уравнений для линейной регрессии всегда может быть записано в этой форме при условии, что $\mathbf{X}'\mathbf{X}$ — неособенная матрица и регрессионная задача сформулирована правильно.

Пользуясь данными нашего примера, мы найдем, что

$$\mathbf{b} = \begin{bmatrix} 0{,}4267941 & -0{,}0073535 \\ -0{,}0073535 & 0{,}0001398 \end{bmatrix} \begin{bmatrix} 235{,}60 \\ 11821{,}4320 \end{bmatrix} = \begin{bmatrix} 13{,}623790 \\ -0{,}079848 \end{bmatrix}.$$

Заметим, что полученные результаты не совпадают — с точностью до шестого знака после запятой — с величинами, найденными в параграфе 1.2. Такие несоответствия встречаются часто, они связаны

с округлением чисел при вычислениях. Пренебрежение этим при определенных условиях может привести к серьезным ошибкам. В данном случае численные расхождения малы с практической точки зрения, но они указывают на то, что, в общем, в регрессионных вычислениях следует учитывать столько знаков, сколько возможно. Иногда числа в расчетах таковы, что из-за округлений смысл расчетов будет вообще полностью потерян.

Некоторые способы проведения вычислений (особенно если они выполняются вручную, т. е. с помощью настольных калькуляторов) лучше, чем остальные, так как они меньше зависят от ошибок округления. В частности, целесообразно откладывать операцию деления на самый конец, если это возможно. Например, если бы мы воспользовались уравнением (2.1.16) вместо (2.1.15) для нахождения $(\mathbf{X}'\mathbf{X})^{-1}$, то мы получили бы

$$(\mathbf{X}'\mathbf{X})^{-1} = \frac{1}{178860{,}5}\begin{bmatrix} 76323{,}42 & -1315 \\ -1315 & 25 \end{bmatrix}.$$

Затем мы могли бы найти $\mathbf{b}$ из уравнения

$$\mathbf{b} = \frac{1}{178860{,}5}\begin{bmatrix} 76323{,}42 & -1315 \\ -1315 & 25 \end{bmatrix} \times \begin{bmatrix} 235{,}60 \\ 11821{,}432 \end{bmatrix} =$$

$$= \frac{1}{178860{,}5}\begin{bmatrix} 2436614{,}672 \\ -14278{,}2 \end{bmatrix} = \begin{bmatrix} 13{,}622989 \\ -0{,}079829 \end{bmatrix},$$

выполняя операцию деления в последнюю очередь.

Вычисляя коэффициенты тремя различными способами, получим:

	По формулам из параграфа 1.2	Обращением матрицы	Обращением матрицы (с делением в конце)
b_0	13,623005	13,623790	13,622989
b_1	—0,079829	—0,079848	—0,079829

Как мы уже говорили, эти расхождения не имеют большого значения в данном примере. Третий метод обычно наиболее точный. Чтобы увидеть, к каким последствиям может привести округление, мы предлагаем читателю найти обратную матрицу вторым способом, а округление элементов выполнить несколькими путями, например округляя до 6, 5, 4 или 3-го знака после запятой. Ошибки округления — основная причина расхождений, если одна и та же задача решается разными людьми с применением настольных калькуляторов.

Если программа регрессионного анализа написана для вычислительной машины, то при этом автоматически сохраняется много значащих цифр. Тем не менее некоторые программы ориентированы на проведение вычислений с удвоенной точностью, хотя этого обычно и не требуется (см. параграф 5.4).

Р е з ю м е. Если мы выразим одномерную линейную модель, подлежащую оцениванию на основе данных нашего примера, в форме

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

отвечающей уравнению (2.1.4), то МНК-оценки для параметров β_0, β_1, т. е. МНК-оценка вектора $\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$, выражаются формулой

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Этот результат является очень важным и его следует запомнить. Заметим, что оценка $\hat{\mathbf{Y}}$ получается из выражения

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}.$$

2.2. ДИСПЕРСИОННЫЙ АНАЛИЗ В МАТРИЧНЫХ ОБОЗНАЧЕНИЯХ

Напомним (см. параграф 1.3), что в таблицу дисперсионного анализа наиболее общего вида мы вписывали

$$\mathrm{SS}(b_1 \mid b_0) = b_1\left[\Sigma X_i Y_i - \frac{(\Sigma X_i)(\Sigma Y_i)}{n}\right] = b_1\left[\Sigma X_i Y_i - n\bar{X}\bar{Y}\right],$$

$$\mathrm{SS}(b_0) = \text{коррекция на среднее} = \frac{(\Sigma Y_i)^2}{n} = n\bar{Y}^2.$$

Каждая из этих сумм квадратов имеет одну степень свободы. Далее

$$\begin{aligned}\mathrm{SS}(b_1 \mid b_0) + \mathrm{SS}(b_0) &= b_1\Sigma X_i Y_i - b_1 n\bar{X}\bar{Y} + n\bar{Y}^2 = \\ &= b_1\Sigma X_i Y_i + n\bar{Y}(\bar{Y} - b_1\bar{X}) = b_1\Sigma X_i Y_i + b_0\Sigma Y_i = \\ &= (b_0,\ b_1)\begin{bmatrix} \Sigma Y_i \\ \Sigma X_i Y_i \end{bmatrix} = \mathbf{b}'\mathbf{X}'\mathbf{Y} \qquad (2.2.1)\end{aligned}$$

в матричных обозначениях с двумя степенями свободы. Следовательно, можно записать таблицу дисперсионного анализа в матричной форме:

Источник	Число степеней свободы	SS	MS
$\mathbf{b}' = (b_0,\ b_1)$ Остаток	2 $n-2$	$\mathbf{b}'\mathbf{X}'\mathbf{Y}$ $\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$	s^2.
Общий (нескорректированный)	n	$\mathbf{Y}'\mathbf{Y}$	

Этим способом мы можем расщепить общую вариацию $\mathbf{Y}'\mathbf{Y}$ на две части. Первая из них обусловлена регрессионной зависимостью, которая оценивается. Вторая связана с остатком и отражает вариации

точек около линии регрессии. Для того чтобы найти, какая часть общей вариации может быть отнесена за счет добавления члена $\beta_1 X_t$ к более простой модели $Y_t = \beta_0 + \varepsilon_i$, необходимо вычесть корректирующий фактор $n\overline{Y}^2$ из суммы квадратов $\mathbf{b}'\mathbf{X}'\mathbf{Y}$, в итоге получим SS $(b_1|b_0)$, как это следует из (2.2.1). Величина $n\overline{Y}^2$ представляет собой $\mathbf{b}'\mathbf{X}'\mathbf{Y}$, если речь идет о подборе модели $Y_i = \beta_0 + \varepsilon_t$. Остаток от $\mathbf{b}'\mathbf{X}'\mathbf{Y}$, следовательно, представляет собой *дополнительную сумму квадратов, обусловленную исключением* из модели слагаемого b_1. Если из параллельных опытов известна оценка «чистой» ошибки, из остаточной суммы квадратов вычитается соответствующая величина и получаются такие же разложения и критерии, которые были описаны в параграфе 1.5.

Пример. Для нашего основного примера мы имели

$$\mathbf{b} = \begin{bmatrix} 13{,}62 \\ -0{,}0798 \end{bmatrix}, \qquad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} 235{,}60 \\ 11821{,}4320 \end{bmatrix}.$$

Следовательно,

$$\text{SS}(\mathbf{b}) = \mathbf{b}'\mathbf{X}'\mathbf{Y} = 2265{,}5217,$$

$$\text{SS}(b_0) = (\Sigma Y_i)^2/n = 2220{,}2944,$$

SS $(b_1 \mid b_0) =$ SS $(b_1,$ после поправки на $b_0) = \mathbf{b}'\mathbf{X}'\mathbf{Y} - (\Sigma Y_i)^2/n = 45{,}2273$

Прежде мы получили 45,59, так что разница составляет 0,36. Это снова связано с округлением чисел при счете и указывает фактически на то, что даже при простых регрессионных вычислениях рационально сохранять столько значащих цифр, сколько возможно.

Заметим, что SS $(b_1|b_0)$ может быть записана в матричной форме:

$$\text{SS}(b_1 \mid b_0) = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{1}\mathbf{1}'\mathbf{Y}/n = \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{1}\mathbf{1}'\mathbf{Y}/n =$$

$$= \mathbf{Y}'(\mathbf{R} - \mathbf{1}\mathbf{1}'/n)\,\mathbf{Y},$$

где $\mathbf{R} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ — есть обычная симметричная идемпотентная матрица (см. с. 164), которая часто встречается в работах по регрессионному анализу. При замене $\mathbf{b}'$ на $\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$, как это сделано выше, мы используем важное правило, состоящее в том, что при транспонировании произведения матриц получается произведение транспонированных матриц, записанных в *обратном порядке*. В общем случае, например, имеем

$$(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}'.$$

Если мы применим это правило к выражению $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, обозначив $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}$, $\mathbf{B} = \mathbf{X}'$ и $\mathbf{C} = \mathbf{Y}$, и учтем при этом, что $(\mathbf{X}'\mathbf{X})^{-1}$ есть симметричная матрица, а также, что $(\mathbf{X}')' = \mathbf{X}$ (транспонирование транспонированной матрицы приводит к исходной матрице), то получим указанный выше результат.

Приведем также некоторые результаты матричного регрессионного анализа, проверяя которые читатель может убедиться в своем умении обращаться с матрицами [7]:

[7] Вторая из трех приведенных ниже формул справедлива лишь при наличии свободного члена в модели. —*Примеч. пер.*

$$\mathbf{e} = (\mathbf{I} - \mathbf{R})\,\mathbf{Y},$$
$$\mathbf{e}'\mathbf{1} = \mathbf{1}'\mathbf{e} = 0,$$
$$\mathbf{e}'\hat{\mathbf{Y}} = \hat{\mathbf{Y}}'\mathbf{e} = 0.$$

2.3. ДИСПЕРСИЯ И КОВАРИАЦИЯ КОЭФФИЦИЕНТОВ НА ОСНОВЕ МАТРИЧНЫХ ВЫЧИСЛЕНИЙ

Напомним, что $V(b_1) = \sigma^2/\Sigma\,(X_i - \bar{X})^2$. Кроме того,

$$V(b_0) = V(\bar{Y} - b_1\bar{X}) = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma\,(X_i - \bar{X})^2}\right] =$$
$$= \frac{\sigma^2\Sigma X_i^2}{n\Sigma\,(X_i - \bar{X})^2},$$

поскольку, как указывалось ранее, $\bar{Y}$ и b_1 имеют нулевую ковариацию, а величины X предполагаются постоянными. В дополнение имеем

$$\operatorname{cov}(b_0,\ b_1) = \operatorname{cov}[(\bar{Y} - b_1\bar{X}),\ b_1] = -\bar{X}V(b_1) = -\bar{X}\sigma^2/\Sigma(X_i - \bar{X})^2.$$

Таким образом, мы можем записать *матрицу дисперсий-ковариаций вектора* $\mathbf{b}$ следующим образом:

$$\mathbf{V}(\mathbf{b}) = \mathbf{V}\begin{pmatrix} b_0 \\ b_1 \end{pmatrix} = \begin{bmatrix} V(b_0) & \operatorname{cov}(b_0,\ b_1) \\ \operatorname{cov}(b_0,\ b_1) & V(b_1) \end{bmatrix} =$$
$$= \begin{bmatrix} \dfrac{\sigma^2\Sigma X_i^2}{n\Sigma\,(X_i - \bar{X})^2} & -\dfrac{\bar{X}\sigma^2}{\Sigma\,(X_i - X)^2} \\ -\dfrac{\bar{X}\sigma^2}{\Sigma\,(X_i - \bar{X})^2} & \dfrac{\sigma^2}{\Sigma\,(X_i - \bar{X})^2} \end{bmatrix}. \qquad (2.3.1)$$

Далее, если все элементы матрицы имеют одинаковый сомножитель, то мы можем вынести его как общий сомножитель перед матрицей. Так мы можем поступить с коэффициентом σ^2. Матрица, которая остается, есть $(\mathbf{X}'\mathbf{X})^{-1}$ из уравнения (2.1.15). Следовательно,

$$\mathbf{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2. \qquad (2.3.2)$$

Это важный результат и его стоит запомнить. Если σ^2 неизвестна, то мы можем использовать вместо нее s^2, т. е. оценку величины σ^2 (получаемую из таблицы дисперсионного анализа, когда модель адекватна), или величину s_e^2, т. е. дисперсию, связанную с «чистой» ошибкой, если модель не адекватна. Это приводит нас к оценке матрицы или *выборочной матрице дисперсий-ковариаций вектора* $\mathbf{b}$ (см. также параграф 2.12).

2.4. ДИСПЕРСИЯ ВЕЛИЧИНЫ $\hat{Y}$ В МАТРИЧНЫХ ОБОЗНАЧЕНИЯХ

Пусть X_0 будет некоторым заранее выбранным значением величины X. Предсказываемое среднее значение величины Y при данном значении величины X есть

$$\hat{Y}_0 = b_0 + b_1X_0.$$

Введем далее вектор

$$\mathbf{X}_0' = (1,\ X_0).$$

Следовательно, мы можем записать

$$\hat{Y}_0 = (1,\ X_0)\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \mathbf{X}_0'\mathbf{b} = \mathbf{b}'\mathbf{X}_0.$$

Так как $\hat{Y}_0$ есть линейная комбинация случайных переменных b_0 и b_1,

$$V(\hat{Y}_0) = V(b_0) + 2X_0 \operatorname{cov}(b_0,\ b_1) + X_0^2 V(b_1).$$

Используя произведения указанных выше матриц и векторов, можно выразить величину $V(\hat{Y}_0)$ в альтернативной форме:

$$V(\hat{Y}_0) = [1,\ X_0]\begin{bmatrix} V(b_0) & \operatorname{cov}(b_0,\ b_1) \\ \operatorname{cov}(b_0,\ b_1) & V(b_1) \end{bmatrix} \times$$

$$\times \begin{bmatrix} 1 \\ X_0 \end{bmatrix} = \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\sigma^2.$$

Несмотря на другой вид, это выражение тождественно величине, получаемой с помощью формулы (1.4.9). Этот важный результат также следует запомнить. С соответствующим переобозначением величин $\mathbf{X}_0$ и $\mathbf{X}$ приведенная формула применима для случая общей линейной регрессии. Оценка дисперсии (выборочная дисперсия) получается, если σ^2 заменить оценкой s^2.

2.5. РЕЗЮМЕ К МАТРИЧНОМУ ПОДХОДУ ПРИ ПОДБОРЕ ПРЯМОЙ

1. Представим модель в форме $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$.
2. Найдем МНК-оценку $\mathbf{b}$ вектора $\boldsymbol{\beta}$, используя имеющиеся данные по формуле $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$.
3. Составим сумму квадратов $\mathbf{b}'\mathbf{X}'\mathbf{Y}$, связанную с коэффициентами, и таким образом получим следующую таблицу дисперсионного анализа:

Источник	Число степеней свободы	SS	MS
Регрессия	2	$\mathbf{b}'\mathbf{X}'\mathbf{Y}$	
Остаток	$n-2$	$\mathbf{Y}'\mathbf{Y}-\mathbf{b}'\mathbf{X}'\mathbf{Y}$	s^2 (оценка σ^2, если модель корректна)
Общий	n	$\mathbf{Y}'\mathbf{Y}$	

Дополнительное разбиение сумм квадратов достигается в результате определения величины $\mathrm{SS}(b_1|b_0)$, т. е. дополнительной суммы

квадратов, обусловленной b_1, и введения «чистой» ошибки. Более детальная таблица дисперсионного анализа будет иметь форму:

Источник	Число степеней свободы	SS	MS
SS (b) Средн (b_0)	1	$n\bar{Y}^2$	
SS ($b_1 \mid b_0$)	1	$\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2$	MS_L
Остаток* 1	$n-2-n_e$	$\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$—SS (р. е.)	s^2
Остаток* 2	n_e	SS (р. е.)	s_e^2
Общий	n	$\mathbf{Y}'\mathbf{Y}$	

* Остаток 1 — неадекватность, остаток 2 — «чистая» ошибка. — *Примеч. пер.*

Вторая таблица нередко записывается так, что в последней строке фигурирует *скорректированная* общая сумма квадратов, при этом сумма квадратов $n\bar{Y}^2$, связанная со средним, опускается. (Между прочим, как мы указывали ранее, можно записать величину $n\bar{Y}^2$ в матричной форме как $\mathbf{Y}'\mathbf{1}\mathbf{1}'\mathbf{Y}/n$, хотя это делать и необязательно. Однако вычисления по этой формуле менее чувствительны к ошибкам округления, чем при использовании выражения $(\Sigma Y_i)^2/n$.) Сокращенная таблица имеет вид:

Источник	Число степеней свободы	SS	MS
SS ($b_1 \mid b_0$)	1	$\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2$	
Остаток: неадекватность	$n-2-n_e$	$\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$—SS (р. е.)	MS_L
Остаток: «чистая» ошибка	n_e	SS (р. е.)	s_e^2
Общий, скорректированный	$n-1$	$\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2$	

Проверка гипотез о неадекватности модели и о параметре β_1 проводится, как описано в гл. 1. В качестве дополнительной меры, характеризующей вклад регрессии, может служить отношение

$$R^2 = \frac{(\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2)}{(\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2)}.$$

4. Если не выявлено отсутствие согласия модели и экспериментальных данных (т. е. модель признается адекватной), то матрица $(\mathbf{X}'\mathbf{X})^{-1}s^2$ дает оценки $V(b_0)$, $V(b_1)$ и $\mathrm{cov}(b_0, b_1)$ и позволяет прове-

рить гипотезы относительно отдельных коэффициентов или провести другие операции, как описано в гл. 1.

5. Могут быть найдены следующие величины:

вектор предсказываемых значений: $\hat{\mathbf{Y}} = \mathbf{Xb}$,

предсказываемое значение Y при X_0: $\hat{Y}_0 = \mathbf{X}_0'\mathbf{b} = \mathbf{b}'\mathbf{X}_0$

с дисперсией

$$V(\hat{Y}_0) = \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\sigma^2.$$

2.6. СЛУЧАЙ ОБЩЕЙ РЕГРЕССИИ

Мы показали, как можно решить проблему подбора уравнения прямой методом наименьших квадратов, используя матрицы. Этот подход важен по следующим соображениям. Если мы хотим подобрать с помощью метода наименьших квадратов *любую* модель, линейную по параметрам, то вычисления необходимо проводить точно по тем же матричным формулам, как и при оценивании уравнения прямой линии, содержащей лишь два параметра: β_0 и β_1. Однако сложность вычислений с увеличением числа параметров резко возрастает. Таким образом, хотя формулы и легко запоминаются, почти во всех случаях приходится прибегать к помощи цифровых вычислительных машин. Исключение составляют случаи, когда:

1) число параметров мало, скажем, меньше пяти;

2) обрабатываемые данные получены на основе заранее спланированных экспериментов, что приводит к матрице $\mathbf{X}'\mathbf{X}$ простого или специального вида.

Дадим теперь общее изложение методов линейной регрессии. Для ознакомления с теоретическим обоснованием этих результатов читатель может обратиться, например, к книге[8]: Plackett R. L. Regression Analysis.— Oxford: Clarendon Press, 1960.

---

[8] А также к кн.: Линник Ю. В. Метод наименьших квадратов и основы теории обработки наблюдений.— М.: Физматгиз, 1962.— 350 с.; Себер Дж. Линейный регрессионный анализ/Пер. с англ. Под ред. М. Б. Малютова.— М.: Мир, 1980.— 456 с.; Демиденко Е. З. Линейная и нелинейная регрессии.— М.: Финансы и статистика, 1981.— 304 с.; Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия/Пер. с англ. Под ред. Ю. П. Адлера.— М.: Финансы и статистика, 1982, вып. 1.— 320 с.; вып. 2.— 239 с.; Петрович М. Л. Регрессионный анализ и его математическое обеспечение на ЕС ЭВМ.— М.: Финансы и статистика, 1982.— 200 с.; Фёрстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа. Пер. с нем.— М.: Финансы и статистика, 1983.— 304 с.; Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика.— М.: Финансы и статистика, 1983.— 472 с.; Wolberg J. R. Prediction Analysis.— Princeton: D. Van Nostrand Co, 1967.— 291 p.; Weisberg S. Applied Linear Regression.— New York; John Wiley and Sons, 1980.— 280 p. Последние достижения в методологии регрессионного анализа нашли отражение в обстоятельном обзоре: Hocking R. R. Developments in Liner Regression Methodology. 1959—1982.— Technometrics, 1983, 25, N 3, p. 219—230.— *Примеч. пер.*

Предположим, что мы имеем модель, подлежащую исследованию, и она может быть представлена в виде

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon},$$

где $\mathbf{Y}$ — $(n \times 1)$-вектор наблюдений; $\mathbf{X}$ — $(n \times p)$-матрица с известными численными элементами; β — $(p \times 1)$-вектор параметров; $\boldsymbol{\varepsilon}$ — $(n \times 1)$-вектор ошибок и где $E(\boldsymbol{\varepsilon}) = 0$, $V(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma^2$, так что элементы вектора $\boldsymbol{\varepsilon}$ некоррелированы.

Поскольку $E(\boldsymbol{\varepsilon}) = 0$, альтернативная форма записи модели имеет вид

$$E(\mathbf{Y}) = \mathbf{X}\beta. \tag{2.6.1}$$

Сумма квадратов ошибок равна:

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{Y}'\mathbf{Y} - \beta'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\beta + \beta'\mathbf{X}'\mathbf{X}\beta =$$
$$= \mathbf{Y}'\mathbf{Y} - 2\beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta. \tag{2.6.2}$$

(Это вытекает из того, что $\beta'\mathbf{X}'\mathbf{Y}$ есть (1×1)-матрица или скаляр; ее транспонирование ничего не изменяет $(\beta'\mathbf{X}'\mathbf{Y})' = \mathbf{Y}'\mathbf{X}\beta$.)

МНК-оценка вектора β есть вектор $\mathbf{b}$, который при подстановке в (2.6.2) доставляет минимум величине $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$. Эту оценку можно найти, дифференцируя выражение (2.6.2) по β и приравнивая результирующее матричное выражение к нулевому вектору. Причем надо заменить β на $\mathbf{b}$. (Дифференцирование $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$ по вектору β эквивалентно ее дифференцированию отдельно по каждому элементу вектора, последовательной записи получаемых выражений (одно под другим) с дальнейшим переписыванием последних в матричном виде.) Отсюда получаются *нормальные уравнения*

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{Y}. \tag{2.6.3}$$

Встречаются два основных случая: либо уравнение (2.6.3) содержит p независимых уравнений относительно p неизвестных параметров, либо некоторые уравнения зависят от других и тогда независимых уравнений будет меньше, чем неизвестных величин, подлежащих определению. Если некоторые нормальные уравнения зависят от других, то матрица $\mathbf{X}'\mathbf{X}$ особенная, и потому $(\mathbf{X}'\mathbf{X})^{-1}$ не существует. В таком случае надо или выразить модель через меньшее число параметров, или выдвинуть дополнительные ограничения на параметры. Некоторые примеры такого рода рассматриваются в гл. 9. Если все нормальные уравнения независимы, то матрица $\mathbf{X}'\mathbf{X}$ неособенная и для нее существует обратная матрица. В этом случае решение нормальных уравнений может быть записано в виде

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}. \tag{2.6.4}$$

Решение $\mathbf{b}$ обладает следующими свойствами:

1. Вектор $\mathbf{b}$ — это оценка вектора β, которая минимизирует сумму квадратов ошибок *независимо* от того, каков характер распределения этих ошибок.

(Примечание. Предположение о том, что ε есть нормально-распределенный вектор, не требуется для отыскания оценки $\mathbf{b}$, но оно

необходимо в дальнейшем для того, чтобы можно было использовать такие статистические критерии, как t- и F-критерии, поскольку они опираются на предположение о нормальности, или для получения доверительных интервалов, которые в свою очередь базируются на t- и F-распределениях.)

2. Элементы вектора $\mathbf{b}$ — линейные функции наблюдений Y_1, $Y_2, \ldots, Y_n$ — представляют собой несмещенные оценки элементов вектора $\boldsymbol{\beta}$, обладающие минимальными дисперсиями (среди любых линейных функций наблюдений, являющихся несмещенными оценками) безотносительно к характеру распределения ошибок.

(Примечание. Предположим, что мы имеем выражение $T = l_1Y_1 + l_2Y_2 + \ldots + l_nY_n$, которое есть линейная функция наблюдений $Y_1, Y_2, \ldots, Y_n$, и что мы используем в качестве оценки параметра θ. Тогда T — случайная величина с распределением, зависящим от распределения величин Y. Если мы будем многократно повторять выборки из совокупности величин Y и вычислять соответствующие значения T, то в результате будем генерировать распределение величин T эмпирически. Независимо от того, есть у нас такое распределение или нет, распределение величин T будет иметь некоторое определенное среднее значение, допускающее запись в виде $E(T)$, и дисперсию, которую можно обозначить как $V(T)$. Если среднее распределения величин T равно параметру θ, оцениваемому с помощью T, т. е. если $E(T) = \theta$, то мы говорим, что T есть несмещенный «оцениватель» θ. Термином *«оцениватель»* обычно пользуются, если речь идет о теоретическом выражении для T исходя из выборки величин Y. Конкретное численное значение величин T следует называть несмещенной *оценкой* параметра θ. Хотя это определение и корректно, оно не всегда применяется в статистических работах. Если мы имеем все возможные линейные функции $T_1, T_2, \ldots$, скажем, от n наблюдений $Y_1, Y_2, \ldots, Y_n$ и если T удовлетворяют условию

$$\theta = E(T_1) = E(T_2) \ldots,$$

т. е. все они суть несмещенные оцениватели величины θ, то одна из них с наименьшей величиной из $V(T_j)$, $j = 1, 2, \ldots$, есть *несмещенный оцениватель параметра* θ *с наименьшей дисперсией* (результат пункта 2 — это теорема Гаусса).)

3. Если ошибки являются независимыми и $\varepsilon_i \sim N(0, \sigma^2)$, то $\mathbf{b}$ есть оценка максимального правдоподобия [9] величины $\boldsymbol{\beta}$. В векторных

---

[9] Метод максимума правдоподобия был предложен Р. Фишером (См.: Фишер Р. Статистические методы для исследователей. Пер. с англ.— М.: Госстатиздат, 1958.— 268 с.) для получения «наилучших» оценок параметров по выборкам данных, закон распределения которых известен. В отличие от метода наименьших квадратов он требует знания закона (плотности) распределения ошибок, но зато позволяет найти не только оцеики параметров модели, но и оценки элементов дисперсионной матрицы откликов. Подробнее с методом максимального правдоподобия можно познакомиться в кн.: Кендалл М. Дж., Стьюарт А. Многомерный статистический анализ и временные ряды/Пер. с англ. Под ред. А. Н. Колмогорова, Ю. В. Прохорова.— М.: Наука,1976, т. 3.— 736 с. (см. гл. 18); Клепиков Н. П., Соколов С. Н. Анализ и планирование экспериментов методом максимума правдоподобия.— М.: Физ-

обозначениях мы можем записать $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$, полагая, что $\boldsymbol{\varepsilon}$ подчиняется n-мерному нормальному распределению с $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ (где $\mathbf{0}$ означает вектор, составляющие которого равны нулю, а размерность та же, что и у $\boldsymbol{\varepsilon}$), т. е. этот вектор имеет матрицу дисперсий-ковариаций, все диагональные элементы которой $V(\varepsilon_i)$, $i = 1, 2, \ldots, n$, равны σ^2, а внедиагональные элементы, представляющие собой ковариации $\operatorname{cov}(\varepsilon_i, \varepsilon_j)$, $i \neq j = 1, 2, \ldots, n$, все равны нулю. Функция правдоподобия для выборки из наблюдений $Y_1, Y_2, \ldots, Y_n$ определяется в этом случае как произведение

$$\prod_{i=1}^{n} \frac{1}{\sigma (2\pi)^{1/2}} e^{-\varepsilon_i^2/2\sigma^2} = \frac{1}{\sigma^n (2\pi)^{n/2}} e^{-\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}/2\sigma^2}. \qquad (2.6.5)$$

Таким образом, при фиксированной величине σ максимизация функции правдоподобия эквивалентна минимизации величины $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}$. Отметим, что этот факт может рассматриваться как обоснование метода наименьших квадратов (т. е. процедуры минимизации суммы квадратов ошибок), поскольку во многих физических ситуациях предположение о нормальном характере распределения ошибок довольно благоразумно. Во всяком случае мы будем выяснять, не нарушается ли это предположение, исследуя остатки в рамках регрессионного анализа. Если, однако, имеются *определенные априорные* сведения о распределении ошибок (из теоретических соображений или из определенных знаний об изучаемом процессе), то использование принципа максимального правдоподобия для отыскания оценок может привести к критерию, отличному от суммы квадратов ошибок. Например, предположим, что ошибки ε_i, $i = 1, 2, \ldots, n$, были бы независимыми и следовали бы двустороннему экспоненциальному распределению

$$f(\varepsilon_i) = (2\sigma)^{-1} e^{-|\varepsilon_i|/\sigma} (-\infty \leqslant \varepsilon_i \leqslant \infty), \qquad (2.6.6)$$

а не нормальному распределению

$$f(\varepsilon_i) = \frac{1}{\sigma (2\pi)^{1/2}} e^{-\varepsilon_i^2/2\sigma^2}, \qquad (2.6.7)$$

которое обычно предполагается. Плотность двустороннего экспоненциального распределения имеет при $\varepsilon_i = 0$ заостренный пик высотой $1/2\,\sigma$ и убывает до нуля, когда ε_i стремится к $+\infty$ или $-\infty$. Тогда применение принципа максимума правдоподобия для оценивания вектора $\boldsymbol{\beta}$ при фиксированном σ свелось бы к минимизации суммы абсолютных значений ошибок $\sum_{i=1}^{n} |\varepsilon_i|$, а не суммы квадратов ошибок $\sum_{i=1}^{n} \varepsilon_i^2$. Для более детального ознакомления с минимизацией суммы абсо-

---

матгиз, 1964.— 185 с.; Худсон Д. Статистика для физиков. Пер. с англ.— М.: Мир, 1970.— 296 с.; Химмельблау Д. Анализ процессов статистическими методами/Пер. с англ. Под ред. В. Г. Горского.— М.: Мир, 1973.— 959 с.— *Примеч. пер.*

лютных значений ошибок см. статью [10]: Gentle J. E. Communications Statistics — Simulated Computations. 1977, B 6 (4), p. 313—328.

Вычислительные аспекты рассмотрены в публикациях: Gentle J. E., Kennedy W. J., Sposito V. A. (Fortran) Algorithm AS 110. Lp norm fit of a straight line.— Applied Statistics, 1977, **26**, p. 114—118; Narula S. C., Wellington J. F. (Fortran) Algorithm AS 108, Multiple linear regression with minimum sum of absolute errors. —Applied Statistics, 1977, **26**, p. 106—111.

Предположения, независимые от распределения

Допустим, что мы используем метод наименьших квадратов для нахождения МНК-оценки $\mathbf{b}$ вектора β. Можно перейти к следующим стадиям (этапам) анализа независимо от того, являются ли ошибки нормально-распределенными.

1. Предсказываемые значения отклика получаются из уравнения

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}.$$

2. Вектор остатков задается выражением $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ (об исследовании остатков см. гл. 3). Верно, что $\sum_{i=1}^{n} e_i \hat{Y}_i = 0$, какой бы ни была модель. В этом можно убедиться, умножая каждое j-е нормальное уравнение на b_j и складывая результаты. Если модель содержит член β_0, то справедливо соотношение $\sum_{i=1}^{n} e_i = 0$. (Здесь e_i и $\hat{Y}_i$, $i = 1, 2, \ldots,$ n,—соответственно i-е элементы векторов $\mathbf{e}$ и $\hat{\mathbf{Y}}$.)

3. $\mathbf{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ дает дисперсии (диагональные элементы) и ковариации (внедиагональные элементы) оценок параметров (получение оценки параметра σ^2 описано ниже).

4. Допустим, что $\mathbf{X}_0'$ есть $(1 \times p)$-вектор, являющийся некоторой строкой матрицы $\mathbf{X}$, так что $\hat{Y}_0 = \mathbf{X}_0'\mathbf{b} = \mathbf{b}'\mathbf{X}_0$ есть предсказываемое значение отклика в точке [11] X_0. Например, если модель имела бы вид $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$, то для данного значения X_0 вектор $\mathbf{X}_0'$ имел бы вид $\mathbf{X}_0' = (1, X_0, X_0^2)$. Тогда $\hat{Y}_0$ есть *величина отклика при X_0, предсказываемая с помощью уравнения регрессии*; она имеет дисперсию

$$V(\hat{Y}_0) = \mathbf{X}_0'\mathbf{V}(\mathbf{b})\mathbf{X}_0 = \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\sigma^2. \qquad (2.6.8)$$

[10] Кроме того, этот метод изложен в работах: Мудров В. И., Кушко В. Л. Метод наименьших модулей.— М.: Знание. Сер. «Математика, кибернетика», 1971, вып. 7.— 61 с.; Мудров В. И., Кушко В. Л. Методы обработки измерений.—М.: Советское радио, 1976.—192 с.; Бородюк В. П. Регрессионные модели с нестандартной ошибкой в задачах идентификации сложных объектов.— М.: Изд-во МЭИ, 1981.— 92 с. (особо с. 74—77); Планирование эксперимента в исследовании технологических процессов/Пер. с нем. Под ред. Э. К. Лецкого.— М.: Мир, 1977.— 547 с. (особо с. 523—531).— *Примеч. пер.*

[11] Не следует путать точку X_0 в факторном пространстве (в данном случае на оси X) с вектор-строкой $\mathbf{X}'_0$.— *Примеч. пер.*

5. Основная таблица дисперсионного анализа может быть составлена так:

Источник	Число степеней свободы	SS	MS
Регрессия	p	$\mathbf{b}'\mathbf{X}'\mathbf{Y}$	MS_R
Остаток	$n-p$	$\mathbf{Y}'\mathbf{Y}-\mathbf{b}'\mathbf{X}'\mathbf{Y}$	MS_E
Общий	n	$\mathbf{Y}'\mathbf{Y}$	

Дальнейшее разбиение таблицы дисперсионного анализа на части может быть выполнено следующим образом:

5а. Если в модели имеется коэффициент β_0, то сумму квадратов, обусловленную регрессией, можно разбить на слагаемые:

$$SS(b_0) = \frac{(\Sigma Y_i)^2}{n} = n\bar{Y}^2, \tag{2.6.9}$$

$$SS(\text{регрессия}|b_0) = SS(R|b_0) = \mathbf{b}'\mathbf{X}'\mathbf{Y} - \frac{(\Sigma Y_i)^2}{n}. \tag{2.6.10}$$

Эти суммы имеют соответственно 1 и $p-1$ степеней свободы. Разбиение суммы квадратов, связанной с регрессией, на составные части будет обсуждаться более детально в параграфе 2.7.

5б. Если имеются повторные наблюдения, то мы можем расщепить остаточную SS на SS («чистой» ошибки), связанную с «чистой» ошибкой и имеющую n_e степеней свободы, которая оценивает $n_e\sigma^2$, и SS (неадекватности) — сумму квадратов, связанную с неадекватностью модели и имеющую $n-p-n_e$ степеней свободы. При проведении повторных опытов должны выдерживаться уровни всех независимых переменных (хотя на практике иногда используются «очень близкие» точки). Это приводит к следующей таблице дисперсионного анализа:

Источник	Число степеней свободы	SS	MS	
b_0	1	SS (b_0)		
Регрессия $\mid b_0$	$p-1$	SS ($R \mid b_0$)	MS ($R \mid b_0$)	MS_R
Неадекватность	$n-p-n_e$	SS (l.o.f.)	MS (l.o.f.)	MS_E
«Чистая» ошибка	n_e	SS (p.e.)	MS (p.e.)	
Общий	n	$\mathbf{Y}'\mathbf{Y}$		

(Примечание. Порядок, в котором расположены члены в этой таблице, не играет роли. Большинство таблиц в данной книге имеет такое расположение, которое зачастую можно видеть в машинных программах.)

R^2-статистика. Отношение

$$R^2 = \frac{\mathrm{SS}(R \mid b_0)}{\mathbf{Y}'\mathbf{Y} - \mathrm{SS}(b_0)} = \frac{\Sigma(\hat{Y}_i - \bar{Y})^2}{\Sigma(Y_i - \bar{Y})^2} \qquad (2.6.11)$$

есть обобщение величины, введенной ранее при рассмотрении линейной регрессии, и представляет собой квадрат множественного коэффициента корреляции. Другое название величины R^2 — *множественный коэффициент детерминации*. Величину R^2 не следует путать с буквой R в выражениях $\mathrm{SS}(R \mid b_0)$ и MS_R, где буква R отражает вклад регрессии. R^2 есть квадрат коэффициента корреляции между $\mathbf{Y}$ и $\hat{\mathbf{Y}}$, при этом $0 \leqslant R^2 \leqslant 1$. Если есть повторные опыты, то R^2 не может достигать 1; см. замечания на с. 54, 61—62, 84, 98—99. $R^2 = 1$ при полном согласии экспериментальных и расчетных данных $\hat{Y}_i = Y_i$, но это маловероятный случай.

Если $Y_i = \bar{Y}$, т. е. $b_1 = b_2 = \ldots = b_{p-1} = 0$ (или адекватна модель $Y = \beta_0 + \varepsilon$), то $R^2 = 0$. Следовательно, R^2 есть мера полезности параметров β_i кроме β_0 в модели. Важно понимать, что величина R^2 может принимать значение 1 только при соответствующем выборе коэффициентов модели, включая β_0, поскольку в этом случае может быть подобрана модель, которая описывает экспериментальные результаты точно. (Например, если мы имеем наблюдения Y для четырех различных значений X, то кубический полином

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$$

пройдет точно через все четыре точки.) Поскольку величина R^2 используется часто в качестве меры эффективности регрессионной модели при объяснении вариации в данных, мы должны быть уверены, что увеличение R^2 благодаря введению новых слагаемых в модель имеет некоторый реальный смысл, а не обусловлено всего лишь тем фактом, что число параметров в модели становится ближе к состоянию насыщения, т. е. к числу наблюдений [12]. Это *особенно* опасно, когда имеются *повторные* наблюдения. Например, если мы имеем сто наблюдений, состоящих из пяти групп, содержащих по двадцать повторных наблюдений, то фактически мы имеем лишь пять величин, несущих содержательную информацию, и они представляются пятью средними значениями, а также 95 степеней свободы для суммы квадратов, связанной с «чистой» ошибкой, по 19 для каждой точки, где проводятся повторные опыты. Следовательно, модель, содержащая пять параметров, дает очень хорошее согласие с пятью средними и может дать величину R^2, очень близкую к 1, особенно, если экспериментальная ошибка мала по сравнению с размахом для пяти средних. В этом случае тот факт, что сотня наблюдений может хорошо предсказываться с помощью модели, содержащей лишь пять параметров, не является удивительным, так как на самом деле модель предсказывает только пять различных определенных экспериментальных то-

[12] См. примечание по поводу R^2-статистики в гл. 1 на с. 54. — *Примеч. пер.*

чек, а не сто, как это могло показаться вначале. Может быть и так, что точных повторений нет, но точки в X-пространстве (для которых имеются наблюдения Y). расположены близко друг к другу. Такая ситуация может быть, однако, не очевидной, хорошо скрытой благодаря определенному подбору экспериментальных данных. Графики данных и остатки (см. гл. 3) обычно позволяют обнаруживать такие «скопления» (кластеры — clasters) точек.

Приведенная R^2-статистика

Предположим, что оцениваемая модель содержит p параметров, включая β_0, и RSS_p есть соответствующая остаточная сумма квадратов. Мы определяем R^2-статистику как меру, выражающую долю вариации относительно среднего, обусловленную оцениваемым уравнением регрессии, в виде

$$R^2 = \frac{\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2}{\mathbf{Y}'\mathbf{Y} - n\bar{Y}^2} = 1 - \frac{RSS_p}{CTSS}, \qquad (2.6.11а)$$

где [13] CTSS обозначает скорректированную общую сумму квадратов, а n — общее число наблюдений. Родственная статистика, использовавшаяся в прошлом и которой отдавалось предпочтение рядом авторов, это так называемая преобразованная R^2-статистика. Она в наших обозначениях имеет вид

$$R_a^2 = 1 - \frac{(RSS_p)/(n-p)}{(CTSS)/(n-1)} = 1 - (1 - R^2)\left(\frac{n-1}{n-p}\right). \qquad (2.6.11б)$$

Такое преобразование производилось для того, чтобы учесть соответствующие числа степеней свободы двух величин: RSS_p и CTSS; его идея заключалась в том, чтобы статистика R_a^2 могла использоваться при сравнении оцениваемых уравнений не только на основании некоторого определенного набора данных, но также при двух и более совершенно различных наборах данных. (Ценность такой статистики для указанных целей, по нашему мнению, невысокая; R_a^2 может быть полезной только как первичный грубый индикатор и не более того.)

Как показано в работе: Kennard R. W. A note on the C_p statistic.— Technometrics, 1971, **13**, p. 899—900, преобразованная R_a^2-статистика тесно связана с C_p-статистикой, используемой в одном из методов выбора наилучшего уравнения регрессии. Этот вопрос обсуждается в гл. 6. Однако мы не рекомендуем пользоваться таким критерием.

Тождество выражений (2.6.11) и (2.6.11а) может быть показано таким образом:

$$\Sigma(\hat{Y}_i - \bar{Y})^2 = \Sigma\hat{Y}_i^2 - (\Sigma Y_i)^2/n$$

---

[13] CTSS — аббревиатура от начальных букв слов «corrected total sum of squares» (скорректированная общая сумма квадратов).— *Примеч. пер.*

и

$$\Sigma \hat{Y}_i^2 = \hat{\mathbf{Y}}'\hat{\mathbf{Y}} = (\mathbf{Xb})'(\mathbf{Xb}) = \mathbf{b}'\mathbf{X}'\mathbf{Xb} = \mathbf{b}'\mathbf{X}'\mathbf{Y},$$

поскольку $\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{Y}$ (см. (2.6.3)).

Предположения, связанные с распределением

Разбиение сумм квадратов в дисперсионном анализе основано на алгебраических (или геометрических, в зависимости от принятой точки зрения, см. параграф 10.6) соотношениях и не зависит от свойств распределения ошибок. Однако если мы предполагаем дополнительно, что $\varepsilon_i \sim N(0, \sigma^2)$ и что ε_i независимы друг от друга, т. е. $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$, то мы можем сделать следующее:

1. Проверить неадекватность модели путем рассмотрения отношения

$$\left[\frac{\text{SS (l.o. f)}/(n-p-n_e)}{\text{SS (p. e.)}/n_e}\right] \tag{2.6.12}$$

как $F[(n-p-n_e), n_e]$-распределенной случайной величины и сравнения ее с $F[(n-p-n_e), n_e, 1-\alpha]$. Если рассогласование экспериментальных и расчетных данных статистически незначимо, то величина $\text{SS (res.)}/(n-p) = \text{MS}_\text{E}$, обычно обозначаемая как s^2, есть несмещенная оценка σ^2. Если нет возможности произвести такую проверку, то, используя s^2 как оценку величины σ^2, мы фактически *предполагаем*, что модель корректна. (Если это не так, то s^2 будет иметь слишком большое численное значение, тогда s^2 — случайная величина, среднее значение которой *больше*, чем σ^2. Однако справедливости ради надо заметить, что благодаря выборочной флуктуации, поскольку величина s^2 есть случайная переменная, она может быть также и слишком малой.)

2. Проверить все уравнение регрессии (более точно — проверить гипотезу $H_0: \beta_1 = \beta_2 = \beta_3 = \ldots = \beta_{p-1} = 0$ против гипотезы H_1: не все $\beta_i = 0$) с помощью отношения средних квадратов

$$\frac{[\text{SS}(R \mid b_0)/(p-1)]}{s^2}, \tag{2.6.13}$$

рассматриваемого как $F(p-1, v)$-распределенная случайная величина, где число степеней свободы v равно $n-p$.

Допустим, что мы задались уровнем риска α. Тот факт, что дисперсионное отношение превосходит значение $F(p-1, v, 1-\alpha)$, означает, что получено «статистически значимое» уравнение регрессии. Другими словами, доля вариаций, наблюдаемых в полученных данных, которая отнесена за счет уравнения регрессии, больше, чем можно было бы ожидать за счет случайных причин в 100 $(1-\alpha)$ % подобных наборов данных с одним и тем же числом наблюдений n и одинаковой матрицей $\mathbf{X}$.

Это не обязательно означает, что такое уравнение полезно для прогнозирования. Если размах величин, предсказываемых уравнением регрессии, не слишком значительно превосходит величину случайной

ошибки, предсказание не будет иметь никакой ценности, хотя и была получена «значимая» величина F, так как уравнение будет «описывать только ошибки».

В работе Дж. М. Ветца (J. M. Wetz) (1964 г., автореферат диссертации «Критерий для суждения об адекватности при оценивании с помощью функции отклика», написанной под руководством доктора Бокса из Висконсинского университета) утверждается: чтобы уравнение можно было считать удовлетворительным для целей предсказания (в том смысле, что размах предсказываемых значений отклика будет значительно больше, чем стандартная ошибка отклика), наблюдаемое значение F-отношения среднего квадрата, обусловленного регрессией, и остаточной дисперсии должно не просто превышать выбранную процентную точку F-распределения, а превосходить ее примерно в 4 раза. Например, пусть $p = 11$, $v = 20$, $\alpha = 0{,}05$, F (10, 20, 0,95) = 2,35. Тогда наблюдаемое значение F-отношения должно превосходить 9,4 для того, чтобы можно было расценивать полученное уравнение как удовлетворительную модель для предсказания. Для более детального ознакомления см. приложение 2В.

Распределение величины R^2. Мы видим, что

$$R^2 = \frac{\text{SS (регрессия} \mid b_0)}{\sum_{i=1}^{n} (Y_i - \bar{Y})^2} = \frac{\text{SS (регрессия} \mid b_0)}{\text{SS (регрессия} \mid b_0) - \text{остаточная SS}} =$$

$$= \frac{\nu_1 F}{\nu_1 F + \nu_2}, \qquad (2.6.13a)$$

где величина

$$F = \frac{\text{SS (регрессия} \mid b_0)/\nu_1}{\text{остаточная SS}/\nu_2}$$

есть наша обычная F-статистика для проверки всей регрессии при наличии b_0, т. е. для проверки гипотезы H_0: все коэффициенты β за исключением β_0 равны нулю против альтернативной гипотезы H_1: по крайней мере хотя бы один коэффициент β_i (кроме β_0) не равен нулю. Величина β_0 не имеет значения. В соответствии с уравнением (2.6.13) следует положить $\nu_1 = p-1$, $\nu_2 = n-p$. Если справедлива гипотеза H_0, то величина F имеет распределение как переменная F (ν_1, ν_2). Из статистики известно, что величина R^2 следует $\beta\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right)$-распределению, т. е. бета-распределению [14], с числами степеней свободы $\frac{\nu_1}{2}$ и $\frac{\nu_2}{2}$. Хотя мы не будем обсуждать бета-распределение, тем не менее ясно, что если мы располагаем подходящими статистическими таблицами, то можно проверить гипотезу H_0 против альтернативы H_1, используя величину R^2. Результат был бы в *точности* тем же, какой мы имеем, применяя стандартный F-критерий.

[14] Таблицы бета-распределения можно найти в справочнике: Хастингс Н., Пикок Дж. Справочник по статистическим распределениям. Пер. с англ.— М.: Статистика, 1980.— 96 с.— *Примеч. пер.*

Значимая точка для R^2 может быть получена с помощью формулы (2.6.13а), в которой следует заменить F на $F(p—1, n—p, 1—\alpha)$. По этой причине, а также поскольку таблицы бета-распределения в статистической литературе встречаются реже, чем таблицы F-распределения, проверка гипотезы с помощью R^2 проводится редко.

3. Если мы используем оценку s_v^2 для σ^2, то 100 $(1—\alpha)$ %-ные доверительные границы для среднего значения $\hat{Y}$ при X_0 можно получить из соотношения

$$\hat{Y}_0 \pm t(v,\ 1-\alpha/2)\, s_v \sqrt{\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}.$$

4. Доказать, что

$$\mathbf{b} \sim N(\boldsymbol{\beta},\ (\mathbf{X}'\mathbf{X})^{-1}\sigma^2). \tag{2.6.14}$$

5. Получить совместную 100 $(1—\alpha)$ %-ную доверительную область для *всех* параметров $\boldsymbol{\beta}$ из выражения

$$(\boldsymbol{\beta}-\mathbf{b})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mathbf{b}) \leqslant ps^2F(p,\ v,\ 1-\alpha), \tag{2.6.15}$$

где $F(p, v, 1—\alpha)$ есть $(1—\alpha)$%-точка (верхняя α-точка) для $F(p, v)$-распределения и где s^2 означает то же самое, что и в пункте 1, причем модель предполагается корректной. Такое представление доверительного множества полезно только в том случае, когда p мало и равно, скажем, 2, 3 или 4, если не предпринимаются какие-либо специальные меры, чтобы представить информацию в форме, доступной пониманию. Неравенство (2.6.15) приводит к уравнениям эллипсоподобных контуров в пространстве, размерность которого равна числу параметров p, т. е. числу составляющих вектора $\boldsymbol{\beta}$. Мы можем получить индивидуальные доверительные интервалы для различных параметров порознь с помощью формулы

$$b_i \pm t(v,\ 1-\alpha/2)\ (\text{оценка стандартного отклонения } (b_i)),$$

где оценка стандартной ошибки в определении параметра b_i есть корень квадратный из i-го диагонального элемента матрицы $(\mathbf{X}'\mathbf{X})^{-1}s^2$. (Вычисления такого рода в случае модели с двумя параметрами β_0 и β_1 проводятся с использованием формулы (2.3.1), если заменить в ней σ^2 на s^2, см. параграф 1.4.) Доверительные интервалы для отдельных параметров рассматриваются в нашей книге, они оказываются часто полезными. Однако мы не придаем им особого значения по следующим причинам. На рис. 2.1 представлены ситуации, которые могут иметь место, когда рассматриваются два параметра. Совместная 95 %-ная доверительная область для двух истинных параметров β_1 и β_2, как было показано выше, представляет собой тонкий, вытянутый эллипс и содержит точки с координатами (β_1, β_2), которые можно рассматривать как *совместно* приемлемые. При этом принимается во внимание корреляция между оценками b_1 и b_2. Индивидуальные 95 %-ные доверительные интервалы для β_1 и β_2 порознь используются при указании диапазонов возможного изменения значений одного параметра безотносительно к значениям другого. Если попытаться интерпретировать эти интервалы одновременно, неправильно трактуя прямоугольник, который они образуют, как совместную доверитель-

ную область, то можно, например, думать, что координаты точки E дают приемлемые значения для (β_1, β_2). Однако из рассмотрения совместной доверительной области ясно, что эта точка не подходит. Если имеются всего лишь два параметра, построение доверительного эллипса несложно. При большем числе параметров необходимые вычисления также не сложны и вполне выполнимы с помощью вычислительной машины, но интерпретация затруднительна. Один из возможных путей разрешения этой трудности состоит в нахождении координат точек, лежащих на концах главных осей области (на рис. 2.1 это точки A, B, C и D). Такая процедура должна включать получение уравнения доверительного контура и приведение его к каноническому виду. Это сделать несложно и здесь пригодны методы, рассмотренные в параграфах 6.9 и 6.10. Однако мы можем сформулировать такую мораль: «одновременное» сообщение индивидуальных доверительных интервалов следует воспринимать с определенной осторожностью и надо обращать внимание как на относительные значения величин $V(b_i)$, так и на ковариации величин b_i и b_j. Если дисперсии коэффициентов b_i и b_j имеют различные значения и коэффициент корреляции между b_i и b_j, а именно

$$\rho_{ij} = \frac{\operatorname{cov}(b_i, b_j)}{[V(b_i)\, V(b_j)]^{1/2}},$$

не является малым (по абсолютной величине), то возникает ситуация, изображенная на рис. 2.1. Если же ρ_{ij} близок к нулю, то пря-

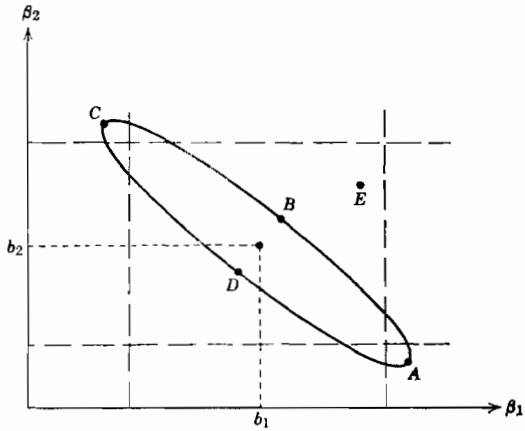


Рис. 2.1. Совместная доверительная область и индивидуальные доверительные интервалы

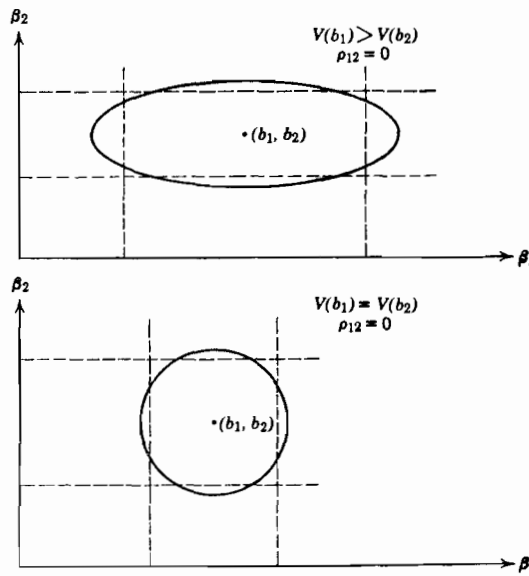


Рис. 2.2. Примеры, показывающие, что прямоугольник, образованный индивидуальными доверительными интервалами, хорошо аппроксимирует совместную доверительную область для двух параметров

моугольник, образованный индивидуальными доверительными интервалами, будет приближенно равен по площади совместной доверительной области. Вытянутость области будет зависеть от соотношения величин $V(b_i)$ и $V(b_j)$, некоторые примеры приведены на рис. 2.2.

(Примечание. Если первоначально записанная модель подбирается в альтернативной форме

$$\mathrm{E}(Y-\bar{Y})=\beta_1(X_1-\bar{X}_1)+\beta_2(X_2-\bar{X}_2)+\ \ldots\ +\beta_k(X_k-\bar{X}_k),$$

где $\bar{Y}, \bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k$ — средние значения по имеющимся наблюдаемым данным, то могут быть получены совместные доверительные интервалы для всех параметров, кроме β_0, который обычно не представляет большого интереса.)

2.7. ПРИНЦИП «ДОПОЛНИТЕЛЬНОЙ СУММЫ КВАДРАТОВ»

В регрессионных задачах нередко возникает вопрос, стоит ли включать в модель определенные члены. Этот вопрос можно исследовать, изучая дополнительную долю или часть суммы квадратов, порожденной регрессией, которая связана с включением в модель рассматриваемых членов. Средний квадрат, который получается из этой дополнительной суммы, может быть затем сопоставлен с оценкой s^2 параметра σ^2, чтобы выяснить, имеется ли значимое различие между ними. Если средний квадрат значимо превышает оценку s^2, то такие члены следует включить в модель. В противном случае их можно рассматривать как излишние и исключить из модели.

Мы уже видели один такой пример при подборе уравнения прямой линии, где величина SS $(b_1|b_0)$ представляла собой дополнительную сумму квадратов, обусловленную включением в модель члена $\beta_1 X$. Теперь мы изложим более общую процедуру. Предположим, что Z_1, $Z_2, \ldots, Z_p$ есть известные функции основных переменных $X_1, X_2, \ldots$, и допустим, что значения X-ов и соответствующих им откликов нам известны. Рассмотрим теперь две модели.

1. $Y=\beta_0+\beta_1 Z_1+\beta_2 Z_2+\ldots+\beta_p Z_p+\varepsilon$. Допустим, что мы получили следующие МНК-оценки: $b_0(1)$, $b_1(1)$, $b_2(1), \ldots, b_p(1)$, и пусть SS $(b_0(1), b_1(1), \ldots, b_p(1)) = S_1$, причем модель адекватна. Пусть далее оценка s^2 параметра σ^2 получается из остатков, соответствующих модели 1.

2. $Y=\beta_0+\beta_1 Z_1+\beta_2 Z_2+\ldots+\beta_q Z_q+\varepsilon\ (q<p)$. Величины Z в модели 2 те же самые функции, что и в модели 1, когда их индексы одинаковые. Заметим, однако, что в модели 2 меньше членов, чем в модели 1.

Предположим далее, что мы получили следующие МНК-оценки: $b_0(2), b_1(2), \ldots, b_q(2)$.

(Примечание. Они могут совпадать или не совпадать с оценками $b_0(1), b_1(1), \ldots, b_q(1)$, приведенными выше. Если они одинаковы, то $b_i(1)$ и $b_j(1)$ являются ортогональными линейными функциями для $1 \leqslant i \leqslant q$, $q+1 \leqslant j \leqslant p$. Это имеет место, когда в модели 1 все первые q столбцов матрицы $\mathbf{X}$ ортогональны последним

p—q столбцам. Подобные случаи мы рассмотрим в следующих главах.)

Обозначим для второй модели SS $(b_0\ (2),\ b_1\ (2), \ldots,\ b_q\ (2)) = S_2$. Тогда S_1—S_2 есть *дополнительная сумма квадратов*, связанная с включением членов $\beta_{q+1}Z_{q+1} + \ldots + \beta_pZ_p$ в модель 1. Так как S_1 имеет $p+1$ степень свободы, а S_2—$q+1$ степень свободы, величина S_1—S_2 имеет соответственно p—q степеней свободы. Можно показать, что если $\beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$, то $E\{(S_1$—$S_2)/(p$—$q)\} = \sigma^2$. Еще укажем, что если ошибки распределены нормально, то величина $(S_1$—$S_2)$ будет распределена как $\sigma^2\chi^2_{p-q}$ и будет независимой от s^2. Это означает, что мы можем сравнить величину $(S_1$—$S_2)/(p$—$q)$ с s^2 при помощи $F\ (p$—$q,\ v)$-критерия, где v — число степеней свободы, с которым определена оценка s^2, и использовать эту процедуру для проверки гипотезы $H_0: \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$.

Для удобства можно записать величину S_1—S_2 как SS $(b_{q+1}, \ldots, b_p | b_0, b_1, \ldots, b_q)$, однако мы должны иметь в виду, что на самом деле здесь рассматриваются две модели, хотя из обозначений это и не очевидно. Эту величину называют суммой квадратов, связанной с $b_{q+1}, \ldots, b_p$, при условии, что коэффициенты $b_0, b_1, \ldots, b_q$ входят в модель. Путем дальнейшего приложения этого принципа мы можем получить последовательно для любой регрессионной модели величины SS (b_0), SS $(b_1|b_0)$, SS $(b_2|b_0, b_1), \ldots,$ SS $(b_p|b_0, b_1, \ldots, b_{p-1})$. Все эти суммы квадратов статистически независимы от s^2 и равны своим средним квадратам, так как имеют по одной степени свободы. Средние квадраты могут быть сопоставлены с величиной s^2 с помощью ряда F-критериев. Такие проверки полезны, когда члены модели имеют логично обоснованный порядок записи, как было бы, например, если бы $Z_j = X^j$. Тогда можно сделать заключение о том, как много членов должно быть в модели.

Если члены, содержащиеся в модели, сгруппированы естественным образом — так, как, скажем, это имеет место в полиномиальных моделях, содержащих: а) параметр β_0, б) члены первого порядка, в) члены второго порядка,— то мы можем построить различные дополнительные суммы квадратов, например SS (b_0), SS (параметры b_i, отвечающие членам первого порядка $|\ b_0$), SS (параметры b_{ij}, отвечающие членам второго порядка $|\ b_0$, параметры b_i, соответствующие членам первого порядка), и сравнить их с величиной s^2. Принцип дополнительной суммы квадратов можно использовать по-разному, чтобы получить такое разложение суммы квадратов, обусловленной регрессией, которое кажется приемлемым для рассматриваемой задачи. Число степеней свободы для каждой такой суммы квадратов будет равно числу параметров, указанных в скобках до вертикальной линии. (Исключая тот случай, когда оценки линейно зависимы, что имеет место, когда $\mathbf{X'X}$ есть особенная матрица и нормальные уравнения оказываются линейно зависимыми. Число степеней свободы при этом равно максимальному числу линейно независимых оценок в рассматриваемом множестве оценок.) Эти дополнительные суммы распределены независимо от s^2. Соответствующие средние квадраты, равные отношению суммы квадратов к числу степеней свободы, можно поделить на s^2 и получить таким образом F-отношения для проверки

гипотез о том, что истинные значения коэффициентов, оценки которых создают дополнительные суммы квадратов, равны нулю. Вопрос о математическом ожидании дополнительной суммы квадратов изложен в приложении 2Б.

Принцип дополнительной суммы квадратов фактически приводит к особому случаю проверки общей линейной гипотезы. При более общем подходе дополнительная сумма квадратов отклонений вычисляется исходя из остаточных сумм квадратов, а не из сумм квадратов, обусловленных регрессией. Поскольку полная сумма квадратов $\mathbf{Y}'\mathbf{Y}$ одинакова при обоих способах регрессионных вычислений, мы получим те же самые численные результаты независимо от того, воспользуемся мы разностью сумм квадратов, обусловленных регрессией, или разностью остаточных сумм квадратов.

Прежде чем обсуждать вопросы, связанные с проверкой гипотез (см. параграф 2.10), рассмотрим один важный случай применения принципа дополнительной суммы квадратов.

2.8. ОРТОГОНАЛЬНЫЕ СТОЛБЦЫ В МАТРИЦЕ X

Допустим, что мы имеем дело с регрессионной задачей, включающей параметры β_0, β_1 и β_2. Используя принцип дополнительной суммы квадратов, можно вычислить ряд таких величин, как:

$$\mathrm{SS}(b_2) \text{ исходя из модели } Y = \beta_2 X_2 + \varepsilon,$$

$$\mathrm{SS}(b_2 \mid b_0) \text{ исходя из модели } Y = \beta_0 + \beta_2 X_2 + \varepsilon,$$

$$\mathrm{SS}(b_2 \mid b_0,\ b_1) \text{ исходя из модели } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

Эти величины имеют обычно совершенно различные численные значения за исключением того случая, когда β_2-столбец матрицы $\mathbf{X}$ ортогонален к β_0- и β_1-столбцам той же матрицы. Если это имеет место, то мы можем однозначно говорить о величине $\mathrm{SS}(b_2)$. Рассмотрим теперь эту ситуацию более подробно.

Предположим, что матрицу $\mathbf{X}$, входящую в модель $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, мы разбиваем на t наборов столбцов. Запишем эту операцию в матричной форме так:

$$\mathbf{X} = \{\mathbf{X}_1,\ \mathbf{X}_2,\ \ldots,\ \mathbf{X}_t\}.$$

Соответствующим образом можно разбить на подвекторы и вектор $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_t \end{bmatrix},$$

где число столбцов в $\mathbf{X}_i$ равно числу элементов в $\boldsymbol{\beta}_i$, $i = 1, 2, \ldots, t$. Тогда модель может быть записана в виде

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \ldots + \mathbf{X}_t\boldsymbol{\beta}_t.$$

Допустим далее, что

$$\mathbf{b}=\begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\\ \cdot\\ \cdot\\ \cdot\\ \mathbf{b}_t\end{bmatrix}$$

есть оценка вектора $\boldsymbol{\beta}$ для такой модели (и имеющихся данных), полученная из нормальных уравнений

$$\mathbf{X}'\mathbf{X}\mathbf{b}=\mathbf{X}'\mathbf{Y}.$$

Вывод. Если столбцы матрицы $\mathbf{X}_i$ ортогональны к столбцам $\mathbf{X}_j$ для всех $i, j=1, 2, \ldots, t$ $(i\neq j)$, т. е. если $\mathbf{X}_i'\mathbf{X}_j=\mathbf{0}$, то справедливо соотношение

$$\mathrm{SS}(\mathbf{b})=\mathrm{SS}(\mathbf{b}_1)+\mathrm{SS}(\mathbf{b}_2)+\ \ldots\ +\mathrm{SS}(\mathbf{b}_t)=$$

$$=\mathbf{b}_1'\mathbf{X}_1'\mathbf{Y}+\mathbf{b}_2'\mathbf{X}_2'\mathbf{Y}+\ \ldots\ +\mathbf{b}_t'\mathbf{X}_t'\mathbf{Y},$$

причем $\mathbf{b}_i$ есть МНК-оценка $\boldsymbol{\beta}_i$ и $\mathrm{SS}(\mathbf{b}_i)=\mathbf{b}_i'\mathbf{X}_i'\mathbf{Y}$ *независимо от того*, будут или не будут содержаться в модели любые другие члены. Таким образом, $\mathrm{SS}(\mathbf{b}_i)=\mathrm{SS}(\mathbf{b}_i\,|$ любой набор $\mathbf{b}_j,\ j\neq i)$. Заметим, что вовсе *не* обязательно, чтобы столбцы матрицы $\mathbf{X}_i$ были ортогональны между собой, надо только, чтобы каждый столбец этой матрицы был ортогональным ко всем прочим столбцам матрицы $\mathbf{X}$.

Рассмотрим случай, когда $t=2$. Здесь

$$\mathbf{X}=(\mathbf{X}_1,\ \mathbf{X}_2),$$

где $\mathbf{X}_1'\mathbf{X}_2=\mathbf{X}_2'\mathbf{X}_1=\mathbf{0}$. (Это означает, что все столбцы матрицы $\mathbf{X}_1$ ортогональны ко всем столбцам матрицы $\mathbf{X}_2$.) Мы можем записать модель так:

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}=\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2+\boldsymbol{\varepsilon},$$

где вектор $\boldsymbol{\beta}'=(\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')$ разбит на два подмножества коэффициентов, каждое из которых соответствует совокупностям столбцов в матрицах $\mathbf{X}_1$ и $\mathbf{X}_2$. Нормальные уравнения $\mathbf{X}'\mathbf{X}\mathbf{b}=\mathbf{X}'\mathbf{Y}$ приводятся к виду

$$\begin{bmatrix}\mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2\\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2\end{bmatrix}\begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{bmatrix}=\begin{bmatrix}\mathbf{X}_1'\mathbf{Y}\\ \mathbf{X}_2'\mathbf{Y}\end{bmatrix},$$

где разбиение вектора $\mathbf{b}$ на подвекторы соответствует разбиению вектора $\boldsymbol{\beta}$. Поскольку внедиагональные матрицы $\mathbf{X}_1'\mathbf{X}_2=\mathbf{0}$, $\mathbf{X}_2'\mathbf{X}_1=\mathbf{0}$, нормальные уравнения могут быть разбиты на две независимые системы уравнений

$$\mathbf{X}_1'\mathbf{X}_1\mathbf{b}_1=\mathbf{X}_1'\mathbf{Y};\qquad \mathbf{X}_2'\mathbf{X}_2\mathbf{b}_2=\mathbf{X}_2'\mathbf{Y},$$

которые имеют решения

$$\mathbf{b}_1=(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y};\quad \mathbf{b}_2=(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{Y}.$$

Последние основаны на предположении, что матрицы, подвергнутые обращению, неособенные. Поэтому $\mathbf{b}_1$ есть МНК-оценка $\boldsymbol{\beta}_1$ независимо от того, содержится $\boldsymbol{\beta}_2$ в модели или нет. Аналогичное утверждение справедливо и для $\mathbf{b}_2$. Затем

$$\mathrm{SS}(\mathbf{b}_1) = \mathbf{b}_1'\mathbf{X}_1'\mathbf{Y}, \quad \mathrm{SS}(\mathbf{b}_2) = \mathbf{b}_2'\mathbf{X}_2'\mathbf{Y}.$$

Таким образом,

$$\mathrm{SS}(\mathbf{b}_1,\ \mathbf{b}_2) = \mathbf{b}'\mathbf{X}'\mathbf{Y}.$$

Отсюда следует, что

$$\mathrm{SS}(\mathbf{b}_1,\ \mathbf{b}_2) = \mathbf{b}'\mathbf{X}'\mathbf{Y} = \mathbf{b}_1'\mathbf{X}_1'\mathbf{Y} + \mathbf{b}_2'\mathbf{X}_2'\mathbf{Y} = \mathrm{SS}(\mathbf{b}_1) + \mathrm{SS}(\mathbf{b}_2),$$

$$\mathrm{SS}(\mathbf{b}_1 \mid \mathbf{b}_2) = \mathrm{SS}(\mathbf{b}_1,\ \mathbf{b}_2) - \mathrm{SS}(\mathbf{b}_2) = \mathrm{SS}(\mathbf{b}_1).$$

Аналогично

$$\mathrm{SS}(\mathbf{b}_2 \mid \mathbf{b}_1) = \mathrm{SS}(\mathbf{b}_2),$$

и это зависит только от ортогональности столбцов $\mathbf{X}_1$ и $\mathbf{X}_2$. Обобщение на случай $t > 2$ следует непосредственно.

2.9. ЧАСТНЫЕ И ПОСЛЕДОВАТЕЛЬНЫЕ F-КРИТЕРИИ

Мы уже видели, как получить дополнительные суммы квадратов для одного или нескольких оцениваемых коэффициентов при наличии других коэффициентов, используя две модели, одна из которых включает рассматриваемые коэффициенты, а другая нет.

Если в регрессионной модели содержится несколько членов, то мы можем полагать, что они вводятся в уравнение в любой желаемой последовательности. Если мы найдем

$$\mathrm{SS}(b_i \mid b_0,\ b_1,\ \ldots,\ b_{i-1},\ b_{i+1},\ \ldots,\ b_k), \qquad i = 1,\ 2,\ \ldots,\ k,$$

то будем иметь одну степень свободы для суммы квадратов, которая измеряет вклад каждого коэффициента b_i в сумму квадратов, обусловленную регрессией при условии, что все члены, не содержащие β_i, уже входят в модель. Другими словами, будем иметь меру для оценки *включения члена* β_i *в модель*, которая первоначально не содержала такого члена. Или, говоря иначе, мы имеем меру важности параметра *как если бы он был добавлен в модель последним*. Соответствующий средний квадрат, равный сумме квадратов, поскольку она имеет одну степень свободы, может сравниваться с величиной s^2 с помощью F-критерия. Такой вариант F-критерия называют *частным F-критерием* для β_i. Если дополнительный член, который мы рассматриваем, есть $\beta_i X_i$, то мы можем свободно говорить о частном F-критерии для переменной X_i, хотя мы и знаем, что на самом деле речь идет о проверке гипотезы относительно β_i.

При построении подходящей модели частный F-критерий — это полезный критерий для решения вопроса о добавлении или исключении членов из модели. Влияние переменной X (например, X_q) на величину отклика может быть большим, если регрессионное уравнение включает только X_q. Однако если та же переменная входит в модель

после других переменных, она может очень мало влиять на отклик благодаря тому, что величина X_q сильно коррелирована с переменными, уже содержащимися в регрессионном уравнении [15]. Мы можем с помощью частного F-критерия выполнить проверку для всех регрессионных коэффициентов, как будто каждая соответствующая им переменная вводилась в уравнение в последнюю очередь, чтобы видеть относительный эффект каждой переменной. Эта информация может использоваться наряду с другой, когда необходимо выбрать переменные. Допустим, например, что только X_1 или X_2 могут подойти для составления регрессионного уравнения с откликом Y. Предположим далее, что использование X_1 приводит к меньшим ошибкам предсказываемых значений отклика, чем использование X_2. Если при этом обеспечена желаемая точность, то в дальнейшем, вероятно, нужно использовать величину X_1. Однако если величина X_2 была бы управляемой переменной, а X_1 только измеряемой, но не управляемой и если бы при этом управление было более важно, чем предсказание, то следовало бы выбрать в качестве независимой переменной величину X_2, а не X_1.

Если переменные добавляются к регрессионному уравнению последовательно одна за другой, то мы говорим о *последовательном* F-критерии. Это как раз и есть частный F-критерий по отношению к переменной, которая вводится на данной стадии.

(Примечание. Некоторые авторы не любят пользоваться терминами *частный* и *последовательный* F-критерий [16] (и считают их непригодными). Мы подчеркиваем, что это просто удобные, короткие названия для конкретных, теоретически обоснованных F-критериев (см. параграф 2.7).)

В некоторых статистических пакетах программ (например, BMDP=UCLA Biomedical Series P, SPSS=Statistical Packages for the Social Sciences, SAS=Statistical Analysis System) частный F-критерий называется «критерий F для исключения», а последовательный F-критерий называют «критерием F для включения».

Случай, когда $t = F^{1/2}$

Частный F-критерий с числами степеней свободы 1 и ν для проверки гипотезы $H_0 : \beta_j = 0$ против альтернативы $H_1 : \beta_j \neq 0$ равен в точности квадрату t-статистики с ν степенями свободы, которая вычисляется по формуле $t = b_j/\{\text{ст. ошибка } (b_j)\}$, где в знаменателе стоит

---

[15] В данной книге предикторы, как правило, предполагаются неслучайными. Поэтому корректнее говорнть не об их корреляции, а о сопряженности. Вместо термина «корреляционная матрица» правильнее в данном случае употреблять термин «матрица сопряженности предикторов». Моменты такого рода обсуждаются в кн.: Демиденко Е. З. Линейная и нелинейная регрессии.— М.: Финансы и статистика, 1981.— 304 с. (см. с. 186).— *Примеч. пер.*

[16] С самыми свежими таблицами F-распределения можно познакомиться в ки.: Мардиа К., Земроч П. Таблицы F-распределений и распределений, связанных с ними. Пер. с англ.— М.: Наука, 1984.— 256 с.— *Примеч. пер.*

стандартная ошибка коэффициента b_j, равная корню квадратному из соответствующего диагонального элемента матрицы $(\mathbf{X}'\mathbf{X})^{-1}s^2$, а s^2 базируется на ν степенях свободы. Проверка гипотезы может выполняться с использованием либо F-, либо t-статистики. Анализируя таблицы процентных точек, можно убедиться, что $F(1, \nu, 1-\alpha) = \left\{t\left(\nu, 1-\frac{\alpha}{2}\right)\right\}^2$ при любых значениях ν и α. Во многих программах используются обе статистики. Заметим, что в силу ошибок округления точное соотношение между указанными статистиками выдерживается не всегда.

2.10. ПРОВЕРКА ОБЩЕЙ ЛИНЕЙНОЙ ГИПОТЕЗЫ В РЕГРЕССИОННЫХ ЗАДАЧАХ

Экспериментаторы нередко постулируют модели, которые оказываются более общими, чем те, которые они надеются использовать фактически. Допустим, например, что экспериментатор включает в однооткликовую модель два предиктора X_1 и X_2 и имеет набор данных (Y_i, X_{1i}, X_{2i}), $i = 1, 2, \ldots, n$. При этом он подозревает, что хотя оба предиктора, X_1 и X_2, влияют на отклик, однако фактически важна лишь разность $X_1 - X_2$. Если использовать обе величины X, то следует оценивать модель

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \tag{2.10.1}$$

если же оправдано указанное выше подозрение, то надо ориентироваться на модель

$$Y = \beta_0 + \beta(X_1 - X_2) + \varepsilon. \tag{2.10.2}$$

Как проверить такое подозрение? По существу, ставится вопрос: «Может ли быть так, что в уравнении (2.10.1) $\beta_1 = -\beta_2$ ($= \beta$, скажем)?» Или по-другому: «Справедливо ли соотношение $\beta_1 + \beta_2 = 0$?» Таким образом, речь идет о проверке нуль-гипотезы $H_0: \beta_1 + \beta_2 = 0$ против альтернативы $H_1: \beta_1 + \beta_2 \neq 0$. Поскольку H_0 включает утверждение о линейной комбинации параметров β, мы называем ее *линейной гипотезой* [17].

Линейные гипотезы обычно вытекают из *знаний экспериментатора или его предположений относительно возможных моделей*. Они могут также появиться в результате консультаций у специалиста-стати-

[17] С теоретическими аспектами проверки общей линейной гипотезы о параметрах можно познакомиться в таких основополагающих руководствах, как: Шеффе Г. Дисперсионный анализ. Пер. с англ.— 2-е изд.— М.: Физматгиз, 1980.— 626 с.; Андерсон Т. Введение в многомерный статистический анализ/Пер. с англ. Под ред. Б. В. Гнеденко.— М.: Физматгиз, 1963.— 500 с.; Леман Э. Проверка статистических гипотез. Пер. с англ.— М.: Наука, 1964.— 500 с.; Рао С. Линейные статистические методы и их применения/Пер. с англ. Под. ред. Ю. В. Линника.— М.: Наука, 1968.— 548 с.; Себер Дж. Линейный регрессионный анализ/Пер. с англ. Под ред. М. Б. Малютова.— М.: Мир, 1980.— 456 с. Популярное изложение некоторых вопросов, связанных с этой проблемой, содержится в кн.: Колкот Э. Проверка значимости. Пер. с англ.— М.: Статистика, 1978.— 128 с.— *Примеч. пер.*

стика, если последний достаточно глубоко вникнет в суть задачи, чтобы понять ее на таком уровне. В идеале статистик и должен быть таким, но на практике это происходит не всегда.

Линейные гипотезы могут включать не одно, а несколько утверждений о параметрах β. Теперь мы приведем дополнительные примеры линейных гипотез, объясним, в общем, как они проверяются, и проиллюстрируем процедуру на простых численных примерах. H_1 всегда будет утверждением, что H_0 в некотором смысле не верна, а в каком — это мы не будем оговаривать особо в примерах.

Подчеркнем снова, что принцип дополнительной суммы квадратов есть специальный частный случай рассматриваемой здесь проблемы.

Пример 1. Модель:

$$\mathrm{E}(Y)=\beta_0+\beta_1X_1+\beta_2X_2.$$

$$H_0:\beta_1=0,$$

$\beta_2=0$ (две линейные функции независимые).

Под «независимыми» мы понимаем *линейно* независимые, так что одно выражение не может быть представлено в виде линейной комбинации остальных выражений, входящих в группу.

Пример 2. Модель:

$$\mathrm{E}(Y)=\beta_0+\beta_1X_1+\beta_2X_2+\ \ldots\ +\beta_kX_k.$$

$$H_0:\beta_1=0,$$

$$\beta_2=0,$$

$$\ldots$$

$\beta_k=0$ (k линейных функций, все независимы).

Пример 3. Модель:

$$\mathrm{E}(Y)=\beta_0+\beta_1X_1+\beta_2X_2+\ \ldots\ +\beta_kX_k.$$

$$H_0:\beta_1-\beta_2=0,$$

$$\beta_2-\beta_3=0,$$

$$\ldots$$

$\beta_{k-1}-\beta_k=0$ ($k-1$ линейных функций независимых).

Заметим, что это соответствует, скажем, такой гипотезе:

$$H_0:\beta_1=\beta_2=\ \ldots\ =\beta_k=\beta.$$

Пример 4 (Общий случай). Модель:

$$\mathrm{E}(Y)=\beta_0+\beta_1X_1+\beta_2X_2+\ \ldots\ +\beta_kX_k.$$

$$H_0: c_{10}\beta_0+c_{11}\beta_1+c_{12}\beta_2+\ \ldots\ +c_{1k}\beta_k=0,$$

$$c_{20}\beta_0+c_{21}\beta_1+c_{22}\beta_2+\ \ldots\ +c_{2k}\beta_k=0,$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$c_{m\,0}\beta_0+c_{m1}\beta_1+c_{m2}\beta_2+\ \ldots\ +c_{mk}\beta_k=0.$$

В данной гипотезе участвует m линейных функций от параметров $\beta_0, \beta_1, \ldots, \beta_k$ и при этом не обязательно, чтобы они были независимыми. Гипотеза H_0 может быть выражена в матричной форме:

$$H_0 : \mathbf{C}\beta = 0,$$

где

$$\mathbf{C} = \begin{bmatrix} c_{10} & c_{11} & c_{12} & \ldots & c_{1k} \\ c_{20} & c_{21} & c_{22} & \ldots & c_{2k} \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ . & . & . & \ldots & . \\ c_{m0} & c_{m1} & c_{m2} & \ldots & c_{mk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{bmatrix}.$$

В дальнейшем мы будем полагать, что m функций в общем случае зависимы и при этом первые из них по порядку q функций независимы, а остальные $m—q$ зависят от них. Следовательно, если мы имеем эти первые q независимых функций, то можем составить линейные комбинации из них так, чтобы получить оставшиеся $m—q$ линейных функций.

Ранее мы видели, как можно проверить гипотезы вида, указанного в примерах 1 и 2. Теперь мы объясним, как можно проверить более общие гипотезы.

Проверка общей линейной гипотезы $\mathbf{C}\beta = 0$

Предположим, что рассматриваемая модель, которая предполагается правильной, имеет вид

$$\mathbf{E}(\mathbf{Y}) = \mathbf{X}\beta,$$

где $\mathbf{Y}$ есть $(n \times 1)$-вектор, $\mathbf{X}$ — $(n \times p)$-матрица, β — $(p \times 1)$-вектор.

Если $\mathbf{X}'\mathbf{X}$ — неособенная матрица, то можем оценить вектор β с помощью соотношения

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Остаточная сумма квадратов для этого случая, как мы видели, задается выражением

$$\mathrm{SSE} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}.$$

Она имеет $n—p$ степеней свободы. Линейная гипотеза, подлежащая проверке, $H_0 : \mathbf{C}\beta = 0$ дает q независимых ограничений на параметры $\beta_0, \beta_1, \ldots, \beta_k$, поскольку условие $\mathbf{C}\beta = 0$ представляет собой m уравнений, из которых только q независимых. Мы можем использовать q независимых уравнений, чтобы выразить q коэффициентов β через остальные $p—q$ коэффициентов. Подстановка этих решений в исходную модель дает преобразованную модель в виде

$$\mathbf{E}(\mathbf{Y}) = \mathbf{Z}\alpha,$$

где α — вектор параметров, подлежащих оцениванию. Число таких параметров равно $p—q$. Правая часть уравнения $\mathbf{Z}\alpha$, где $\mathbf{Z}$ есть

$n \times (p-q)$-матрица, а $\boldsymbol{\alpha} - (p-q) \times 1$-вектор, представляет собой результат подстановки в $\mathbf{X}\boldsymbol{\beta}$ параметров β, на которые наложены ограничения.

Теперь мы можем оценить вектор параметров $\boldsymbol{\alpha}$, входящий в новую модель,

$$\mathbf{a} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y},$$

если матрица $\mathbf{Z}'\mathbf{Z}$ неособенная. Новая остаточная сумма квадратов будет иметь вид

$$\mathrm{SSW} = \mathbf{Y}'\mathbf{Y} - \mathbf{a}'\mathbf{Z}'\mathbf{Y}.$$

Эта сумма квадратов имеет $n-p+q$ степеней свободы.

Так как во второй форме записи участвует меньше параметров, SSW всегда будет больше, чем SSE. Разность сумм SSW — SSE называется *суммой квадратов, обусловленной гипотезой* $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, и имеет $(n-p+q) - (n-q) = q$ степеней свободы. Проверка гипотезы $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ может быть выполнена с помощью отношения

$$\frac{(\mathrm{SSW} - \mathrm{SSE})/q}{\mathrm{SSE}/(n-p)}.$$

Последнее имеет $F(q, n-p)$-распределение. Если ошибки независимы и распределены нормально, то проверка с помощью этого критерия будет корректной.

Соответствующая проверка применительно к примерам 1 и 2 (уже реализованная в виде соотношения (2.6.13), где $k = p-1$) представляет собой частный случай общей процедуры. Преобразованная модель в обоих случаях имеет вид

$$\mathrm{E}(\mathbf{Y}) = \mathbf{1}\beta_0,$$

где $\mathbf{1}' = (1, 1, \ldots, 1)$ — вектор, образованный из единиц. Другой способ записи этой модели:

$$\mathrm{E}(Y_i) = \beta_0, \qquad i = 1, 2, \ldots, n.$$

Поскольку $b_0 = \bar{Y}$, $\mathrm{SSW} = \mathbf{Y}'\mathbf{Y} - n\bar{Y}^2$ с $(n-1)$ степенями свободы, тогда как $\mathrm{SSE} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$ и имеет $(n-k-1)$ степеней свободы. Следовательно, отношение для проверки гипотезы $\beta_1 = \beta_2 = \ldots = \beta_k = 0$ (для примера 2; при $k = 2$ имеем пример 1) есть просто

$$\frac{\mathbf{b}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2}{k} \Big/ \frac{\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}}{n-k-1},$$

и это отношение — случайная величина, подчиняющаяся $F(k, n-k-1)$-распределению, что в точности совпадает с процедурой, выраженной формулой (2.6.13), где

$$k = p-1, \qquad v = n-k-1, \qquad s^2 = \mathrm{MS_E} = \mathrm{SSE}/v.$$

Проиллюстрируем теперь применение процедуры на простом, но не таком уж типичном случае.

Рабочий пример. Дана модель $E(\mathbf{Y}) = \mathbf{X}\beta$, проверить гипотезу $H_0 : \mathbf{C}\beta = \mathbf{0}$, где

$$\mathbf{Y}' = (1,\ 4,\ 8,\ 9,\ 3,\ 8,\ 9),$$

$$\beta' = (\beta_0,\ \beta_1,\ \beta_2,\ \beta_{11}),$$

$$\mathbf{X} = \begin{matrix} 1 & X_1 & X_2 & X_1^2 \end{matrix}\begin{bmatrix} 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 1 \\ 0 & 2 & -2 & 3 \end{bmatrix}.$$

Решение. Сначала найдем остаточную сумму квадратов при условии, что подбирается модель

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2.$$

Имеем

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 7 & 0 & 3 & 4 \\ 0 & 4 & 0 & 0 \\ 3 & 0 & 9 & 0 \\ 4 & 0 & 0 & 4 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{2} & 0 & -\frac{1}{6} & -\frac{1}{2} \\ 0 & \frac{1}{4} & 0 & 0 \\ -\frac{1}{6} & 0 & \frac{1}{6} & \frac{1}{6} \\ -\frac{1}{2} & 0 & \frac{1}{6} & \frac{3}{4} \end{bmatrix},$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 42 \\ 4 \\ 38 \\ 22 \end{bmatrix},$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \frac{11}{3} \\ 1 \\ 3 \\ \frac{11}{6} \end{bmatrix}, \qquad \mathbf{b}'\mathbf{X}'\mathbf{Y} = 312{,}33,$$

$$\mathbf{Y}'\mathbf{Y} = 316.$$

$$\mathrm{SSE} = 316 - 312{,}33 = 3{,}67.$$

Уравнения, соответствующие нулевой гипотезе $H_0 : \mathbf{C\beta} = \mathbf{0}$, имеют вид

$$\beta_{11} = 0,$$

$$\beta_1 - \beta_2 = 0,$$

$$\beta_1 - \beta_2 + \beta_{11} = 0,$$

$$2\beta_1 - 2\beta_2 + 3\beta_{11} = 0.$$

Гипотеза H_0 может быть записана более просто в форме $H_0 : \beta_{11} = 0$, $\beta_1 = \beta_2 = \beta$, поскольку третье и четвертое уравнения являются линейными комбинациями первого и второго.

Подстановка этих выражений преобразует модель

$$E(Y) = \beta_0 + \beta(X_1 + X_2) = \alpha_0 + \alpha Z,$$

где

$$\alpha_0 = \beta_0, \qquad \alpha = \beta, \qquad Z = X_1 + X_2.$$

Таким образом,

$$\mathbf{Z} = \begin{bmatrix} 1 & (-1-1) \\ 1 & (1-1) \\ 1 & (-1+1) \\ 1 & (1+1) \\ 1 & (0+0) \\ 1 & (0+1) \\ 1 & (0+2) \end{bmatrix} = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix},$$

$$\mathbf{Z'Y} = \begin{bmatrix} 42 \\ 42 \end{bmatrix}, \qquad (\mathbf{Z'Z})^{-1} = \begin{bmatrix} 7 & 3 \\ 3 & 13 \end{bmatrix}^{-1} = \frac{1}{82} \begin{bmatrix} 13 & -3 \\ -3 & 7 \end{bmatrix},$$

$$\mathbf{a} = (\mathbf{Z'Z})^{-1}\mathbf{Z'Y} = \frac{21}{41} \begin{bmatrix} 10 \\ 4 \end{bmatrix}, \qquad \mathbf{a'Z'Y} = 301{,}17,$$

$$\mathrm{SSW} = 316 - 301{,}17 = 14{,}83.$$

Далее $p = 4$, $n = 7$, $q = 2$, $n-p=3$ и SSW—SSE=14,83—3,67 = =11,16=SS, обусловленная гипотезой. Соответствующая статистика для проверки гипотезы H_0 имеет, таким образом, значение (11,16/2) ÷ ÷(3,67/3) = 4,56.

Поскольку F (2, 3, 0,95) = 9,55, мы *не отвергаем* гипотезу H_0. Так как исходная модель выражалась уравнением $E(Y) = \beta_0 + + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2$ и гипотеза *не* отвергнута, принятие гипотезы $\beta_{11} = 0$, $\beta_1 = \beta_2 = \beta$ приводит к более правдоподобной модели

$$E(Y) = \beta_0 + \beta(X_1 + X_2).$$

2.11. ВЗВЕШЕННЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

Иногда случается так, что часть наблюдений, используемых в регрессионном анализе, менее надежна, чем остальные. Обычно это означает, что не все дисперсии наблюдений равны. Другими словами, матрица $\mathbf{V}(\varepsilon)$ не имеет вида $\mathbf{I}\sigma^2$, а оказывается диагональной матрицей с неравными элементами. В некоторых задачах может оказаться также, что не равны нулю еще и внедиагональные элементы матрицы $\mathbf{V}(\varepsilon)$, т. е. наблюдения коррелированы.

Если имеет место тот или другой случай, обычная формула $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ для отыскания МНК-оценок неприменима и для их получения необходимо изменить процедуру. Основная идея базируется на преобразовании наблюдений $\mathbf{Y}$ в новые переменные $\mathbf{Z}$, которые удовлетворяют обычно используемым предположениям ($\mathbf{Z} = \mathbf{Q}\beta + \mathbf{f}$, $E(\mathbf{f}) = 0$, $\mathbf{V}(\mathbf{f}) = \mathbf{I}\sigma^2$). Кроме того, для использования F-критерия и построения доверительных интервалов важно, чтобы выполнялось условие $\mathbf{f} \sim N(0, \mathbf{I}\sigma^2)$. К полученным таким образом переменным можно теперь применить обычный (невзвешенный) метод наименьших квадратов. Затем оценки [18] можно снова выразить через исходные переменные $\mathbf{Y}$. Рассмотрим теперь, как в этом случае изменится обычная регрессионная процедура.

Допустим, что исследуется модель

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \tag{2.11.1}$$

где

$$E(\varepsilon) = 0, \quad \mathbf{V}(\varepsilon) = \mathbf{V}\sigma^2 \quad \text{и} \quad \varepsilon \sim N(0, \mathbf{V}\sigma^2). \tag{2.11.2}$$

Покажем, что может быть найдена единственная неособенная симметричная матрица $\mathbf{P}$, такая, что

$$\mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P} = \mathbf{P}^2 = \mathbf{V}. \tag{2.11.3}$$

Запишем:

$$\mathbf{f} = \mathbf{P}^{-1}\varepsilon, \text{ так что } E(\mathbf{f}) = 0. \tag{2.11.4}$$

Если $\mathbf{f}$ — случайный вектор, такой, что $E(\mathbf{f}) = 0$, то $E(\mathbf{f}\mathbf{f}') = \mathbf{V}(\mathbf{f})$, где операция математического ожидания распространяется отдельно на каждый элемент квадратной $(n \times n)$-матрицы $\mathbf{f}\mathbf{f}'$. Следовательно,

$$\begin{aligned}\mathbf{V}(\mathbf{f}) = E(\mathbf{f}\mathbf{f}') = E(\mathbf{P}^{-1}\varepsilon\varepsilon'\mathbf{P}^{-1}) = (\text{так как } (\mathbf{P}^{-1})' = \mathbf{P}^{-1}) = \\ = \mathbf{P}^{-1}E(\varepsilon\varepsilon')\mathbf{P}^{-1} = \mathbf{P}^{-1}\mathbf{P}\mathbf{P}\mathbf{P}^{-1}\sigma^2 = \mathbf{I}\sigma^2.\end{aligned} \tag{2.11.5}$$

Верно также, что $\mathbf{f} \sim N(0, \mathbf{I}\sigma^2)$, т. е. $\mathbf{f}$ — вектор с нормальным распределением, поскольку элементы вектора $\mathbf{f}$ состоят из линейных ком-

---

[18] Взвешенный метод наименьших квадратов подробно рассматривается в книгах, посвященных регрессионному анализу (см. примечание на с. 121). Некоторые важные вопросы, связанные с проверкой статистических гипотез в рамках взвешенного МНК, затронуты в статье: Горский В. Г., Адлер Ю. П. О методологии регрессионного и дисперсионного анализа при планировании эксперимента с неравномерным дублированием опытов.— Заводская лаборатория, 1971, т. 37, № 3, с. 319—325.— *Примеч. пер.*

бинаций элементов вектора ε, которые сами по себе распределены нормально.

Поэтому, если мы умножим обе части уравнения (2.11.1) слева на матрицу $\mathbf{P}^{-1}$, то получим новую модель

$$\mathbf{P}^{-1}\mathbf{Y} = \mathbf{P}^{-1}\mathbf{X}\beta + \mathbf{P}^{-1}\varepsilon \tag{2.11.6}$$

или

$$\mathbf{Z} = \mathbf{Q}\beta + \mathbf{f}. \tag{2.11.7}$$

Обозначения в формуле (2.11.7) очевидны. Понятно, что мы можем теперь применить обычную теорию метода наименьших квадратов к уравнению (2.11.7), так как $E(\mathbf{f}) = \mathbf{0}$ и $\mathbf{V}(\mathbf{f}) = \mathbf{I}\sigma^2$. Остаточная сумма квадратов равна:

$$\mathbf{f}'\mathbf{f} = \varepsilon'\mathbf{V}^{-1}\varepsilon = (\mathbf{Y} - \mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\beta). \tag{2.11.8}$$

Нормальные уравнения $\mathbf{Q}'\mathbf{Q}\mathbf{b} = \mathbf{Q}'\mathbf{Z}$ можно записать так:

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Y} \tag{2.11.9}$$

с решением

$$\mathbf{b} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}, \tag{2.11.10}$$

если матрица $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ неособенная. Сумма квадратов, обусловленная регрессией, имеет вид

$$\mathbf{b}'\mathbf{Q}'\mathbf{Z} = \mathbf{Y}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}, \tag{2.11.11}$$

а полная сумма квадратов выражается соотношением

$$\mathbf{Z}'\mathbf{Z} = \mathbf{Y}'\mathbf{V}^{-1}\mathbf{Y}. \tag{2.11.12}$$

Разность между суммами (2.11.12) и (2.11.11) дает остаточную сумму квадратов. Сумма квадратов, обусловленная средним, есть $(\Sigma Z_i)^2/n$, где Z_i—n элементов вектора $\mathbf{Z}$. Заметим, что если мы вычтем эту величину из уравнения (2.11.11), то разность не будет дополнительной суммой квадратов в обычном смысле, поскольку преобразованная модель не содержит больше параметра β_0. Следовательно, подходящая сумма для вычитания есть та, что связана с первым компонентом уравнения (2.11.7).

Матрица дисперсий-ковариаций вектора $\mathbf{b}$ есть

$$\mathbf{V}(\mathbf{b}) = (\mathbf{Q}'\mathbf{Q})^{-1}\sigma^2 = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\sigma^2. \tag{2.11.13}$$

Совместная доверительная область для всех параметров может быть получена из неравенства

$$(\mathbf{b} - \beta)\,\mathbf{Q}'\mathbf{Q}\,(\mathbf{b} - \beta) \leqslant \left[\frac{p}{(n-p)}\right](\mathbf{Z}'\mathbf{Z} - \mathbf{b}'\mathbf{Q}'\mathbf{Z})\,F(p,\ n-p,\ 1-\alpha). \tag{2.11.14}$$

При желании это выражение можно преобразовать, используя уравнения (2.11.11), (2.11.12) и подстановку $\mathbf{Q} = \mathbf{P}^{-1}\mathbf{X}$.

Остатки во взвешенном методе наименьших квадратов

Остатки, которые должны анализироваться, представляют собой оценки составляющих вектора $\mathbf{f} = \mathbf{P}^{-1}\boldsymbol{\varepsilon}$. Они выражаются вектором $\mathbf{P}^{-1}(\mathbf{Y} - \hat{\mathbf{Y}})$, где $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$ и $\mathbf{b}$ берутся из уравнения (2.11.11). Следовательно, их можно выразить в виде

$$\mathbf{P}^{-1}\{\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\}\mathbf{Y}.$$

Аналогичная формула применяется, если оценивается матрица $\mathbf{V}$ (см. параграф 3.9).

Общие замечания

Наиболее простой случай приложения взвешенного метода наименьших квадратов имеет место тогда, когда наблюдения независимы, но дисперсии различны, так что

$$\mathbf{V}\sigma^2 = \begin{vmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{vmatrix},$$

где некоторые из величин σ_i^2 могут быть и равными.

В практических задачах получить вначале определенную информацию о форме матрицы $\mathbf{V}$ зачастую трудно. Поэтому иногда стоит (имея в виду возможность ошибиться) делать предположение, что $\mathbf{V} = \mathbf{I}$, а затем попытаться определить форму матрицы $\mathbf{V}$, исследуя остатки (см. гл. 3).

Если по существу задачи требуется использовать взвешенный метод наименьших квадратов, но фактически применяется обычный метод наименьших квадратов, то получаемые с его помощью оценки параметров будут все же несмещенными, хотя они и не будут иметь наименьшие дисперсии, поскольку оценки с минимальными дисперсиями получаются на основе метода наименьших квадратов с правильными весами.

Когда применяется нормальный метод наименьших квадратов, оценки получаются из формулы $\mathbf{b}_0 = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ и при этом

$$E(\mathbf{b}_0) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta},$$

но

$$\mathbf{V}(\mathbf{b}_0) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{V}(\mathbf{Y})]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\sigma^2.$$

Напомним: из уравнения (2.11.13) следует, что при правильной обработке

$$\mathbf{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\sigma^2$$

и вообще элементы этой матрицы порождают меньшие значения дис-

персий как для оценок отдельных коэффициентов, так и для их линейных функций.

Пример использования взвешенного метода наименьших квадратов

Это чрезвычайно простой, но интересный пример. Предположим, что мы желаем оценить модель

$$E(\mathbf{Y}) = \boldsymbol{\beta}\mathbf{X}.$$

Допустим, что

$$\mathbf{V}\sigma^2 = \mathbf{V}(\mathbf{Y}) = \begin{bmatrix} 1/w_1 & & 0 \\ & 1/w_2 & \\ 0 & & 1/w_n \end{bmatrix} \sigma^2,$$

где w — веса, которые должны быть определены. Отсюда имеем:

$$\mathbf{V}^{-1} = \begin{bmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \cdot & \\ & & & \cdot \\ & & & \cdot \\ 0 & & & w_n \end{bmatrix}.$$

Применяя общие формулы, приведенные выше, мы найдем после некоторых преобразований

$$b = \frac{\Sigma w_i X_i Y_i}{\Sigma w_i X_i^2},$$

где суммирование ведется по i, $i = 1, 2, \ldots, n$.

Случай 1. Допустим, что $\sigma_1^2 = V(Y_i) = kX_i$, т. е. дисперсия величины Y_i пропорциональна соответствующему значению величины X_i. Тогда $w_i = \sigma^2/kX_i$. Следовательно,

$$b = \frac{\Sigma Y_i}{\Sigma X_i} = \frac{\overline{Y}}{\overline{X}}.$$

Поэтому, если дисперсия отклика Y_i пропорциональна X_i, то наилучшая оценка коэффициента регрессии есть среднее от Y_i, деленное на среднее от X_i. Кроме того,

$$V(b) = \frac{\sigma^2}{\Sigma w_i X_i^2} = \frac{k}{\Sigma X_i}.$$

Случай 2. Пусть $\sigma_i^2 = V(Y_i) = kX_i^2$, т. е. дисперсия отклика Y_i пропорциональна квадрату соответствующего значения X_i. Тогда $w_i = \sigma^2/kX_i^2$. Следовательно,

$$b = \frac{\Sigma (Y_i/X_i)}{\Sigma 1} = \frac{\Sigma (Y_i/X_i)}{n}.$$

Таким образом, если дисперсия отклика Y_i пропорциональна X_i^2, то наилучшая оценка регрессионного коэффициента есть среднее из n отношений $\frac{Y_i}{X_i}$, получаемых на основе каждой пары наблюдений. Кроме того,

$$V(b) = \frac{\sigma^2}{\Sigma w_i X_i^2} = \frac{k}{n}.$$

(Примечание. Подбор уравнения прямой линии, проходящей через начало координат $(X, Y) = (0, 0)$, базируется на излишне строгих предположениях, которые, в общем, не оправдываются. Даже если априори известно, что прямая должна проходить через начало координат (как было бы, например, в случае, если Y — тормозной путь, а X — скорость движения транспорта), то это не означает, что построенная прямая обязательно точно пройдет через начало координат. Согласно имеющимся данным прямая может не пройти через начало координат, однако при большем объеме данных может оказаться адекватной модель более высокого порядка, согласно которой линия должна проходить через начало координат. Вообще лучше исходить из предположения, что модель может содержать коэффициент β_0, а потом, после нахождения оценки b_0, проверять гипотезу о его незначимости. Это замечание справедливо как для взвешенного, так и для обычного метода наименьших квадратов.)

Численный пример использования взвешенного метода наименьших квадратов

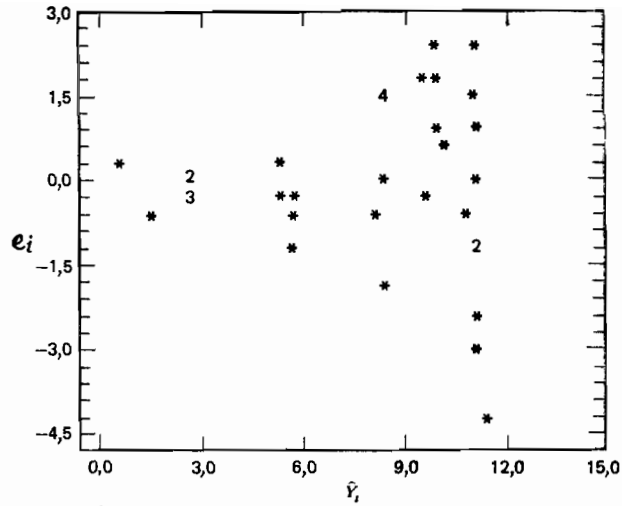
В табл. 2.1 представлены данные, некоторым образом упорядоченные для того, чтобы было удобнее проводить их анализ. Они включают 35 наблюдений, точнее, результатов опытов (X_i, Y_i), часть из которых строго повторные, тогда как другие лишь приблизительно повторные. И те и другие выделены с помощью соответствующих группировок. Обработка данных с помощью обычного метода наименьших квадратов приводит к модели

$$\hat{Y} = -0{,}5790 + 1{,}1354X,$$

графики остатков показаны на рис. 2.3. Из этих данных ясно видно *, что наблюдения имеют неодинаковые дисперсии. Полный график остатков (на рисунке не показан) представляет собой полосу, отчасти скошенную по направлению к отрицательным значениям. Никакой обычный, невзвешенный метод наименьших квадратов в данном случае не подходит, и, по-видимому, целесообразно воспользоваться взвешенным методом наименьших квадратов.

---

* Мы еще не касались исследования остатков, об этом речь пойдет в гл. 3. Здесь мы просто укажем, что размер вертикального размаха остатков свидетельствует о дисперсии, соответствующей «истинным» ошибкам в модели. Следовательно, «раструбообразное» расположение точек на рис. 2.3 указывает на то, что дисперсии не постоянны. Это нарушение одной из основных предпосылок невзвешенного метода наименьших квадратов: $V(Y_i) = \sigma^2$ (для всех i).



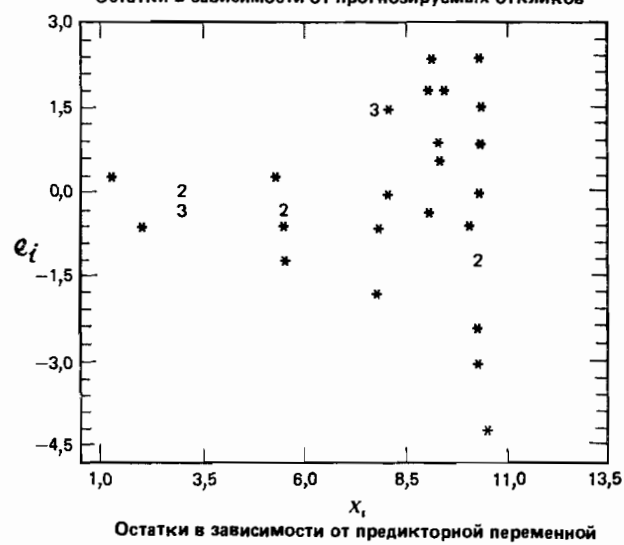


Рис. 2.3. Графики остатков при обычном методе наименьших квадратов. (Две неразличимые точки помечены цифрой 2 и т. д.)

Таблица 2.1. Данные примера использования взвешенного метода наименьших квадратов

X	Y	$\widehat{w}_i$	X	Y	$\widehat{w}_i$
1,15	0,99	1,24028	7,94	8,50	0,78342
1,90	0,98	2,18224	9,03	9,47	0,47385
3,00	2,60	7,84930	9,07	11,45	0,46621
3,00	2,67	7,84930	9,11	12,14	0,45878
3,00	2,66	7,84930	9,14	11,50	0,45327
3,00	2,78	7,84930	9,16	10,65	0,44968
3,00	2,80	7,84930	9,37	10,64	0,41435
5,34	5,92	7,43652	10,17	9,78	0,31182
5,38	5,35	6,99309	10,18	12,39	0,31079
5,40	4,33	6,78574	10,22	11,03	0,30672
5,40	4,89	6,78574	10,22	8,00	0,30672
5,45	5,21	6,30514	10,22	11,90	0,30672
7,70	7,68	0,89204	10,18	8,68	0,31079
7,80	9,81	0,84420	10,50	7,25	0,28033
7,81	6,52	0,83963	10,23	13,46	0,30571
7,85	9,71	0,82171	10,03	10,19	0,32680
7,87	9,82	0,81296	10,23	9,93	0,30571
7,91	9,81	0,79588			

Источник. Wanda M. Hinshaw.

Мы предполагаем (пока нет никаких противопоказаний), что Y_i независимы, так что матрица $\mathbf{V}$ диагональна с различными значениями дисперсий, о чем говорилось ранее. Теперь важно получить информацию о значениях этих дисперсий. Для каждой группы повторных или почти повторных опытов вычислим средние значения величин X, обозначенные, скажем, как $\overline{X}_j$, а также средние квадраты, связанные с «чистой» ошибкой, s_{ej}^2. Они оказались следующими:

$\overline{X}_j$	3,0	5,4	7,8	9,1	10,2
s_{ej}^2	0,0072	0,3440	1,7404	0,8683	3,8964

Зависимость $\widehat{s}_e^2$ от $\overline{X}$ была постулирована в виде квадратичного полинома; с использованием метода наименьших квадратов получили

$$\widehat{s}_e^2 = 1{,}5329 - 0{,}7334\overline{X} + 0{,}0883\overline{X}^2.$$

Теперь можно оценить величины s_{ei}^2, $i = 1, 2, \ldots, 35$, подставляя соответствующие значения X_i в приведенное уравнение. Затем можно найти обратные значения от этих величин и таким образом получить оценки весов $\widehat{w}_i$, указанные в табл. 2.1. Матрица $\mathbf{P}$ в нашем случае диагональна с элементами $\widehat{w}_i^{-\frac{1}{2}}$. Использование этих весов приводит к регрессии

$$\widehat{Y} = -0{,}8891 + 1{,}1468X,$$

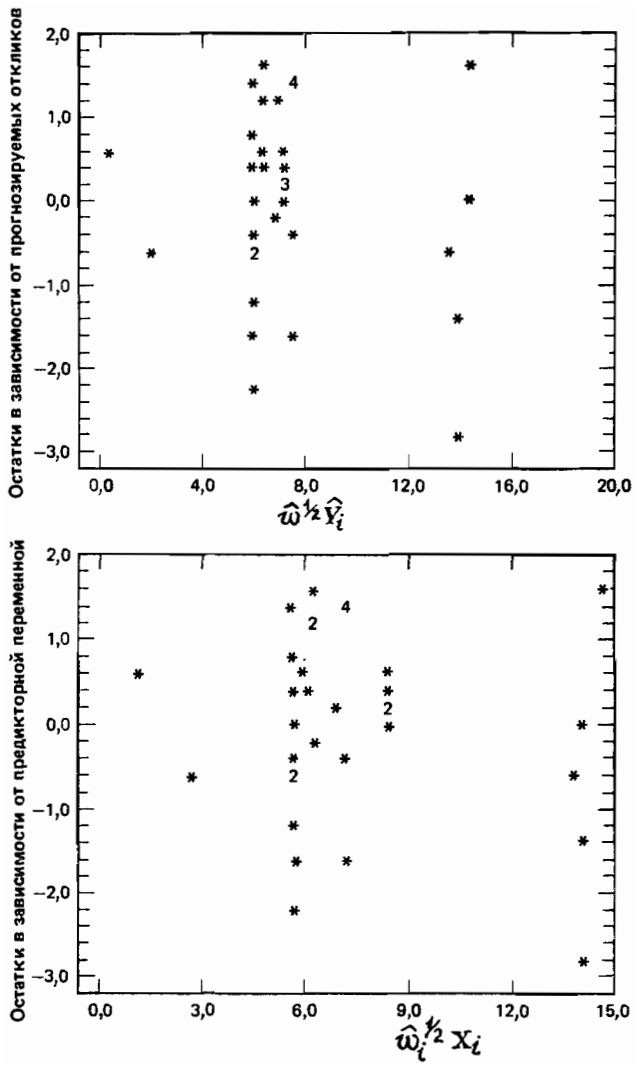


Рис. 2.4. Графики остатков при взвешенном методе наименьших квадратов

а таблица дисперсионного анализа имеет вид:

Источник	Число степеней свободы	SS	MS
$b_1 \mid b_0$	1	496,96	496,96
Остаток	33	42,66	1,29
Общий, скорректированный	34	539,62	

Соответствующие «наблюдаемые значения откликов» и «предсказываемые значения откликов» есть теперь $\hat{w}_i^{1/2} Y_i$ и $\hat{w}_i^{1/2} \hat{Y}_i$, а «остатки», подлежащие исследованию, имеют вид $\hat{w}_i^{1/2} (Y_i - \hat{Y}_i)$. Полный график остатков по-прежнему обнаруживает некоторую скошенность, но картина выглядит лучше. Графики остатков на рис. 2.4 показывают, что вертикальные размахи остатков теперь примерно одинаковые на двух основных уровнях преобразованных откликов. На нижних уровнях имеются только два наблюдения, что не слишком много для оценки размаха. Применение взвешенного метода наименьших квадратов, по-видимому, оправдано и полезно. Прекрасные программы взвешенного метода наименьших квадратов для расчетов на ЭВМ приведены в Minitab series (см.: Ryan T. A., Joiner B. L., Ryan B. F. MINITAB Student Handbook.— Duxbury: Duxbury Press, MA, 1976).

Вычислительные аспекты

Читатель, интересующийся вычислительными аспектами этого метода, может ознакомиться с ними в работе: Gentleman W. M. (Algol 60) Algorithm AS 75. Basic procedures for large, sparse or weighted linear least squares problems. Applied Statistics, 1974, 23, p. 448—454 и в ссылках, приведенных там.

2.12. СМЕЩЕНИЕ РЕГРЕССИОННЫХ ОЦЕНОК

Мы уже указывали ранее (см. параграф 2.6), что МНК-оценка $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ вектора $\boldsymbol{\beta}$, входящего в модель $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, есть несмещенная оценка. Это означает, что

$$E(\mathbf{b}) = \boldsymbol{\beta}.$$

Значит, если мы рассмотрим распределение случайного вектора $\mathbf{b}$ (которое можно получить, повторяя выборки одинакового объема из совокупности величин Y при фиксированной матрице $\mathbf{X}$ и определяя оценку вектора $\boldsymbol{\beta}$ по каждой выборке), то среднее для этого распределения как раз и будет равно $\boldsymbol{\beta}$.

Теперь подчеркнем, что это верно *только тогда, когда постулированная модель «правильна»*. Если модель *не правильна*, то оценки будут *смещенными*, т. е. $E(\mathbf{b}) \neq \boldsymbol{\beta}$. Величина смещения зависит, как мы уже видели, не только от постулированной и «правильной» модели, но

также и от значений переменных X, которые используются в регрессионных вычислениях. Если проводится заранее спланированный (активный) эксперимент, то величина смещения зависит как от экспериментального плана, так и от модели.

С самого начала условимся, что будем иметь дело с неособенной регрессионной моделью общего вида, и, поскольку мы имеем необходимые формулы в матричной форме, они могут применяться всюду. В отдельных случаях, если есть желание, можно для упражнения сделать также выкладки в скалярной форме. Предположим, что мы постулируем модель

$$E(\mathbf{Y}) = \mathbf{X}_1\beta_1. \tag{2.12.1}$$

Это приводит к МНК-оценкам

$$\mathbf{b}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}. \tag{2.12.2}$$

Если постулируемая модель правильна, то, поскольку $\mathbf{X}_1$ есть неслучайная матрица, а $\mathbf{b}_1$ и $\mathbf{Y}$ — случайные, будем иметь

$$E(\mathbf{b}_1) = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'E(\mathbf{Y}) = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_1\beta_1 = \beta_1. \tag{2.12.3}$$

Таким образом, $\mathbf{b}_1$ есть несмещенная оценка вектора β_1. Допустим опять-таки, что постулируется модель (2.12.1), так что $\mathbf{b}_1$ в соответствии с (2.12.2) есть вектор оценок регрессионных коэффициентов. Однако предположим теперь, что «истинное» соотношение между вектором математических ожиданий отклика и переменными есть не (2.12.1), а

$$E(\mathbf{Y}) = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2, \tag{2.12.4}$$

т. е. имеются члены $\mathbf{X}_2\beta_2$, которые мы не учитывали при определении оценок. Отсюда следует, что

$$\begin{aligned} E(\mathbf{b}_1) &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'E(\mathbf{Y}) = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'(\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2) = \\ &= (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_1\beta_1 + (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2\beta_2 = \beta_1 + \mathbf{A}\beta_2, \end{aligned} \tag{2.12.5}$$

где

$$\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 \tag{2.12.6}$$

называется *матрицей смещения*. Заметим, что члены вектора $\mathbf{A}\beta_2$ зависят не только от постулированной и «истинной» модели, но также и от плана эксперимента, который определяется матрицами $\mathbf{X}_1$ и $\mathbf{X}_2$. Таким образом, хороший выбор плана может привести к получению оценок с меньшим смещением, даже если была постулирована и оценивалась неправильная модель [19]. Теперь мы проиллюстрируем применение уравнения (2.12.5) на простых численных примерах.

---

[19] Анализ и планирование эксперимента в условиях неадекватности постулированной модели рассматриваются в ряде работ: Налимов В. В., Чернова Н. А. Статистические методы планирования экстремальных экспериментов.— М.: Наука, 1965.— 340 с. (см. гл. 10, параграф 5); Седунов Е. В. Оптимальное планирование и анализ регрессионных экспериментов с учетом систематической ошибки (обзор).— Заводская лаборатория, 1979, № 1, с. 55—62. Этот вопрос обсуждается также в наших примечаниях к гл. 5.— *Примеч. пер.*

Пример 1. Допустим, что мы постулируем модель в виде

$$E(Y)=\beta_0+\beta_1 X,$$

тогда как фактически модель выражается соотношением

$$E(Y)=\beta_0+\beta_1 X+\beta_{11}X^2,$$

но это нам неизвестно. Если мы используем наблюдения за величиной Y при $X=-1$, 0 и 1, чтобы оценить β_0 и β_1 в постулированной модели, то каким будет смещение? Иными словами, что будут представлять собой оценки b_0 и b_1? «Истинная» модель должна быть выражена в виде

$$E(\mathbf{Y})=E\begin{bmatrix}Y_1\\Y_2\\Y_3\end{bmatrix}=\underset{\mathbf{1}\quad\mathbf{X}\quad\mathbf{X}^2}{\begin{bmatrix}1&-1&1\\1&0&0\\1&1&1\end{bmatrix}}\begin{bmatrix}\beta_0\\\beta_1\\\beta_{11}\end{bmatrix}=\underset{\mathbf{1}\quad\mathbf{X}}{\begin{bmatrix}1&-1\\1&0\\1&1\end{bmatrix}}\begin{bmatrix}\beta_0\\\beta_1\end{bmatrix}+$$

$$+\underset{\mathbf{X}^2}{\begin{bmatrix}1\\0\\1\end{bmatrix}}\beta_{11}=\mathbf{X}_1\boldsymbol{\beta}_1+\mathbf{X}_2\boldsymbol{\beta}_2.$$

Отсюда следует, что

$$(\mathbf{X}_1'\mathbf{X}_1)^{-1}=\begin{bmatrix}3&0\\0&2\end{bmatrix}^{-1}=\begin{bmatrix}\frac{1}{3}&0\\0&\frac{1}{2}\end{bmatrix},$$

$$\mathbf{X}_1'\mathbf{X}_2=\begin{bmatrix}1&1&1\\-1&0&1\end{bmatrix}\begin{bmatrix}1\\0\\1\end{bmatrix}=\begin{bmatrix}2\\0\end{bmatrix}.$$

Таким образом,

$$\mathbf{A}=\begin{bmatrix}\frac{1}{3}&0\\0&\frac{1}{2}\end{bmatrix}\begin{bmatrix}2\\0\end{bmatrix}=\begin{bmatrix}\frac{2}{3}\\0\end{bmatrix}.$$

Применяя уравнение (2.12.5), мы получим

$$E\begin{bmatrix}b_0\\b_1\end{bmatrix}=\begin{bmatrix}\beta_0\\\beta_1\end{bmatrix}+\begin{bmatrix}2/3\\0\end{bmatrix}\beta_{11}=\begin{bmatrix}\beta_0+2/3\beta_{11}\\\beta_1\end{bmatrix},$$

$$E(b_0)=\beta_0+\frac{2}{3}\beta_{11},\quad E(b_1)=\beta_1.$$

Выходит, что b_0 — смещенная на $2/3\beta_{11}$ оценка, а b_1 — несмещенная оценка.

Пример 2. Пусть постулируемая модель имеет вид

$$E(Y) = \beta_0 + \beta_1 X,$$

тогда как «фактически правильная» есть

$$E(Y) = \beta_0 + \beta_1 X + \beta_{11} X^2 + \beta_{111} X^3.$$

Какие смещения оценок будут в случае обработки наблюдений при $X = -3, -2, -1, 0, 1, 2, 3$?

Мы находим

$$\mathbf{X}_1 = \begin{bmatrix} 1 & -3 \\ 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} 9 & -27 \\ 4 & -8 \\ 1 & -1 \\ 0 & 0 \\ 1 & 1 \\ 4 & 8 \\ 9 & 27 \end{bmatrix},$$

$$(\mathbf{X}_1'\mathbf{X}_1)^{-1} = \begin{bmatrix} 7 & 0 \\ 0 & 28 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{7} & 0 \\ 0 & \frac{1}{28} \end{bmatrix},$$

$$\mathbf{X}_1'\mathbf{X}_2 = \begin{bmatrix} 28 & 0 \\ 0 & 196 \end{bmatrix},$$

$$\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = \begin{bmatrix} 4 & 0 \\ 0 & 7 \end{bmatrix}.$$

Следовательно,

$$E(b_0) = \beta_0 + 4\beta_{11},$$
$$E(b_1) = \beta_1 + 7\beta_{111}.$$

При использовании формулы (2.12.5) можно найти смещение любой регрессионной оценки, если только установлены постулируемая и «точная» модели, а также план. Это позволяет нам в конкретных случаях судить о том, какие эффекты отражаются на наших оценках, если имеются определенные отклонения от предполагаемой модели. Если постулируется полиномиальная модель, рациональная процедура часто состоит в том, чтобы исходить из предположения об ошибочности постулированной модели. Она ведь не включает слагаемые, показатели степени которых на единицу выше старших членов ряда, включенных в модель.

(Примечание. Только в этой части параграфа мы будем обозначать буквой $\mathbf{X}$ матрицу, ранее обозначавшуюся как $\mathbf{X}_1$, а вместо $\boldsymbol{\beta}$ будем использовать обозначение $\boldsymbol{\beta}_1$. Выражение $\mathbf{X}_2\boldsymbol{\beta}_2$ остается неизменным, оно будет применяться для обозначения дополнительных слагаемых в модели. Теперь выясним, какой эффект оказывает смещение на обычную процедуру метода наименьших квадратов.)

Примем следующие предположения:

1. Постулируемая модель $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$ содержит p параметров; $\mathbf{V}(\mathbf{Y}) = \mathbf{I}\sigma^2$.

2. «Истинная» модель есть $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\beta}_2$, где $\boldsymbol{\beta}_2$ может быть равным $\mathbf{0}$, и в этом случае постулируемая модель правильна.

3. Общее число наблюдений равно n и имеется f степеней свободы для оценки неадекватности и e степеней свободы для оценки «чистой» ошибки, так что $n = p + f + e$. (Это значит, что план содержит $p + f$ различных точек.)

4. Оценки $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ и вектор предсказаний $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$ получаются, как обычно.

5. $\mathbf{A} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_2$.

Тогда будут верны следующие результаты.

1. Матрица $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ есть всегда правильная матрица дисперсий-ковариаций $\mathbf{V}(\mathbf{b})$ оценок коэффициентов $\mathbf{b}$.

2. $E(\mathbf{b}) = \boldsymbol{\beta} + \mathbf{A}\boldsymbol{\beta}_2$.

3. $E(\hat{\mathbf{Y}}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{X}\mathbf{A}\boldsymbol{\beta}_2$.

4. Таблица дисперсионного анализа имеет вид:

Источник	Число степеней свободы	SS	Математическое ожидание среднего квадрата, MS
b_0	1	$\mathbf{Y}'\mathbf{1}\mathbf{1}'\mathbf{Y}/n$	$\sigma^2 + (\mathbf{X}\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\beta}_2)'\mathbf{1}\mathbf{1}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\beta}_2)/n$
Другие оценки | b_0	$p-1$	$\mathbf{b}'\mathbf{X}'\mathbf{Y} - \mathbf{Y}'\mathbf{1}\mathbf{1}'\mathbf{Y}/n$	$\sigma^2 + (\mathbf{X}\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\beta}_2)'\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}\mathbf{1}'/n\}(\mathbf{X}\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\beta}_2)/(p-1)$
Неадекватность	f	по разности	$\sigma^2 + \boldsymbol{\beta}_2'(\mathbf{X}_2 - \mathbf{X}\mathbf{A})'(\mathbf{X}_2 - \mathbf{X}\mathbf{A})\boldsymbol{\beta}_2/f$
«Чистая» ошибка	e	es_e^2	σ^2
Общий	n	$\mathbf{Y}'\mathbf{Y}$	

5. Если $\beta_2 = 0$, т. е. если постулируемая модель правильна, то приведенные выше результаты сводятся к

$$E(\mathbf{b}) = \boldsymbol{\beta}, \quad E(\hat{\mathbf{Y}}) = \mathbf{X}\boldsymbol{\beta}.$$

E (средний квадрат, обусловленный прочими оценками $| b_0) = \sigma^2 + \boldsymbol{\beta}'\mathbf{X}'(\mathbf{I} - \mathbf{1}\mathbf{1}'/n)\mathbf{X}\boldsymbol{\beta}/(p-1)$. Вот почему средний квадрат, обусловленный оценками, сравнивается с оценкой σ^2, чтобы проверить гипотезу $H_0 : \boldsymbol{\beta} = 0$, если оцениваемая модель не отклоняется сразу при проверке гипотезы о неадекватности модели. Если же критерий неадекватности свидетельствует об отсутствии согласия между экспериментальными и расчетными данными, причем $\beta_2 \neq 0$, то бесполезно производить проверку, используя средний квадрат, обусловленный регрессией, даже если в качестве оценки параметра σ^2 используется средний квадрат, обусловленный «чистой» ошибкой, s_e^2, а не остаточный средний квадрат.

В этом случае, если справедлива гипотеза $H_0 : \boldsymbol{\beta} = 0$, E (средний квадрат, обусловленный прочими оценками $| b_0) = \sigma^2 + \boldsymbol{\beta}_2'\mathbf{X}_2'\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - \mathbf{1}\mathbf{1}'/n\}\mathbf{X}_2\boldsymbol{\beta}_2/(p-1)$, и F-отношение, которое мы могли бы использовать, имеют нецентральное F-распределение, а не обычное центральное распределение, которое мы предполагаем, когда выполняем (ошибочную) проверку гипотезы в рамках обычной процедуры.

Определение математического ожидания средних квадратов

Для нахождения математического ожидания средних квадратов оказываются полезными определенные матричные результаты. Предположим, что $\mathbf{Q}$ — $(n \times n)$-матрица, такая, что $\mathbf{Y}'\mathbf{Q}\mathbf{Y}$ есть квадратичная форма элементов $\mathbf{Y}$. В таком случае, если E, как и ранее,— оператор математического ожидания, то

$$E(\mathbf{Y}'\mathbf{Q}\mathbf{Y}) = E(\mathbf{Y})'\mathbf{Q}E(\mathbf{Y}) + \text{trace}\ (\mathbf{Q}\boldsymbol{\Sigma}),$$

где «trace» (след) означает сумму диагональных элементов указанной квадратной матрицы. Примем обозначение: $\boldsymbol{\Sigma} = \mathbf{V}(\mathbf{Y})$ — $(n \times n)$-матрица дисперсий-ковариаций элементов вектора $\mathbf{Y}$. Кроме того, если $\mathbf{M}_1$ и $\mathbf{M}_2$ — любые квадратные матрицы одинакового размера, то

$$\text{trace}\ (\mathbf{M}_1 + \mathbf{M}_2) = \text{trace}\ \mathbf{M}_1 + \text{trace}\ \mathbf{M}_2.$$

Далее, если $\mathbf{T}$ — $(t \times s)$-матрица, а $\mathbf{S}$ — $(s \times t)$-матрица и при этом могут существовать оба произведения, $\mathbf{TS}$ и $\mathbf{ST}$, то

$$\text{trace}\ (\mathbf{TS}) = \text{trace}\ (\mathbf{ST}).$$

Последний результат очень полезен и подчас ведет к значительным

упрощениям. Так, например, если $\mathbf{X}$ — $(n \times p)$-матрица и мы обозначим $\mathbf{T} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$, $\mathbf{S} = \mathbf{X}'$, то получим

$$\text{trace}\,\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\} = \text{trace}\,\{\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\} = \text{trace}\,\{\mathbf{I}_p\} = p.$$

Мы воспользуемся этим соотношением в следующем примере.

Пример. Найти $E(\mathbf{b}'\mathbf{X}'\mathbf{Y}/p)$, если $E(\mathbf{Y}) = \mathbf{X}\beta + \mathbf{X}_2\beta_2$ и $\mathbf{V}(\mathbf{Y}) = \mathbf{I}\sigma^2$.

$$E(\mathbf{b}'\mathbf{X}'\mathbf{Y}) = E(\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}) = (\mathbf{X}\beta+\mathbf{X}_2\beta_2)'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta+\mathbf{X}_2\beta_2)+$$
$$+\text{trace}\,(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{I}\sigma^2) = \{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_2\beta_2\}' \times$$
$$\times(\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_2\beta_2\} + p\sigma^2 =$$
$$= (\beta + \mathbf{A}\beta_2)'\mathbf{X}'\mathbf{X}(\beta + \mathbf{A}\beta_2) + p\sigma^2.$$

Поделив обе части на p, получим результат, приведенный в таблице на с. 157 — 158. Заметим, что в показанных выкладках введение единичных матриц $\mathbf{I} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}$ и $\mathbf{I} = (\mathbf{X}'\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}$ ведет к желаемому результату. Другой пример содержится в приложении 2Б.

2.13. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ ПРИ НАЛИЧИИ ОГРАНИЧЕНИЙ

Для ознакомления с методом наименьших квадратов при наличии ограничений на параметры можно рекомендовать, например, работы: Waterman M. S. A restricted least squares problem.— Technometrics, 1974, **16**, p. 135—136; Judge G. G., Takayama T. Inequality restrictions in regression analysis. — Journal of the American Statistical Association, 1966, **61**, p. 166 — 181[20]. Если ограничения выражаются равенствами вида $\mathbf{C}\beta = \mathbf{d}$, то можно использовать метод неопределенных множителей Лагранжа и минимизировать функцию Лагранжа (см. приложение 2Г)

$$(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) + \lambda'(\mathbf{d} - \mathbf{C}\beta) \qquad (2.13.1)$$

по отношению к β и λ. Решение для β имеет вид:

$$\hat{\beta} = \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}'[\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}']^{-1}(\mathbf{d} - \mathbf{C}\mathbf{b}), \qquad (2.13.2)$$

где $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ есть обычный МНК-оцениватель при отсутствии ограничений.

2.14. НЕКОТОРЫЕ ЗАМЕЧАНИЯ ОТНОСИТЕЛЬНО ОШИБОК В ПРЕДИКТОРАХ (ОДНОВРЕМЕННО С ОШИБКАМИ В ОТКЛИКАХ)

Рассмотрим сначала обычную ситуацию, когда мы имеем отклик Y и один предиктор X, и при этом постулируется функция отклика $Y = \beta_0 + \beta_1 X + \varepsilon$ в виде уравнения прямой. В таких случаях пред-

[20] МНК-оценивание параметров линейных моделей при ограничениях на параметры рассматривается во многих руководствах по регрессионному анализу, в частности, например, в кн.: Рао С. Линейные статистические методы и их применения/Пер. с англ. Под ред. Ю. В. Линника.— М.: Наука, 1968.— 548 с.; Себер Дж. Линейный регрессионный анализ/Пер. с англ. Под ред. М. Б. Малютова.— М.: 1980.— 456 с.— *Примеч. пер.*

полагается, что отклик Y содержит ошибку, тогда как X нет. А что если переменная X также подвержена ошибке? Пусть η_i — истинное значение переменной Y_i, а ξ_i — истинная величина X_i, $i = 1, 2, \ldots, n$. Тогда наблюдаемые значения Y_i и X_i выражаются соотношениями

$$Y_i = \eta_i + \varepsilon_i, \tag{2.14.1}$$

$$X_i = \xi_i + \delta_i, \tag{2.14.2}$$

где ε_i и δ_i есть случайные ошибки, добавляемые к η_i и ξ_i соответственно, а ε_i независима от ξ_i и δ_i; δ_i независима от η_i и ε_i. Постулированная нами модель имеет вид

$$\eta_i = \beta_0 + \beta_1 \xi_i. \tag{2.14.3}$$

Как мы теперь должны поступить при оценивании такой модели? Подставляя выражения (2.14.1) и (2.14.2) в (2.14.3), получим

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i^*, \tag{2.14.4}$$

где

$$\varepsilon_i^* = (\varepsilon_i - \beta_i \delta_i). \tag{2.14.5}$$

На этой стадии было бы соблазнительно использовать для оценивания уравнения (2.14.4) обычные приемы регрессионного анализа, приводящие к уравнениям (1.2.9) и (1.2.10). Однако здесь есть некоторый подвох. Он состоит в следующем. Если мы предположим, что

$$\varepsilon_i \sim N(0,\ \sigma^2), \quad \delta_i \sim N(0,\ \sigma_\delta^2)$$

и независимы, и если мы введем обозначения

$$\sigma_\xi^2 = \Sigma\,(\xi_i - \bar{\xi})^2/n,$$

$$\sigma_{\xi\delta} = \operatorname{cov}(\xi,\ \delta),$$

то, используя уравнения (1.2.9), (2.14.4) и (2.14.5), получим

$$E(b_1) = E\left\{\sum_{i=1}^{n}(X_i - \bar{X})\,(\beta_0 + \beta_1\xi_i + \varepsilon_i)/\sum_{i=1}^{n}(X_i - \bar{X})^2\right\} =$$

$$= E\left\{\beta_1\sum_{i=1}^{n}(\xi_i - \bar{\xi} + \delta_i - \bar{\delta})\,\xi_i + \sum_{i=1}^{n}(X_i - \bar{X})\,\varepsilon_i\right\}/E\left\{\sum_{i=1}^{n}(\xi_i - \right.$$

$$\left. - \bar{\xi} + \delta_i - \bar{\delta})^2\right\} = \beta_1\,(\sigma_\xi^2 + \sigma_{\xi\delta})/(\sigma_\xi^2 + \sigma_\delta^2 + 2\sigma_{\xi\delta}) =$$

$$= \beta_1\,(1 + \rho r)/(1 + 2\rho r + r^2), \tag{2.14.6}$$

где $\rho = \sigma_{\xi\delta}/(\sigma_\xi\sigma_\delta)$ и $r = \sigma_\delta/\sigma_\xi$. Таким образом, b_1 есть в общем случае смещенный оцениватель β_1. Фактор смещения < 1, если $\sigma_\delta^2 + \sigma_{\xi\delta} > 0$, т. е. если $r + \rho > 0$, однако смещение отсутствует вообще при $\sigma_\delta^2 + \sigma_{\xi\delta} = 0$. Последний вариант мы будем обсуждать более подробно ниже. Если $\sigma_{\xi\delta} = 0$, то мы имеем

$$E(b_1) = \beta_1/(1 + \sigma_\delta^2/\sigma_\xi^2) = \beta_1/(1 + r^2). \tag{2.14.7}$$

Тогда фактор смещения всегда меньше 1, за исключением того случая,

когда отсутствуют ошибки в переменных X_t. Тогда оценка будет несмещенной. Основная проблема заключается в том, что ε_t^* и X_t не являются обычно независимыми. Фактически

$$\operatorname{cov}(X_t, \varepsilon_t^*) = \operatorname{cov}(\xi_t + \delta_t,\ \varepsilon_t - \beta_1\delta_t) = -\beta_1(\sigma_{\xi\delta} + \sigma_\delta^2). \quad (2.14.8)$$

Непосредственная подгонка уравнения (2.14.4), однако, приемлема, если:

1. Величина σ_δ^2 мала по сравнению с σ_ξ^2, что в свою очередь означает малость r в уравнениях (2.14.6).

2. Переменные X_t являются детерминированными (см.: Berkson (1950) в библиографии). В этом случае $\xi_t = X_t - \delta_t$ и $\sigma_{\xi\delta} = \operatorname{cov}(\delta, X - \delta) = -\sigma_\delta^2$, так что $\sigma_\delta^2 + \sigma_{\xi\delta} = 0$. В результате фактор смещения в уравнении (2.14.6) равен 1, правая часть (2.14.8) обращается в 0.

3. Постулируемая модель есть $\eta_t = \beta_0 + \beta_1 X_t$, а не (2.14.3).

Если ситуация не совпадает ни с одной из указанных выше, то надо применять альтернативный метод анализа. Один из приемов, предложенный Вальдом и усовершенствованный Бартлеттом (см. библиографию), состоит в следующем. Необходимо разделить исходные данные на три непересекающиеся группы одинакового объема (или почти одинакового объема, если n не кратно 3) таким образом, чтобы:

1) в первую группу попали данные с наименьшими значениями X. Обозначим точку $P_1 \equiv (\bar{X}_1, \bar{Y}_1)$, где $\bar{X}_1$ и $\bar{Y}_1$ — соответствующие средние величин X и Y, входящих в эту треть;

2) во вторую группу попали данные с наибольшими значениями X. Обозначим по аналогии с предыдущим $P_3 \equiv (\bar{X}_3, \bar{Y}_3)$;

3) в третью группу попали остальные данные. В дальнейшем анализе они не участвуют.

Теперь проведем линию через точки P_1 и P_3. Это «наилучшая» линия. Она имеет тангенс угла наклона, равный $(\bar{Y}_3 - \bar{Y}_1)/(\bar{X}_3 - \bar{X}_1)$. Обоснование этой процедуры дано в работе Вальда (см. библиографию). Дополнительная литература, посвященная этой проблеме, также приведена в библиографии [21].

Выводы из этого параграфа

Если переменные X подвержены ошибкам, как и переменные Y, то регрессионная задача заметно усложняется, даже когда модель однофакторная, как мы могли видеть из приведенного выше мате-

[21] Регрессионный анализ при наличии случайных ошибок в предикторных переменных детально рассматривается в ряде работ: Кендалл М., Стьюарт А. Статистические выводы и связи/Пер. с англ. Под ред. А. Н. Колмогорова.— М.: Наука, 1973.— 899 с.; Федоров В. В. Теория оптимального эксперимента.— М.: Наука, 1971.— 312 с.; Жилинская Е. И., Товмаченко Н. Н., Федоров В. В. Методы регрессионного анализа при наличии ошибок в предикторных переменных.— М., 1979.— 34 с. (Препринт/АН СССР, Науч. совет «Кибернетика»). Этой проблеме посвящено много публикаций в журнале «Заводская лаборатория».— *Примеч. пер.*

риала. По этой причине всегда полезно по возможности так организовать эксперимент, чтобы отношение $\sigma_\delta^2/\sigma_\xi^2$ было малой величиной. Практически это означает, что разброс величин ξ, мерой которого служит σ_ξ^2, должен существенно превышать разброс ошибок, содержащихся, вероятно, в переменных X. То же самое верно и для других предикторов. После того как это обеспечено, ошибками в переменных X можно пренебречь и применять обычный метод наименьших квадратов. Если этого сделать нельзя, то при использовании обычного анализа могут возникнуть осложнения.

Дополнение

Для ознакомления с детальной трактовкой регрессионного анализа функциональных и структурных соотношений, включая рассмотрение идентифицируемости параметров при использовании принципа максимального правдоподобия в условиях нормального распределения ошибок, см. работу М. Кендалла и А. Стьюарта «Статистические выводы и связи» (М.: Наука, 1973, гл. 29). Экономические аспекты проблемы освещены, например, в монографии: Goldberger A. S. Econometric Theory.— New York: Wiley, 1964. Решение такого типа задач с помощью двухступенчатого метода наименьших квадратов рассмотрено в кн: Kmenta Jan. Elements of Econometrics.— New York: MacMillan, 1971, p. 559—565. Некоторые журнальные публикации, относящиеся к данной проблеме, указаны в библиографии.

2.15. ОБРАТНАЯ РЕГРЕССИЯ (В СЛУЧАЕ МНОГОМЕРНОГО ПРЕДИКТОРА)

Пусть имеется регрессионное уравнение

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k,$$

построенное по экспериментальным данным, и известно некоторое истинное среднее значение величины Y, обозначаемое буквой Y_0. Требуется определить [22] «фидуциальную область» для соответствующей точки $(X_1, X_2, \ldots, X_k)$. Используя уравнение (1.7.8), получим следующее уравнение, описывающее границы искомой области:

$$\{-Y_0 + b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k\}^2 = t^2s^2 \left\{ (1, X_1, X_2, \ldots, X_k)(\mathbf{X}'\mathbf{X})^{-1} \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_k \end{bmatrix} \right\}. \quad (2.15.1)$$

Это поверхность типа гиперболоида. На рис. 2.5 представлена такая поверхность для случая, когда $k = 2$. Если Y_0 есть среднее из q на-

[22] См. примечание на с. 70. — *Примеч. пер.*

блюдений, то необходимо добавить к содержимому фигурных скобок в правой части уравнения (2.15.1) слагаемое $1/q$. Если $\hat{Y}$ есть полином, то не составляет труда внести необходимые очевидные изменения в обе части уравнения (2.15.1).

(Примечание. Описанный выше метод может быть использован также для решения другого типа задач. Так, например, значения переменной X, обращающие в максимум или минимум функцию $\hat{Y} = b_0 + b_1X + b_2X^2 + b_3X^3 + b_4X^4$, представляют собой корни урав-

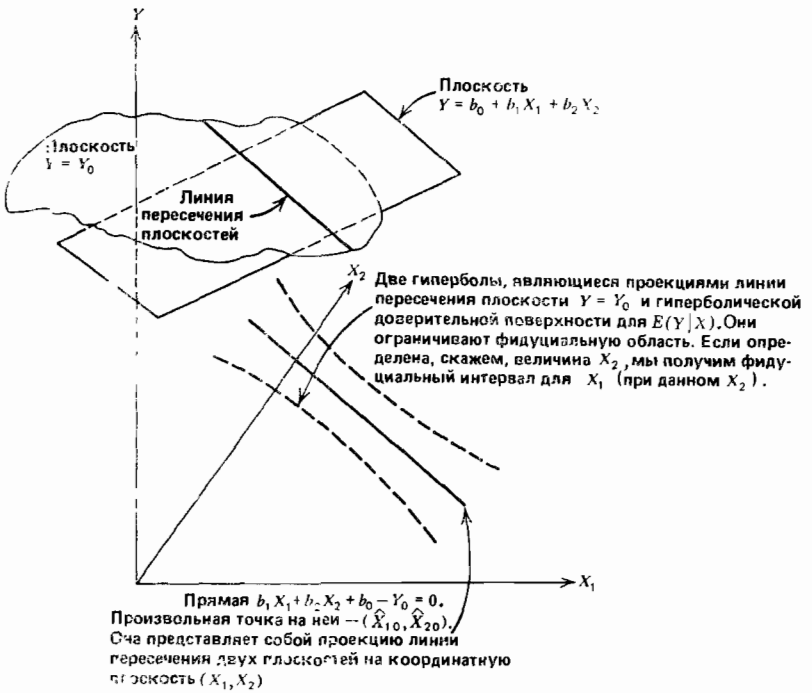


Рис. 2.5. Обратная регрессия для двух предикторов

нения $f \equiv b_1 + 2b_2X + 3b_3X^2 + 4b_4X^3 = 0$. Фидуциальные границы для корней могут быть найдены из уравнения

$$f^2 = t^2s^2\{V(f)/\sigma^2\}, \qquad (2.15.2)$$

где $V(f)$ есть дисперсия функции f, которая имеет сомножитель σ^2. Более полные сведения на этот счет, включая возможные осложнения, связанные с мнимыми корнями, приведены в кн.: Williams E. J. Regression Analysis.— New York: Wiley, 1959, p. 108—109, 114—116. По этому поводу см. также: Box G. E. P., Hunter J. S. A confidence region for the solution of a set of simultaneous equations with an application to experimental design.— Biometrika, 1954, **41**, p. 190—199.)

(Более полный перечень подобных результатов указан, например, в монографиях: Graybill F. A. An Introduction to Linear Statistical Models.— New York: McGraw-Hill, 1961; Рао С. Р. Линейные статистические методы и их применения.— М.: Наука, 1968.)

1. $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$.
2. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$.
3. Квадратную матрицу $\mathbf{C}$ называют ортогональной, если $\mathbf{C}'\mathbf{C} = \mathbf{I}$.
4. Квадратная матрица $\mathbf{M}$ называется идемпотентной, если $\mathbf{MM} = \mathbf{M}$.
5. Если $\mathbf{M}$ симметричная и идемпотентная, то

$$(\mathbf{I} - 2\mathbf{M})'(\mathbf{I} - 2\mathbf{M}) = \mathbf{I}.$$

Следовательно, любая матрица вида $\mathbf{I}-2\mathbf{M}$, где $\mathbf{M}$ симметричная и идемпотентная, будет ортогональной.

6. Trace $(\mathbf{AB})$ = Trace $(\mathbf{BA})$, где «trace» (след) обозначает сумму диагональных элементов квадратной матрицы (например, $\mathbf{A} = (p \times q)$, $\mathbf{B} = (q \times p)$).

7. Если

$$\mathbf{M} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \quad \text{и} \quad \begin{cases} \mathbf{P} = \mathbf{A} - \mathbf{BD}^{-1}\mathbf{C}, \\ \mathbf{Q} = \mathbf{D} - \mathbf{CA}^{-1}\mathbf{B}, \end{cases}$$

то

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{P}^{-1} & -\mathbf{A}^{-1}\mathbf{BQ}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CP}^{-1} & \mathbf{Q}^{-1} \end{bmatrix},$$

при этом предполагается, что все матрицы, которые здесь обращаются, неособенные. С другой стороны,

$$\mathbf{M}^{-1} = \begin{bmatrix} \mathbf{A}^{-1} + \mathbf{A}^{-1}\mathbf{BQ}^{-1}\mathbf{CA}^{-1} & -\mathbf{A}^{-1}\mathbf{BQ}^{-1} \\ -\mathbf{Q}^{-1}\mathbf{CA}^{-1} & \mathbf{Q}^{-1} \end{bmatrix}.$$

Если $\mathbf{M}$ симметричная, то $\mathbf{C} = \mathbf{B}'$.

8. Если $\mathbf{E}$ и $\mathbf{F}$ есть матрицы с размерами $n \times p$ и $p \times n$, то

$$(\mathbf{I}_n + \mathbf{EF})^{-1} = \mathbf{I}_n - \mathbf{E}(\mathbf{I}_p + \mathbf{FE})^{-1}\mathbf{F}.$$

Этот результат особенно полезен, если p намного меньше, чем n.

Частный случай 1. Если $\mathbf{X}$ есть $(n \times p)$-матрица, то

$$[\mathbf{I}_n + \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']^{-1} = \mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}[\mathbf{I}_p + \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}]^{-1}\mathbf{X}' =$$

$$= \mathbf{I}_n - \frac{1}{2}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'.$$

Частный случай 2. Если $\mathbf{A}$ есть $(n \times n)$-матрица, а $\mathbf{u}$ и $\mathbf{v}$ — $(n \times 1)$-векторы, то

$$(\mathbf{A} + \mathbf{uv}')^{-1} = (\mathbf{I} + \mathbf{A}^{-1}\mathbf{uv}')^{-1}\mathbf{A}^{-1} = \mathbf{A}^{-1} - (\mathbf{A}^{-1}\mathbf{u})(\mathbf{v}'\mathbf{A}^{-1})/\{1 + \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}\},$$

отсюда появляется возможность обращать матрицу $\mathbf{A} + \mathbf{uv}'$, зная $\mathbf{A}^{-1}$. (Для этого надо положить $\mathbf{E} = \mathbf{A}^{-1}\,\mathbf{u}$, $\mathbf{F} = \mathbf{v}'$.)

9. Если $\mathbf{A}$ есть $(p \times p)$-, $\mathbf{B}$ — $(p \times q)$-, $\mathbf{C}$ — $(q \times p)$- и $\mathbf{D}$ — $(q \times q)$-матрица, то

$$\begin{vmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{vmatrix} = |\mathbf{A}||\mathbf{D}-\mathbf{C}\mathbf{A}^{-1}\mathbf{B}| = |\mathbf{A}-\mathbf{B}\mathbf{D}^{-1}\mathbf{C}||\mathbf{D}|.$$

Доказательство. Умножим исходную матрицу слева на матрицу

$$\begin{bmatrix} \mathbf{I}_p & \mathbf{0} \\ -\mathbf{C}\mathbf{A}^{-1} & \mathbf{I}_q \end{bmatrix}$$

и запишем затем детерминант полученной матрицы. В итоге получим указанное выше выражение.

Частный случай 1. Положим, что $\mathbf{C} = -\mathbf{B}'$, $\mathbf{D} = \mathbf{I}$, получим следующий результат:

$$|\mathbf{A}||\mathbf{I}+\mathbf{B}'\mathbf{A}^{-1}\mathbf{B}| = |\mathbf{A}+\mathbf{B}\mathbf{B}'|.$$

Частный случай 2. Положим, что блочная матрица симметрична, т. е. $\mathbf{C} = \mathbf{B}'$. Некоторые полезные результаты по специальным случаям обращения матриц см. в статье: Roy S. N., Sarhan A. E. On inverting a class of patterned matrices.— Biometrika, 1956, **43**, p. 227—231.

Приложение 2Б. МАТЕМАТИЧЕСКОЕ ОЖИДАНИЕ ДОПОЛНИТЕЛЬНОЙ СУММЫ КВАДРАТОВ

Запишем уравнения для модели 1 и для модели 2:

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \quad \text{для модели 1,}$$

$$\mathbf{Y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon} \quad \text{для модели 2,}$$

где $\mathbf{X}_1$ — $(n \times q)$, $\mathbf{X}_2$ — $[n \times (p-q)]$. Заметим, что столбцы $[n \times (p-q)]$-матрицы

$$\mathbf{X}_{2\cdot 1} = \mathbf{X}_2 - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = \{\mathbf{I}-\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\}\mathbf{X}_2,$$

которая является матрицей остатков при построении регрессии $\mathbf{X}_2$ от $\mathbf{X}_1$, ортогональны к столбцам $\mathbf{X}_1$. (*Доказательство.* $\mathbf{X}_1'\mathbf{X}_{2\cdot 1} = \{\mathbf{X}_1'-\mathbf{X}_1'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\}\mathbf{X}_2 = \{\mathbf{0}\}\mathbf{X}_2 = \mathbf{0}$). Если мы обозначим произведение матриц

$$\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2,$$

что представляет собой матрицу смещения, то мы можем выразить матрицу $\mathbf{X}_{2\,1}$ в виде

$$\mathbf{X}_{2\cdot 1} = \mathbf{X}_2 - \mathbf{X}_1\mathbf{A}.$$

Используя этот результат, можно записать модель 1 так:

$$\mathbf{Y} = \mathbf{X}_1(\boldsymbol{\beta}_1 + \mathbf{A}\boldsymbol{\beta}_2) + \mathbf{X}_{2\cdot 1}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon},$$

где мы просто добавили и вычли вектор $\mathbf{X}_1\mathbf{A}\boldsymbol{\beta}_2$ и произвели перегруп-

пировку членов. Полагая $\alpha_1 = \beta_1 + \mathbf{A}\beta_2$, можно переписать модель 1:

$$\mathbf{Y} = \mathbf{X}_1\alpha_1 + \mathbf{X}_{2\cdot1}\beta_2 + \varepsilon,$$

где обе части модели «взаимно ортогональны», поскольку $\mathbf{X}_1'\mathbf{X}_{2\cdot1} = 0$. Пусть $\mathbf{a}_1$, $\mathbf{b}_2$ есть МНК-оценки параметров α_1, β_2 модели 1. Тогда сумма квадратов, обусловленная регрессией, для модели 1, соответствующая величине $\mathbf{b}'\mathbf{X}'\mathbf{Y}$, равна:

$$S_1 = (\mathbf{a}_1'\mathbf{b}_2')\,[\mathbf{X}_1,\ \mathbf{X}_{2\cdot1}]'\,\mathbf{Y} = \mathbf{Y}'\,[\mathbf{X}_1,\ \mathbf{X}_{2\cdot1}] \times$$
$$\times \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_{2\cdot1} \\ \mathbf{X}_{2\cdot1}'\mathbf{X}_1 & \mathbf{X}_{2\cdot1}'\mathbf{X}_{2\cdot1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_{2\cdot1}' \end{bmatrix} \mathbf{Y}.$$

В силу ортогональности столбцов матриц $\mathbf{X}_1$ и $\mathbf{X}_{2\cdot1}$ внедиагональные члены в матрице, подлежащей обращению, равны нулевым матрицам. Поэтому достаточно обратить лишь диагональные матрицы порознь и можно получить

$$S_1 = \mathbf{Y}'\mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y} + \mathbf{Y}'\mathbf{X}_{2\cdot1}(\mathbf{X}_{2\cdot1}'\mathbf{X}_{2\cdot1})^{-1} = \mathbf{X}_{2\cdot1}'\mathbf{Y} =$$
$$= \mathbf{S}_2 + \mathbf{Y}'\mathbf{Q}\mathbf{Y},$$

где S_2, очевидно, есть сумма квадратов, обусловленная регрессией, для модели 2, а матрица $\mathbf{Q}$ непосредственно вытекает из приведенных выкладок. Мы можем, таким образом, записать «дополнительную сумму квадратов для b_2 при наличии b_1»:

$$S_1 - S_2 = \mathbf{Y}'\mathbf{Q}\mathbf{Y}.$$

Для того чтобы получить математическое ожидание этой суммы, применим общую формулу

$$E(\mathbf{Y}'\mathbf{Q}\mathbf{Y}) = \{E(\mathbf{Y})\}'\,\mathbf{Q}\,\{E(\mathbf{Y})\} + \operatorname{trace}\{\mathbf{Q}\Sigma\},$$

где $\Sigma = \mathbf{V}(\mathbf{Y})$. В нашем случае

$$\Sigma = \mathbf{I}\sigma^2 \text{ и } E(\mathbf{Y}) = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2,$$

так что

$$E(S_1 - S_2) = (\beta_1'\mathbf{X}_1' + \beta_2'\mathbf{X}_2')\{\mathbf{X}_{2\cdot1}(\mathbf{X}_{2\cdot1}'\mathbf{X}_{2\cdot1})^{-1}\mathbf{X}_{2\cdot1}'\} \times$$
$$\times (\mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2) + \operatorname{trace}\{\mathbf{X}_{2\cdot1}(\mathbf{X}_{2\cdot1}'\mathbf{X}_{2\cdot1})^{-1}\mathbf{X}_{2\cdot1}'\mathbf{I}\sigma^2\}.$$

Вспомним, что $\mathbf{X}_1'\mathbf{X}_{2\cdot1} = 0$. Запишем далее: $\mathbf{U} = \mathbf{X}_2'\mathbf{X}_{2\cdot1}(\mathbf{X}_{2\cdot1}'\mathbf{X}_{2\cdot1})^{-1}\mathbf{X}_{2\cdot1}'\mathbf{X}_2$. След произведения матриц можно преобразовать, используя известную формулу $\operatorname{trace}(\mathbf{ST}) = \operatorname{trace}(\mathbf{TS})$. Обозначив $\mathbf{S} = \mathbf{X}_{2\cdot1}$ и $\mathbf{T} = (\mathbf{X}_{2\cdot1}'\mathbf{X}_{2\cdot1})^{-1}\mathbf{X}_{2\cdot1}'$, получим

$$\operatorname{trace}\{\mathbf{X}_{2\cdot1}(\mathbf{X}_{2\cdot1}'\mathbf{X}_{2\cdot1})^{-1}\mathbf{X}_{2\cdot1}'\mathbf{I}\sigma^2\} =$$
$$= \operatorname{trace}(\mathbf{TS})\sigma^2 = \sigma^2 \operatorname{trace}(\mathbf{I}_{p-q}) = (p-q)\sigma^2.$$

В итоге имеем

$$E(S_1 - S_2) = \beta_2'\mathbf{U}\beta_2 + (p-q)\sigma^2.$$

Отсюда следует, что при *соблюдении нуль-гипотезы* $H_0: \beta_2 = 0$

$$E\{(S_1 - S_2)/(p-q)\} = \sigma^2.$$

Р е з ю м е. Приложение работы Бокса и Ветца (см. параграф 2.6) к регрессионной ситуации вкратце состоит в следующем. Для «полезной» в отличие от «значимой» регрессии наблюдаемая величина F-отношения для регрессии должна в несколько раз превосходить обычную процентную точку. Однако точно указать во сколько раз нельзя. Эта величина произвольна, поскольку произволен выбор уровня значимости. Однако по этому поводу все же можно дать некоторые указания.

Критерий γ_m

В регрессионных задачах, когда неадекватность не обнаруживается, проверка значимости регрессионных параметров обычно проводится с помощью F-отношения, в котором в числителе стоит сумма квадратов, обусловленная регрессией при наличии b_0, а в знаменателе — остаточный средний квадрат s^2. Эта величина сравнивается с соответствующей верхней α %-ной точкой $F(\nu_m, \nu_r, 1-\alpha)$, где ν_m и ν_r — соответственно числа степеней свободы для числителя и знаменателя F-статистики. Если F-отношение статистически значимо, т. е. если $F > F(\nu_m, \nu_r, 1-\alpha)$, предполагается, что большая часть вариаций в данных относительно среднего отклика обусловлена регрессионным уравнением. Отсюда не следует, однако, что полученное уравнение приемлемо для предсказания в том смысле, что размах предсказываемых откликов заметно превосходит стандартную ошибку отклика. При этом возникает вопрос, как можно различить статистически значимые и ценные для предсказания уравнения среди статистически значимых уравнений, имеющих ограниченную ценность?

Некоторые работы, где даются ответы на этот вопрос, существенно опираются на появившуюся в 1964 г. в Висконсинском университете диссертацию Ветца «Критерий для суждения об адекватности при оценивании с помощью функции отклика». (Существует также одноименный отчет: Box G. E. P., Wetz J. U. W. Statistics Department Technical Report, No 9, 1973.) Суть этого подхода сводится к следующему.

Предположим, что с помощью метода наименьших квадратов мы подбираем модель

$$\mathbf{Y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\psi} + \boldsymbol{\varepsilon}, \qquad (2\text{В}.1)$$

где $\mathbf{X}\boldsymbol{\beta}$ — часть модели, подлежащая проверке с помощью «теста для регрессии», а $\mathbf{Z}\boldsymbol{\psi}$ описывает такие эффекты, как среднее, блоковые переменные, временные дрейфы и т. д., которые мы хотим исключить из вариации данных, но которые в остальном не представляют интереса. Предположим также, что $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathbf{V}(\boldsymbol{\varepsilon}) = \mathbf{I}\sigma^2$. Изменения величин откликов η_i в n экспериментальных точках можно охаракте-

ризовать с помощью величины

$$\sum_{i=1}^{n} (\eta_i - \tilde{\eta}_i)^2/n, \tag{2B.2}$$

где η_i — истинный отклик в i-м наблюдении, а $\tilde{\eta}_i$ — элемент вектора $\tilde{\boldsymbol{\eta}} = \mathbf{Z}\boldsymbol{\psi}$. Если исключается только коэффициент β_0, то $\tilde{\eta}_i = \bar{\eta}$, среднему из η_i.

Мы можем сравнить величину (2B.2) с ошибками, которые можно совершить при оценивании разностей $\eta_i - \tilde{\eta}_i$. МНК-оцениватель для величины $\eta_i - \tilde{\eta}_i$ есть $\hat{Y}_i - \tilde{Y}_i$, т. е. i-й элемент вектора (см. выражение (2B.1))

$$\hat{\mathbf{Y}} - \tilde{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{R}\mathbf{Y}; \tag{2B.3}$$

матрица дисперсий-ковариаций для этого вектора имеет вид

$$E\{\mathbf{R}\mathbf{Y} - E(\mathbf{R}\mathbf{Y})\}\{\mathbf{R}\mathbf{Y} - E(\mathbf{R}\mathbf{Y})\}' = \mathbf{R}\sigma^2\mathbf{R}' = \mathbf{R}\sigma^2, \tag{2B.4}$$

где $\mathbf{V}(\mathbf{Y}) = \mathbf{I}\sigma^2$, ввиду того что $\mathbf{R}$ — симметричная и идемпотентная матрица. Следовательно, $V(\hat{Y}_i - \tilde{Y}_i)$ есть i-й диагональный элемент матрицы, а среднее значение по этим дисперсиям, которое может служить полной мерой того, как мы оцениваем величины $\eta_i - \tilde{\eta}_i$, выражается соотношениями

$$\sigma^2_{\hat{Y}-\tilde{Y}} = \text{trace}(\mathbf{R}\sigma^2)/n = \nu_m\sigma^2/n. \tag{2B.5}$$

Это соотношение справедливо, поскольку

$$\text{trace } \mathbf{R} = \text{trace } \mathbf{X}\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\} = \text{trace}\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\mathbf{X} = \text{trace } \mathbf{I}_{\nu_m},$$

где ν_m — число параметров, или число элементов вектора $\boldsymbol{\beta}$ (т. е. число степеней свободы для суммы, обусловленной регрессией). Отсюда вытекает, что разумное сравнение размеров вариаций $\eta_i - \tilde{\eta}_i$ и ошибок их оценок можно выполнить с помощью корня квадратного из отношения (2B.2) и (2B.5), а именно:

$$\gamma_m = \left\{\sum_{i=1}^{n} (\eta_i - \tilde{\eta}_i)^2/(\nu_m\sigma)^2\right\}^{\frac{1}{2}}. \tag{2B.6}$$

На рис. 2B.1 показана ситуация для одной предикторной переменной. Критерий γ_m позволяет сравнить отклонения жирной линии $\eta_i - \tilde{\eta}_i$ от среднего разброса их оценок, распределения которых показаны при разных значениях X_i. Насколько большой должна быть величина γ_m для того, чтобы построенная регрессия была практически полезной в отличие от регрессии, только статистически значимой? Эта величина в значительной степени произвольна, так как произволен выбираемый статистический уровень значимости. (Однако для того, чтобы появились какие-либо идеи, возьмем $\gamma_m = 2, 3, 4$, так чтобы мы смогли исследовать ряд значений и выбрать подходящее. Допустим, что γ_0 есть минимально приемлемый уровень для γ_m. В таком случае Бокс и Ветц показали, что надо найти определенное значение

F_0, зависящее от γ_0, и если обычное регрессионное отношение F превосходит эту величину F_0, то мы будем считать, что γ_m достаточно велико для того, чтобы считать регрессию полезной в практическом отношении. Бокс и Ветц показали далее, что критическое значение F_0 приблизительно равно

$$F_0 \cong (1+\gamma_0^2)\, F(\nu_0,\ \nu_r,\ 1-\alpha), \tag{2B.7}$$

где ν_r — число степеней свободы для остаточной дисперсии и где

$$\nu_0 = \nu_m (1+\gamma_0^2)^2 / (1+2\gamma_0^2). \tag{2B.8}$$

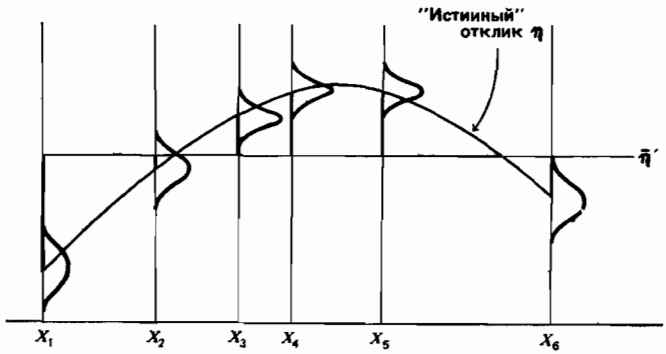


Рис. 2B.1. Отклонения «истинных» величин η относительно их среднего по сравнению с размахами оценок $\hat{Y}_i - \tilde{Y}_i$ для одной переменной X

Иными словами, для того, чтобы регрессия была стóящей в практическом отношении, надо, чтобы выполнялось условие $F > F_0$. Конечно, нетрудно определить величины F_0 для конкретных случаев, но более целесообразно анализировать отношение

$$F_0/F\,(\nu_m,\ \nu_r,\ 1-\alpha) \tag{2B.9}$$

для данных значений γ_0 и различных значений ν_m, ν_n и α. В табл. 2B.1, 2B.2, 2B.3 показаны значения этого отношения, округленные до целых чисел для $\gamma_0 = 2$, 3 и 4 соответственно при $\alpha = 0{,}05$. Из этих таблиц видно следующее. Если при данном уровне вероятности принять $\gamma_0 = 2$ за критическое значение, при котором регрессия может считаться достаточно информативной для наших целей, то необходимо, чтобы наблюдаемое F-отношение было по крайней мере в 4 раза больше, чем обычная процентная точка F-распределения.

Если же мы согласимся принять за критическое значение $\gamma_0 = 3$, то величина F должна быть по крайней мере в 6—10 раз больше, чем обычная процентная точка. Переходя к табл. 2B.3, мы видим, что с увеличением выбранного значения γ_0 отношения также увеличиваются, но при этом наблюдаются большие расхождения между ними. (Для $\alpha = 0{,}01$ картина почти та же, значения отношений или такие же, или на 1 или 2 единицы меньше приведенных.)

Таблица 2В.1. **Отношения F_0/F (ν_m, ν_r, 0,95) для $\gamma_0 = 2$**

Число степеней свободы остаточной суммы, ν_r	Число степеней свободы для суммы квадратов, обусловленной регрессией, ν_m								
	1	2	3	4	5	6	10	15	21
3	5	5	5	5	5	5	5	5	5
4	4	4	5	5	5	5	5	5	5
5	4	4	4	5	5	5	5	5	5
10	4	4	4	4	4	4	5	5	5
15	4	4	4	4	4	4	4	5	5
20	4	4	4	4	4	4	4	5	5
30	4	4	4	4	4	4	4	4	5
40 до ∞	3	4	4	4	4	4	4	4	4

Таблица 2В.2. **Отношения F_0/F (ν_m, υ_r, 0,95) для $\gamma_0 = 3$**

Число степеней свободы остаточной суммы, ν_r	Число степеней свободы для суммы квадратов, обусловленной регрессией, ν_m								
	1	2	3	4	5	6	10	15	21
3	9	9	9	9	10	10	10	10	10
4	8	9	9	9	9	9	10	10	10
5	8	8	9	9	9	9	9	10	10
10	7	7	8	8	8	8	9	9	9
15	6	7	7	8	8	8	9	9	9
20	6	6	7	7	8	8	8	9	9
30	6	6	7	7	7	8	8	9	9
40	6	6	7	7	7	7	8	8	9
60	6	6	7	7	7	7	8	8	8
120	6	6	7	7	7	7	8	8	8
∞	6	6	6	7	7	7	7	8	8

Таблица 2В.3. **Отношения F_0/F (ν_m, ν_r, 0,95) для $\gamma_0 = 4$**

Число степеней свободы остаточной суммы, ν_r	Число степеней свободы для суммы квадратов, обусловленной регрессией, ν_m								
	1	2	3	4	5	6	10	15	21
3	15	15	16	16	16	16	17	17	17
4	13	14	15	15	16	16	16	16	17
5	12	13	14	15	15	15	16	16	16
10	10	12	13	13	14	14	15	15	16
15	10	11	12	12	13	13	14	15	15
20	9	11	11	12	12	13	14	15	15
30	9	10	11	11	12	12	14	14	15
40	9	10	11	11	12	12	13	14	14
60	9	10	10	11	11	12	13	14	14
120	9	9	10	11	11	11	13	13	14
∞	8	9	10	10	11	11	12	12	12

В общем, ясно, что наблюдаемое значение F-отношения должно быть по меньшей мере в 4—5 раз больше обычной процентной точки. На практике, вероятно, целесообразно ориентироваться на цифры, приведенные в табл. 2В.2. Чаще всего их желательно достигать или превосходить. Во всяком случае надо гарантировать, чтобы $\gamma_m \geqslant 3$. Однако выбор подходящего доверительного уровня в значительной мере зависит от вкусов исследователей, поэтому приведенные таблицы надо расценивать как указания для такого выбора. Эти результаты были получены, исходя из F-статистики для полной регрессии. Однако аналогичные результаты справедливы для подмножества коэффициентов подгоняемой модели. Такое же правило пригодно для F-статистики, составленной для любого подмножества коэффициентов. (См.: Ellerton R. R. W. Is the regression equation adequate — a generalization.— Technometrics, 1978, **20**, p. 313—315.)

Приложение 2Г. НЕОПРЕДЕЛЕННЫЕ МНОЖИТЕЛИ ЛАГРАНЖА

Обозначения. Поскольку метод неопределенных множителей Лагранжа исключительно широко применяется, мы приняли вполне «нейтральные» обозначения переменных $\theta_1, \theta_2, \ldots, \theta_m$, входящих в функции f и g_j ниже. Когда этот метод применялся в параграфе 2.13, под величинами θ понимались все параметры в векторе $\boldsymbol{\beta}$. При рассмотрении гребневой регрессии в параграфе 6.7 под величинами θ подразумеваются все регрессионные коэффициенты β, кроме β_0. В других приложениях величины θ могут играть роль предикторных переменных, т. е. величин $\mathbf{X}$.

Основной метод

Предположим, что мы хотим найти стационарную или экстремальную точку функции $f(\theta_1, \theta_2, \ldots, \theta_m)$ от m переменных $\theta_1, \theta_2, \ldots, \theta_m$, на которые наложены ограничения типа

$$g_j(\theta_1, \theta_2, \ldots, \theta_m) = 0 \quad (j = 1, 2, \ldots, q).$$

Сформируем функцию

$$F = f - \sum_{j=1}^{q} \lambda_j g_j, \tag{2Г.1}$$

где $\lambda_1, \lambda_2, \ldots, \lambda_q$ неизвестны. Найдем частные производные от (2Г.1) по отношению к каждой величине θ_i и приравняем полученные выражения к нулю. В результате получим m уравнений

$$\frac{\partial F}{\partial \theta_i} = \frac{\partial f}{\partial \theta_i} - \sum_{j=1}^{q} \lambda_j \frac{\partial g_j}{\partial \theta_i} = 0 \qquad (i = 1, 2, \ldots, m). \tag{2Г.2}$$

Эти m уравнений совместно с дополнительными q уравнениями

$$g_j = 0 \qquad (j = 1, 2, \ldots, q) \tag{2Г.3}$$

образуют совместную систему из $(q + m)$ уравнений, которые могут

быть решены относительно неизвестных $\theta_1, \theta_2, \ldots, \theta_m, \lambda_1, \lambda_2, \ldots, \lambda_q$. Нередко величины λ_j сразу исключаются и фактически не определяются. По этой причине их называют неопределенными множителями. В некоторых случаях, однако, решение для переменных $\theta_1, \theta_2, \ldots, \theta_m$ получить легче, если сначала определить λ_j. В других случаях может оказаться проще определить величины λ_j из уравнений (2Г.2) и рассматривать затем другие величины в (2Г.3) как неизвестные.

Является ли решение точкой максимума или минимума?

Предположим теперь, что $(\theta_1, \theta_2, \ldots, \theta_m)=(a_1, a_2, \ldots, a_m)$— решение уравнений (2Г.2) и (2Г.3), после того как из них были исключены величины λ_j. Пусть

$$\mathbf{M}(\theta) = \mathbf{M}(\theta_1, \theta_2, \ldots, \theta_m) =$$

$$= \begin{bmatrix} \dfrac{\partial^2 F}{\partial \theta_1^2} & \dfrac{\partial^2 F}{\partial \theta_1 \partial \theta_2} & \cdots & \dfrac{\partial^2 F}{\partial \theta_1 \partial \theta_m} \\ \dfrac{\partial^2 F}{\partial \theta_2 \partial \theta_1} & \dfrac{\partial^2 F}{\partial \theta_2^2} & \cdots & \dfrac{\partial^2 F}{\partial \theta_2 \partial \theta_m} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \dfrac{\partial^2 F}{\partial \theta_m \partial \theta_1} & \dfrac{\partial^2 F}{\partial \theta_m \partial \theta_2} & \cdots & \dfrac{\partial^2 F}{\partial \theta_m^2} \end{bmatrix} \qquad (2Г.4)$$

— матрица частных производных второго порядка. И пусть далее $\mathbf{M}(a_1, a_2, \ldots, a_m) = \mathbf{M}(\mathbf{a})$ — матрица, получаемая из (2Г.4) после подстановки в нее решения $\mathbf{a}' = (a_1, a_2, \ldots, a_m)$. В таком случае если $\mathbf{M}(\mathbf{a})$ есть:

1) положительно определенная, т. е. $\mathbf{u}'\mathbf{M}\mathbf{u} > 0$ для всех $\mathbf{u}$,

2) отрицательно определенная, т. е. $\mathbf{u}'\mathbf{M}\mathbf{u} < 0$ для всех $\mathbf{u}$, где $\mathbf{u}' = (u_1, u_2, \ldots, u_m)$ — произвольный $(1 \times m)$-вектор с действительными элементами, то функция $f(\theta_1, \theta_2, \ldots, \theta_m)$ достигает:

1) локально наименьшего значения при $\boldsymbol{\theta} = \mathbf{a}$;

2) локально наибольшего значения при $\boldsymbol{\theta} = \mathbf{a}$ соответственно.

Если мы разложим функцию F в ряд Тейлора в окрестности $\mathbf{a}$, используя частные производные и учитывая, что частные производные первого порядка от функции F в точке $\boldsymbol{\theta} = \mathbf{a}$ равны нулю, то мы увидим, что справедливо соотношение

$$F(\mathbf{a}+\mathbf{h}) - F(\mathbf{a}) = \frac{1}{2}\mathbf{h}'\mathbf{M}(\mathbf{a})\mathbf{h} + o(h^3),$$

где $\mathbf{h}$ представляет собой вектор малых приращений h_i одинакового порядка, а $o(h^3)$ — остаточный член, содержащий слагаемые, зави-

сящие от приращений в третьей и более высоких степенях. Поэтому если, например, $\mathbf{M}(\mathbf{a})$ — положительно определенная матрица, то

$$F(\mathbf{a}+\mathbf{h})>F(\mathbf{a}) \quad \text{для всех малых}^{23}\ \mathbf{h}.$$

Если, однако, $\mathbf{h}$ варьируется только таким образом, что все ограничения соблюдаются, это означает, что

$$f(\mathbf{a}+\mathbf{h})>f(\mathbf{a}),$$

т. е. $f(\mathbf{a})$ — локально минимальное значение при соблюдении указанных ограничений. Может случиться так, что

$$F(\mathbf{a}+\mathbf{h}) \not> F(\mathbf{a})$$

для всех малых $\mathbf{h}$, но

$$f(\mathbf{a}+\mathbf{h})>f(\mathbf{a})$$

для всех $\mathbf{h}$, которые удовлетворяют ограничениям. Поэтому условие положительной определенности матрицы $\mathbf{M}(\mathbf{a})$ достаточно, но не необходимо для того, чтобы в условиях указанных ограничений при $\boldsymbol{\theta}=\mathbf{a}$ имел место локальный минимум. Аналогичные замечания можно сделать и в случае отрицательной определенности матрицы $\mathbf{M}$. Если $\mathbf{M}(\mathbf{a})$ не является ни положительно, ни отрицательно определенной, то для определения типа стационарной точки необходимы дополнительные исследования функции в окрестности точки $\mathbf{a}$.

Упражнения

1. В этих вопросах приняты следующие обозначения:

$$\mathbf{A}=\begin{bmatrix} 4 & 1 \\ 3 & -2 \\ 1 & 7 \end{bmatrix}, \qquad \mathbf{B}=\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \mathbf{C}=\begin{bmatrix} 3 & -1 & 2 \\ -2 & 3 & 1 \\ 1 & 4 & 1 \end{bmatrix}.$$

Проверьте, правильны или ошибочны соотношения:

1) $\mathbf{B}+\mathbf{C}=\begin{bmatrix} 5 & 0 & 2 \\ -1 & 5 & 1 \\ 1 & 4 & 1 \end{bmatrix}$. 2) $\mathbf{AC}=\begin{bmatrix} 7 & 9 & 12 \\ 14 & 21 & 7 \end{bmatrix}$.

3) $\mathbf{AB}=\begin{bmatrix} 9 & 6 \\ 4 & -1 \\ 9 & 15 \end{bmatrix}$. 4) $\mathbf{B}^{-1}=\frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$.

5) $\mathbf{A}^{-1}=\frac{1}{11}\begin{bmatrix} 2 & 1 & 0 \\ 3 & -4 & 0 \end{bmatrix}$. 6) $(\mathbf{A}'\mathbf{A})^{-1}\mathbf{A}'\mathbf{ABB}^{-1}=\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$.

---

[23] Надо было бы пояснить, что значит «для всех малых $\mathbf{h}$». Малость может быть выражена, например, ограничением на норму вектора $\mathbf{h}$, т. е. $\mathbf{h}'\mathbf{h}\leqslant\varepsilon$, где ε — некоторое малое действительное число.— *Примеч. пер.*

2. Примем обозначения:

$$\mathbf{A}=\begin{bmatrix}4&0&3\\0&4&0\\3&0&2\end{bmatrix},\qquad \mathbf{B}=\begin{bmatrix}-1&1\\2&3\\3&2\end{bmatrix},\qquad \mathbf{C}=\begin{bmatrix}4&3\\3&2\end{bmatrix}.$$

Вычислите матрицы, приведенные ниже, или обоснуйте, что это невозможно сделать:

1) $\mathbf{B}+\mathbf{C}$. 2) $\mathbf{BB}'$. 3) $\mathbf{A}+\mathbf{B}'\mathbf{B}$.

4) $\mathbf{BC}$. 5) $\mathbf{AA}^{-1}\mathbf{BC}$. 6) $\mathbf{CB}'$.

7) $\mathbf{CAB}$. 8) $\mathbf{BC}^{-1}$, где $\mathbf{C}^{-1}=\begin{bmatrix}-2&3\\3&-4\end{bmatrix}$. 9) $\mathbf{A}^{-1}$.

10) $\mathbf{A}'\mathbf{A}(\mathbf{A}')^{-1}\mathbf{A}^{-1}$.

3. В следующих задачах:

$$\mathbf{A}=\begin{bmatrix}1&1&1\\1&2&3\end{bmatrix},\qquad \mathbf{b}=\begin{bmatrix}-1\\1\end{bmatrix},\qquad \mathbf{C}=\begin{bmatrix}1&0\\2&3\end{bmatrix},$$

$$\mathbf{D}=\begin{bmatrix}1&1&1\\1&2&0\\-1&1&0\end{bmatrix}.$$

Правильны или ошибочны приведенные ниже результаты? Если неправильны, то объясните, почему.

1) $\mathbf{Ab}=[0\quad 1\quad 2]$. 2) $\mathbf{A}'\mathbf{C}=\begin{bmatrix}3&3\\5&6\\7&9\end{bmatrix}$.

3) $\mathbf{AD}=\begin{bmatrix}1&4&1\\6&8&1\end{bmatrix}$. 4) $\mathbf{C}^{-1}=\begin{bmatrix}1&0\\-\frac{2}{3}&\frac{1}{3}\end{bmatrix}$.

5) $\mathbf{bC}=\begin{bmatrix}1\\3\end{bmatrix}$. 6) $\mathbf{A}-\mathbf{D}=\begin{bmatrix}0&0&0\\0&0&3\\-1&1&0\end{bmatrix}$.

7) $\mathbf{C}^{-1}\mathbf{CCC}^{-1}\ \mathbf{C}^{-1}\mathbf{CCC}^{-1}\mathbf{C}=\mathbf{C}$. 8) $\mathbf{D}^{-1}=\frac{1}{3}\begin{bmatrix}0&1&-2\\0&1&1\\3&-2&1\end{bmatrix}$.

9) $\mathbf{b}'\mathbf{Cb}=2$.

4. Используя матричный подход, по данным, приведенным ниже, постройте регрессионное уравнение. Исходя из модели $Y=\beta_0+\beta_1 X+\varepsilon$ найдите оценки b_0 и b_1. Вычертите график с экспериментальными данными и подобранную прямую. Найдите предсказываемые значения отклика и остатки с точностью до 0,1. Составьте таблицу дисперсионного анализа и проверьте адекватность модели. Найдите матрицу дисперсий-ковариаций оцениваемых параметров, а также матричное выражение для $V(\hat{Y})$. Найдите $V(\hat{Y})$ при $X=65$ и постройте 95 %-ный доверительный интервал для $E(Y\mid X=65)$.

X:	30	40	50	80	30	40	60	70	70	70	30	80	70	70
Y:	13	17	20	29	12	15	22	25	23	27	15	27	24	26

5. Используя матричный подход, постройте уравнение прямой линии исходя из модели $Y = \beta_0 + \beta_1 X + \varepsilon$ согласно данным, приведенным ниже. Выполните полный анализ (т. е. все относящиеся к делу операции согласно параграфу 2.6).

X:	1	1	2	3	4	4	4	5	6	6
Y:	4,2	3,8	3,0	2,3	1,8	2,0	2,2	2,0	2,5	2,7

6. Используя приведенные ниже данные, выполните следующие операции:
1) Постройте регрессионное уравнение прямой согласно модели

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \mathbf{I}\sigma^2).$$

2) Нанесите на график экспериментальные данные и построенную прямую линию.
3) Составьте основную таблицу дисперсионного анализа.
4) Пополните эту таблицу дополнительными сведениями.
5) Проверьте гипотезу об адекватности модели.
6) Проведите проверку всей регрессии с помощью F-критерия. Полезна ли такая проверка?
7) Используя обычную процедуру, постройте 95 %-ный доверительный интервал для истинного среднего величины Y при $X_0 = \sqrt{122} + 6$. Полезна ли такая операция?
8) Вычислите предсказываемые значения отклика и остатки. Постройте график величин e_i в зависимости от соответствующих значений $\hat{Y}_i$.
9) Выпишите выражение матрицы дисперсий-ковариаций коэффициентов b.
10) Рассчитайте величину R^2.
11) Вычислите r^2_{xy}. Каково соотношение между этой величиной и R^2 для данной модели?
12) Каково ваше общее заключение?

X	Y	XY		X	Y	XY
0	—2	0		9	0	0
2	0	0		9	0	0
2	2	4		9	1	9
5	1	5		10	—1	—10
5	3	15	Сумма	60	5	32
9	1	9	Сумма квадратов	482	21	528

(Примечание. Лучше всего работать с целыми числами или простыми дробями, если это возможно. К десятичным дробям следует обращаться по возможности в последний момент.)

7. Используя данные, приведенные в табл. 1.1, найдите совместную 90 %-ную доверительную область для (β_0, β_1) [F (2, 23, 0,90)=2,55]. На точном рисунке типа рис. 1 к этому упражнению укажите:
1) Оцениваемую точку (b_0, b_1).
2) 90 %-ный доверительный контур для (β_0, β_1).
3) 95 %-ные доверительные интервалы для β_0 и β_1 порознь и прямоугольник, образованный этими интервалами.
Прокомментируйте кратко ваши результаты.
Подсказка. Выписанное полностью уравнение (2.6.15) есть квадратное уравнение относительно β_0 и β_1. Для того чтобы построить эллипс, можно поступить так. Надо зафиксировать величину β_0 и затем решать полученное квадратное уравнение относительно β_1. Найти два корня, соответствующие верхней и нижней точкам на вертикали (см. рис. 2 к этому упражнению). Если корни получились мнимые, то это означает, что вертикаль, соответствующая выбранному значению β_0, не пересекается с контуром эллипса (см. рис. 3 к этому упражнению). Повторяя такие операции при нескольких подходящих значениях β_0, получим точки, формирующие контур эллипса. Процедуру такого рода можно успешно выполнить с помощью ЭВМ.

8. На основе материала, изложенного в параграфе 2.12, покажите, что:
1) математическое ожидание среднего квадрата, связанного с неадекватностью, выражается формулой, приведенной в табл. на с. 157, 158;
2) E (остаточный средний квадрат) $= \sigma^2$, если модель корректна, т. е. если $\beta_2 = 0$.

9. Покажите, что величина R^2, определяемая по формуле (2.6.11), равна квадрату коэффициента корреляции между $\mathbf{Y}$ и $\hat{\mathbf{Y}}$.

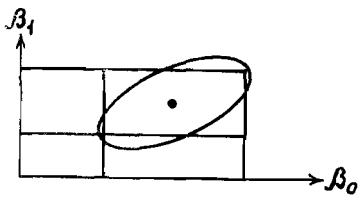


Рис. 1 к упражнению 7

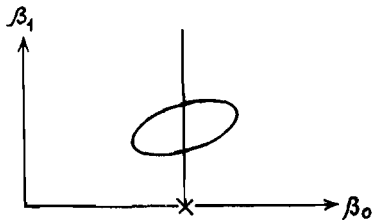


Рис. 2 к упражнению 7

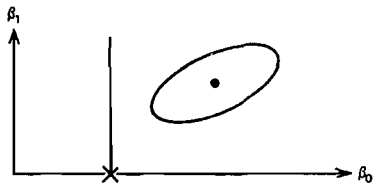


Рис. 3 к упражнению 7

10. Рассмотрите возможность описания взаимосвязи остатков e_i и предсказываемых значений отклика $\hat{Y}_i$ с помощью формальной квадратичной регрессии: $e_i = \alpha_0 + \alpha_1 \hat{Y}_i + \alpha_2 \hat{Y}_i^2$. Покажите, что все три оцениваемых коэффициента регрессии зависят от $T_{12} = \Sigma e_i \hat{Y}_i^2$. Что из этого следует?

11. При подгонке уравнения прямой линии мы пользуемся формулами $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ и $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$. Используя обозначение $\mathbf{R} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, покажите, что

$$\begin{aligned}\mathbf{SS}\ (\text{обусловленная регрессией}) &= \mathbf{Y}'\mathbf{R}\mathbf{Y},\\ &= \hat{\mathbf{Y}}'\hat{\mathbf{Y}},\\ &= \hat{\mathbf{Y}}'\mathbf{R}^2\mathbf{Y}.\end{aligned}$$

12. Покажите, что $\mathbf{X}'\mathbf{e} = 0$.

13. Применяя формулу (2.6.8) к n экспериментальным данным (см. также параграф 1.8), покажите, что для любой линейной модели справедливо соотношение

$$\sum_{i=1}^{n} V(\hat{Y}_i)/n = \text{trace}\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\sigma^2/n = p\sigma^2/n.$$

Используйте при этом преобразования, описанные в параграфе 2.12.

14. Подставьте в формулу (2.6.13а) определенные значения ν_1, ν_2 и $F(\nu_1, \nu_2, 1-\alpha)$ для выбранного значения α (например, $\alpha = 0{,}05$). Величина R^2, которую вы получите, соответствует той F, которая «в точности значима на уровне

100 α %». Вы будете удивлены тем, насколько низкое значение имеет величина R^2, и это даст вам дополнительные аргументы для более детального ознакомления с приложением 2В. Применение данных, указанных в табл. 2В.1—2В.3, при условии, что $\nu_m = \nu_1$, $\nu_r = \nu_2$, приведет к получению более высоких значений R^2. Проделайте несколько простых примеров такого рода.

15. Можно ли распространить результат упражнения 25 из гл. 1 на случаи, когда имеется большее число переменных X? (Да.)

16. Предположим, что регрессионная модель имеет вид $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ и при этом она содержит коэффициент β_0 «в первой позиции». Примем обозначение $\mathbf{1} = (1, 1, \ldots, 1)' - (n \times 1)$-вектор из единиц. Покажите, что $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{1} = (1, 0, \ldots, 0)'$ и, следовательно, что $\mathbf{1}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{1} = n$. (Подсказка. $\mathbf{X}'\mathbf{1}$ есть первый столбец матрицы $\mathbf{X}'\mathbf{X}$). Эти результаты могут быть полезными при проведении регрессионного анализа с помощью матриц. В связи с этим см. письма в Amer. Statist., April 1972, p. 47—48.

17. Имея в виду, что $\mathbf{X}_0 = (1, \bar{X}_1, \bar{X}_2, \ldots)'$ можно записать в виде $\mathbf{X}'\mathbf{1}/n$, и, применяя результаты из параграфа 2.4, а также из упражнения 16, покажите, что $V(\hat{Y})$ в точке $(\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_n)$ есть σ^2/n.

Ответы к упражнениям

1. 1) Неправильно. Матрицы имеют разные размеры. 2) Неправильно. **A** есть 3×2-матрица, **C** — 3×3-матрица.
3) Правильно. 4) Правильно.
5) Неправильно. Только квадратные матрицы имеют обратные.
6) Неправильно. В правой части должна быть матрица 2×2, а не 3×3.

2. 1) Невозможно, так как **B** — 3×3-матрица, а **C** — 2×2-матрица.

2) $$\mathbf{BB}' = \begin{bmatrix} -1 & 1 \\ 2 & 3 \\ 3 & 2 \end{bmatrix}\begin{bmatrix} -1 & 2 & 3 \\ 1 & 3 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 1 & -1 \\ 1 & 13 & 12 \\ -1 & 12 & 13 \end{bmatrix}.$$

3) Невозможно, так как **A** есть 3×3-матрица, $\mathbf{B}'$ **B** есть 2×2-матрица.

4) $$\mathbf{BC} = \begin{bmatrix} -1 & 1 \\ 2 & 3 \\ 3 & 2 \end{bmatrix}\begin{bmatrix} 4 & 3 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} -1 & -1 \\ 17 & 12 \\ 18 & 13 \end{bmatrix}.$$

5) $\mathbf{AA}^{-1}\mathbf{BC} = \mathbf{IBC} = \mathbf{BC}$, см. выше.

6) $$\mathbf{CB}' = (\mathbf{BC}')' = (\mathbf{BC})' = \begin{bmatrix} -1 & 17 & 18 \\ -1 & 12 & 13 \end{bmatrix}.$$

7) Невозможно. **C** есть 2×2-матрица. **A** — 3×3-матрица.

8) $$\begin{bmatrix} -1 & 1 \\ 2 & 3 \\ 3 & 2 \end{bmatrix}\begin{bmatrix} -2 & 3 \\ 3 & -4 \end{bmatrix} = \begin{bmatrix} 5 & -7 \\ 5 & -6 \\ 0 & 1 \end{bmatrix}.$$

9) Разбейте матрицу на блоки. Найдите обратное значение от элемента в середине матрицы, получите 1/4. Элементы, стоящие в углах, образуют матрицу **C**. Получите $\mathbf{C}^{-1}$. Расположите эти результаты необходимым образом и получите

$$\begin{bmatrix} -2 & 0 & 3 \\ 0 & 1/4 & 0 \\ 3 & 0 & -4 \end{bmatrix}.$$

10) $\mathbf{A}$ симметрична, так что $\mathbf{A}' = \mathbf{A}$. Следовательно, $\mathbf{A}'\mathbf{A}(\mathbf{A}')^{-1}\mathbf{A}^{-1} = \mathbf{AAA}^{-1}\mathbf{A}^{-1} = \mathbf{AIA}^{-1} = \mathbf{I}_{3\times 3}$.

3. $\mathbf{A}$ есть 2×3-матрица, $\mathbf{B}$ — 2×1-матрица, $\mathbf{C}$ — 2×2-матрица, $\mathbf{D}$ — 3×3-матрица. Так что 1), 5) и 6) неправильны. Указанные операции при данных размерах матриц неосуществимы. 3) также неправильно. Элемент (2,1) равен не 6, а 0, т. е.

$$\mathbf{AD} = \begin{bmatrix} 1 & 4 & 1 \\ 0 & 8 & 1 \end{bmatrix}.$$

Уравнения 2), 4), 7), 8) и 9) правильны. Однако в уравнении 7) необходимо проверить, что $\mathbf{C}^{-1}$ существует.

4. $$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 14 & 790 \\ 790 & 49300 \end{bmatrix}^{-1} = \frac{1}{66100}\begin{bmatrix} 49300 & -790 \\ -790 & 14 \end{bmatrix},$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 295 \\ 18030 \end{bmatrix},$$

$$\mathbf{b} = \begin{bmatrix} 299800/66100 \\ 19370/66100 \end{bmatrix} = \begin{bmatrix} 4{,}535552 \\ 0{,}293041 \end{bmatrix}.$$

(График данных, линия, остатки).

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
Общий	14	6641,00		
b_0	1	6216,07		
$b_1 \mid b_0$	1	405,45	405,45	250, значимо
Остаток	12	19,48	$s^2 = 1{,}62$	
Неадекватность	4	0,81	0,20	не значимо
«Чистая» ошибка	8	18,67	2,33	

$$\mathbf{V}(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}s^2 = \begin{bmatrix} 1{,}208 & -0{,}019 \\ -0{,}019 & 0{,}000343 \end{bmatrix}.$$

Если $X = 65$,

$$V(\hat{Y}) = (1{,}65)\begin{bmatrix} 1{,}208 & -0{,}019 \\ -0{,}019 & 0{,}000343 \end{bmatrix}\begin{bmatrix} 1 \\ 65 \end{bmatrix} = 0{,}1872,$$

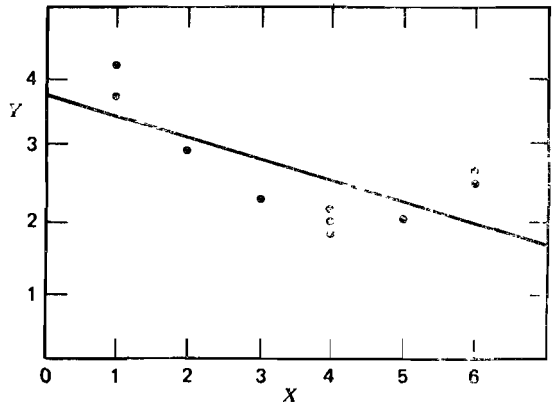
$$\hat{Y}(65) = 23{,}583, \qquad t(12,\ 0{,}975) = 2{,}179,$$

то 95 %-ный доверительный интервал для $E\ (Y \mid X = 65)$ есть $23{,}583 \pm 2{,}179\,(0{,}1872)^{1/2}$ = от 22,640 до 24,526.

5. $$\mathbf{b} = \begin{bmatrix} 10 & 36 \\ 36 & 160 \end{bmatrix}^{-1}\begin{bmatrix} 26{,}5 \\ 86{,}1 \end{bmatrix} = \frac{1}{304}\begin{bmatrix} 160 & -36 \\ -36 & 10 \end{bmatrix}\begin{bmatrix} 26{,}5 \\ 86{,}1 \end{bmatrix} = \begin{bmatrix} 3{,}75132 \\ -0{,}305921 \end{bmatrix}.$$

Y_i	4,2	3,8	3,0	2,3	1,8	2,0	2.2	2,0	2,5	2,7
$\hat{Y}_i$	3,45	3,45	3,14	2,83	2,53	2,53	2,53	2,22	1,92	1,92
$e_i = Y_i - \hat{Y}_i$	0,75	0,35	−0,14	−0,53	−0,73	−0,53	−0,33	−0,22	0,58	0,78

$\Sigma e_i = -0{,}02$



К решению упражнения 5

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
b_0	1	70,225		
$b_1 \mid b_0$	1	2,845		
Неадекватность	4	2,740	$MS_L = 0{,}685$	15,222*
«Чистая» ошибка	4	0,180	$s_e^2 = 0{,}045$	
Общий	10	75,990		

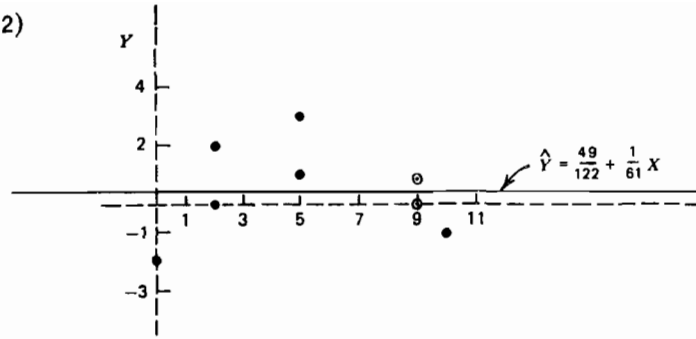
* Прежде всего необходимо проверить, является ли модель адекватной. Если она неадекватна, то большинство вычислений (например, F-тест для регрессии, доверительные интервалы, доверительные границы) не обосновано и их не следует проводить совсем. Теперь находим табличное значение $F\ (4, 4; 0{,}95) = 6{,}39$ Модель неадекватна, поскольку $15{,}222 > 6{,}39$. $R^2 = \dfrac{2{,}845}{5{,}765} = 0{,}4935$. Прямая линия объясняет только 49,35 % вариаций около среднего, т. е. немного.

Из анализа остатков мы отчетливо видим чередование положительных, затем отрицательных и снова положительных остатков. Это означает, что данные имеют характер квадратичной зависимости, которую нельзя описать с помощью прямой линии. Это видно, конечно, из графика.

Мы заключаем, что линейная модель неадекватна и не пригодна для использования Необходимо далее пытаться использовать модель $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$. (Если мы построим такую модель, то получим $\hat{Y} = 5{,}462 - 1{,}6380X + 0{,}192840\ X^2$ при хорошем согласии с экспериментальными данными.)

6. 1) $\mathbf{b} = \begin{bmatrix} 10 & 60 \\ 60 & 482 \end{bmatrix}^{-1} \begin{bmatrix} 5 \\ 32 \end{bmatrix} = \begin{bmatrix} 49/122 \\ 1/61 \end{bmatrix}$, $\quad \hat{Y} = \dfrac{49}{122} + \dfrac{1}{61} X.$

2)



К решению упражнения 6

3) Основа дисперсионного анализа:

$$\mathbf{b}'\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \frac{49}{122}, & \frac{1}{61} \end{bmatrix} \begin{bmatrix} 5 \\ 32 \end{bmatrix} = \frac{245+64}{122} = 2\frac{65}{122}.$$

Дисперсионный анализ

Источник	Число степеней свободы	SS
b_0, b_1	2	$2\frac{65}{122}$
Остаток	8	$18\frac{57}{122}$
Общий	10	21

4) $\mathrm{SS}(b_0) = (\Sigma Y)^2/n = 2\frac{1}{2}$.

X	Вклад в «чистую» ошибку	Число степеней свободы
2	2	1
5	2	1
9	1	3
—	5 = SS, обусловленная «чистой» ошибкой	5

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS
b_0	1	$2\frac{1}{2}$	
$b_1 \mid b_0$	1	$\frac{2}{61} = 0{,}033$	
Неадекватность	3	$13\frac{57}{122} = 13{,}467$	$4\frac{179}{366}$
«Чистая» ошибка	5	5	
Общий	10	21	

5) $F = \dfrac{\mathrm{MS_L}}{s_e^2} = 4\dfrac{179}{366} < F\,(3;\ 5;\ 0{,}95) = 5{,}41.$

Следовательно, модель адекватна. Заменим SS, обусловленную неадекватностью, на SS остаточную и получим

$$s^2 = 18\,\frac{57}{122}\,/8 = 2\,\frac{301}{976} = 2{,}308525.$$

6) $F = \dfrac{2/61}{2\dfrac{301}{976}} < 1.$

F-отношение незначимо. Да, тест полезен, поскольку модель адекватна.

7) $V(\hat{Y}_0) = \left\{\dfrac{1}{n} + \dfrac{(X_0 - \overline{X})^2}{\Sigma\,(X_i - \overline{X})^2}\right\}\sigma^2 = \left\{\dfrac{1}{10} + \dfrac{(X_0 - 6)^2}{122}\right\}\sigma^2.$

Если $X_0 = \sqrt{122} + 6$, то эта величина будет равна 1,1 σ^2. Ее оценкой является

$$1{,}1s^2 = (1{,}1)\left(2\frac{301}{976}\right) = (1{,}1)\,(2{,}308525) = 2{,}5393775,$$

$$t\,(8;\ 0{,}975) = 2{,}306.$$

Так что требуемый интервал есть

$$\hat{Y}_0 \pm 2{,}306\sqrt{2{,}5393775}, \qquad \text{или} \qquad \hat{Y}_0 \pm 3{,}6747.$$

Да, это обоснованно, поскольку модель адекватна. Однако точка X_0 довольно далеко выходит за пределы области экспериментирования, и мы не можем гарантировать здесь работоспособность модели.

8)

X_i	Y_i	$\hat{Y}_i$	$e_i = Y_i - \hat{Y}_i$	X_i	Y_i	$\hat{Y}_i$	$e_i = Y_i - \hat{Y}_i$
0	—2	0,40	—2,40	9	1	0,54	0,46
2	0	0,43	—0,43	9	0	0,54	—0,54
2	2	0,43	1,57	9	0	0,54	—0,54
5	1	0,48	0,52	9	1	0,54	0,46
5	3	0,48	2,52	10	—1	0,56	—1,56

$$0{,}06 = \Sigma e_i$$

(Два знака после запятой достаточно, даже и одного хватило бы.) График остатков e_i в зависимости от Y_i похож на график в пункте 2) с изменением осей, конечно. (Ответьте сами, почему.)

$$9)\ \mathbf{V(b)} = (\mathbf{X'X})^{-1}\sigma^2 = \begin{bmatrix} 10 & 60 \\ 60 & 482 \end{bmatrix}^{-1}\sigma^2 = \frac{\sigma^2}{1220}\begin{bmatrix} 482 & -60 \\ -60 & 10 \end{bmatrix} =$$

$$= \frac{\sigma^2}{610}\begin{bmatrix} 241 & -30 \\ -30 & 5 \end{bmatrix}.$$

$$10)\ \mathbf{R}^2 = \frac{2/61}{18\frac{1}{2}} = \frac{2}{61}\cdot\frac{2}{37} = \frac{4}{2257} = 0{,}18\%\ .$$

11) Идентично R^2 для модели в виде прямой линии.

12) Регрессия незначима, модель адекватна. Тем не менее графики остатков указывают на определенные систематические отклонения от случайного поведения, так что необходимо дальнейшее исследование, несмотря на адекватность модели. Полученная модель не представляет особой ценности.

7. 1) Оценки параметров $(b_0,\ b_1)$ из параграфа 1.2 имеют вид

$$(b_0,\ b_1) = (13{,}623005,\ -0{,}079829).$$

2) 90 %-ная доверительная область для $(\beta_0,\ \beta_1)$ в соответствии с (2.6.15) выражается неравенством

$$(\boldsymbol{\beta} - \mathbf{b})'\,\mathbf{X'X}\,(\boldsymbol{\beta} - \mathbf{b}) \leqslant ps^2F\,(p,\ \nu,\ 1-\alpha).$$

Пусть

$$\begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix} = \boldsymbol{\gamma} = \boldsymbol{\beta} - \mathbf{b} = \begin{pmatrix} \beta_0 - b_0 \\ \beta_1 - b_1 \end{pmatrix}.$$

Это эквивалентно приведению к новому началу координат в точке $\mathbf{b}$. Следовательно, (2.6.15) приобретает вид

$$\boldsymbol{\gamma}'\mathbf{X'X}\boldsymbol{\gamma} \leqslant ps^2F\,(p,\ \nu,\ 1-\alpha)\ .\ .\ . \qquad (1)$$

$$\gamma'\mathbf{X}'\mathbf{X}\gamma = (\gamma_0,\ \gamma_1)\begin{bmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{bmatrix}\begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix} = \left[n\gamma_0 + \gamma_1 \Sigma X_i,\ \gamma_0 \Sigma X_i + \right.$$

$$\left. + \gamma_1 \Sigma X_i^2\right]\begin{bmatrix} \gamma_0 \\ \gamma_1 \end{bmatrix} = \left(n\gamma_0 + \gamma_1 \Sigma X_i\right)\gamma_0 + \left(\gamma_0 \Sigma X_i + \gamma_1 \Sigma X_i^2\right)\gamma_1 =$$

$$= n\gamma_0^2 + 2\gamma_0\gamma_1 \Sigma X_i + \gamma_1^2 \Sigma X_i^2.$$

Таким образом, соотношение (1) можно записать, например, так:

$$n\gamma_0^2 + 2\gamma_0\gamma_1 \Sigma X_i + \gamma_1^2 \Sigma X_i^2 \leqslant ps^2 F\ (p,\ \nu,\ 1-\alpha) = c.$$

Доверительная область включает внутренние точки эллипса, граница которого выражается уравнением

$$n\gamma_0^2 + 2\gamma_0\gamma_1 \Sigma X_i + \gamma_1^2 \Sigma X_i^2 = c\ \ldots \qquad (2)$$

в пространстве переменных $(\gamma_0,\ \gamma_1)$.

Чтобы найти точки, лежащие на этой границе, выберем некоторое значение β_0, а следовательно, и γ_0. Решим квадратное уравнение (2) относительно γ_1. Оно имеет два корня (которые дают верхнюю и нижнюю точки на эллипсе).

$$\gamma_{11} = \left\{-\gamma_0 \Sigma X_i - \sqrt{\left(\gamma_0 \Sigma X_i\right)^2 - \left(\Sigma X_i^2\right)(n-c)}\right\}/\Sigma X_i^2,$$

$$\gamma_{12} = \left\{-\gamma_0 \Sigma X_i + \sqrt{\left(\gamma_0 \Sigma X_i\right)^2 - \left(\Sigma X_i^2\right)(n-c)}\right\}/\Sigma X_i^2.$$

(Мнимые корни означают, что мы выбрали величину γ_0, соответствующую точке, лежащей вне эллипса.) Затем, используя формулы $\beta_0 = \gamma_0 + b_0$, $\beta_{1j} = \gamma_{1j} + b_1$, получим координаты двух точек $(\beta_0,\ \beta_{11})$, $(\beta_0,\ \beta_{12})$ в пространстве истинных параметров $(\beta_0,\ \beta_1)$. Эти точки наносятся на график, затем выбирается новое значение β_0, и весь цикл вычислений повторяется до тех пор, пока не получится рисунок всего эллипса. Нам известно, что

$$\Sigma X_i = 1315, \qquad \Sigma X_i^2 = 76323{,}42, \qquad c = ps^2 F\ (p,\ \nu,\ 1-\alpha) = 2\,(0{,}7926)\,(2{,}55) =$$

$$= 4{,}04225.$$

С помощью программы на Фортране мы получили следующие результаты:

β_0	β_{11}	β_{12}	β_0	β_{11}	β_{12}
12,4	—0,0614	—0,0561	13,7	—0,0884	—0,0739
12,5	—0,0643	—0,0567	13,8	—0,0901	—0,0757
12,6	—0,0668	—0,0576	13,9	—0,0917	—0,0775
12,7	—0,0691	—0,0587	14,0	—0,0933	—0,0794
12,8	—0,0713	—0,0600	14,1	—0,0948	—0,0813
12,9	—0,0734	—0,0613	14,2	—0,0963	—0,0832
13,0	—0,0755	—0,0627	14,3	—0,0977	—0,0853
13,1	—0,0775	—0,0641	14,4	—0,0991	—0,0873
13,2	—0,0794	—0,0657	14,5	—0,1004	—0,0895
13,3	—0,0813	—0,0672	14,6	—0,1015	—0,0918
13,4	—0,0832	—0,0688	14,7	—0,1025	—0,0942
13,5	—0,0850	—0,0705	14,8	—0,1033	—0,0969
13,6	—0,0867	—0,0722	14,9	—0,1035	—0,1001

3) 95 %-ные доверительные пределы для β_1 имеют вид

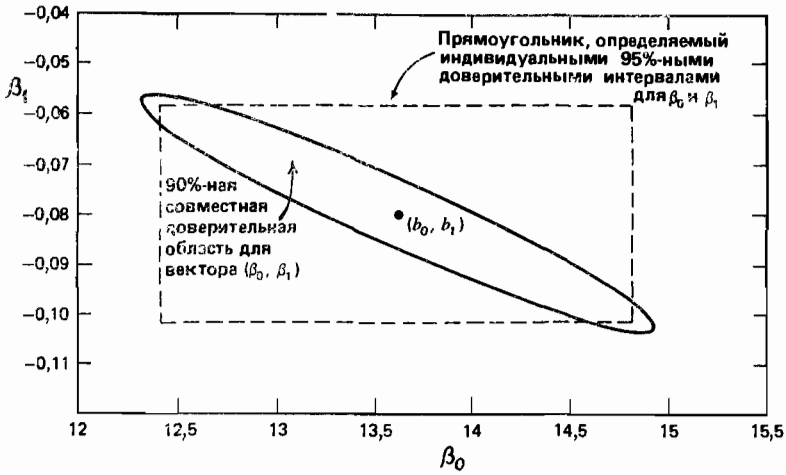
$$b_1 \pm t\,(23;\ 0{,}975)\, s/\{\Sigma\,(X_i - \overline{X})^2\}^{1/2},$$

или $-0{,}0798 \pm (2{,}069)\,(0{,}0105)$; они образуют интервал $-0{,}1015 \leqslant \beta_1 \leqslant -0{,}0581$.

95 %-ные доверительные пределы для β_0 выражаются соотношением

$$b_0 \pm t\,(23;\ 0{,}975)\, s/\{\Sigma X_i^2 / n\Sigma\,(X_i - \overline{X})^2\}^{1/2},$$

или $13{,}623 \pm (2{,}069)\,(0{,}5814)$; они образуют интервал $12{,}420 \leqslant \beta_0 \leqslant 14{,}826$.



К решению упражнения 7

Прямоугольник, порождаемый этими двумя интервалами, показан на рисунке к решению данного упражнения.

Примечания. (1) Совместная 90 %-ная доверительная область для параметров β_0 и β_1 изображена в виде тонкого вытянутого эллипса. Она накрывает значения (β_0, β_1), которые при наших данных выглядят как совместно приемлемые для параметров. (2) Если мы интерпретируем 95 %-ные доверительные интервалы для β_0 и β_1 одновременно (по ошибке) как «совместную 90,25 %-ную доверительную область» (прямоугольник), то ясно, что мы будем до некоторой степени заблуждаться.

8. 1) Либо выразите средний квадрат, отвечающий неадекватности, в виде квадратичной формы и примените результаты, содержащиеся в параграфе 2.12 непосредственно, либо вычислите

$$E\,(\mathbf{Y}'\mathbf{Y}) = E\,(\mathbf{Y}'\mathbf{I}\mathbf{Y}) = E\,(\mathbf{Y}')\,E\,(\mathbf{Y}) + \operatorname{trace}\mathbf{I}\sigma^2 = (\mathbf{X}\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\beta}_2)'\,(\mathbf{X}\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\beta}_2) + n\sigma^2,$$

а затем найдите математическое ожидание суммы квадратов, обусловленной неадекватностью, с помощью разности, полагая, что другие результаты имеются в таблице. Не забудьте при этом, где необходимо, умножить на соответствующие числа степеней свободы. Напомним, что $\mathbf{I} - \mathbf{R} = (\mathbf{I} - \mathbf{R})^2$, где $\mathbf{R} = \mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'$ и что $(\mathbf{I} - \mathbf{R})\,\mathbf{X} = \mathbf{0}$,

$$(\mathbf{I} - \mathbf{R})\,\mathbf{X}_2\boldsymbol{\beta}_2 = (\mathbf{X}_2 - \mathbf{X}\mathbf{A})\,\boldsymbol{\beta}_2.$$

2) E (остаточный средний квадрат $|\ \boldsymbol{\beta}_2 = \mathbf{0}) = (f\sigma^2 + e\sigma^2)/(f + e) = \sigma^2$.

9. Мы имеем $\hat{Y}_i - \overline{\hat{Y}}_i = \hat{Y}_i - \overline{Y}$; $\quad Y_i - \overline{Y} = (\hat{Y}_i - \overline{Y}) + e_i$.

Числитель в выражении для $r_{Y\hat{Y}}$ есть сумма смешанных произведений двух

указанных выше величин. Она сводится к $\Sigma(\hat{Y}_t - \overline{Y})^2$, другие члены исчезают благодаря тому, что (а) остатки ортогональны к величинам $\hat{Y}_t$ и (б) поскольку $\Sigma e_t = 0$. (Чтобы доказать (а), выполним следующие выкладки:

$$\hat{\mathbf{Y}}'\mathbf{e} = (\mathbf{RY})'(\mathbf{I} - \mathbf{R})\,\mathbf{Y} = \mathbf{Y}'\mathbf{R}'(\mathbf{I} - \mathbf{R})\,\mathbf{Y} = \mathbf{0},$$

поскольку $\mathbf{R}$ симметрична ($\mathbf{R}' = \mathbf{R}$) и идемпотентна ($\mathbf{R} = \mathbf{R}^2$).) Корень квадратный из числителя сокращается теперь с одним из сомножителей в знаменателе. То, что остается, как раз и есть $\mathbf{R}$, как определено соотношением (2.6.11).

10. Вектор, образованный из правых частей соответствующих нормальных уравнений, включает три элемента $\Sigma e_t = 0$, $\Sigma e_t \hat{Y}_t = 0$ и $T_{12} = \Sigma e_t \hat{Y}_t^2$. Следовательно, все три оценки коэффициентов зависят от T_{12}. Эта величина является мерой квадратичного тренда графика зависимости остатков e_t от $\hat{Y}_t$.

11. SS (регрессия) $= \mathbf{b}'\mathbf{X}'\mathbf{Y} = ((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y})'\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{Y}'\mathbf{R}\mathbf{Y}$.

Напомним, что $\mathbf{R} = \mathbf{R}^2 = \mathbf{R}^3 = \ldots = \mathbf{R}^m \ldots$ Следовательно,

$$\hat{\mathbf{Y}} = \mathbf{Xb} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{RY}.$$

Так что

$$\hat{\mathbf{Y}}'\hat{\mathbf{Y}} = \mathbf{Y}'\mathbf{R}'\mathbf{RY} = \mathbf{Y}'\mathbf{RRY} = \mathbf{Y}'\mathbf{RY} = \text{SS (регрессия)}.$$

$$\hat{\mathbf{Y}}'\mathbf{R}^3\mathbf{Y} = \hat{\mathbf{Y}}'\mathbf{RY} = \hat{\mathbf{Y}}'\hat{\mathbf{Y}}.$$

12. $\mathbf{X}'\mathbf{e} = \mathbf{X}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\,\mathbf{Y} = (\mathbf{X}' - \mathbf{X}')\,\mathbf{Y} = 0.$

13. Метод решения содержится в вопросе.

14—17. Решение не приводится.

Глава 3 ● ИССЛЕДОВАНИЕ ОСТАТКОВ

«Почти все величайшие открытия в астрономии вытекают из рассмотрения того, что мы уже раньше назвали качественными или численными ОСТАТОЧНЫМИ ФЕНОМЕНАМИ, иначе говоря, они вытекают из анализа той части числовых или качественных результатов наблюдения, которая «торчит» и остается необъясненной после выделения и учета всего того, что согласуется со строгим применением известных методов».

Сэр Джон Ф. В. Гершель[1]. *Основы астрономии.* (*Sir John F. W. Herschel, Bart K. H. in: Outlines of Astronomy. Philadelphia: Lea and Blanchard, 1849, p. 548.*)

3.0. ВВЕДЕНИЕ

(Примечание. Материал этой главы полезен и имеет силу не только для линейных, но и для нелинейных регрессионных моделей, а также для моделей дисперсионного анализа. Фактически выводы этой главы приложимы к *любой* ситуации, где речь идет о подборе модели и есть пригодные для исследования меры необъясненных вариаций (в форме остатков).)

Остатки были определены как n разностей $e_i = Y_i - \hat{Y}_i$, $i = 1, 2, \ldots, n$, где Y_i — наблюдаемая величина, а $\hat{Y}_i$ — соответствующая прогнозируемая величина, получаемая при помощи найденного уравнения регрессии.

Из этого определения можно видеть, что остатки e_i есть разности между тем, что фактически наблюдалось, и тем, что предсказывается

[1] Сэр Джон Фредерик Вильям Гершель, баронет, английский придворный королевский астроном (1792—1871), сын знаменитого Вильяма Гершеля (1738—1822), человека яркой судьбы, музыканта, механика-строителя телескопов, астронома, открывшего планету Уран. Вильям Гершель стоял у оснований современных представлений о строении Вселенной. Он один из первых применил статистические методы в астрономии (см.: Еремеева А. И. Вселенная Гершеля.— М.: Наука, 1966, 319 с.; Курт Р. Введение в звездную статистику/Пер. с англ. Под ред. А. С. Шарова.— М.: Мир, 1969, с. 23). Сны продолжил дело отца и тоже стал крупным астрономом. Его книги выдержали ряд изданий. Есть и русские переводы: Гершель Д. Изложение астрономии.— Спб.: 1838, ч. 1—2 и Д. Гершель. Очерки астрономии Д. Гершеля/Пер. с 6-го англ. изд.— М.: Изд. А. Драшусова, 1861—1862, ч. 1—2. Мысль о роли остаточных эффектов отчетливо прослеживается в истории астрономии, см., например: Гребенников Е. А., Рябов Ю. А. Поиски и открытия планет.— 2-е изд.—М.: Наука, 1984, 224 с. (особенно рис. 16 и 17, с. 70 и др.).— *Примеч. пер.*

с помощью регрессионного уравнения. Иными словами — это величины, которые нельзя *объяснить* с помощью регрессионного уравнения. Таким образом, мы можем считать, что e_i — *наблюдаемые проявления ошибок, если модель правильна.* (Однако величины e_i связаны ограничениями, см. параграф 3.7.) При проведении регрессионного анализа мы делали некоторые предположения относительно ошибок. Обычные предположения состоят в признании ошибок независимыми, имеющими нулевые средние, одинаковую (постоянную) дисперсию σ^2 и подчиняющимися нормальному распределению. Последнее предположение необходимо для применения F-критерия. Таким образом, если подбираемая нами модель правильна, то остатки будут проявлять тенденцию к подтверждению сделанных предположений или по меньшей мере не будут противоречить им. Именно эта идея лежит в основе исследования остатков; мы должны сформулировать вопрос: «Не показывают ли остатки, что наши предположения ошибочны?». А после того как остатки исследованы, мы можем прийти к одному из следующих выводов:

1) предположения, по-видимому, нарушены (в определенном смысле, который можно оговорить);

2) предположения, по-видимому, не нарушены.

Утверждение (2) не означает, что мы пришли к выводу о правильности предположений; это означает только, что на основе данных, которые рассматривались, мы не имеем оснований для утверждения о неправильности. Такое же положение возникает при проверке гипотез, когда мы либо их отвергаем, либо не отвергаем (но не принимаем)[2]. Теперь обсудим способы исследования остатков для проверки модели. Все они имеют графический характер, легко выполнимы и обычно позволяют без всяких затруднений обнаружить нарушения предположений.

Основные виды графиков остатков:

1) общий;

---

[2] Затронутая здесь проблема «верифицируемости—фальсифицируемости» научных гипотез — одна из центральных в современной науке. Вывод, полученный на основании выборки конечного объема, всегда имеет некоторые шансы (пусть сколь угодно малые) оказаться ложным. Поэтому «положительный» ответ может означать, например, лишь то, что объем выборки слишком мал. Следовательно, эмпирические факты не могут доказывать что-либо, зато они могут вступать в противоречие с чем-то, например, с априорной моделью, и тем самым ее «губить». Природа, как хорошо сказал Г. Пойя, говорит «нет» громовым голосом, а вместо «да» что-то «невнятно попискивает». Авторы близки к позиции одного из крупнейших специалистов по данному вопросу — К. Поппера (см.: Поппер К. Логика и рост научного знания/Пер. с англ. Под ред. В. Н. Садовского.— М.: Прогресс, 1983, 605 с.). Интерпретацией этой проблемы неоднократно занимался В. В. Налимов (см., например: Налимов В. В. Вероятностная модель языка.— 2-е изд.— М.: Наука, 1979, с. 69—70; Налимов В. В. Логика принятия гипотез в развитии научного познаиия.— В кн.: Наука в социальных, гносеологических и ценностных аспектах.— М.: Наука, 1980, с. 139—176; Nalimov V. V. Faces of Science. Philadelphia: ISI Press, PA, 1981.— 297 p.) В контексте истории математики верификация обсуждается, например, в работе: Клайн М. Математика. Утрата определенности/Пер. с англ. Под ред. И. М. Яглома.— М.: Мир, 1984, с. 355—377 (особенно с. 369).— *Примеч. пер.*

2) в зависимости от времени, если известна последовательность реализации опытов;

3) в зависимости от предсказываемых значений $\hat{Y}_i$;

4) в зависимости от независимых переменных X_{ji} для $j = 1, 2, \ldots, k$.

Кроме того, графики остатков могут быть также вычерчены:

5) любым способом, который целесообразен для данной конкретной задачи.

Теперь рассмотрим эти графики подробно. Для иллюстрации возьмем следующий простой пример.

Пример. Регрессионный анализ дает одиннадцать остатков e_1, $e_2, \ldots, e_{11}$ со значениями 5, — 2, — 4, 4, 0, — 6, 9, — 2, — 5, 3, — 2.

[Примечание. Обычно остатки стоит записывать с тем же числом знаков после запятой, что и в исходных наблюдаемых откликах. Иногда выписывают еще один лишний знак, но, как правило, это просто «пустые хлопоты». (Приводимые в нашей книге машинные распечатки содержат больше знаков, чем это необходимо, поскольку так обычно выдает машина, но, конечно, их ничего не стоит округлить, если данные описываются для целей публикации.)]

3.1. ОБЩИЙ ГРАФИК

Если вычертить все остатки, приведенные выше, то мы получим диаграмму, показанную на рис. 3.1. Если наша модель правильна, то эти остатки должны иметь сходство с одиннадцатью наблюдениями из нормального распределения со средним, равным нулю. Противоречит ли общий график этим представлениям?

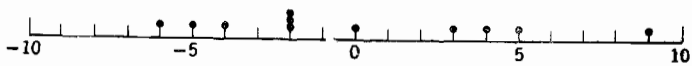


Рис. 3.1. Общий график остатков

Прежде всего заметим, что среднее остатков равно нулю. Но так обстоит дело в случае регрессионной модели со свободным членом β_0. Это хорошо видно из первого нормального уравнения, получаемого дифференцированием суммы квадратов ошибок по β_0. Если модель имеет вид $E(Y) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k$, то первое уравнение можно записать так:

$$-2\Sigma (Y_i - b_0 - b_1 X_{1i} - \ldots - b_k X_{ki}) = 0,$$

где суммирование проводится по $i = 1, 2, \ldots, n$. Это приводит к выражению

$$\Sigma (Y_i - \hat{Y}_i) = 0.$$

Таким образом,

$$\Sigma e_i = \frac{\Sigma e_i}{n} = 0.$$

Несмотря на то что график обнаруживает некоторую нерегулярность, она не кажется аномальной для одиннадцати наблюдений из нормального распределения. Но как это можно выразить? Для того чтобы установить какую-либо меру, позволяющую судить о графиках, подобных данному, можно использовать таблицу случайных нормальных отклонений. (Обширная таблица опубликована корпорацией Рэнд (Rand Corporation [3]), более короткие таблицы приведены в некоторых учебниках по статистике [4].) Ряд выборок заданного объема (здесь — из 11 наблюдений) можно взять и нанести на график так же, как это сделано выше. Даже небольшой опыт выполнения подобных графиков позволяет получить хорошее «представление» о том, как выглядел бы нормальный график. И его полезно приобрести прежде, чем выносить суждения о противоречии данных предположению о нормальности.

Другая процедура состоит в построении либо *нормального*, либо *полунормального* графика остатков на стандартной вероятностной бумаге (см. приложение 3А). Точки должны ложиться приблизительно на прямую линию. Но и здесь опять-таки надо иметь некоторую меру для оценки графика. Следовательно, нет никаких особых преимуществ для использования той или другой процедуры, хотя отдельные авторы и отдают предпочтение некоторым из них.

Когда число остатков очень велико, общему графику лучше придавать вид гистограммы, а не точечной диаграммы [5]. В этом случае нормальные или полунормальные графики стоит строить, только отбирая последовательно самые маленькие наблюдения. Пусть, например, 200 наблюдений наносятся на полунормальный график. Мы должны нанести сначала 10 наименьших результатов, затем 20 самых маленьких и действовать в том же духе до (скажем) 180 самых маленьких

---

[3] Rand Corporation — аббревиатура от слов Research and Development. Одна из крупнейших научно-исследовательских организаций США; расположена в Санта-Монике, штат Калифорния. Основу ее деятельности составляет проведение научной работы и исследований в области проблем национальной безопасности.— *Примеч. пер.*

[4] Наиболее полный фрагмент таких таблиц опубликован на русском языке в кн.: Большев Л. Н., Смирнов Н. В. Таблицы математической статистики.— 3-е изд.— М.: Наука, 1983, с. 100, 366—375. Фрагменты меньших размеров приводятся во многих книгах по прикладной статистике и планированию эксперимента, см., например: Налимов В. В. Применение математической статистики при анализе вещества.— М: Физматгиз, 1960, с. 386—396. К сожалению, проблема «случайности» случайных чисел оказалась не столь простой, как того хотелось бы. Краткую сводку основных результатов и соответствующую библиографию см., например, в предисловии Ю. П. Адлера к кн.: Иванова В. М. Случайные числа и их применение.— М.: Финансы и статистика, 1984, с. 3—10.— *Примеч. пер.*

[5] Существует и еще одна возможность.— это предложеиный Дж. Тьюки метод «опора и консоль», см.: Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия/Пер. с англ. Под ред. Ю. П. Адлера.— М.: Финансы и статистика, 1982, вып. 1, гл. 3, с. 58—91. В переводе другой книги одного из этих авторов вместо термина «опора и консоль» используется термин «стебель с листьями», см.: Тьюки Дж. Анализ результатов наблюдений/Пер. с англ. Под ред. В. Ф. Писаренко.— М.: Мир, 1981, гл. 1, с. 17—43. Различие, конечно, не принципиально, но о нем важно помнить. Можно думать, что «опора и консоль» станет постоянным конкурентом гистограммы.— *Примеч. пер.*

ких значений, после чего остается нанести на график все неиспользованные точки, по которым очень хорошо видно поведение правого хвоста распределения.

Выражение остатков через «единичные нормальные отклонения»

Обычно предполагают, что $\varepsilon_i \sim N(0, \sigma^2)$, так что $\varepsilon_i/\sigma \sim N(0, 1)$. Тогда если модель правильна, то средний квадрат остатков

$$s^2 = \frac{\sum_{i=1}^{n} (e_i - \bar{e})^2}{(n-p)} = \frac{\sum_{i=1}^{n} e_i^2}{(n-p)}$$

служит оценкой величины σ^2.

(Примечания: 1. Если не учитывать ошибки округления, то $\bar{e} = \Sigma e_i/n = 0$. Величину e_i/s часто называют *единичным нормальным отклонением*, образованным из остатка e_i. Величины e_i/s, $i = 1, 2, \ldots, n$, можно исследовать с помощью общего графика и оценить, ошибочно ли предположение $\varepsilon_i/\sigma \sim N(0, 1)$. Поскольку с вероятностью 90 % значения случайной величины $N(0, 1)$ заключены в пределах (— 1,96, 1,96), мы можем ожидать, что примерно 95 % величин e_i/s будут в пределах (— 2, 2). Иногда удобно исследовать остатки этим способом, например при проверке выпадающих наблюдений («выбросов»), см. параграф 3.8. Если число $(n—p)$ мало, то при установлении 95 %-ных пределов вместо нормального распределения можно использовать $t\,(n—p)$-распределение.

2. Тот совет, что мы дали раньше, в некотором смысле не вполне точен. Ведь теоретически все случайные ошибки ε_i предполагаются независимыми и имеющими одну и ту же дисперсию σ^2, конкретные остатки отнюдь не независимы и, следовательно, не имеют одинаковых дисперсий. Как показано в параграфе 3.7, $V(e_i) = (1 - r_{ii})\sigma^2$, где r_{ii} — i-й диагональный элемент матрицы $\mathbf{R} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, который, таким образом, зависит от вида матрицы $\mathbf{X}$. Отсюда следует, что рассмотренный выше общий график может вводить в заблуждение, когда в $V(e_i)$ имеют место большие вариации, и что, вообще говоря, было бы лучше строить график не для e_i/s, а для $e_i/\{(1 - r_{ii})s^2\}^{1/2}$, если бы это было возможно. (Конечно, для обеспечения свободного выбора между тем и другим способом построения графиков пришлось бы написать новые программы.) Дело, однако, в том, что для подавляющего большинства наборов данных как в графиках для e_i, так и в графиках для e_i/s в общем проявляются свойства (хорошие или плохие), которые обнаруживаются и в более правильном графике для $e_i/\{(1 - r_{ii})s^2\}^{1/2}$. А это означает, что гораздо проще продолжать пользоваться простейшими вариантами. Таким образом, хотя мы и советуем работать с величинами $e_i/\{(1 - r_{ii})s^2\}^{1/2}$, если только это возможно, мы чувствуем, что в большинстве реальных задач совершенно доста-

точно[6] ограничиться величинами e_i и e_i/s; см.: Behnken D. W., Draper N. R. Residuals and their variance patterns.— Technometrics, 1972, 14, p. 101—111.)

Максимальный нормированный остаток

Для некоторых частных видов экспериментальных планов критические значения поддаются вычислению. Это дает возможность использовать их для проверки того, не «слишком ли велик» наибольший нормированный остаток. (Подробности и ссылки см. в статье: Stefansky W. Rejecting outliers in factorial designs. — Technometrics, 1972, 14, p. 469—479.)

3.2. ГРАФИК ВРЕМЕННО́Й ПОСЛЕДОВАТЕЛЬНОСТИ

Предположим, что остатки в примере, приведенном выше, получены именно в той последовательности по времени, как они приведены. Тогда график временной зависимости станет таким, как пока-

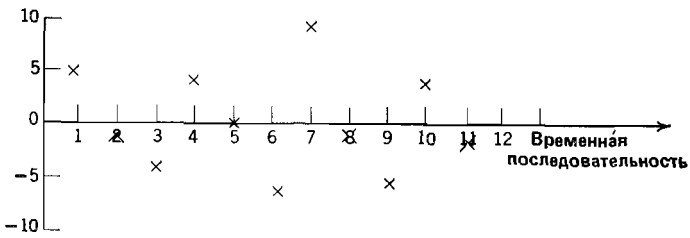


Рис. 3.2. Остатки, представленные во временно́й последовательности

зано на рис. 3.2. Если «отойти» от этой диаграммы, то изображение остатков сольется в горизонтальную «полосу» вроде той, которая представлена на рис. 3.3. Она показывает, что эффект времени не

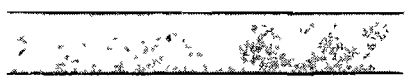


Рис. 3.3. Подходящий график остатков должен иметь такой общий вид

влияет на данные (или, если он и влияет, то это можно отнести за счет переменных X, подверженных еще и действию времени). Однако если полоса остатков напоминает по виду одну из диаграмм на рис. 3.4, то мы должны заключить, что не учтен эффект времени. Он может быть связан с такими обстоятельствами:

1) дисперсия не постоянна, а растет со временем, поэтому надо применить взвешенный метод наименьших квадратов;

[6] Существует и иная точка зрения. См., например: Бородюк В. П. Регрессионные модели с нестандартной ошибкой в задачах идентификации сложных объектов.— М.: Изд-во МЭИ, 1981, с. 17—34 (особенно с. 32—34). В этой работе выяснено, когда возникает то «меньшинство» реальных задач для которых советом авторов пользоваться опасно.— *Примеч. пер.*

2) в модель следовало бы включить линейный член от времени;

3) в модель должны быть включены линейный и квадратичный члены от времени.

Конечно, могут иметь место сочетания или вариации этих эффектов (например, в случае 2 возможен наклон в противоположную сторону и т. д.).

Мы говорили о том, что остатки на нашем графике не обнаруживают никакой явной тенденции, свидетельствующей о временной зависимости. Более детальное исследование этого графика показывает, что если это и верно для долгосрочной тенденции, то на краткосрочные тенденции наши утверждения распространить нельзя. Если сгруппировать остатки по три, а именно (1, 2, 3), (4, 5, 6), (7, 8, 9), (10, 11), то можно наблюдать тенденцию к уменьшению их величин в каждой группе, напоминающую некоторый вид регулярных («сезонных») изменений. Эти изменения должны быть отражены в пересматриваемой модели. Если предположить, например, что это линейные тренды, имеющие одинаковые наклоны, то можно добавить к модели член вида [7] $\delta\{(t-1) \bmod 3\}$, где δ — коэффициент регрессии, подлежащий оцениванию, а $(t-1) \bmod 3$ есть переменная, представляющая собой остаток от деления величины $(t-1)$ на 3. Значения этой новой переменной таковы:

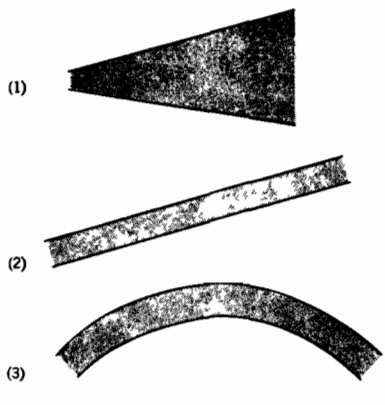


Рис. 3.4. Примеры ситуаций, характеризующих неудовлетворительное поведение остатков

Старая переменная:	t: 1 2 3 4 5 6 7 8 9 10 11
$(t-1) \bmod 3$:	0 1 2 0 1 2 0 1 2 0 1

(Другая альтернатива состоит в использовании в качестве переменной величины $\{(t-1) \bmod 3-1\}$. Это дает уровни -1, 0, $+1$ вместо 0, 1, 2, что иногда более удобно.)

Хотя работа с такими фиктивными переменными обеспечивает прослеживание за имеющимися вариациями, следует также предпринимать усилия для нахождения их причин. Легко представить себе образцы таких графиков остатков и фиктивных переменных, которые не стоит рассматривать без каких-либо основательных причин.

Мы предполагали, что на нашем графике (см. рис. 3.2) остатки разделяют равные промежутки времени. Если это не так и известны действительные отрезки времени, то, конечно, при построении графиков следует воспользоваться этой информацией.

---

[7] Использованная здесь операция в теории чисел называется сравнением. Говорят, что a сравнимо с b по модулю m, если разность $a-b$ делится на m. Обозначение mod заменяет слово «модуль». Как работает эта операция, хорошо видно из обсуждаемого примера.— *Примеч. пер.*

Временные графики остатков можно еще обработать в соответствии с предложениями У. Кливленда и Б. Кляйнера (см.: Cleveland W. S., Kleiner B. A graphical technique for enhancing scatterplots with proving statistics.— Technometrics, 1975, **17**, p. 447—454).

На график наносятся три кривые текущих значений следующих статистик: 1) *срединное* (*усеченное*) *среднее* (среднее арифметическое значение всех наблюдений, лежащих между квартилями, для данных, полученных к рассматриваемому моменту времени); 2) *нижнее* полусрединное среднее (срединное среднее всех наблюдений, лежащих *ниже* медианы всех данных, полученных к рассматриваемому моменту времени); 3) верхнее полусрединное среднее (все то же самое, что и в предыдущем случае, только *выше* медианы). «Эти три статистики представляют собой сводки оценок центра рассеяния, размаха и асимметрии наших данных» (см. с. 449 указанной выше работы).

3.3. ГРАФИК ЗАВИСИМОСТИ ОСТАТКОВ ОТ $\hat{Y}_i$

Допустим, что $\hat{Y}_i$, которые соответствуют приведенным выше величинам e_i, примут значения 44, 8, 10, 62, 22, 48, 56, 30, 24, 16, 34. Тогда мы будем иметь график, показанный на рис. 3.5. «Горизонтальная полоса» не указывает на какую-либо ненормальность, и МНК-анализ, по-видимому, вполне оправдан.

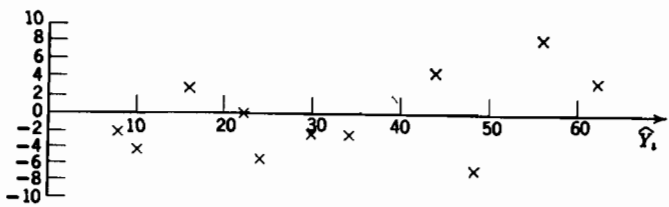


Рис. 3.5. Остатки в зависимости от предсказанных значений

Ненормальность проявилась бы на графиках в формах, подобных тем, что показаны на рис. 3.4 под номерами (1), (2) и (3). Такие графики могли бы сказать о следующем:

1) дисперсия не постоянна, как это предполагалось; надо прибегнуть к взвешенному методу наименьших квадратов или произвести преобразование наблюдений Y_i до регрессионного анализа;

2) ошибочен сам анализ; отклонения от данного уравнения регрессии носят систематический характер; отрицательные остатки соответствуют низким значениям $\hat{Y}$, положительные остатки — высоким значениям $\hat{Y}$. Этот результат может получиться еще и от того, что в модели ошибочно пропущен член β_0;

3) модель неадекватна, надо вводить дополнительные члены (например, квадратичные и взаимодействия) или сначала провести преобразование наблюдений Y_i, а затем уже анализ.

Вопрос. Почему для обычной линейной модели мы строим график зависимости остатков $e_i = Y_i - \hat{Y}_i$ от $\hat{Y}_i$, а не от Y_i?

Ответ. Потому, что остатки e_i обычно коррелируют со значениями Y_i и, как правило, не коррелируют со значениями $\hat{Y}_i$. Один из возможных способов убедиться в этом заключается в построении графиков остатков e_i на ординате (а) от значений Y_i и (б) от значений $\hat{Y}_i$, а затем в нахождении методом наименьших квадратов углового коэффициента прямой, проходящей через эти точки. Для (а) он должен быть равен $1-R^2$, а для (б) — нулю. Иначе говоря, мы можем просто найти (а) r_{eY} и (б) $r_{e\hat{Y}}$ следующим образом:

(а) $\Sigma (e_i - \bar{e})(Y_i - \overline{Y}) = \Sigma e_i (Y_i - \overline{Y})$ (поскольку $\bar{e} = 0$, если член β_0 включен в модель),

$= \Sigma e_i Y_i$ $(\bar{e} = 0)$,

$= \mathbf{e}'\mathbf{Y}$,

$= \mathbf{e}'\mathbf{e}$ (поскольку $\mathbf{e}'\mathbf{e} = \mathbf{Y}'(\mathbf{I}-\mathbf{R})'(\mathbf{I}-\mathbf{R})\mathbf{Y}$, $= \mathbf{Y}'(\mathbf{I}-\mathbf{R})\mathbf{Y}$, $= \mathbf{Y}'\mathbf{e}$, $= \mathbf{e}'\mathbf{Y}$, где $\mathbf{R} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$),

$= $ Остаточная SS

$$\Sigma (e_i - \bar{e})^2 = \Sigma e_i^2 = \mathbf{e}'\mathbf{e},$$

$$\Sigma (Y_i - \overline{Y})^2 = \text{общая скорректированная SS},$$

$$r_{eY} = \frac{\mathbf{e}'\mathbf{e}}{\{(\mathbf{e}'\mathbf{e})\Sigma (Y_i - \overline{Y})^2\}^{1/2}} = \left\{\frac{\text{Остаточная SS}}{\text{общая скорректированная SS}}\right\}^{1/2} = \{1 - R^2\}^{1/2}.$$

Здесь нуль может получиться, *только* если модель — само совершенство. Во всех других случаях график зависимости остатков от Y_i будет иметь угол наклона, равный $1-R^2$.

(б) $\Sigma (e_i - \bar{e})(\hat{Y}_i - \overline{\hat{Y}}) = \Sigma e_i \hat{Y}_i$ (по аналогии со сверткой, приведенной выше),

$\mathbf{e}'\hat{\mathbf{Y}}$,

$= \mathbf{Y}'(\mathbf{I}-\mathbf{R})'\mathbf{R}\mathbf{Y}$ (поскольку $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{R}\mathbf{Y}$),

$= \mathbf{Y}'(\mathbf{R}-\mathbf{R}^2)\mathbf{Y} = 0$, таким образом, $r_{e\hat{Y}} = 0$.

3.4. ГРАФИК ЗАВИСИМОСТИ ОСТАТКОВ ОТ ПРЕДИКТОРНЫХ ПЕРЕМЕННЫХ X_{ji}, $i = 1, 2, \ldots, n$

Эти графики имеют такую же форму, как и графики зависимости остатков от $\hat{Y}_i$, за исключением того, что мы используем (вместо величин, соответствующих $\hat{Y}_i$) величины, отвечающие переменным X_{ji}, а именно, $X_{j1}, X_{j2}, \ldots, X_{jn}$. Общее представление остатков в виде горизонтальной полосы снова вполне удовлетворительно. Аномалии, проиллюстрированные рис. 3.4, указывают на следующее:

1) дисперсия не постоянна; нужно привлечь взвешенный метод наименьших квадратов или предварительно преобразовать наблюдения Y;

2) ошибка в вычислениях; линейный эффект X_j исключен неверно;

3) надо ввести в модель дополнительные члены от X_j, например квадратичные, или произвести преобразование Y-ов.

В малых регрессионных задачах, которые включают только две или три переменные X, зависимость остатков от переменных можно изобразить графически в двух- или трехмерном пространстве. В таком случае можно изобразить и точки, в которых были выполнены наблюдения, и точки, отвечающие остаткам. Если подобные построения возможны, то зачастую они дают хорошее визуальное представление о ситуации. Если же число переменных больше трех, то можно воспользоваться аналогичными графиками для подмножеств переменных, что иногда целесообразно [8].

Двумерный график представлен на рис. 7.6 (см. кн. 2).

3.5. ДРУГИЕ ГРАФИКИ ОСТАТКОВ

Специальные знания в изучаемой области нередко позволяют нам предложить и другие типы графиков остатков, подлежащих исследованию. Предположим, например, что одиннадцать наблюдений, на основании которых получены указанные выше одиннадцать остатков, проводились на трех машинах — A, B и C и что сгруппированные остатки выглядят следующим образом:

A: — 2, — 4, — 6;
B: — 2, — 5, — 2;
C: 5, 4, 0, 9, 3.

Рис. 3.6. График остатков, указывающий из блоковые эффекты, не связанные с подобранной моделью

На рис. 3.6 представлены графики остатков для разных машин. Они позволяют предположить, что имеется заметное различие в величине отклика Y для машины C по сравнению с машинами A и B. Эти различия можно было бы учесть путем введения в модель фиктивных переменных (см. гл. 5, рис. 5.4).

Возможен еще один пример «других графиков остатков» — когда в рассмотрение вводится новая переменная. Пусть мы подозреваем, что окружающая температура влияет на содержимое большого сосуда.

[8] Методы наглядного представления остатков и их интерпретации хорошо развиты в работах Тьюки, см., например: Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия/Пер. с англ. Под ред. Ю. П. Адлера.— М.: Финансы и статистика, 1982, вып. 1, с. 189—192; вып. 2, гл. 16, с. 144—182. Последняя глава этой книги представляет собой оригинальную трактовку анализа остатков, которую можно рассматривать как развитие результатов, полученных в данной главе настоящей книги.— *Примеч. пер.*

Хотя температура внутри сосуда регистрировалась в определенной защищенной точке, на температуру снаружи, вероятно, влияло состояние окружающего воздуха. Если окружающая температура регистрировалась в течение того периода, когда снимались экспериментальные данные, то можно построить график зависимости остатка от наблюдаемых значений температуры, чтобы видеть, имеется ли какая-либо зависимость откликов от окружающей температуры. Если зависимость обнаруживают, то можно добавить к модели новые члены соответствующего вида, чтобы ее учесть.

Эти два примера «других графиков остатков» могут оказаться полезными для практики. Вообще остатки можно представить графически любым рациональным способом, который придет в голову экспериментатору или статистику и который основан на специальных сведениях об изучаемой задаче. Однако те графики, что приведены в параграфах 3.1—3.4, — основные, и для полноты анализа их стоит вычерчивать всегда.

3.5. СТАТИСТИКИ ДЛЯ ИССЛЕДОВАНИЯ ОСТАТКОВ

На графиках, рекомендованных в предыдущих параграфах, основаны визуальные методы проверки некоторых основных предположений регрессионного анализа. Был предложен ряд статистик, дающих количественную меру для оценки некоторых из описанных выше расхождений. Рассмотрим их совсем кратко. Мы не акцентируем на них внимание умышленно, поскольку в практических регрессионных задачах подробное исследование соответствующих графиков остатков обычно гораздо более информативно, и графики почти наверняка будут обнаруживать любые достаточно существенные нарушения предположений, требующие корректировки.

Вернемся к графику зависимости e_i от $\hat{Y}_i$, описанному в параграфе 3.3. Обсуждались три особых вида расхождений, проиллюстрированные на рис. 3.4. Мы можем измерить каждый из этих дефектов с помощью соответствующей статистики. Введем величину

$$T_{pq} = \sum_{i=1}^{n} e_i^p \hat{Y}_i^q. \tag{3.6.1}$$

Тогда

1. $T_{21} = \sum_{i=1}^{n} e_i^2 \hat{Y}_i$ — мера дефекта, представленного на рис. 3.4 (1) из параграфа 3.2 (но в смысле параграфа 3.3). Она связана с более общей статистикой, приведенной Ф. Энскамби в работе, посвященной исследованию остатков (см.: Anscombe F. J. Examination of residuals. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1961, 1, p. 1—36).

2. $T_{11} = \sum_{i=1}^{n} e_i \hat{Y}_i$. Эта величина должна быть всегда равна нулю. Она представляет собой меру дефекта, отраженного на рис. 3.4 (2)

из параграфа 3.2 (но в смысле параграфа 3.3). При желании вычисление этой статистики можно сделать стандартным тестом.

3. $T_{12}=\sum_{i=1}^{n} e_i \hat{Y}_i^2$ представляет собой меру дефекта, изображенного на рис. 3.4 (3) из параграфа 3.2 (но в смысле параграфа 3.3). Она связана со статистикой «одна степень свободы для неаддитивности», предложенной Дж. Тьюки (см. также упражнение 10 из гл. 2).

Существуют также и другие типы статистик. Читатели, желающие ознакомиться с этим материалом более подробно, могут обратиться к указанной выше статье Ф. Энскамби, а также к статье: Anscombe F. J., Tukey J. W. The examination and analysis of residuals.— Technometrics, 1963, **5**, p. 141—160.

Еще можно пользоваться преобразованными остатками. О том, как работать с некоторыми преобразованными остатками, предложенными Г. Тейлом, можно прочесть [9] в: Farebrother R. W. (Algol 60) Algorithm AS 104, BLUS residuals.— Applied Statistics, 1976, **25**, p. 317—322, там же содержатся ссылки на соответствующую литературу. О статистиках, вычисляемых по остаткам, см.: Cox D. R., Snell E. J. On test statistics calculated from residuals.— Biometrika, 1971, **58**, p. 589—594. Замечание о рекурсивных [10] остатках можно найти в: Farebrother R. W., Applied Statistics, 1976, **25**, p. 323—324.

3.7. КОРРЕЛЯЦИЯ МЕЖДУ ОСТАТКАМИ

В общей регрессионной ситуации, когда по n наблюдениям оцениваются p параметров, n остатков связаны лишь с $n-p$ степенями свободы. Следовательно, остатки не могут быть независимыми и между ними существует корреляция. Если постулируется модель $\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$ и если матрица $\mathbf{X}'\mathbf{X}$ неособенная, то можно записать остатки в матричной форме:

$$\mathbf{e}=\mathbf{Y}-\hat{\mathbf{Y}}=\mathbf{Y}-\mathbf{X}\mathbf{b}=\mathbf{Y}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}=(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y}=$$
$$=(\mathbf{I}-\mathbf{R})\mathbf{Y},$$

---

[9] Предложенные Г. Тейлом преобразования остатков получили дальнейшее развитие в оригниальных работах В. П. Бородюка и В. Е. Кузнецова, которые предложили прибавлять к вектору остатков некий случайный вектор с нулевым математическим ожиданием и отиосительно малыми дисперсиями элементов. Тогда за счет некоторого смещения оценок можно избавиться от коррелированности остатков и применять для проверки гипотез обычные формулы. (См.: Бородюк В. П. Принцип «малости ошибок» в задачах синтеза алгоритмов идентификации и управления.— В кн.: Опыт создания и внедрения АСУ ТП, ч. 2/Ред. В. В. Солодовникова.— Фрунзе: Изд-во ИЛИМ, 1979, с. 169—180; Бородюк В. П., Кузнецов В. Е. Анализ остатков в активном эксперименте. Тезисы докладов IV Всесоюзной конференции по планированию и автоматизации эксперимента в иаучных исследованиях. Ч. 1.— М.: Изд-во МЭИ, 1973, с. 64; Кузиецов В. Е. Построеиие и анализ регрессиоиной модели промышленных объектов. Автореф. канд. дис.— М.: Изд-во МЭИ, 1974.— *Примеч. пер.*

[10] Рекурсивными называются остатки, вычисляемые с помощью некоторого алгоритма.— *Примеч. пер.*

где $\mathbf{R} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. (Матрица $\mathbf{R}$ — важнейшая матрица, между прочим, многократно встречающаяся в регрессионных исследованиях.) Поскольку $E(\mathbf{Y}) = \mathbf{X}\beta$,

$$\mathbf{e} - E(\mathbf{e}) = (\mathbf{I} - \mathbf{R})(\mathbf{Y} - \mathbf{X}\beta) = (\mathbf{I} - \mathbf{R})\varepsilon,$$

и матрица дисперсий-ковариаций вектора $\mathbf{e}$ определяется выражением

$$\mathbf{V}(\mathbf{e}) = E\{[\mathbf{e} - E(\mathbf{e})][\mathbf{e} - E(\mathbf{e})]'\} = (\mathbf{I} - \mathbf{R})E(\varepsilon\varepsilon')(\mathbf{I} - \mathbf{R})'.$$

Теперь $E(\varepsilon\varepsilon') = \mathbf{V}(\varepsilon) = \mathbf{I}\sigma^2$, если $E(\varepsilon) = 0$, как мы обычно предполагаем, и если используется невзвешенный (обычный) метод наименьших квадратов. Кроме того,

$$(\mathbf{I} - \mathbf{R})' = (\mathbf{I}' - \mathbf{R}') = \mathbf{I} - [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']' = \mathbf{I} - [\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{I} - \mathbf{R}.$$

Следовательно, матрица $\mathbf{I} - \mathbf{R}$ симметрична и

$$\mathbf{V}(\mathbf{e}) = (\mathbf{I} - \mathbf{R})\mathbf{I}\sigma^2(\mathbf{I} - \mathbf{R})' = (\mathbf{I} - \mathbf{R})(\mathbf{I} - \mathbf{R})\sigma^2 =$$
$$= (\mathbf{I} - \mathbf{R} - \mathbf{R} + \mathbf{R}\mathbf{R})\sigma^2 = (\mathbf{I} - \mathbf{R})\sigma^2,$$

так как $\mathbf{R}\mathbf{R} = \mathbf{R}^2 = \mathbf{R}$, что легко проверить *. Значит, дисперсия $V(e_i)$ представляет собой i-й диагональный элемент, а covar $(e_i e_j)$ задается (i, j)-элементом матрицы $(\mathbf{I} - \mathbf{R})\sigma^2$. Коэффициент корреляции между e_i и e_j равен:

$$\rho_{ij} = \frac{\text{covar}(e_i, e_j)}{\{V(e_i) \cdot V(e_j)\}^{1/2}}.$$

Таким образом, значения этих коэффициентов полностью определяются элементами матрицы $\mathbf{X}$, поскольку σ^2 исключается.

Возникает вопрос, не обесценивают ли эти корреляции графики остатков? Замечания по этому поводу см. в работе: Анскомбе F. J., Тьюки J. W. The examination and analysis of residuals.— Technometrics, 1963, 5, p. 144. При обсуждении дисперсионного анализа двусторонних классификаций (где имеется несколько ограничений, накладываемых на остатки) Ф. Энскамби и Дж. Тьюки отмечают, что хотя корреляции и ограничения влияют на функции распределения от остатков, « ... соответствующим влиянием их на графические процедуры можно обычно пренебречь ... Это связано, главным образом, со способом, лежащим в основе получения графика остатков, но отчасти и с отсутствием точно определенных уровней значимости. (Это верно также для большинства других сбалансированных планов.)» В заключение Ф. Энскамби и Дж. Тьюки говорят, что в таблице сопряженности [11] (таблице с двумя входами) при четырех или более строках и четырех или более столбцах « ... влияние корреляции на графические процедуры обычно пренебрежимо мало ...». По-видимому, в общей регрессионной ситуации, когда строятся графики ос-

* Если $\mathbf{R}\mathbf{R} = \mathbf{R}$, то $\mathbf{R}$ называют идемпотентной матрицей. Как $\mathbf{R}$, так и $\mathbf{I} - \mathbf{R}$ — симметричные и идемпотентные матрицы.

[11] Анализ остатков в таблицах сопряженности обсуждается в уже упомиавшихся работах Дж. Тьюки. Еще см.: Аптон Г. Анализ таблиц сопряженности. Пер. с англ.— М.: Финансы и статистика, 1982, 143 с., особенно с. 86—88.— *Примеч. пер.*

татков, нет необходимости учитывать эффект корреляции между остатками, за исключением того случая, когда отношение числа степеней свободы для остатков к общему числу остатков, $(n—p)/n$, весьма мало.

3.8. ВЫБРОСЫ

Выбросом среди остатков называется остаток, который по абсолютной величине значительно превосходит остальные и отличается от среднего по остаткам на три, четыре или даже более стандартных отклонений. Выброс означает определенную особенность и показывает экспериментальную точку, которая совсем не типична по отношению к остальным данным. Отсюда следует, что выброс должен подвергаться особо тщательному исследованию с целью выяснения причин его возникновения.

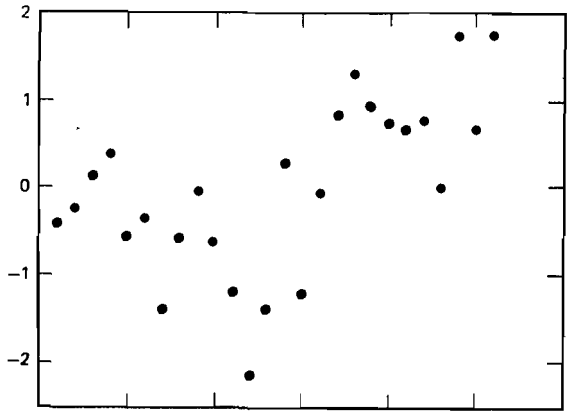
Были предложены правила отбрасывания выбросов (т. е. правила, согласно которым исключают соответствующее наблюдение (или наблюдения) и затем снова анализируют данные без этих наблюдений). Автоматическое исключение выбросов — это отнюдь не всегда наиболее целесообразная процедура. Иногда выброс дает такую информацию, которую другие данные не могут дать благодаря тому, что он связан с необычной комбинацией условий, являющейся жизненно важной. В этом случае требуется, скорее, дальнейшее углубление исследования, а не механическое отбрасывание выбросов (см., например, параграф 4.1). Общее правило таково: выбросы должны исключаться сразу, если только выяснится, что они вызваны такими причинами, как ошибки в регистрации результатов наблюдений или в настройке аппаратуры. В противном случае требуется тщательное исследование [12] (см. Anscombe F. J. Rejection of outliers. — Technometrics, 1960, 2, p. 123—147).

3.9. СЕРИАЛЬНАЯ КОРРЕЛЯЦИЯ ОСТАТКОВ

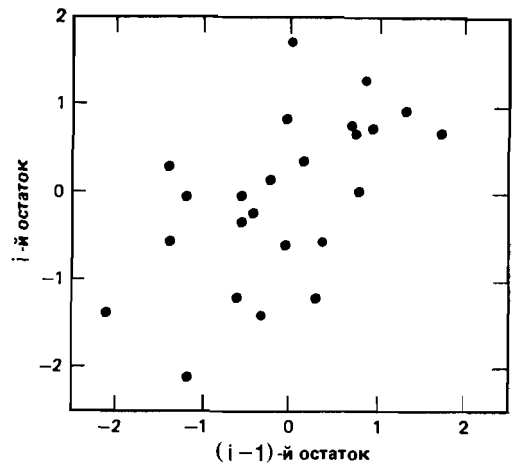
В регрессионных исследованиях мы, как правило, предполагаем, что ошибки наблюдений попарно некоррелированы. Если бы это предположение оказалось фактически нарушенным, то можно было бы ожидать, что график зависимости остатков от времени или какой-нибудь другой, чувствительный к порядку график, который мы бы построили с учетом наших обстоятельств, поможет установить этот факт. Конечно, ошибки могут коррелировать самыми разнообразными спо-

[12] Одна из последних сводок по выбросам: Barnett V., Lewis T. Outliers in Statistical Data. 2-nd. ed.— New York: J. Wiley, 1984. На русском языке см., например: Кендалл М., Стьюарт А. Статистические выводы и связи/Пер. с англ. Под ред. А. Н. Колмогорова.— М.: Наука, 1973, с. 707—712; Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. Методы обработки данных/Пер. с англ. Под ред. Э. К. Лецкого.— М.: Мир, 1980, с. 285—297; Закс Л. Статистическое оценивание/Пер с нем. Под ред. Ю. П. Адлера и В. Г. Горского.— М.: Статистика, 1976, с. 256—261; Химмельблау Д. Анализ процессов статистическими методами/Пер. с англ. Под ред. В. Г. Горского.— М.: Мир, 1973, с. 177—181.— *Примеч. пер.*

собами. Но, как правило, все сводится к тому, что они могут проявить сериальную корреляцию, т. е. такую корреляцию, где зависимость между ошибками, отстоящими друг от друга на s шагов, всегда остается одинаковой. Мы примем для такой корреляции обозначение ρ_s, где $s = 1, 2, \ldots$.



а



б

Рис. 3.7. (а) Серия остатков, проявляющих локально положительную корреляцию. (б) График сериальной корреляции с единичным сдвигом для тех же данных

В тех частных случаях, когда остатки проявляют локально положительную сериальную корреляцию, соседние остатки во временной последовательности становятся более похожими друг на друга, чем на другие остатки. Их временной график будет в общем похожим на рис. 3.7 (а) со взлетами и падениями, но с соседними точками, более

близкими между собой, чем с остальными. Корреляция между остатками, которые разделяет один шаг (или два, или три) называется сериальной корреляцией с единичным сдвигом (или со сдвигом на два,

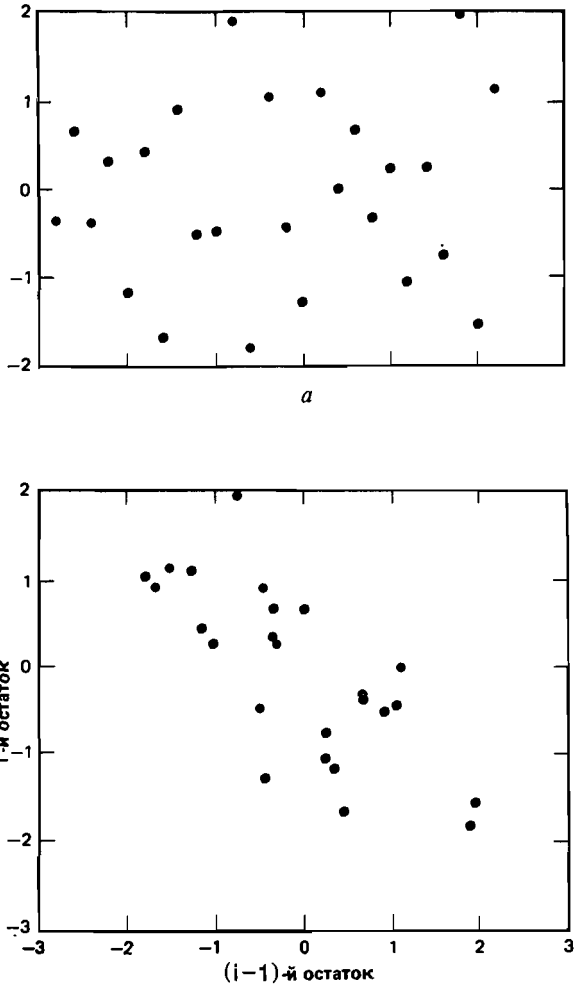


Рис. 3.8. (а) Серия остатков, проявляющих локально отрицательную сериальную корреляцию, обусловленную, быть может, транспортным запаздыванием. (б) График сериальной корреляции с единичным сдвигом для тех же данных

три и т. д. шага). Эмпирически сериальную корреляцию с единичным сдвигом можно изучить на графике, где каждый остаток, кроме первого, сдвинут на один шаг назад. Положительная с единичным сдви-

гом сериальная корреляция, имеющаяся в данных на рис. 3.7 (а), сама собой проявляется на рис. 3.7 (б) в тенденции «из левого нижнего в правый верхний угол». Чтобы увидеть корреляции для больших сдвигов, мы можем точно так же построить графики остатков со сдвигом на два шага, на три шага и т. д.

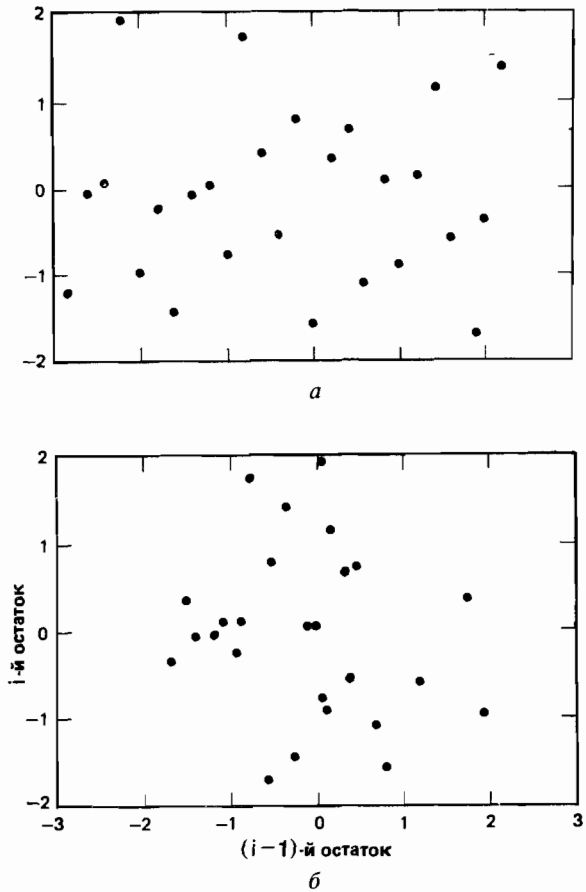


Рис. 3.9. (а) Серия некоррелированных остатков. (б) График сериальной корреляции с единичным сдвигом для тех же данных

Могут встретиться и отрицательные сериальные корреляции между соседними остатками. Одна из причин этого явления, которое встречается в периодических или циклических процессах, известна как транспортное запаздывание. Вот как это может происходить. Положим, что некоторая конкретная порция продукта, введенная в процесс, переработана не полностью, поскольку часть продукта осталась в трубопроводах и насосах, обеспечивающих данный реактор. Тогда зафиксированный выход продукта для этой порции ока-

зался бы необычно низким. Зато следующая порция способствовала бы переработке этого запасенного материала, давая тем самым необычно высокий выход. В итоге структура остатков может оказаться похожей на ту, что приведена на рис. 3.8 (а), где положительные значения обычно следуют за отрицательными, и наоборот. Существование в этих данных отрицательной сериальной корреляции с единичным сдвигом видно из структуры рис. 3.8 (б), где точки группируются вдоль направления «с правого нижнего в левый верхний угол».

На рис. 3.9 (а) показана некоторая случайная серия остатков, а на рис. 3.9 (б) — соответствующий график сериальной корреляции с единичным сдвигом, на котором вообще не проявляется никакой тенденции к тренду.

Исследование структур сериальных корреляций — один из методов анализа временных рядов. Такой анализ коррелированных данных — часто дело стоящее. Заинтересованный читатель может посмотреть например, такие книги[13]: Box G. E. P., Jenkins G. M. Time Series Analysis, Forecasting and Control.— Holden-Day, San Francisco: 1970; Jenkins G. M., Watts D. G. Spectral Analysis and its Applications.— San Fracisco: Holden-Day, 1968.

Использование взвешенного метода наименьших квадратов для сериально коррелированных данных

Основной метод анализа, которым можно воспользоваться при наличии сериальной корреляции остатков, — это взвешенный метод наименьших квадратов, описанный в параграфе 2.11. А главная трудность в его применении — отыскание матрицы $\mathbf{V}$ из уравнения (2.11.2). Когда наблюдения упорядочены во времени, элементом V_{ij} матрицы $\mathbf{V}$ будет служить ρ_l, где $l = |i-j|$, причем $\rho_0 = 1$. Для оценки ρ_l нам надо сдвигать наблюдения на l шагов и вычислять коэффициент корреляции по формуле (1.6.5), отбрасывая те наблюдения, для которых при сдвиге не найдется пары. Полученные таким образом оценки подставляются в матрицу $\mathbf{V}$, что дает матрицу $\hat{\mathbf{V}}$, которая в свою очередь используется в таких уравнениях, как (2.11.10) и (2.11.11). Для анализа остатков *такой* взвешенной модели нам нужны оценки матрицы $\mathbf{f} = \mathbf{P}^{-1}\boldsymbol{\varepsilon}$; см. уравнения (2.11.3) и (2.11.4). Вот эти оценки:

$$\hat{\mathbf{f}} = \hat{\mathbf{P}}^{-1}(\mathbf{Y} - \hat{\mathbf{Y}}), \qquad (3.9.1)$$

[13] Обе эти книги переведены на русский язык: Бокс Дж., Дженкинс Г. Анализ временных рядов: Прогноз и управление. Пер. с англ. Под ред. В. Ф. Писаренко.— М.: Мир, 1974, вып. 1, 408 с.; Дженкинс Г., Уаттс Д. Спектральный анализ и его приложения. Пер. с англ.— М.: Мир, 1971, вып. 1, 316 с.; 1972, вып. 2, 287с. Интерес к анализу коррелированных рядов характерен для эконометрии. См., например: Маленво Э. Статистические методы эконометрии/Пер. с фр. Под ред. Б. Н. Михалевского и И. Ш. Амирова.— М.: Статистика, вып. 2, ч. 4, с. 8—180; Джонстон Дж. Экономические методы. Пер. с англ.— М.: Статистика, 1980, гл. 8, с. 242—265; Дрэймз Ф. Распределениые лаги/Пер. с англ. Под ред. Э. Б. Ершова.—М.: Финансы и статистика, 1982.— 383 с.; Песаран М., Слейтер Л. Динамическая регрессия: теория и алгоритмы/Пер. с англ. Под ред. Э. Б. Ершова.— М.: Финансы и статистика, 1984, 310 с. — *Примеч. пер.*

где

$$\hat{\mathbf{P}}'\hat{\mathbf{P}} = \hat{\mathbf{V}}, \tag{3.9.2}$$

а $\hat{\mathbf{Y}}$ — предсказание взвешенного метода наименьших квадратов, т. е.

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}. \tag{3.9.3}$$

Иными словами, элементы матрицы $\hat{\mathbf{f}}$ есть

$$\hat{\mathbf{f}} = \hat{\mathbf{P}}^{-1}\{\mathbf{I} - \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\}\mathbf{Y} \tag{3.9.4}$$

и при исследовании годятся методы, описанные в этой главе.

(Примечание. Фактически это та же формула для $\mathbf{e} = \{\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\}\mathbf{Y}$ из обычного невзвешенного метода наименьших квадратов, но только с подстановками $\hat{\mathbf{P}}^{-1}\mathbf{X}$ и $\hat{\mathbf{P}}^{-1}\mathbf{Y}$ вместо $\mathbf{X}$ и $\mathbf{Y}$ соответственно.)

Две проверки для сериальной корреляции

Существуют два хорошо известных способа проверки того, есть ли в остатках признаки сериальной корреляции. Это критерий серий и критерий Дарбина—Уотсона. К их описанию мы теперь и переходим.

3.10. ИССЛЕДОВАНИЕ СЕРИЙ НА ГРАФИКАХ ВРЕМЕННÓЙ ПОСЛЕДОВАТЕЛЬНОСТИ ОСТАТКОВ

При известной временной последовательности множеств остатков иногда удается наблюдать необычные скопления положительных или отрицательных остатков. Возьмем для примера крайний случай. Если временная последовательность из тридцати остатков содержит сначала десять отрицательных, а затем двадцать положительных значений, то мы вправе ожидать, что есть какая-то не принимавшаяся во внимание переменная, которая изменила уровень между десятым и одиннадцатым опытами. Значит, мы можем исследовать причины, вызывающие подобное поведение. Когда встречается такая последовательность знаков, полезно иметь метод, позволяющий выносить решение об отклонении чередования знаков в последовательности опытов от случайного, т. е. выносить решение о «ненормальности» последовательности.

Допустим, мы имеем следующую последовательность знаков:

$$+\,+\,-\,+\,-\,-\,-\,-\,+\,+\,-\,+\,+\,+\,.$$

Это знаки остатков во временной последовательности (которыми мы воспользуемся). Знаки «плюс» и «минус» могут означать также «мужской» и «женский», «головы» и «хвосты», «лучше» и «хуже», «обработка A» и «обработка B» или два уровня любой другой дихотомической классификации. Предположим, что здесь всего n знаков, из которых n_1 знаков плюс и n_2 знаков минус, а u — число серий. В при-

веденном примере $n_1 = 8$, $n_2 = 6$ и $u = 7$ серий, показанных круглыми скобками:

$$(+\,+)(-)(+)(-\,-\,-\,-)(+\,+)(-)(+\,+\,+).$$

Мы вправе поставить вопрос, является ли данное расположение знаков «выделяющимся» расположением? Если, например, из шести знаков два знака плюс, а остальные минус, то возможны следующие расположения:

Расположение	Число серий
+ + — — — —	2
+ — + — — —	4
+ — — + — —	4
+ — — — + —	4
+ — — — — +	3
— + + — — —	3
— + — + — —	5
— + — — + —	5
— + — — — +	4
— — + + — —	3
— — + — + —	5
— — + — — +	4
— — — + + —	3
— — — + — +	4
— — — — + +	2

Распределение этих серий таково:

$u =$	2	3	4	5	
Частота =	2	4	6	3	(Всего = 15)
Накопленная вероятность =	0,133	0,400	0,800	1,000	

Следовательно, пять серий могут встречаться в $^3/_{15}$, или 20 % возможных случаев, т. е. с вероятностью 0,2. С другой стороны, две серии могут иметь место в $^2/_{15}$, или 13,3 % возможных случаев, или с вероятностью 0,133. При исследовании остатков обычно нас интересует только случай с малым числом серий, поэтому мы не тревожимся, может ли быть их «слишком много». Если мы имеем только $u = 2$ в наборе из шести остатков, из которых два остатка положительны, то вероятность существования такого случая составляет 0,133. Для любой данной последовательности знаков можно найти вероятность того, что наблюдаемое значение u (или меньшее) может иметь место.

(Пример. При $n_1 = 2$, $n_2 = 4$ вероятность $(u \leqslant 3) = (2 + 4)/15 = 0{,}4$, и нет ничего необычного в том, что происходит в 40 % случаев.)

На основе таких уровней вероятности можно решать, правомерно ли полагать, что мы имеем дело со случайным расположением знаков. Мы можем, например, сравнить вероятность с заранее заданной величиной, скажем $\alpha = 0{,}05$, и отвергнуть предположение о случайном характере расположения, если вероятность ($u \leqslant u$ наблюдаемое) $\leqslant$ $\leqslant 0{,}05$. В табл. 3.1 приведены кумулятивные распределения вероят-

Таблица 3.1. **Распределение накопленной вероятности**

(n_1, n_2) / u	2	3	4	5	6	7	8	9	10
(2,3)	0,200	0,500	0,900	1,000					
(2,4)	0,133	0,400	0,800	1,000					
(2,5)	0,095	0,333	0,714	1,000					
(2,6)	0,071	0,286	0,643	1,000					
(2,7)	0,056	0,250	0,583	1,000					
(2,8)	0,044	0,222	0,533	1,000					
(2,9)	0,036	0,200	0,491	1,000					
(2,10)	0,030	0,182	0,455	1,000					
(3,3)	0,100	0,300	0,700	0,900	1,000				
(3,4)	0,057	0,200	0,543	0,800	0,971	1,000			
(3,5)	0,036	0,143	0,429	0,714	0,929	1,000			
(3,6)	0,024	0,107	0,345	0,643	0,881	1,000			
(3,7)	0,017	0,083	0,283	0,583	0,833	1,000			
(3,8)	0,012	0,067	0,236	0,533	0,788	1,000			
(3,9)	0,009	0,055	0,200	0,491	0,745	1,000			
(3,10)	0,007	0,045	0,171	0,455	0,706	1,000			
(4,4)	0,029	0,114	0,371	0,629	0,886	0,971	1,000		
(4,5)	0,016	0,071	0,262	0,500	0,786	0,929	0,992	1,000	
(4,6)	0,010	0,048	0,190	0,405	0,690	0,881	0,976	1,000	
(4,7)	0,006	0,033	0,142	0,333	0,606	0,833	0,954	1,000	
(4,8)	0,004	0,024	0,109	0,279	0,533	0,788	0,929	1,000	
(4,9)	0,003	0,018	0,085	0,236	0,471	0,745	0,902	1,000	
(4,10)	0,002	0,014	0,068	0,203	0,419	0,706	0,874	1,000	
(5,5)	0,008	0,040	0,167	0,357	0,643	0,833	0,960	0,992	1,000
(5,6)	0,004	0,024	0,110	0,262	0,522	0,738	0,911	0,976	0,998
(5,7)	0,003	0,015	0,076	0,197	0,424	0,652	0,854	0,955	0,992
(5,8)	0,002	0,010	0,054	0,152	0,347	0,576	0,793	0,929	0,984
(5,9)	0,001	0,007	0,039	0,119	0,287	0,510	0,734	0,902	0,972
(5,10)	0,001	0,005	0,029	0,095	0,239	0,455	0,678	0,874	0,958
(6,6)	0,002	0,013	0,067	0,175	0,392	0,608	0,825	0,933	0,987
(6,7)	0,001	0,008	0,043	0,121	0,296	0,500	0,733	0,879	0,966
(6,8)	0,001	0,005	0,028	0,086	0,226	0,413	0,646	0,821	0,937
(6,9)	0,000	0,003	0,019	0,063	0,175	0,343	0,566	0,762	0,902
(6,10)	0,000	0,002	0,013	0,047	0,137	0,288	0,497	0,706	0,864
(7,7)	0,001	0,004	0,025	0,078	0,209	0,383	0,617	0,791	0,922
(7,8)	0,000	0,002	0,015	0,051	0,149	0,296	0,514	0,704	0,867
(7,9)	0,000	0,001	0,010	0,035	0,108	0,231	0,427	0,622	0,806
(7,10)	0,000	0,001	0,006	0,024	0,080	0,182	0,355	0,549	0,743
(8,8)	0,000	0,001	0,009	0,032	0,100	0,214	0,405	0,595	0,786
(8,9)	0,000	0,001	0,005	0,020	0,069	0,157	0,319	0,500	0,702
(8,10)	0,000	0,000	0,003	0,013	0,048	0,117	0,251	0,419	0,621
(9,9)	0,000	0,000	0,003	0,012	0,044	0,109	0,238	0,399	0,601
(9,10)	0,000	0,000	0,002	0,008	0,029	0,077	0,179	0,319	0,510
(10,10)	0,000	0,000	0,001	0,004	0,019	0,051	0,128	0,242	0,414

* Адаптированные данные из статьи: Swed Frieda S., Eisenhart C. Tables matical Statistics. 1943. 14. p. 66—87.

для общего числа серий u в выборках объема (n_1, n_2)*

11	12	13	14	15	16	17	18	19	20
1,000									
1,000									
1,000									
1,000									
1,000									
0,998	1,000								
0,992	0,999	1,000							
0,984	0,998	1,000							
0,972	0,994	1,000							
0,958	0,990	1,000							
0,975	0,996	0,999	1,000						
0,949	0,988	0,998	1,000	1,000					
0,916	0,975	0,994	0,999	1,000					
0,879	0,957	0,990	0,998	1,000					
0,900	0,968	0,991	0,999	1,000	1,000				
0,843	0,939	0,980	0,996	0,999	1,000	1,000			
0,782	0,903	0,964	0,990	0,998	1,000	1,000			
0,762	0,891	0,956	0,988	0,997	1,000	1,000	1,000		
0,681	0,834	0,923	0,974	0,992	0,999	1,000	1,000	1,000	
0,586	0,758	0,872	0,949	0,981	0,996	0,999	1,000	1,000	1,000

for testing randomness of grouping in a sequence of alternatives. —Annals of Mathe-

ностей для случаев $n_1 \leqslant 10$ и $n_1 \leqslant n_2 \leqslant 10$. (Если $n_1 > n_2$, то надо поменять местами n_1 и n_2.) Эти распределения первоначально были даны Ф. Свед и С. Эйзенхартом в работе: S w e d Frieda S., E i s e n h a r t C. Tables for testing randomness of grouping in a sequence of alternatives. — Annals of Mathematical Statistics, 1943, **14**, p. 66—87. В этих таблицах приводится большее число знаков после запятой и иное расположение материала.

Если $n_1 > 10$ и $n_2 > 10$, то отпадает необходимость знать точные значения вероятностей, так как можно достаточно точно аппроксимировать истинное распределение с помощью нормального. Пусть

$$\mu = \frac{2n_1n_2}{n_1 + n_2} + 1, \tag{3.10.1}$$

$$\sigma^2 = \frac{2n_1n_2\,(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2\,(n_1 + n_2 - 1)}\,. \tag{3.10.2}$$

Можно показать, что это истинные среднее и дисперсия дискретного распределения величины u. Тогда величина

$$z = \frac{\left(u - \mu + \frac{1}{2}\right)}{\sigma} \tag{3.10.3}$$

приближенно может считаться единичным (нормированным) нормальным отклонением, причем—$^1/_2$ — обычная поправка на непрерывность, которая позволяет компенсировать тот факт, что непрерывное распределение используется для аппроксимации дискретного.

П р и м е р. Исследуется множество из двадцати семи остатков (пятнадцать из которых имеют один знак, а двенадцать — противоположный), размещенных во временной последовательности, содержащей $u = 7$ серий. Можно ли считать, что в таком расположении знаков «слишком мало серий»?

Здесь $n_1 = 15$, $n_2 = 12$, $u = 7$. Из уравнений (3.10.1) и (3.10.2) имеем $\mu = 43/3$, $\sigma^2 = 740/117$. Следовательно, наблюдаемая величина z в соответствии с уравнением (3.10.3) есть

$$z = \frac{\left(7 - \frac{43}{3} + \frac{1}{2}\right)}{\left(\frac{740}{117}\right)^{1/2}} = -2{,}713.$$

Вероятность того, что нормированное нормальное отклонение будет иметь значение —2,713 или меньше, составляет 0,0033 (или 0,33 %). Так что, по-видимому, здесь необычно малое число серий. Мы должны отвергнуть предположение, что данное расположение знаков носит случайный характер. Модель следует взять под подозрение и исследовать причины такого расположения остатков.

(П р и м е ч а н и я: 1. Слишком малое число серий указывает на возможную положительную сериальную корреляцию в остатках. А слишком большое число серий указывает на возможную отрицательную сериальную корреляцию. В этом случае нужно пользоваться

верхним «хвостом» критерия серий и уровнями вероятности, накопленными с правого конца, а не с левого, как в табл. 3.1. Вероятности правого хвоста можно считать и из табл. 3.1, если иметь в виду, что накопленная вероятность того, что $u \geqslant u_0$, обозначенная, допустим, СР $(u \geqslant u_0)$, равна 1 — СР $(u \leqslant u_0 - 1)$. Так, например, накопленная вероятность того, что в случае (5, 5) число серий $u \geqslant 7$, равна 1— —СР $(u \leqslant 6) = 1 - 0{,}643 = 0{,}357$, и т. д. Для значений (n_1, n_2), превосходящих табличные, вполне подходит верхний «хвост» нормированного нормального отклонения, для которого $z = \left(u - \mu - \frac{1}{2}\right)/\sigma$, т. е. поправку на непрерывность надо вычитать. Конечно, теперь требуется не нижний, а верхний «хвост» вероятности $N(0, 1)$.

2. Строго говоря, критерий для серий применим только в том случае, когда причины, приводящие к данному расположению серий, независимы. Для временно́й последовательности остатков это не так из-за корреляций, которые имеют место между ними (см. параграф 3.7), и уровень вероятности, получаемый по этой процедуре, будет меняться в зависимости от конкретной структуры данных. В большинстве регрессионных ситуаций, встречающихся на практике, если отношение $(n-p)/n$ не слишком мало, таким эффектом можно пренебречь.)

3.11. КРИТЕРИЙ ДАРБИНА—УОТСОНА ДЛЯ НЕКОТОРЫХ ВИДОВ СЕРИАЛЬНОЙ КОРРЕЛЯЦИИ

Есть широко распространенный критерий для выявления сериальной корреляции определенного вида, который называется *критерием Дарбина—Уотсона*. (Он назван именами двух исследователей, которые обсуждали его применение к анализу регрессионных остатков и в 1951 г. построили удобные таблицы. Первоначально он был предложен еще в 1941 г. фон Нейманом[14] для нерегрессионных задач. Некоторые работы, посвященные этим вопросам, приведены в библиографии к гл. 3, см. кн. 2.)

Пусть мы хотим подобрать постулированную линейную модель

$$y_u = \beta_0 + \sum_{i=1}^{k} \beta_i X_{iu} + \varepsilon_u \qquad (3.11.1)$$

методом наименьших квадратов по наблюдениям $(Y_u, X_{1u}, X_{2u}, \ldots, X_{ku})$, $u = 1, 2, \ldots, n$. Обычно мы должны предполагать, что ошибки ε_u — независимые случайные величины с распределением $N(0, \sigma^2)$, т. е. что все сериальные корреляции $\rho_s = 0$. С помощью критерия

[14] Джон фон Нейман (1903—1957) — один из выдающихся математиков XX века, внесший огромный вклад во многие различные области математики и ее приложений — от квантовой механики и теории автоматов, до экономики и теории игр. См.: Данилов Ю. А. Джон фон Нейман. — М.: Знание, 1981, 62 с.; Вигнер Е. Джон фон Нейман.—В кн.: Этюды о симметрии/Пер. с англ. Под ред. Я. А. Смородинского.— М.: Мир, 1971, с. 204—208.— *Примеч. пер.*

Дарбина—Уотсона можно проверить нуль-гипотезу H_0 о том, что все $\rho_s = 0$ против альтернативы

$$H_1 : \rho_s = \rho^s$$

($\rho \neq 0$ и $|\rho| < 1$). Такая альтернатива появляется из предположения о том, что ошибки ε_u подчиняются условию

$$\varepsilon_u = \rho\varepsilon_{u-1} + z_u,$$

где $z_u \sim N(0, \sigma^2)$, а независимы

$$\varepsilon_{u-1}, \varepsilon_{u-2}, \ldots \quad \text{и} \quad z_{u-1}, z_{u-2}, \ldots$$

При этом еще предполагается, что и среднее, и дисперсия ошибок ε_u постоянны и не зависят от u, откуда с необходимостью следует, что $\varepsilon_u \sim N\{0, \sigma^2/(1-\rho^2)\}$. Заметим, что когда нуль-гипотеза верна и $\rho = 0$, это условие сводится к $\varepsilon_u \sim N(0, \sigma^2)$, т. е. к нашим обычным предположениям для всех $u = 1, 2, \ldots, n$.

Для проверки H_0 против альтернативы H_1 мы строим модель по уравнению (3.11.1) и находим набор остатков $e_1, e_2, \ldots, e_n$. Теперь можно построить статистику

$$d = \sum_{u=2}^{n} (e_u - e_{u-1})^2 / \sum_{u=1}^{n} e_u^2 \qquad (3.11.2)$$

и определить на ее основе, можно ли отвергнуть нуль-гипотезу. Такое определение несколько сложнее, чем то, с которым мы имели дело раньше, поскольку вместо одного критического значения теперь приходится использовать два. Кроме того, d применяют только для проверки нижнего хвоста, т. е. против альтернатив $\rho > 0$. А для проверки против «обратных» альтернатив $\rho < 0$, вообще говоря, требуется критерий для верхнего хвоста. К счастью, его можно легко заменить на критерий для нижнего хвоста статистики $(4-d)$.

В табл. 3.2, 3.3 и 3.4, опубликованных Дарбином и Уотсоном в 1951 г., содержатся пары (обозначенные d_L, d_U) точек для уровней значимости 5, 2,5 и 1 %, т. е. критические значения для уровней вероятностей $\alpha = 0{,}05$, 0,025 и 0,01 соответственно. Они приводятся для различных чисел наблюдений n и для предикторов $k = 1, 2, \ldots, 5$ (см. k в уравнении (3.11.1)). Проверка сводится к следующему.

1. *Односторонний критерий против альтернатив* $\rho > 0$. Если $d < d_L$, то заключают, что d значимо и отбрасывают H_0 на уровне α.

Если $d > d_U$, то заключают, что d не значимо и не отбрасывают H_0.

Если же $d_L \leqslant d \leqslant d_U$, то проверка не позволяет сделать никакого вывода.

2. *Односторонний критерий против альтернатив* $\rho < 0$. Повторить (1), используя $(4-d)$ вместо d.

3. *Двусторонний критерий с равными хвостами против альтернатив* $\rho \neq 0$. Если $d < d_L$ или $4-d < d_L$, то заключают, что d значимо и отбрасывают гипотезу H_0 на уровне 2α.

Если $d > d_U$ и $4-d > d_U$, то заключают, что d незначимо и не отбрасывают гипотезу H_0 на уровне 2α.

Таблица 3.2. **Критические точки d_L и d_U для уровня 5 %**

n	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

Источник: Durbin J., Watson G. S. Testing for serial correlation in least squares regression II. — Biometrika, 1951, **38**, p. 159—178.

Таблица 3.3. **Критические точки d_L и d_U для уровня 2,5 %**

n	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0,95	1,23	0,83	1 40	0,71	1,61	0,59	1,84	0,48	2,09
16	0,98	1,24	0,86	1,40	0,75	1,59	0,64	1,80	0,53	2,03
17	1,01	1,25	0,90	1,40	0,79	1,58	0,68	1,77	0,57	1,98
18	1,03	1,26	0,93	1,40	0,82	1,56	0,72	1,74	0,62	1,93
19	1,06	1,28	0,96	1,41	0,86	1,55	0,76	1,72	0,66	1,90
20	1,08	1,28	0,99	1,41	0,89	1,55	0,79	1,70	0,70	1,87
21	1,10	1,30	1,01	1,41	0,92	1,54	0,83	1,69	0,73	1,84
22	1,12	1,31	1,04	1,42	0,95	1,54	0,86	1,68	0,77	1,82
23	1,14	1,32	1,06	1,42	0,97	1,54	0,89	1,67	0,80	1,80
24	1,16	1,33	1,08	1,43	1,00	1,54	0,91	1,66	0,83	1,79
25	1,18	1,34	1,10	1,43	1,02	1,54	0,94	1,65	0,86	1,77
26	1,19	1,35	1,12	1,44	1,04	1,54	0,96	1,65	0,88	1,76
27	1,21	1,36	1,13	1 44	1,06	1,54	0,99	1,64	0,91	1,75
28	1,22	1,37	1,15	1,45	1,08	1.54	1,01	1,64	0,93	1,74
29	1,24	1,38	1,17	1,45	1,10	1,54	1,03	1,63	0,96	1,73
30	1,25	1,38	1,18	1,46	1,12	1,54	1,05	1,63	0,98	1,73
31	1,26	1,39	1,20	1,47	1,13	1,55	1,07	1,63	1,00	1,72
32	1,27	1,40	1,21	1,47	1,15	1,55	1,08	1,63	1,02	1,71
33	1,28	1,41	1,22	1,48	1,16	1,55	1,10	1,63	1,04	1,71
34	1,29	1,41	1,24	1,48	1,17	1,55	1,12	1,63	1,06	1,70
35	1,30	1,42	1,25	1,48	1,19	1,55	1,13	1,63	1,07	1,70
36	1,31	1,43	1,26	1,49	1,20	1,56	1,15	1,63	1,09	1,70
37	1,32	1,43	1,27	1,49	1,21	1,56	1,16	1,62	1,10	1,70
38	1,33	1,44	1,28	1,50	1,23	1,56	1,17	1,62	1,12	1,70
39	1,34	1,44	1,29	1,50	1,24	1,56	1,19	1,63	1,13	1,69
40	1.35	1,45	1,30	1,51	1,25	1,57	1,20	1,63	1,15	1,69
45	1,39	1,48	1,34	1,53	1,30	1,58	1,25	1,63	1,21	1,69
50	1,42	1,50	1,38	1,54	1,34	1,59	1,30	1,64	1,26	1,69
55	1,45	1,52	1,41	1,56	1,37	1,60	1,33	1,64	1,30	1,69
60	1,47	1,54	1,44	1,57	1,40	1,61	1,37	1,65	1,33	1,69
65	1,49	1,55	1,46	1,59	1,43	1,62	1,40	1,66	1,36	1,69
70	1,51	1,57	1,48	1,60	1,45	1,63	1,42	1,66	1,39	1,70
75	1,53	1,58	1,50	1,61	1,47	1,64	1,45	1,67	1,42	1,70
80	1,54	1,59	1,52	1,62	1,49	1,65	1,47	1,67	1,44	1,70
85	1,56	1,60	1.53	1,63	1,51	1,65	1,49	1,68	1,46	1,71
90	1,57	1,61	1,55	1,64	1,53	1,66	1,50	1,69	1,48	1,71
95	1,58	1,62	1.56	1,65	1,54	1,67	1,52	1,69	1,50	1,71
100	1,59	1,63	1,57	1,65	1,55	1.67	1,53	1,70	1,51	1,72

Источник. Durbin J., Watson G. S. Testing for serial correlation in least squares regression II. — Biometrika, 1951, **38**, p. 159—178.

Таблица 3.4. **Критические точки d_L и d_U для уровня 1 %**

n	$k=1$		$k=2$		$k=3$		$k=4$		$k=5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0,81	1,07	0,70	1,25	0,59	1,46	0,49	1,70	0,39	1,96
16	0,84	1,09	0,74	1,25	0,63	1,44	0,53	1,66	0,44	1,90
17	0,87	1,10	0,77	1,25	0,67	1,43	0,57	1,63	0,48	1,85
18	0,90	1,12	0,80	1,26	0,71	1,42	0,61	1,60	0,52	1,80
19	0 93	1,13	0,83	1,26	0,74	1,41	0,65	1,58	0,56	1,77
20	0,95	1,15	0,86	1,27	0,77	1,41	0,68	1,57	0,60	1,74
21	0,97	1,16	0,89	1,27	0,80	1,41	0,72	1,55	0,63	1,71
22	1,00	1,17	0,91	1,28	0,83	1,40	0,75	1,54	0,66	1,69
23	1,02	1,19	0,94	1,29	0,86	1,40	0,77	1,53	0,70	1,67
24	1,04	1,20	0,96	1,30	0,88	1,41	0,80	1,53	0,72	1,66
25	1,05	1,21	0,98	1,30	0,90	1,41	0,83	1,52	0,75	1,65
26	1,07	1,22	1,00	1,31	0,93	1,41	0,85	1,52	0,78	1,64
27	1,09	1,23	1,02	1,32	0,95	1,41	0,88	1,51	0,81	1,63
28	1,10	1,24	1,04	1,32	0,97	1,41	0,90	1,51	0,83	1,62
29	1,12	1,25	1,05	1,33	0,99	1,42	0,92	1,51	0,85	1,61
30	1,13	1,26	1,07	1,34	1,01	1,42	0,94	1,51	0,88	1,61
31	1,15	1,27	1,08	1,34	1,02	1,42	0,96	1,51	0,90	1,60
32	1,16	1,28	1,10	1,35	1,04	1,43	0,98	1,51	0,92	1,60
33	1,17	1,29	1,11	1,36	1,05	1,43	1,00	1,51	0,94	1,59
34	1,18	1,30	1,13	1,36	1,07	1,43	1,01	1,51	0,95	1,59
35	1,19	1,31	1,14	1,37	1,08	1,44	1,03	1,51	0,97	1,59
36	1,21	1,32	1,15	1,38	1,10	1,44	1,04	1,51	0,99	1,59
37	1,22	1,32	1,16	1,38	1,11	1,45	1,06	1,51	1,00	1,59
38	1,23	1,33	1,18	1,39	1,12	1,45	1,07	1,52	1,02	1,58
39	1 24	1,34	1,19	1,39	1,14	1,45	1,09	1,52	1,03	1,58
40	1,25	1,34	1,20	1,40	1,15	1,46	1,10	1,52	1,05	1,58
45	1,29	1,38	1,24	1,42	1,20	1,48	1,16	1,53	1,11	1,58
50	1,32	1,40	1,28	1,45	1,24	1,49	1,20	1,54	1,16	1,59
55	1,36	1,43	1,32	1,47	1,28	1,51	1,25	1,55	1,21	1,59
60	1,38	1,45	1,35	1,48	1,32	1,52	1,28	1,56	1,25	1,60
65	1,41	1,47	1,38	1,50	1,35	1,53	1,31	1,57	1,28	1,61
70	1,43	1,49	1,40	1,52	1,37	1,55	1,34	1,58	1,31	1,61
75	1,45	1,50	1,42	1,53	1,39	1,56	1,37	1,59	1,34	1,62
80	1,47	1,52	1,44	1,54	1,42	1,57	1,39	1,60	1,36	1,62
85	1,48	1,53	1,46	1,55	1,43	1,58	1,41	1,60	1,39	1,63
90	1,50	1,54	1,47	1,56	1,45	1,59	1,43	1,61	1,41	1,64
95	1,51	1,55	1,49	1,57	1,47	1,60	1,45	1,62	1,42	1,64
100	1,52	1,56	1,50	1,58	1,48	1,60	1,46	1,63	1,44	1,65

Источник. Durbin J., Watson G. S. Testing for serial correlation in least squares regression II. — Biometrika, 1951, 38, p. 159—178.

В противном случае критерий не позволяет сделать каких-либо заключений.

Невозможность иногда принимать решение, что характерно для описанных выше критериев, совсем не привлекательна, однако эта ситуация оказалась далеко не простой. В более позднем исследовании действительно были предложены средства, позволяющие избавиться от случаев неразрешимости, но они гораздо сложнее того, что мы рассматриваем, и мы здесь не будем о них более говорить. Тем не менее было установлено, что во многих случаях работа с тестом так, как будто d_L не существует, а d_U остается единственным подходящим критическим значением, дает прекрасное приближение к действительному положению вещей *. Такая упрощенная приближенная процедура проверки гипотез сводится к следующему.

1′. *Упрощенный односторонний критерий против альтернатив* $\rho > 0$. Если $d < d_U$, то отвергнуть гипотезу H_0 на уровне α, в противном случае не отвергать.

2′. *Упрощенный односторонний критерий против альтернатив* $\rho < 0$. Если $4-d < d_U$, то отвергнуть гипотезу H_0 на уровне α, в противном случае не отвергать.

3′. *Упрощенный двусторонний критерий против альтернатив* $\rho \neq 0$. Если $d < d_U$ или $4-d < d_U$, то отвергнуть гипотезу H_0 на уровне 2α.

Для практических целей на этом уровне сложности мы предлагаем прежде всего воспользоваться (d_L, d_U)-критерием, чтобы увидеть, можно ли получить ясное решение. Невозможность получить результат в такой проверке была бы, конечно, окончательным приговором упрощенному критерию, но в нашем решении второго уровня она могла бы указывать либо на появление «сигнала опасности», либо, возможно, на несколько более высокий уровень α-риска, чем тот, что соотносится с упрощенным критерием. Ниже в примере 2 приведены рассуждения такого рода.

Пример 1. Остатки при подборе прямой по $n = 50$ парам значений (X, Y) привели к значению d-статистики, равному $d = 0{,}625$. Сначала мы проверим двусторонним критерием гипотезу $H_0 : \rho = 0$ против двусторонних альтернатив $\rho \neq 0$. Для этого сравним значения d и $4-d = 3{,}375$ с соответствующими значениями d_L и d_U из табл. 3.2—3.4. Для $\alpha = 0{,}01$ при $k = 1$ и $n = 50$ мы находим

$$d = 0{,}625 < d_L = 1{,}32$$

Отсюда следует, что, пользуясь методом (3), мы отвергнем гипотезу H_0 на уровне $2\alpha = 0{,}02$ и придем к выводу, что, по-видимому, сериальная корреляция проверяемого вида *действительно* существует в наших данных. Такое предположение ставит под сомнение подобранную модель и требует пересмотра данных в свете новой инфор-

* Можно обратиться к работе: Durbin J., Watson G. S. Testing for serial correlation in least squares regression III.— Biometrika, 1971, 58, p. 1—19, где обсуждается точность приближения d_U и рассматриваются альтернативные варианты.

мации. (См., например, книги [15]: Jenkins G. M., Watts D. G. Spectral Analysis and Its Applications.— San Francisco: Holden-Day, 1968; Box G. E. P., Jenkins G. M. Time Series, Forecasting and Control.— San Francisco: Holden-Day, 1970.)

Пример 2. При построении линейной модели с четырьмя предикторными переменными получилось 70 остатков, для которых мы нашли значение d-статистики, равное 1,51. Надо проверить гипотезу $H_0 : \rho_s = 0$ против односторонней альтернативы $H_1 : \rho_s = \rho^s$, где $\rho > 0$.

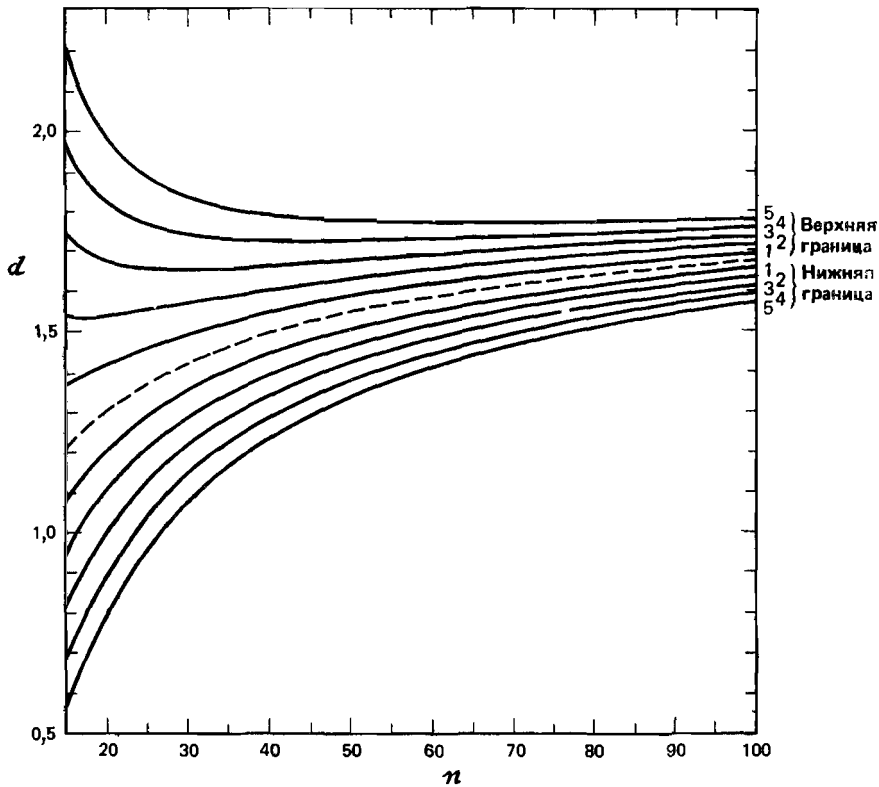


Рис. 3.10. Зависимость 5 %-ных значений d_U и d_L от n при $k = 1, 2, 3, 4, 5$

Из табл. 3.2—3.4 находим следующие критические точки:

	d_L	d_U
$\alpha = 0{,}05$	1,49	1,74
$\alpha = 0{,}025$	1,42	1,66
$\alpha = 0{,}01$	1,34	1,58

[15] О переводах этих книг на русский язык см. примечание 13 на с. 203. — *Примеч. пер.*

Мы видим, что воспользовавшись методом (1), на всех уровнях получим неразрешимый результат первичной проверки. Обратившись к методу 1′, мы придем к вторичному выводу, что гипотеза H_0 должна быть отвергнута на том основании, что величина d оказалась меньше, чем d_U, на уровне $\alpha = 0{,}01$. Фактический уровень отбрасывания будет, возможно, не так низок, как $\alpha = 0{,}01$, поскольку мы воспользовались упрощенным критерием. Мы также видим, однако, что и при обычном критерии мы должны были бы почти отвергнуть нашу гипотезу на уровне $\alpha = 0{,}05$, поскольку 1,51 очень близко к $d_L = 1{,}49$. Значит, можно с некоторым основанием думать, что уровень, при котором отбрасывается H_0, лежит где-то между $\alpha = 0{,}05$ и $\alpha = 0{,}01$ *. Это ставит под сомнение подобранную модель и делает целесообразным повторный анализ данных, учитывающий выявленную сериальную корреляцию (см. ссылки, приведенные в примере 1).

На рис. 3.10 представлены графики зависимости d_L и d_U от числа наблюдений n при уровне значимости 5 %. Заметим, что вертикальные расстояния между парами кривых с соответствующими номерами образуют область отсутствия решения при применении стандартного критерия и что с ростом n эта область сжимается. Мораль ясна: чем больше наблюдений, тем более вероятно, что мы сможем принять определенное решение с помощью критерия Дарбина—Уотсона. Специалисты по анализу временных рядов на основании своего опыта знают, что для получения полезных выводов в их исследованиях надо иметь $n \geq 50$. Как видно из рис. 3.10, такое рабочее правило не стоит забывать при применении критерия Дарбина—Уотсона.

3.12. ОПРЕДЕЛЕНИЕ ВЛИЯЮЩИХ НАБЛЮДЕНИЙ

Сначала мы рассмотрим пример (в значительной мере искусственный), в котором уравнение прямой подбирается по множеству данных, включающему 5 наблюдений, причем 4 при значении $X = a$, а одно при $X = b$. Если $V(Y_i) = \sigma^2$, то можно показать, что для $X = a$ $V(e_i) = 0{,}75\,\sigma^2$, $i = 1, 2, 3, 4$, тогда как для $X = b$ $V(e_5) = 0$. На первый взгляд нулевая дисперсия кажется весьма желательной, но на самом деле это совсем не так, поскольку подбираемая прямая определяется совместно средним уровнем Y при $X = a$ и единственным наблюдаемым значением Y при $X = b$. Остаток при $X = b$ равен нулю при любом значении Y, так что фактически параметры оцениваются в зависимости от «веса» этого единственного наблюдения. Сколь угодно большая ошибка в подобном наблюдении неопределима в процессе построения модели, да и исследование остатков не сможет ее обнаружить, даже если она и существует. Наблюдение при $X = b$ оказывает огромное влияние на результат, безотносительно к тому, верно оно или нет.

Тот факт, что какое-то наблюдение имеет большой «выброс», конечно, совсем не положителен, но из этого вовсе не обязательно сле-

* Возможна и альтернативная интерпретация: первичный тест был бы значим на уровне, близком к 0,06.

дует, что именно это наблюдение влияет на построение выбранной модели. Так, на рис. 3.11 мы видим, что наблюдение под номером 19 было бы, конечно, выбросом для большинства наиболее простых моделей, подходящих к имеющимся данным, хотя оно и не влияет в том смысле, что «перевес» в точках для соседних значений X не позволяет ему существенно сдвинуть оценку, так что данное значение не сможет оказать существенного влияния на оценки коэффициентов модели. С другой стороны, ясно, что наблюдение под номером 18 оказывает

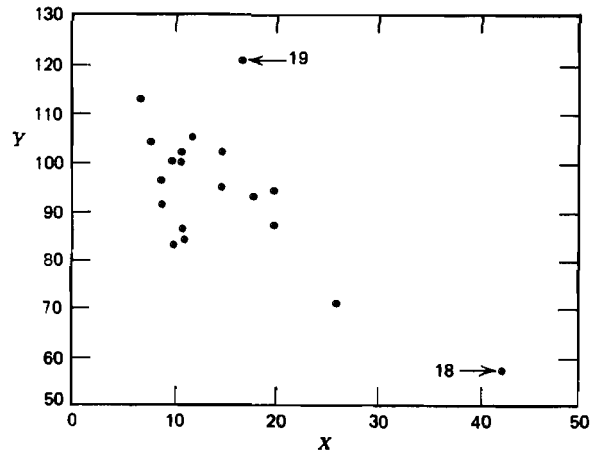


Рис. 3.11. Регрессия с наблюдением под номером 19, которое ни на что не влияет, и с наблюдением под номером 18, которое может оказать очень сильное влияние. Значения X соответствуют возрасту ребенка в месяцах к моменту, когда он сказал первое слово, а значения Y представляют оценки тех же детей по одному из тестов способностей. Воспроизводятся с разрешения авторов и издателей из работы: Andrews D. F., Pregibon D. Finding the outliers that matter. Journal of the Royal Statistical Society, 1978, **B4**, p. 84—93. Исходные данные были получены доктором Линдом (L. M. Linde) из Калифорнийского университета в Лос-Анджелесе и использованы в работе: Mickey M. R., Dunn O. J., Clark V. Note on the use of stepwise regression in detecting out liers.— Computers and Biomedical Research, 1967, 1, p. 105—111. [Речь идет в данном случае не об определении общих способностей, интеллекта или уровня развития ребенка, а о некотором специальном тесте, выявляющем «способность к говорению» («вербальную способность»).— *Примеч. пер.*]

влияние. Оно стоит одиноко, вдали от остальных точек, и вполне можно ожидать, что оно оказывает наиболее сильное влияние на положение подбираемой здесь модели. В этой точке может быть большой остаток или его может не быть в зависимости от того, какая именно модель подбирается и сколько остается степеней свободы. В любых наборах данных, где оценивание одного или нескольких параметров сильно зависит от очень малого числа наблюдений, такие проблемы могут появиться. Один из путей их решения заключается в проверке, нельзя ли вычеркиванием одного-двух критических наблюдений резко повлиять на подбор модели и последующие выводы. Если это удается, то наши выводы оказываются шаткими и требуются допол-

нительные данные. Метод ПРЕСС (PRESS), описанный в параграфе 6.8, принадлежит к методам такого рода. А вот другие предложения, высказанные в литературе.

1. Р. Кук в работе, посвященной определению влияющих наблюдений в линейной регрессии (см.: Cook R. D. Detection of influential observations in linear regression.— Technometrics, 1977, **19**, p. 15—18) предположил, что влияние i-й точки в данных можно измерить расстоянием

$$D_i = \{\mathbf{b} - \mathbf{b}(i)\}' \mathbf{X}'\mathbf{X} \{\mathbf{b} - \mathbf{b}(i)\}/(ps^2), \qquad (3.12.1)$$

где $\mathbf{X}$ — матрица размера $n \times p$, $\mathbf{b}$ — обычный вектор МНК-оценок, а $\mathbf{b}(i)$ — вектор МНК-оценок, полученный *после того*, как из данных исключена i-я точка. Расстояния D_i сравниваются с помощью F-критерия при $F(p, n-p, 1-\alpha)$ для выбранного α. Большие значения D_i воспринимаются как указания на влияние i-го наблюдения. Расстояние D_i, может быть, легче оценить, если переписать его в следующей эквивалентной форме:

$$D_i = \left\{\frac{e_i}{s(1-r_{ii})^{1/2}}\right\}^2 \left\{\frac{r_{ii}}{1-r_{ii}}\right\} \frac{1}{p}, \qquad (3.12.2)$$

где e_i — i-й остаток для случая, когда используются все данные; s^2 — оценка дисперсии $V(Y_i) = \sigma^2$, обусловленной остаточным средним квадратом, когда используются все данные, а r_{ii}—i-й диагональный элемент матрицы $\mathbf{R} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Мы видим, что первый сомножитель в уравнении (3.12.2) — «стьюдентизированный» остаток, т. е. остаток, деленный на свою стандартную ошибку (см. параграф 3.7), тогда как второй член представляет собой отношение (дисперсия i-го предсказанного значения)/(дисперсия i-го остатка). Заметим, что $0 \leqslant r_{ii} \leqslant 1$. Расстояние D_i может быть большим, когда велик либо первый, либо второй сомножитель. А эти сомножители служат мерами двух различных характеристик каждой точки.

В примере, который мы привели раньше,

$$D_i = \left\{\frac{e_i^2}{0{,}75s^2}\right\}\left(\frac{1}{3}\right)\left(\frac{1}{2}\right), \quad i = 1, 2, 3, 4 \qquad (3.12.3)$$

и

$$D_5 \quad \text{неопределенно,}$$

где при $i = 1, 2, 3, 4$ каждое $e_i = Y_i - \bar{Y}_a$, причем $\bar{Y}_a = (Y_1 + Y_2 + Y_3 + Y_4)/4$, а $e_5 = 0$. Пятое наблюдение при $X = b$, таким образом, «сигнализирует» о том, что здесь есть какая-то особенность, а исследование выявляет в этих обстоятельствах огромное влияние этой точки.

Для вычисления D_i рекомендуется использовать машинные программы (например, такие, как BMDP9R). Вычисления привлекают своей простотой, а сама статистика обладает свойством эффективности. За дальнейшими подробностями обращайтесь к работе, указанной выше, а также к статье: Cook R. D. Influential observations

in linear regression.— Journal of the American Statistical Association, 1979, **74**, p. 169—174.

2. Д. Эндрьюс и Д. Прегибон (см.: Andrews D. F., Pregibon D. Finding the outliers that matter.— Journal of the Royal Statistical Society, 1978, **B—40**, p. 84—93) предложили статистику (называемую ниже AP):

$$R_{ij}^{(k)} \ldots (\mathbf{X}^*) = |D_{ij}^{(k)} \ldots |\mathbf{X}^{*\prime}\mathbf{X}^*|| / |\mathbf{X}^{*\prime}\mathbf{X}^*|, \qquad (3.12.4)$$

где $\mathbf{X}^* = (\mathbf{X}, \mathbf{Y})$, т. е. это обычная матрица $\mathbf{X}$, к которой справа присоединена матрица-столбец $\mathbf{Y}$, и где оператор $D_{ij}^{(k)} \ldots$ означает «выполнение операции, указанной в конце, но после исключения элементов, связанных с k элементами $ij \ldots$.». Например, $D_{ij}^{(2)} (\mathbf{X}'\mathbf{X})^{-1}$ означает «получение обратной матрицы от произведения матриц в скобках после исключения из матрицы $\mathbf{X}$ строк, связанных с X_i и X_j». Можно показать, что решение уравнения (3.12.4) сводится к вычислению определителя размером $k \times k$, полученного следующим образом. Вычислите $\mathbf{W}^* = \mathbf{I} - \mathbf{X}^* (\mathbf{X}^{*\prime}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}$. Вычеркните из матрицы $\mathbf{W}^*$ все строки и столбцы, кроме тех, что связаны с k наблюдениями, которые требуются для вычисляемой статистики, и найдите определитель той матрицы, которая останется после вычеркиваний. Используя $\bar{D}_{ij}^{(k)} \ldots$ для обозначения операции «вычеркнуть все, кроме элементов, связанных с k наблюдениями $i, j, \ldots$», мы можем переписать:

$$R_{ij}^{(k)} \ldots (\mathbf{X}^*) = \bar{D}_{ij}^{(k)} \ldots |\mathbf{W}^*|. \qquad (3.12.5)$$

Такой определитель размера $k \times k$ — это безразмерная величина. Представляют интерес его малые значения, поскольку они указывают на «связь с особенностью и/или с *влияющими* наблюдениями» (см. с. 88 работы Д. Эндрьюса и Д. Прегибона). График функции

$$\log \{R_{ij}^{(k)} \ldots (\mathbf{X}^*)/R_k^0\} \qquad (3.12.6)$$

для самых маленьких k часто оказывается горизонтальным. В уравнении (3.12.6) величина R_k^0 обозначает минимальное значение $R_{ij}^{(k)} \ldots \ldots (\mathbf{X}^*)$, какое только удалось наблюдать при всех возможных вычерчиваниях по k. На этих горизонтальных графиках мы видим то значение k, для которого самое маленькое значение в уравнении (3.12.6) отличается от всех остальных. Это и есть значение k, рассматриваемое как число влияющих наблюдений, нуждающихся в дополнительном исследовании. Относительно влияющие наблюдения — это наблюдения, приводящие к выделяющимся значениям в уравнении (3.12.6).

Легко назвать причины, по которым статистики AP вполне разумны. Величина $1 - \{R_{ij}^{(k)} \ldots (\mathbf{X}^*)\}^{1/2}$ «соответствует той доле объема, образованного матрицей $\mathbf{X}^*$, который обеспечивается k наблюдениями ($ij \ldots$). Если это подмножество наблюдений значительно удалено от остальных в факторном пространстве, то можно ожидать, что оно даст большую долю объема в пространстве, образованном матрицей $\mathbf{X}^*$. Это позволяет получить естественную интерпретацию термина «выброс». Следовательно, малые значения в уравнении

(3.12.4) связаны с особенностью и/или влияющими наблюдениями. Какова бы ни была действительная причина этого явления, стоит выделить подмножества наблюдений, дающих малые значения в уравнении (3.12.4) для дальнейшего тщательного изучения» (см. с. 88 указанной выше работы).

Можно показать, что при $k = 1$ AP-статистика из уравнения (3.12.4) сводится к выражению: $1 - r_{ii} - e_i^2 / \Sigma e_i^2$. Прилагая его к нашему примеру из этого параграфа, мы найдем такие значения AP-статистики:

$$1 - \frac{1}{4} - e_i^2 / \Sigma e_i^2, \quad i = 1, 2, 3, 4$$

и

$$1 - 1 - 0, \quad i = 5.$$

И снова пятое наблюдение «сигнализирует» о том, что есть какая-то аномалия и/или что оно сильно влияет на значения МНК-оценок.

3. В статье Н. Дрейпера и Дж. Джона о влияющих наблюдениях и выбросах в регрессии (D r a p e r N. R., J o h n J. A. Influential observations and outliers in regression. — Technometrics, 1981, 23, p. 21—26) выявляются роли статистик Кука и AP. Запишем исходную регрессионную модель для n наблюдений и p параметров в виде

$$E(\mathbf{Y}) = E\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \boldsymbol{\beta}. \tag{3.12.7}$$

Все наблюдения разделены на две группы. В одной K наблюдений ($\mathbf{Y}_2$), подлежащих исследованию как подозреваемые в том, что они представляют собой выбросы или влияющие наблюдения. Во второй остальные $n-K$ наблюдений, которые ни в чем не заподозрены. Естественно, для такого разделения в уравнении (3.12.7) может понадобиться перестановка строк. Пользуясь обычным МНК-анализом, получим остатки для построенной модели в виде

$$\mathbf{e} = \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} = (\mathbf{I} - \mathbf{R})\,\mathbf{Y} = \begin{pmatrix} \mathbf{I} - \mathbf{R}_{11} & -\mathbf{R}_{12} \\ -\mathbf{R}_{21} & \mathbf{I} - \mathbf{R}_{22} \end{pmatrix} \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix}, \tag{3.12.8}$$

где

$$\mathbf{R}_{ij} = \mathbf{X}_i (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_j' \tag{3.12.9}$$

— это подматрица матрицы $\mathbf{R} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Вычеркивая подозрительные наблюдения $\mathbf{Y}_2$, получаем модель $E(\mathbf{Y}_1) = \mathbf{X}_1\boldsymbol{\beta}$. С другой стороны, можно было бы воспользоваться моделью

$$E\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & 0 \\ \mathbf{X}_2 & \mathbf{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \boldsymbol{\gamma} \end{pmatrix}, \tag{3.12.10}$$

где $\boldsymbol{\gamma}$ — вектор размера $K \times 1$, состоящий из дополнительных пара-

метров (см. Draper N. R. Missing values in response surface designs.— Technometrics, 1961, 3, p. 389—398). Вот окончательные оценки векторов $\mathbf{b}$ и $\mathbf{c}$, соответствующие «истинным» значениям $\boldsymbol{\beta}$ и $\boldsymbol{\gamma}$:

$$\mathbf{b} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}_1, \tag{3.12.11}$$

$$\mathbf{c} = (\mathbf{I} - \mathbf{R}_{22})^{-1}\mathbf{e}_2. \tag{3.12.12}$$

Подставив $\mathbf{Y}_2$ в уравнение (3.12.7) в качестве оценок «пропущенных значений»

$$\mathbf{Y}_2 - \mathbf{c} = (\mathbf{I} - \mathbf{R}_{22})^{-1}\mathbf{R}_{21}\mathbf{Y}_1, \tag{3.12.13}$$

и пересчитав уравнение (3.12.7), найдем новые остатки, компоненты которых $\mathbf{u}$ выражаются так:

$$\mathbf{u}_1 = (\mathbf{I} - \mathbf{R}_{11} - \mathbf{R}_{12}(\mathbf{I} - \mathbf{R}_{22})^{-1}\mathbf{R}_{21})\,\mathbf{Y}_1, \tag{3.12.14}$$

$$\mathbf{u}_2 = 0,$$

причем размерности $\mathbf{u}_i$ те же, что и размерности $\mathbf{Y}_i$ в уравнении (3.12.7). Описанная выше процедура корректировки вектора $\mathbf{Y}_2$ с необходимостью требует, чтобы выполнялось условие $\mathbf{u}_2 = 0$, тогда как $\mathbf{u}_1$ есть не что иное, как остатки для модели $E(\mathbf{Y}_i) = \mathbf{X}_1\boldsymbol{\beta}$. Эти остатки u_i называют «пересмотренными остатками».

Дополнительная сумма квадратов, обусловленная включением в модель параметров $\boldsymbol{\gamma}$ из уравнения (3.12.10), в отличие от модели уравнения (3.12.7) равна:

$$Q_K = \mathbf{e}_2'(\mathbf{I} - \mathbf{R}_{22})^{-1}\mathbf{e}_2. \tag{3.12.15}$$

Такая статистика может применяться как критерий для «выбросов», см.: Gentleman J. F., Wilk M. B. Detecting outliers in a two-way table I. Statistical behavior of residuals.— Technometrics, 1975, 17, p. 1—14 и Detecting outliers II. Supplementing the direct analysis of residuals.— Biometrics, 1975, 31, p. 387—410, а также John J. A., Draper N. R. On testing for two outliers or one outlier in two-way tables.— Technometrics, 1978, 20, p. 69—78. Самое последнее, что появилось в печати, это работа: Draper N. R., John J. A. Testing for three or fewer outliers in two-way tables.— Technometrics, 1980, 22, p. 9—15.

В качестве приближенного критерия для одного выброса на уровне α мы вычисляем статистику

$$F = Q_1/(\text{остаточный средний квадрат})_1, \tag{3.12.16}$$

где знаменатель представляет собой остаточный средний квадрат с ν_2 степенями свободы, полученный для модели такого типа, как (3.12.10) в предположении, что возможен один-единственный выброс в некотором заранее определенном месте. Величину F надо сравнивать с α/n %-ной точкой $F(1, \nu_2, 1-\alpha/n)$, *а не* с α %-ной точкой $F(1, \nu_2, 1-\alpha)$. Чтобы узнать, почему это так, обратитесь к работе о критических значениях критерия для обнаружения выбросов в факторных экспериментах: John J. A., Prescott P. Applied Statistics, 1975, 24, p. 56—59.

Разложение статистики Эндрьюса—Прегибона (АР) на множители

Можно показать, что статистика АР допускает следующее разложение на множители:

$$R_{ij}^{(K)} \ldots = (1 - Q_K/\mathrm{RSS})\,|\mathbf{I} - \mathbf{R}_{22}|, \qquad (3.12.17)$$

где RSS — остаточная сумма квадратов, получаемая при подборе полной модели в виде (3.12.7), величина Q_K получается по уравнению (3.12.15), а $\mathbf{R}_{22}$ определяется по уравнению (3.12.9). Таким образом, первый сомножитель несет ту же информацию, что и Q_K, причем он тем меньше, чем больше само Q_K. Вместе с тем второй сомножитель становится малым, когда K точек отбираются среди данных так, что они оказываются удаленными в факторном пространстве (это можно доказать).

Рекомендации

Н. Дрейпер и Дж. Джон в цитированной выше работе о влияющих наблюдениях и выбросах в регрессии рекомендуют выводить на печать Q_K, статистику Кука и второй множитель АР-статистики вот по каким соображениям:

1. Значения Q_K служат мерами для остатков, их большие величины указывают на особенности.

2. Вид статистики Кука гарантирует, что она будет чувствительна к изменениям в модели при пропуске наблюдений. Значит, статистика Кука будет показывать, какие наблюдения влияют, и это влияние проявляется, в частности, в том, что изменяются коэффициенты подбираемого уравнения.

Общий вид статистики Кука таков:

$$C_{ij\ldots} = (\mathbf{b} - \mathbf{b}^*)' \mathbf{X}'\mathbf{X} (\mathbf{b} - \mathbf{b}^*)/ps^2,$$

где $\mathbf{b}$ — оценка метода наименьших квадратов (МНК-оценка) для вектора $\boldsymbol{\beta}$ в уравнении (3.12.7); $\mathbf{b}^*$ — МНК-оценка для $\boldsymbol{\beta}$ в уравнении (3.12.10); $s^2 = \mathrm{RSS}/(n-p)$, а $ij \ldots$ обозначают K индексов, отобранных для образования подвектора $\mathbf{Y}_2$.

3. Второй сомножитель АР-статистики $|\mathbf{I} - \mathbf{R}_{22}|$ — это пространственная мера, показывающая, какие наблюдения «влиятельны» в том смысле, что они отделены от массы наблюдений в пространстве, образованном столбцами матрицы $\mathbf{X}$. Отметим, что такие наблюдения *могут быть, а могут и не быть* влияющими в том смысле, который обсуждался в предыдущем параграфе.

Использование «оценок» для пропущенных значений

Обычно, когда наблюдения пропущены или забракованы, параметры можно оценивать с помощью уравнения (3.12.11). А уравнение (3.12.13) может работать нормально только в ситуациях спланированного эксперимента, когда матрица $\mathbf{X}'\mathbf{X}$ имеет очень простую структуру, т. е. когда она легко обращается. Этот случай мы проиллюстрируем числовым примером.

Числовая иллюстрация

Данные в табл. 3.5 воспроизводят полный факторный эксперимент 2^4 из книги: Cochran W. G., Cox G. M. Experimental Designs.— New York : Wiley, 1957. План эксперимента задается столбцами X_1, X_2, X_3 и X_4. Два существовавших значения (равные 19 и 30 соответственно) были якобы потеряны и заменены буквенными обозначениями (m_1 и m_2) специально для данного примера. Это уже было сделано раньше в иллюстративных целях в работах: Hasemen J. K., Gaylor D. W. An algorithm for noniterative estimation of multiple missing values for crossed classifications. — Technometrics, 1973, **15**, p. 631—636 и John J. A., Prescott P. Estimating missing values in experiments.— Applied Statistics, 1975, **24**, p. 190—192.

Таблица 3.5. **Факторный эксперимент 2^4 с двумя «пропущенными» значениями**

X_0	X_1	X_2	X_3	X_4	X_1X_2	X_1X_3	X_1X_4	X_2X_3	X_2X_4	X_3X_4	Y
1	−1	−1	−1	−1	1	1	1	1	1	1	m_1
1	1	−1	−1	−1	−1	−1	−1	1	1	1	45
1	−1	1	−1	−1	−1	1	1	−1	−1	1	18
1	1	1	−1	−1	1	−1	−1	−1	−1	1	64
1	−1	−1	1	−1	1	−1	1	−1	1	−1	28
1	1	−1	1	−1	−1	1	−1	−1	1	−1	48
1	−1	1	1	−1	−1	−1	1	1	−1	−1	m_2
1	1	1	1	−1	1	1	−1	1	−1	−1	63
1	−1	−1	−1	1	1	1	−1	1	−1	−1	35
1	1	−1	−1	1	−1	−1	1	1	−1	−1	53
1	−1	1	−1	1	−1	1	−1	−1	1	−1	67
1	1	1	−1	1	1	−1	1	−1	1	−1	66
1	−1	−1	1	1	1	−1	−1	−1	−1	1	36
1	1	−1	1	1	−1	1	1	−1	−1	1	72
1	−1	1	1	1	−1	−1	−1	1	1	1	73
1	1	1	1	1	1	1	1	1	1	1	101

Мы хотим подобрать по данным следующую модель:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{12} X_1 X_2 + \ldots + \beta_{34} X_3 X_4 + \varepsilon,$$

так что матрица $\mathbf{X}$ как раз приведена в табл. 3.5. Следовательно, матрица $\mathbf{X}_2$, включающая первую и седьмую строки матрицы $\mathbf{X}$, соответствующие значениям m_1 и m_2, равна:

$$\mathbf{X}_2 = \begin{bmatrix} 1 & -1 & -1 & -1 & -1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 \end{bmatrix},$$

а матрица $\mathbf{X}_1$ образуется из той части матрицы $\mathbf{X}$, которая остается после вычеркивания строк, образующих матрицу $\mathbf{X}_2$. Вектор $\mathbf{Y}_1$

соответствует всем значениям Y-ов, кроме пропущенных m_1 и m_2. Теперь мы видим, что

$$(\mathbf{X}'\mathbf{X})^{-1} = (16)^{-1}\mathbf{I} = \frac{1}{16}\,\mathbf{I},$$

$$\mathbf{R}_{21} = \frac{1}{16} \times$$

$$\times \begin{bmatrix} 3 & 3 & -1 & 3 & -1 & -1 & 3 & -1 & -1 & -1 & -1 & -1 & -1 & 3 \\ -1 & 3 & -1 & 3 & -1 & 3 & -1 & 3 & -1 & -1 & -1 & -1 & 3 & -1 \end{bmatrix},$$

$$\mathbf{I} - \mathbf{R}_{22} = \frac{1}{16}\begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}.$$

Таким образом,

$$\hat{\mathbf{Y}}_2 = \mathbf{Y}_2 - \mathbf{c} = (\mathbf{I} - \mathbf{R}_{22})^{-1}\mathbf{R}_{21}\mathbf{Y}_1 = \frac{2}{3}\begin{bmatrix} 5 & -1 \\ -1 & 5 \end{bmatrix}\frac{1}{16}\begin{bmatrix} 139 \\ 171 \end{bmatrix} =$$

$$= \begin{bmatrix} 21{,}833 \\ 29{,}833 \end{bmatrix}.$$

Эти найденные значения можно снова подставить вместо m_1 и m_2 и вычислить вектор $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Заметим, что в данном примере $\mathbf{b} = \mathbf{X}'\mathbf{Y}/16$, так что подобные вычисления крайне просты. Другой способ подсчета $\mathbf{b} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{Y}_1$ должен привести к тому же ответу, но он более сложен [16].

Приложение 3А. НОРМАЛЬНЫЕ И ПОЛУНОРМАЛЬНЫЕ ГРАФИКИ

Нормальные графики

Площадь под кривой нормированного нормального распределения N (0, 1) от $-\infty$ до некоторой точки x дается интегралом

$$y = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} exp\left(-\frac{1}{2}t^2\right) dt. \qquad (3А.1)$$

Если мы отложим на ординате $100y$, а на абсциссе — x, то получим «S-образную» кривую, называемую кривой накопленной (или кумулятивной) вероятности распределения N (0, 1). Множество точек принадлежит этой кривой, например точки: $(x, y) = (-1{,}96, 2{,}5), (0, 50)$

[16] Наиболее полно и систематично проблема влияющих наблюдений изложена в монографии: Belsley D. A., Kuh E., Welsch R. E. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity.— New York: J. Wiley, 1980. Еще см.: Weisberg S. Applied Linear Regression.— New York: J. Wiley, 1980. Ch. 5, 6, p. 97—149. Оригинальный подход к проблеме, основанный на «функциях (кривых) влияния», рассмотрен в работе: Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия/Пер. с англ. Под ред. Ю. П. Адлера.— М.: Финансы и статистика, 1982, вып. 2, с. 92—94, 110—113.— *Примеч. пер.*

и (1,96, 97,5). Все они легко находятся по таблицам кумулятивного распределения N (0, 1), см. рис. 3А.1 и 3А.2.

Нормальная вероятностная бумага — это специальный вид миллиметровки, который продается в большинстве магазинов технической книги [17]. Горизонтальная ось размечена, как обычно, равномерной сеткой без чисел, а на вертикальной оси нанесена специальная шкала. Эта вертикальная шкала размечена от 0,01 до 99,99, причем ее деления расширяются по мере удаления от точки 50 как к точке 99,99, так и к точке 0,01 симметрично относительно горизонтали со значением 50. Эта шкала устроена так, что если значения y из уравнения (3А.1) умножить на 100 и отложить на графике в зависимости от значений x, то в результате должна получиться «кривая», называемая прямой линией. Таким образом, получается, что вертикальная шкала определяется функцией, обратной к уравнению (3А.1) и равной $x = F^{-1}(y)$, которая «спрямляет» верхнюю и нижнюю ветви S-образной кривой, представленной на рис. 3А.2. Отметим, что, поскольку точки $(-\infty, 0)$ и $(\infty, 100)$ принадлежат графику прямой линии, значения 0 и 100 нельзя нанести на координатную ось, так как она имеет конечную длину, а не простирается от $-\infty$ до ∞. Еще одна точка, принадлежащая нашей прямой, имеет координаты (1, 84,13). Нам будет интересно сейчас ее отыскать.

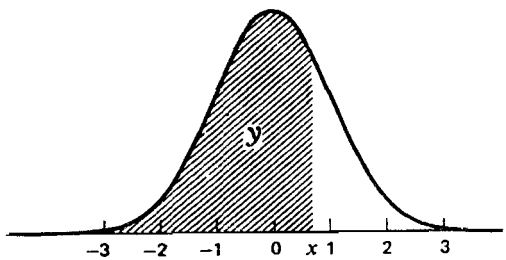


Рис. 3А.1. Площадь, накопленная под кривой нормального распределения до точки x (заштрихована)

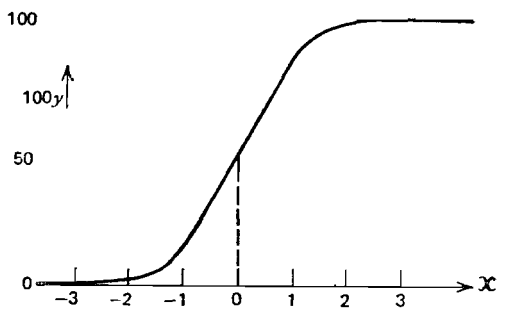


Рис. 3А.2. Накопленная нормальная кривая

Если точки накопленного распределения $N(\mu, \theta^2)$ (а не точки распределения N (0, 1)) нанести на нормальную вероятностную бумагу, то прямая будет проходить через такие точки, как $(x, y) = (\mu - 1{,}96\ \theta,\ 2{,}5)$, $(\mu, 50)$, $(\mu + \theta, 83{,}13)$, $(\mu + 1{,}96\ \theta, 97{,}5)$ и т. п. Этот факт очень полезен, когда у нас есть некая выборка $x_1, x_2, \ldots, x_m$, а мы хотим

[17] У нас вероятностная бумага бывает в специализированных магазинах и отделах канцелярских товаров и школьно-письменных принадлежностей. Ее применение регламентируется стандартом: ГОСТ 11.008-75. Прикладная статистика. Правила построения и применения вероятностных сеток.— М.: Изд-во стандартов, 1977.— *Примеч. пер.*

знать, подчиняется ли она нормальному распределению, и если действительно подчиняется, то мы хотим быстро оценить стандартное отклонение θ. Сначала упорядочим выборку по возрастанию с учетом величин и знаков. Положим, что это уже сделано, т. е. что $x_1, x_2, \ldots, x_m$ как раз и есть нужный порядок. Теперь построим график зависимости x_i от ординаты *, имеющей следующее выражение:

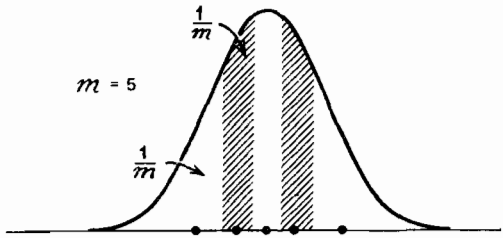


Рис. 3А.3. Разделение площади под кривой нормального распределения на m равных частей. Мы можем «ожидать» по одному наблюдению в каждой такой части, которая делит площадь на равные доли

$$100\left(i-\frac{1}{2}\right)/m. \qquad (3А.2)$$

Основанием для такого выбора служит следующее: если мы разделим единичную площадь под кривой нормального распределения на m равных площадей, то можно «ожидать», что одно наблюдение при-

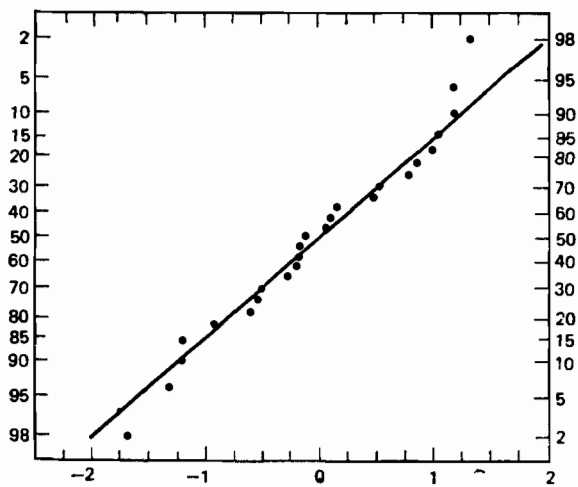


Рис. 3А.4. Нормальный график для остатков из табл. 1.2

---

* О возможных альтернативах к $100\left(i-\frac{1}{2}\right)/m$ см. работу: Barnett V. Probability plotting methods and order statistics.— Applied Statistics, 1975, 24, p. 95—108. Обратите особое внимание на последний абзац на с. 101 и на первый абзац на с. 104. В программе из пакета BMDP используется $100\,(3i-1)/(3m+1) = 100\left(i-\frac{1}{3}\right)\left(m+\frac{1}{3}\right)$ и тоже получается «несдвинутый нормальный вероятностный график», наклон которого можно менять.

дется на каждую из размеченных частей. Значит, i-му наблюдению в упорядоченном ряду, x_i, на графике будет соответствовать накопленная площадь до середины i-й части, которая равна $\left(i - \frac{1}{2}\right)/m$. Ну а умножение на 100 приводит данные в соответствие со шкалой нормальной вероятностной бумаги (см. рис. 3А.3).

Если наша выборка *действительно* принадлежит нормальному распределению, то мы сможем провести (на глаз) хорошую прямую через все множество экспериментальных точек, нанесенных на график, хотя, быть может, и не найдется ни одной точки, которая бы легла точно на прямую. Тогда мы сможем воспользоваться хорошо подобранной прямой для оценки θ. Найдем x_{50} и $x_{84,13}$, т. е. те значения x, при которых наша прямая пересекает горизонтали, соответствующие уровням 50 и 84,13. Тогда разность $x_{84,13} - x_{50}$ как раз и будет искомой оценкой $[(\mu + \theta) - \mu] = \theta$ (см. рис. 3А.4).

Очень полезный способ приобретения опыта принятия решений по графикам такого рода заключается в получении выборок различного объема из таблицы нормальных случайных отклонений [18] и нанесении их на нормальную вероятностную бумагу. Тогда возникнут представления о таких отклонениях от линейности, которые *могут* встретиться, *не будучи* ненормальными.

Половинные нормальные графики

Когда «известно», что выборка подчиняется некоторому (быть может, нормальному) распределению с *нулевым* средним, удобной альтернативой полному нормальному графику служит половинный нормальный график. Если

$$x \sim \frac{1}{\theta\sqrt{2\pi}} exp\left\{-\frac{x^2}{2\theta^2}\right\} \quad (-\infty \leqslant x \leqslant \infty),$$

то

$$|x| \sim \frac{1}{\theta\sqrt{\frac{1}{2}\pi}} exp\left\{-\frac{x^2}{2\theta^2}\right\} \quad (|x| \geqslant 0).$$

Случайная величина $|x|$ подчиняется *половинному* нормальному распределению, имеющему в точности ту же *форму*, что и правая половина распределения $N(0, \theta^2)$, только каждая ордината у него вдвое больше.

Положим

$$y = \int_0^{|x|} \frac{1}{\theta\sqrt{\frac{1}{2}\pi}} exp\left\{-\frac{x^2}{2\theta^2}\right\} dx.$$

Если теперь мы возьмем 50 $(1 + y)$ на ординате, а $|x|$ на абсциссе

[18] Ссылки на таблицы случайных отклонений см. в примечании 4 на с. 189. — *Примеч. пер.*

нормальной вероятностной бумаги при $|x| \geqslant 0$, то в результате должна получиться прямая, проходящая через точку с координатами (абсцисса, ордината) = (0, 50). Фактически это была бы верхняя половина теоретического «полного» нормального графика. Действительно, мы так сложили полный нормальный график, что его нижняя часть теперь совместилась с верхней. Пусть мы имеем выборку из какого-то (быть может, нормального) распределения, про истинное среднее которого «известно», что оно должно быть нулем. Тогда можно проверить нормальность и равенство среднего нулю, нанося на график *половинного* нормального распределения наши выборочные точки. При этом будем наносить на график не сами точки, а их *абсолютные* значения (модули). Если, например, наша выборка включает наблюдения — 17, — 4, 1, 2, 3, 6, 23 (отметим, что они приведены в возрастающем порядке, как и должно быть при использовании «полного» нормального графика), то их знаки надо отбросить, а сами числа снова упорядочить по возрастанию: 1, 2, 3, 4, 6, 17, 23. А если z_1, $z_2, \ldots, z_i, \ldots, z_m$ — это числа, полученные после такого переупорядочения, то z_i можно нанести на нормальную вероятностную бумагу, где на горизонтали будут последовательные значения

$$50 + 50\left(i - \frac{1}{2}\right)/m \qquad (3А.3)$$

при $i = 1, 2, \ldots, m$.

Основания этого метода подобны тем, что были у полного нормального графика. Мы делим равную единице площадь под половинной нормальной кривой на m равных областей и «ожидаем», что одно из выборочных наблюдений попадет в каждую часть. Наблюдение с номером i, z_i, соответствует на графике площади, накопленной к середине i-й части, которая равна $\left(i - \frac{1}{2}\right)/m$. Для приспособления этих значений к 50—100 долям на шкале ординат вероятностной бумаги мы на самом деле откладываем на ординате значения $50 + 50\left(i - \frac{1}{2}\right)/m$. В сущности, мы накладываем нижнюю половину полного нормального графика на верхнюю и с учетом этого меняем цену делений верхней части вертикальной шкалы. Если наша выборка действительно принадлежит нормальному распределению со средним нуль, то окажется возможным провести хорошую прямую через все множество точек, причем так, чтобы она проходила через точку с координатами (абсцисса, ордината) = (0, 50). Прямая обязана проходить через эту точку, если только верно предположение о нулевом среднем. В этом случае снова разность $x_{84,13} - x_{50}$ дает оценку θ стандартного отклонения нанесенных на график наблюдений. Однако поскольку $x_{50} = 0$ по построению линии, требуется только одно значение $x_{84,13}$ (см. рис. 3А.5).

Причина, по которой снова можно воспользоваться $x_{84,13}$, заключается в том, что если распределение $N(0, \theta^2)$, то

$$P\{-\theta \leqslant x \leqslant 0\} + P\{0 \leqslant x \leqslant \theta\} = P\{|x| \leqslant \theta\}.$$

Следовательно, для теоретической прямой на половинном нормальном графике величина $x_{84,13}$ лежит в θ единицах от значения $x_{50} = 0$. Причем мы *не можем* сказать, что «она лежит в одном стандартном отклонении от среднего», поскольку среднее половинного нормального распределения *не равно* $x_{50} = 0$, да и θ — не стандартное отклонение.

(Примечания: 1. Некоторые специалисты при использовании половинного нормального графика перестраивают интервал 50—100 на нормальной вероятностной бумаге. Для этого они берут преобразование $p' = 2p - 100$, при котором $0 < p' \leqslant 100$, когда $50 \leqslant p \leqslant 100$. В таком случае θ оценивает $x_{68,26}$, поскольку $2\,(84{,}13) - 100 = 68{,}26$. В этом преобразовании нет никакой необходимости, хотя оно и используется повсеместно. Но если все-таки преобразование сделано, то i-е наблюдение в упорядоченном ряду, z_i, должно наноситься на график против

$$p' = 100\left(i - \frac{1}{2}\right)/m, \qquad (3A.4)$$

поскольку теперь размах включает значения от 0 до 100.

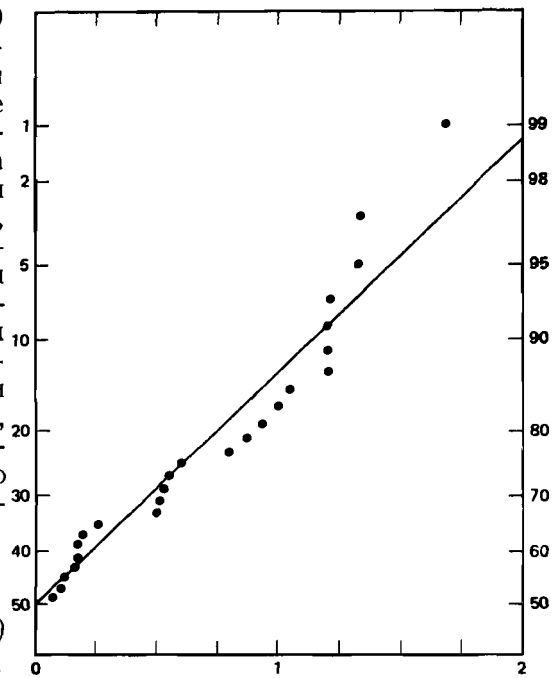


Рис. 3А.5. Половинный нормальный график для остатков из табл. 1.2

2. Вероятностную бумагу, где кривые накопленных распределений превращаются в прямые линии, можно аналогичным образом построить для любого непрерывного распределения. Для этого надо провести горизонтали с равным шагом на вертикальной шкале вероятностей от 0 до 1. В тех точках, где горизонтали пересекут нашу кривую, надо опустить перпендикуляры на произвольную горизонтальную прямую l, которая разделится основаниями этих перпендикуляров на 100 частей, определяемых теми значениями на вертикальной шкале вероятностей, что образуют перпендикуляры. Теперь на горизонтали l получилась новая шкала. Она и должна работать на вертикали вероятностной бумаги. Фактически мы применяем обратное преобразование $x = F^{-1}(y)$, где $y = F(x)$ — накопленная функция распределения при равных интервалах по y. При разметке новой

вертикальной оси мы, как договаривались, умножаем результаты на 100.)

Некоторые модификации стандартных половинных нормальных графиков предлагались и обсуждались Д. Заном в работах: Z a h n D. A. Modifications of and revised critical values for the half-normal plot.— Technometrics, 1975, 17, p. 189—200; Z a h n D. A. An empirical study of the half-normal plot.— Technometrics, 1975, 17, 201—211.

Д. Спаркс составил программу на языке Фортран «Половинные нормальные графики» (см.: S p a r k s D. N. Algorithm AS30.— Applied Statistics, 1970, **19**, p. 192—196, см. также [19] заметку Манфорда: M u n f o r d A. G. Remark ASR5.— Applied Statistics, 1972, **21**, p. 351).

П р и м е р. Вернемся к $m = 25$ остаткам, приведенным в табл. 1.2. Сначала мы упорядочим их с учетом знаков: — 1,68;—1,32; —1,20; —1,20; —0,93; —0,60; — 0,53; — 0,51; — 0,26; — 0,19; — 0,17; — 0,16; — 0,12; 0,08; 0,11; 0,17; 0,50; 0,55; 0,80; 0,87; 1,00; 1,05; 1,20; 1,20; 1,34. Для получения полного нормального графика по этим значениям мы положим $m = 25$ в уравнении (3А.2) и, последовательно меняя $i = 1, 2, \ldots, m$, найдем такие ординаты: 2, 6, 10, 14, 18, 22, 26, 30, 34, 38, 42, 46, 50, 54, 58, 62, 66, 70, 74, 78, 82, 86, 90, 94, 98. Эти значения ординат соответствуют упорядоченным остаткам. Отсюда нижняя точка на рис. 3А.4 имеет абсциссу —1,68, а ордината = 2, т. е. (— 1,68; 2). Читателям полезно иметь в виду, что при таком построении надо пользоваться самой нижней точкой и строить на вероятностной бумаге шкалу, растущую *слева направо* (*правостороннюю*). Если бы шкала была левосторонней, то наша точка имела бы координаты (—1,68; 98), поскольку для левосторонней шкалы справедливо преобразование (100 — правосторонняя шкала). Это характерная особенность вероятностной бумаги, и с ней всегда приходится считаться. Вторая точка на графике имеет координаты (—1,32; 6) и т. д. Линия на графике проведена «на глаз». Она представляет собой попытку грубого приближения к большинству точек, причем с несколько большим весом, придаваемым центральным точкам. Обычно абсцисса, отсекающая на ординате значение 50, должна давать оценку выборочного среднего, но на самом деле наша линия всегда проходит через точку (0, 50), поскольку сумма остатков равна нулю теоретически. (Конечно, на практике, как мы отмечали, могут проявиться ошибки округления.) Приближенная оценка стандартного отклонения равна: $x_{84,13} - x_{50} = 0{,}97 - 0 = 0{,}97$. Она хорошо согласуется с величиной $s = (0{,}7923)^{1/2} = 0{,}89$, приведенной в табл. 1.5. Этот нормальный график отнюдь не нетипичен для графи-

---

[19] Подробности использования различных вариантов шкал на вероятностной бумаге можно найти, например, в гл. 8 монографии: Х а н Г., Ш а п и р о С. Статистические модели в инженерных задачах/Пер. с англ. Под ред. В. В. Налимова.— М.: Мир, 1969, с. 300—351. О применении вероятностной бумаги при анализе остатков в задачах планирования эксперимента см.: Д э н и е л К. Применение статистики в промышленном эксперименте/Пер. с англ. Под ред. Э. К. Лецкого.— М.: Мир, 1979, 300 с. — *Примеч. пер.*

ков, получаемых по нормальным выборкам такого объема. Два самых маленьких и два самых больших значения несколько «выпадают», но это совсем не необычно для остатков метода наименьших квадратов. Очевидно, здесь нет выбросов, которые бы резко отклонялись от графика влево внизу или вправо наверху.

Для построения по тем же остаткам половинного нормального графика сначала надо взять все остатки по модулю и заново их упорядочить: 0,08; 0,11; 0,12; 0,16; . . . ; 1,34; 1,68. Воспользовавшись уравнением (3А.3) для $m = 25$ и $i = 1, 2, \ldots, 25$, найдем соответствующие значения ординат: 51, 53, 55, 57, . . . , 97, 99. Наша первая точка на таком графике имеет координаты (0,08; 51), что видно и на рис. 3А.5. Снова надо использовать эту нижнюю точку и правостороннюю шкалу. Проведенная «на глаз» прямая проходит через точку (0, 50). Она «обгоняет» самые нижние точки (если сразу ясно, что эти точки не связаны с началом координат, то предположение о нулевом среднем в исходной выборке остается под вопросом). Снова мы видим, что график получился не из ряда вон выходящий. Нет выбросов, которые лежали бы правее верхней части прямой. Стандартное отклонение теперь оценивается величиной $x_{84,13}$, чрезвычайно близкой к $s = 0,89$. Она получена в таблице дисперсионного анализа 1.5.

Упражнения

1. Для описания одного конкретного процесса использовалась модель вида

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$$

Было проведено 25 опытов. Вот сводка полученных данных:

1) $n = 25$;

2) $R^2 = 88,5\%$;

3) $s = 0,5915$;

4) $\overline{Y} = 9,42$;

5) **Дисперсионный анализ**

Источник вариации	Число степеней свободы	SS	MS	F
Регрессия	3	56,4875	18,8292	53,81
Остаток	21	7,3477	0,3499	
Общий (скорректированный)	24	63,8352		

6) Расчетное уравнение:

$$\hat{Y} = -2,952790 - 0,073932 X_1 + 0,198999 X_2 + 0,401528 X_3;$$

7) Остатки:

Номер наблюдений	Наблюдаемый Y	Предсказываемый $\hat{Y}$	Остаток	Номер наблюдений	Наблюдаемый Y	Предсказываемый $\hat{Y}$	Остаток
1	10,98	10,86	0,12	14	9,57	9,18	0,39
2	11,13	10,48	0,65	15	10,94	10,61	0,33
3	12,51	11,79	0,72	16	9,58	9,49	0,09
4	8,40	8,73	—0,33	17	10,09	9,49	0,60
5	9,27	9,13	0,14	18	8,11	8,30	—0,19
6	8,73	8,20	0,53	19	6,83	6,51	0,32
7	6,36	6,18	0,18	20	8,88	8,56	0,32
8	8,50	8,40	0,10	21	7,68	7,74	—0,06
9	7,82	8,05	—0,23	22	8,47	9,38	—0,91
10	9,14	9,22	—0,08	23	8,86	9,78	—0,92
11	8,24	9,64	—1,40	24	10,36	11,01	—0,65
12	12.19	11,54	0,65	25	11,08	11,76	—0,68
13	11,88	11,60	0,28				

1) Постройте следующие графики остатков:
   а) общий,
   б) в зависимости от номера наблюдения,
   в) в зависимости от $\hat{Y}$.

2) Проинтерпретируйте графики и сформулируйте выводы относительно нарушения предположений, которые обычно делаются при применении методов множественного регрессионного анализа.

3) Проверьте гипотезу: «слишком мало серий», применяя критерий серий к остаткам во временном порядке, представленном в таблице.

4) Подсчитайте статистику Дарбина—Уотсона и используйте ее для проверки нуль-гипотезы $H_0 : \rho = 0$ против двусторонних альтернатив $H_1 : \rho \neq 0$ в предположении, что $\rho_s = \rho^s$ в принятых обозначениях.

2. Для производителей мыла важное значение имеет такой показатель качества, как высота слоя пены («мылкость»). Проведен эксперимент, в котором варьировалось количество мыла и измерялась высота пены в стандартном лотке при определенной степени перемешивания. Получены следующие данные:

Колнчество мыла в граммах, X	Высота пены, Y	Количество мыла в граммах, X	Высота пены, Y
4,0	38	6,0	53
4,5	42	6,5	61
5,0	45	7,0	62
5,5	51		

Допустим, что модель имеет вид: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$.

1) Получите наилучшее уравнение для предсказания.

2) Проверьте его статистическую значимость.

3) Найдите остатки и оцените, имеются ли какие-нибудь предпосылки, указывающие на то, что здесь лучше подходит более сложная модель.

3. При проведении эксперимента, аналогичного описанному в упражнении 2, исследователь заметил, что «модель бессмысленна, если не принять, что $\beta_0 = 0$, поскольку каждый знает, что если не положить сколько-нибудь мыла в лоток, то не будет и пены». Таким образом, он настаивал на модели $Y = \beta_1 X + \varepsilon$.

Им получены следующие данные (в обозначениях упражнения 2):

X	Y	X	Y
3,5	24,4	6,0	51,4
4,0	32,1	6,5	61,9
4,5	37,1	7,0	66,1
5,0	40,4	7,5	77,2
5,5	43,3	8,0	79,2

1) Исходя из модели, предложенной экспериментатором, найдите наилучшую оценку коэффициента β_1.

2) Используя это уравнение, найдите $\hat{Y}$ для каждого значения X.

3) Исследуйте остатки.

4) Сделайте выводы и сформулируйте рекомендации для экспериментатора.

4. Приведенные ниже данные отражают соотношение между количеством β-эритроидина [20] в водном растворе и прозрачностью раствора по показаниям колориметра.

Концентрация (мг/миллилитр), X	Показания колориметра, Y	Концентрация (мг/миллилитр), X	Показания колориметра, Y
40	69	90	415
50	175	40	72
60	272	60	265
70	335	80	492
80	490	50	180

Постройте регрессионную зависимость на основе модели $Y = \beta_0 + \beta_1 X + \varepsilon$, получите остатки, исследуйте их и рассмотрите вопрос об адекватности модели.

5. Синтетическое волокно благодаря тому, что оно похоже на волос, было признано подходящим для производства париков, но до использования его надо декатировать (подвергнуть усадке). Эта операция проводится в два этапа:

1) волокно вымачивается в разбавленном растворе химиката А, что необходимо для сохранения его блеска на втором этапе;

2) волокно подвергается термической обработке (спеканию) в больших печах при очень высокой температуре в течение одного часа.

Предполагается, что температура, при которой производится термическая обработка волокна, может влиять на эффективность процесса декатировки. Был проведен эксперимент, в ходе которого менялась температура для разных партий волокна. Полученное волокно затем промывалось в чистой воде в течение соответствующего периода времени и сушилось на солнце. Величина дальнейшей усадки Y (в процентах), получающаяся при обработке чистой водой, записывалась вместе со значением температуры для каждой партии:

Номер партии	T	Y	Номер партии	T	Y
1	280	2,1	6	280	3,9
2	250	3,0	7	320	1,3
3	300	3,2	8	300	3.4
4	320	1,4	9	320	2,8
5	310	2,6			

[20] Эритроидии — фирменное название одного из известных наркотиков — кокаина. Он представляет собой кристаллический алколоид состава $C_{16}H_{19}NO_3$, добываемый из растений рода Erythrina. Как все алколоиды, хорошо растворим в воде. Колоритмия водного раствора — экспрессный метод определения концентрации кокаина. Следовательно, в данном примере обсуждается возможность применения уравнения прямой в качестве градуировочного графика в таком анализе. См.: Астахова В. Г. Загадки ядовитых растений. — М.: Лесная промышленность, 1977. — *Примеч. пер.*

1) На основе имеющихся данных с помощью метода наименьших квадратов подберите уравнение линии регрессии

$$\hat{Y} = b_0 + b_1 T.$$

(Примечание. Переходя к кодированной переменной, можно упростить вычисления, но мы об этом вспомнили поздно и поэтому выразим уравнение регрессии в зависимости от исходной переменной T.)

2) Проведите дисперсионный анализ и проверьте:

а) неадекватность модели,

б) значимость уравнения регрессии. Какая доля вариации может быть отнесена за счет регрессии?

3) Какова оценка стандартного отклонения b_1? Постройте 95 %-ный доверительный интервал для коэффициента регрессии β_1.

4) Найдите предсказанные значения отклика $\hat{Y}_i$ и остатки $Y_i - \hat{Y}_i$, соответствующие каждому опыту (партии).

5) В окрестности предсказанного значения $\hat{Y}_0$ при $T_0 = 315$ постройте интервал, который с вероятностью 0,95 накроет новое экспериментальное значение отклика Y.

6) Можно ли применить полученное регрессионное уравнение для предсказания величины Y при $T = 360$? Обсудите свой ответ.

6. Предположим, что в экспериментальной ситуации, описанной в упражнении 5, в каждом опыте записывались данные по концентрации химиката А. Допустим, что вариации этого фактора могут вызывать значительные вариации в отклике, описанные ранее. Измеренные концентрации (фактор C, %) были следующими:

Партия	1	2	3	4	5	6	7	8	8
C	6	6	8	7	9	8	5	9	11

где номера партий те же, что и в упражнении 5.

1) Вычертите график остатков, полученных в упражнении 5, в зависимости от С. Заметили ли вы что-нибудь особенное?

2) Оцените * модель $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ на основании указанных данных, причем $X_1 = (T - 300)/10$ и $X_2 = C - 8$.

3) Проведите дисперсионный анализ и проверьте:

а) неадекватность модели;

б) значимость эффекта включения в модель β_1 и β_2, а не только β_0;

в) значимость включения в модель β_2, а не только β_0 и β_1.

4) Какую долю общей вариации (скорректированной) можно отнести за счет включения эффекта концентрации в нашу модель?

5) Каковы оценки стандартных отклонений коэффициентов $\tilde{b}_1$ и $\tilde{b}_2$? ($\tilde{b}_1$ и $\tilde{b}_2$ — регрессионные коэффициенты в выражении для $\hat{Y}$ в зависимости от исходных переменных T и C.)

6) Выпишите предсказываемые значения $\hat{Y}_i$ и остатки для каждой партии. Обратили ли вы внимание на какую-нибудь особенность?

7) Чему равна оценка дисперсии предсказанного значения $\hat{Y}$ в точке $T = 315$, $C = 8$?

7. Исследуйте остатки из табл. 4.3 следующими способами:

---

* Если вы работаете с кодированными переменными: $X_1 = \dfrac{(T - 300)}{10}$, $X_2 = C - 8$, то можете воспользоваться следующим соотношением:

$$\begin{bmatrix} 9 & -2 & -3 \\ -2 & 46 & 13 \\ -3 & 13 & 29 \end{bmatrix}^{-1} = \frac{1}{10111} \begin{bmatrix} 1165 & 19 & 112 \\ 19 & 252 & -111 \\ 112 & -111 & 410 \end{bmatrix}.$$

1) Примените критерий для гипотезы «слишком мало серий», воспользовавшись критерием серий для остатков, упорядоченных во времени так, как это представлено в таблице.

2) Вычислите статистику Дарбина—Уотсона и используйте ее для проверки нуль-гипотезы $H_0: \rho = 0$ против двусторонней альтернативы $H_1: \rho \neq 0$, предполагая, что $\rho_s = \rho^s$ в принятых обозначениях.

8. Управляющий небольшой конторой, принимающей заказы на почтовые пересылки, нанимает дополнительных сотрудников всякий раз, когда пик заказов превосходит возможности трех его постоянных служащих. Для проверки полезности этого метода он записывает дневные отправления, обеспечиваемые всем наличным персоналом в разные дни различных периодов, как при пиковой нагрузке, так и в затишье. Собранные им данные приведены ниже. Подберите по имеющимся данным уравнение прямой $Y = \beta_0 + \beta_1 X + \varepsilon$ методом наименьших квадратов, затем проверьте адекватность и (если неадекватность не проявится) проверьте всю регрессию. В любом случае исследуйте остатки и сделайте выводы из вашего исследования.

Вам пригодятся следующие полезные факты:

$$n = 13, \qquad \Sigma Y_i = 2990, \qquad \Sigma X_i Y_i = 19120,$$

$$\Sigma X_i = 65, \qquad \Sigma Y_i^2 = 857500, \qquad \Sigma Y_i^2 - (\Sigma Y_i)^2/n = 169800.$$

$$\Sigma X_i^2 = 437,$$

Число посылок, отосланных в день Y	Число служащих X	Число мужчин Z **	Число посылок, отосланных в день Y	Число служащих X	Число мужчин Z **
50	1*		310	6	
110	2*		330	6	
90	2*		340	7	
150	3		360	8	
140	3		380	10	
180	3		360	10	
190	4				

* Постоянный сотрудник (или сотрудники) болел или место было вакантным.

** Этот столбец понадобится только в упражнении 8 из гл. 4. Пока не обращайте на него внимания.

9. Возьмите остатки, которые вы получили в упражнении 11 из гл. 1, и постройте для них графики всеми разумными способами. Сделайте выводы.

10. (Источник: Watts D. G., Bacon D. W. Using an hyperbola as a transition model to fit two-regime straight-line data.— Technometrics, 1974, **16**, p. 369—373.) Набор данных по седиментации (осаждению) осадочной породы трижды использовался в регрессии с тремя различными моделями. В таблице к этому упражнению представлены остатки для трех разных моделей, умноженные на 1000 и записанные в том порядке, в каком данные появлялись, а их фактическое время наблюдений приведено в первом столбце. Постройте для каждого набора остатков графики их зависимостей от времени и проанализируйте их поведение с помощью двустороннего критерия серий. Каковы ваши выводы?

11. (Источник: Andrews D. F. Car accidents — environmental aspects.— Int. Stat. Rev., 1973, **41**, p. 235—239.) Данные в таблице к этому упражнению содержат пятьдесят наблюдений над значениями отклика Y = «число дорожных происшествий со смертельным исходом» и значениями шести возможных предикторов $X_1, X_2, \ldots, X_6$ в сорока девяти штатах и округе Колумбия. На рис. 1 приведена зависимость Y от X = число водителей

автомобилей $\times 10^{-6}$, а на рис. 2 представлены $y = \log Y$ в зависимости от $Z_1 = \log X_1$. Для этих последних данных подобрана модель

$$\hat{y} = -0{,}101 + 0{,}938 Z_1,$$

остатки которой нанесены на рис. 3. При исследовании остатков, отмеченных на последнем рисунке названиями штатов, возникает предположение, что есть какая-то переменная, влияющая на наши данные. Что могло бы стать наиболее логичным кандидатом на включение в регрессию на следующем этапе?

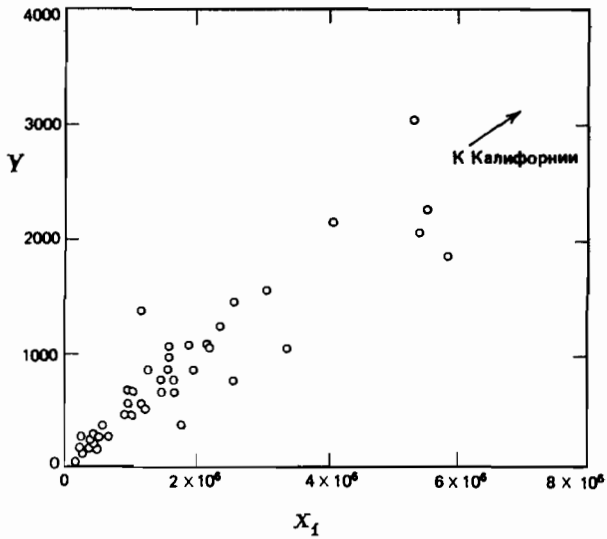


Рис. 1 к упражнению 11. Дорожные происшествия со смертельным исходом и водители по штатам

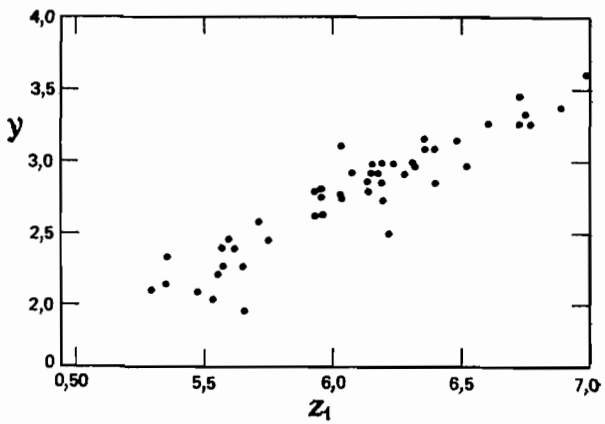


Рис. 2 к упражнению 11. Логарифмы числа дорожных происшествий со смертельным исходом и числа водителей по штатам

Рис. 3 к упражнению 11. Остатки с некоторыми пояснительными ярлычками

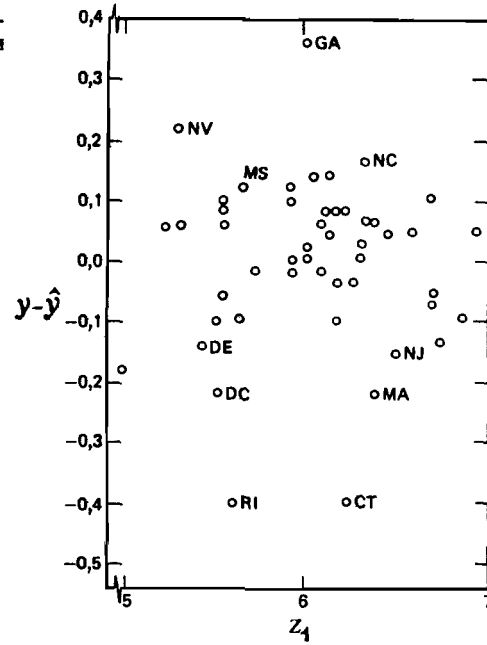


Таблица к упражнению 10. **Три набора остатков в зависимости от времени** t (умноженные на 1000)

Время, t	Набор 1	Набор 2	Набор 3	Время, t	Набор 1	Набор 2	Набор 3
0,5	−19	*	*	44	3	−9	−25
1	−19	0	2	46	3	1	−16
1,5	−18	0	2	48	4	2	−16
2	−28	−10	−8	50	−4	−6	−26
2,5	−27	0	2	52	−9	−5	−25
3	−27	0	2	54	−11	−3	−23
4,5	−45	−20	−15	56	−11	0	−21
6	−23	19	25	58	−17	−6	−27
9	−19	2	10	60	−17	−2	−23
12	−5	12	20	62	−11	5	−17
14	18	23	27	64	−6	3	−20
16	−9	−25	−23	66	−12	−6	−31
18	4	13	16	68	−7	5	−22
20	−3	−5	−4	70	−9	−2	−32
22	−9	−6	−5	72	−8	1	−31
24	−6	4	4	74	4	14	−20
26	8	14	13	76	7	6	−29
28	3	−4	−7	78	8	5	−27
30	7	6	2	80	17	12	−17
32	22	17	11	82	31	19	−6
34	27	8	0	90	45	34	−16
36	33	9	−1	106	−11	−17	11
40	26	1	−24	120	−23	−4	47
42	14	−9	−24	150	−36	−9	33

* В этом случае проведите анализ без использования первого остатка.

Таблица к упражнению 11. **Пятьдесят наблюдений над дорожными происшествиями со смертельным исходом и некоторыми возможными объясняющими предикторными переменными**

Штат *	Y, 1964, смерти	X_1, 1964, водители $\times 10^{-4}$	X_2, 1960, число жителей на квадратную милю	X_3, 1963, длина проселочных дорог в милях	X_4, 1960, больше мужчин, чем женщин	X_5, нормальная максимальная температура января	X_6, 1964, расход топлива на шоссе в галлонах $\times 10^7$
1	2	3	4	5	6	7	8
AL	968	158	64	66	Н (Нет)	62	119
AK	43	11	0,4	5,9	Д (Да)	30	6,2
AZ	588	91	12	33	Д	64	65
AK	640	92	34	73	Н	51	74
CA	4743	952	100	118	Н	65	105
CO	566	109	17	73	Н	42	78
CT	325	167	518	5,1	Н	37	95
DE	118	30	226	3,4	Н	41	20
DC	115	35	12 524	—	Н	44	23
FL	1545	298	91	57	Н	67	216
GA	1302	203	68	83	Н	54	162
ID	262	41	8,1	40	Д	36	29
IL	2207	544	180	102	Н	33	350
IN	1410	254	129	89	Н	37	196
IA	833	150	49	100	Н	30	109
KS	669	136	27	124	Н	42	94
KY	911	147	76	65	Н	44	104
LA	1037	146	72	40	Н	65	109
ME	196	46	31	19	Н	30	37
MD	616	157	314	29	Н	44	113
MA	766	255	655	17	Н	37	166
MI	2120	403	137	95	Н	33	306
MN	841	189	43	110	Н	22	132
MS	648	85	46	59	Н	57	77
MO	1289	234	63	100	Н	40	180
MT	259	38	4,6	72	Д	29	31
NB	450	89	18,4	97	Н	32	61
NV	215	23	2,6	44	Д	40	24
NH	158	37	67	13	Н	32	23
NJ	1071	329	807	21	Н	43	231
NM	387	54	7,8	62	Д	46	48
NY	2745	744	350	84	Н	31	439
NC	1580	226	93	71	Н	51	177
ND	185	38	9,1	102	Д	20	24
OH	2096	530	237	84	Н	41	358
OK	785	137	34	94	Н	46	107
OR	575	108	18	73	Н	45	81
PA	1889	570	252	89	Н	39	353
RI	100	46	812	1,3	Н	38	27
SC	870	122	79	52	Н	61	86
SD	270	40	9	87	Д	23	28
TN	1059	177	85	67	Н	49	135
TX	3006	515	37	196	Н	50	448
UT	295	57	10,8	32	Н	37	38
VT	131	20	42	13	Н	25	15
VA	1050	208	100	50	Н	50	150
WA	730	160	43	59	Д	46	109
WV	467	88	77	32	Н	43	54
WI	1059	207	72	87	Н	26	141
WY	148	22	3,4	67	Д	37	20

* AL — Алабама, AK — Аляска, AZ — Аризона, AK — Арканзас, CA — Калифорния, CO — Колорадо, CT — Коннектикут, DE — Делавэр, DC — Округ Колумбия (где расположена столица США Вашингтон), FL — Флорида, GA — Джорджия, ID — Айдахо, IN — Индиана, IA — Айова, KS — Канзас, KY — Кентукки, LA — Луизиана, IL — Иллинойс, ME — Мэн, MD — Мэриленд, MA — Массачусетс, MI — Мичиган, MN — Миннесота, MS — Миссисипи, MO — Миссури, MT — Монтана, NB — Небраска, NV — Невада, NH — Нью-Гэмпшир, NJ — Нью-Джерси, NM — Нью-Мексико, NY — Нью-Йорк, NC — Северная Каролина, ND — Северная Дакота, OH — Огайо, OK — Оклахома, OR — Орегон, PA — Пенсильвания, RI — Род-Айленд, SC — Южная Каролина, SD — Южная Дакота, TN — Теннесси, TX — Техас, UT — Юта, VT — Вермонт, VA — Виргиния, WA — Вашингтон, WV — Западная Виргиния, WI — Висконсин, WY — Вайоминг. — *Примеч. пер.*

12. Проверьте с помощью выражения $\mathbf{R} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ вычисления для примера, приведенного в начальных абзацах параграфа 3.12.

13. Покажите, что статистика, лежащая в основе критерия M, предложенного Л. Нельсоном (см.: Nelson L. S. The mean square successive difference test.— Journal of Quality Technology, 1980, **12**, p. 174—175), это то же самое, что и статистика Дарбина—Уотсона для модели $Y = \beta_0 + \varepsilon$ без предикторов (т. е. при $k = 0$). Нанесите приведенные Нельсоном процентные точки при $\alpha = 0{,}05$ на график такого типа, как рис. 3.10, или сверьте их с соответствующими числами из табл. 3.2. Попали ли найденные значения на те места, где вы и ожидали их обнаружить? Какой свет проливает это (если вообще проливает) на упрощенный вариант критерия Дарбина—Уотсона, в котором используется только d_U. (Аналогичное сравнение возможно и для $\alpha = 0{,}01$.) Можно ли воспользоваться таблицей Нельсона для приближенного продолжения, рис. 3.10 в сторону больших значений n?

Ответы к упражнениям

1. График остатков в зависимости от $\hat{Y}_i$ показывает, что предположение о равенстве дисперсий нарушено. График остатков говорит еще о наличии от одного до пяти выбросов, требующих дальнейших исследований.

Критерий серий: $n_1 = 15$, $n_2 = 10$, $u = 8$. Следовательно, $\mu = 13$, $\sigma^2 = 5{,}5$, $z = -1{,}919$. Вероятность того, что z столь мало или еще меньше, равна 0,0275. Значит, есть указание на некоторую особенность.

Статистика Дарбина—Уотсона: $k = 3$, $n = 25$, $d = 10{,}2618/7{,}3477 = 1{,}397$, что не убедительно.

2. 1) $\hat{Y} = -2{,}679 + 9{,}5X$.

2) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий	6	651,714		
Регрессия: $b_1 \mid b_0$	1	631,750	631,750	158,33
Остаток	5	19,964	3,99	

Так как $158{,}33 > 6{,}61$, регрессия значима при $\alpha = 0{,}05$.

3) Нет оснований предполагать, что требуется более сложная модель.

3. 1) $b_1 = 9{,}13$, $\hat{Y} = 9{,}13X$.

2)

Номер наблюдения	X	Y	$\hat{Y}$	Остатки
1	3,5	24,4	31,955	—7,555
2	4,0	32,1	36,520	—4,420
3	4,5	37,1	41,085	—3,985
4	5,0	40,4	45,650	—5,250
5	5,5	43,3	50,215	—6,915
6	6,0	51,4	54,780	—3,380
7	6,5	61,9	59,345	2,555
8	7,0	66,1	63,910	2,190
9	7,5	77,2	68,475	8,725
10	8,0	79,2	73,040	6,160

3) График остатков в зависимости от $\hat{Y}$ указывает на пропуск в модели члена β_0.

4) Рекомендуется модель $Y = \beta_0 + \beta_1 X_1 + \varepsilon$, причем область ее применения ограничивается диапазоном X_1 от 3,5 до 8,0, что следует из данных.

Если истинная модель действительно имеет $\beta_0 = 0$, то отклик должен получаться ближе к нулю по мере уменьшения значений переменной X и возрастать с их увеличением.

4. 1) $\hat{Y} = -252{,}298 + 8{,}529X$.

2) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий	9	219 270,5000		
$b_1 \mid b_0$	1	200 772,3188		
Остаток	8	18 498,1812		
неадекватность	4	18 454,6812	4613,6703	424,2455
«чистая» ошибка	4	43,5000	10,8750	

Проверка неадекватности показывает, что модель неадекватна. График остатков в зависимости от $\hat{Y}$ указывает на определенную тенденцию к переходу от отрицательных остатков к положительным при увеличении значений $\hat{Y}$.

Имеется очевидный выброс, а именно: $Y = 415$ при $X = 90$. Эту точку надо проверить.

5. 1) $\hat{Y} = 7{,}950 - 0{,}0179T$.

2) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F расчетное	$F_{0,95}$
Общий	8	6,260			
Регрессия	1	1,452	1,452	2,114	5,59
Остаток	7	4,808	0,687		
неадекватность	3	1,763	0,588	<1	
«чистая» ошибка	4	3,045	0,761		

Регрессия незначима, $R^2 = 23{,}2\ \%$.

3) $s_{b_1} = 0{,}01228$, $-0{,}04689 \leqslant \beta_1 \leqslant 0{,}01119$.

4)

Номер партии	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	Номер партии	Y_i	$\hat{Y}_i$	$Y_i - \hat{Y}_i$
1	2,10	2,95	—0,85	6	3,90	2,95	0,95
2	3,00	3,49	—0,49	7	1,30	2,24	—0,94
3	3,20	2,59	0,61	8	3,40	2,59	0,81
4	1,40	2,24	—0,84	9	2,80	2,24	0,56
5	2,60	2,42	0,18				

Никаких различий в «картинках» остатков.

5) $0{,}193 \leqslant \hat{Y}_0 \leqslant 4{,}453$.

6) Нет. Наклон подобранной линии незначим, и в любом случае 360 далеко выходит за температурный диапазон, что делает рискованным применение данного уравнения.

6. 1) Отрицательные остатки имеют место при низких уровнях концентрации, а положительные — при высоких.

2) $\hat{Y} = 2{,}693374 - 0{,}277361 X_1 + 0{,}365028 X_2$.

3) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F расчетное	$F_{0,95}$
Общий (скорректированный)	8	6,2600			
Регрессия	2	4,7381	2,3690	9,34	5,15
$b_1 \mid b_0$	1	1,4521	1,4521	5,73	5,99
$b_2 \mid b_0, b_1$	1	3,2860	3,2860	12,96	5,99
Остаток	6	1,5219	0,2536		

а) Поскольку нет никаких повторных опытов, определение неадекватности не могло быть выполнено.

б) Модель $\hat{Y} = \bar{Y}$ объясняет 62,41/68,67 = 90,88 % исходного разброса данных, измеренных при Y = 0. Из оставшейся части разбросов модель $\hat{Y} = b_0 + b_1 X_1 + b_2 X_2$ объясняет 75,69 % или в итоге 97,78 % исходного разброса.

в) Добавление β_2 к модели улучшает адекватность, о чем свидетельствует и возрастание величины R^2 от 23,30 до 75,69 %.

4) $R^2 = 75{,}69$ %.

5) Стандартное отклонение [21] $\tilde{b}_1 = 0{,}00795$.

Стандартное отклонение $\tilde{b}_2 = 0{,}10141$.

6)

Номер партии	Y	$\hat{Y}$	$Y - \hat{Y}$	Номер партии	Y	$\hat{Y}$	$Y - \hat{Y}$
1	2,100	2,518	—0,418	6	3,900	3,248	0,652
2	3,000	3,350	—0,350	7	1,300	1,044	0,256
3	3,200	2,693	0,507	8	3,400	3,058	0,342
4	1,400	1,774	—0,374	9	2,800	3,234	—0,434
5	2,600	2,781	—0,181				

7) Дисперсия $\hat{Y}$ (кодированного) = 0,044871.

---

[21] Здесь авторы, видимо, забыли исправить термин «стандартная ошибка» из первого издания на «стандартное отклонение», что противоречит и гл. 1, и условию данного упражнения. Поэтому мы восстановили старую запись.— *Примеч. пер.*

7. 1) Критерий серий: $n_1 = 14$, $n_2 = 11$, $u = 14$. Следовательно, $\mu = 13{,}32$, $\sigma^2 = 5{,}810933$, $z = 0{,}4895$, а область под левым хвостом распределения приблизительно равна 0,688. Нет смысла верить в то, что нам встретилась какая-то неслучайная структура остатков.

2) Статистика Дарбина—Уотсона: $k = 2$, $n = 25$, $d = 2{,}1925{,}4 - d = 1{,}8075$. Различие не значимо. Сериальная корреляция не проявилась.

8. Подобранное уравнение: $\hat{Y} = 43{,}84 + 37{,}23X$.

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
Общий	13	857 500		
b_0	1	687 700		
$b_1 \mid b_0$	1	155 258	155 258	
Остаток	11	14 542	1 322	
неадекватность	6	13 075	2 179	7,41*
«чистая» ошибка	5	1 467	293	

* Неадекватность значима на 5 %-ном уровне.

Неупорядоченный ряд остатков: — 31, — 8, — 28, — 6, — 16, 24, — 3, 43, 63, 36, 18, — 36, — 56.

Выводы. И упорядоченный график остатков, и график самих данных ясно говорят о «квадратичной кривой», к тому же, как мы уже отмечали, есть значимая неадекватность. Нам надо совершенствовать модель. Одним из путей такого совершенствования могло бы стать построение по нашим данным квадратичной кривой относительно X. Если мы это сделаем, то получим уравнение

$$\hat{Y} = -49{,}05 + 83{,}18X - 4{,}07X^2.$$

Дальнейший анализ показывает, что неадекватность не обнаруживается и что вся регрессия весьма значима при $F(2,10) = 237{,}4$. Эта модель объясняет $100\ R^2 = 97{,}94$ % всей вариации относительно среднего. Другой вариант анализа, в котором используется дополнительная информация, см. в упражнении 8 из гл. 4.

9. Какие бы то ни было особенности, настолько сильные, чтобы стоило предпринимать корректировку, не проявились. (Есть, правда, слабая тенденция малых остатков связываться с малыми значениями X-ов, но лишь дальнейшее исследование сможет прояснить этот момент.)

10. См. в указанном источнике с. 372—373, где на рис. 3, 4 и 5 графики, построенные Уаттсом и Беконом, соответствуют нашим графикам для 1-, 2- и 3-го наборов данных. Результаты приложения критерия серий похожи на следующие:

Набор данных	1	2	3	Набор данных	1	2	3
u	9	22	13	n	47	46	46
n_1	19	29	16	z (нижний)	—4,33	—	—2,76
n_2	28	17	30	z (верхний)	—	—	—

Подходящими оказались только два значения z, все прочие не нашли себе места в областях под хвостами распределения, что служит недвусмысленным указанием на наличие положительной сериальной корреляции в первом и третьем наборах данных. Когда этот критерий применяется к данным, неравномерно расположенным в пространстве, результат может оказаться неточным. Однако здесь сильно вытянутые, равномерно расположенные данные гарантируют нужную точность.

11. См. источник, приведенный в этом упражнении.

12. Сверьтесь с параграфом 3.12.

13. Решение не приводится

Глава 4 ● ДВЕ ПРЕДИКТОРНЫЕ ПЕРЕМЕННЫЕ

4.0. ВВЕДЕНИЕ

До сих пор мы подробно рассматривали линейную регрессионную модель первого порядка от одной переменной X

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

и показали, как просто можно выразить анализ в матричных терминах. Обычно в практике требуются более сложные модели. Они возникают во многих задачах, где не обойтись одной единственной независимой (или предикторной) переменной, чтобы лучше понять и (или) лучше предсказать данный отклик. Матричный подход, обсуждавшийся в конце гл. 2, позволяет нам получить общую методику, распространяющую результаты гл. 1 на более сложные линейные модели. В этой главе мы будем применять матричный анализ к линейной модели первого порядка в виде

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

Вернемся к примеру из гл. 1 (с данными из приложения А, см. кн. 2) и добавим теперь к модели еще фактор 6. А чтобы было ясно, какая из переменных модели рассматривается, будем использовать подстрочные индексы исходных переменных. Тогда наша модель примет вид

$$Y = \beta_0 X_0 + \beta_8 X_8 + \beta_6 X_6 + \varepsilon, \qquad (4.0.1)$$

где Y — отклик, или количество пара, расходуемого в месяц; X_0 — фиктивная переменная, которая всегда равна единице; X_8 — среднемесячная температура воздуха (°F); X_6 — число рабочих дней в месяце.

Теперь можно построить следующие матрицы (полные данные для вектора **Y**, а также второго и третьего столбцов матрицы **X** приведены в приложении А и в табл. 4.3):

$$\mathbf{Y} = \begin{bmatrix} 10{,}98 \\ 11{,}13 \\ 12{,}51 \\ 8{,}4 \\ \vdots \\ 10{,}36 \\ 11{,}08 \end{bmatrix}, \qquad \mathbf{X} = \begin{matrix} X_0 & X_8 & X_6 \\ \begin{bmatrix} 1 & 35{,}3 & 20 \\ 1 & 29{,}7 & 20 \\ 1 & 30{,}8 & 23 \\ 1 & 58{,}8 & 20 \\ \vdots & \vdots & \vdots \\ 1 & 33{,}4 & 20 \\ 1 & 28{,}6 & 22 \end{bmatrix} \end{matrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_8 \\ \beta_6 \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \vdots \\ \varepsilon_{24} \\ \varepsilon_{25} \end{bmatrix},$$

где $\mathbf{Y}$ — (25×1)-вектор, $\mathbf{X}$ — (25×3)-матрица, $\boldsymbol{\beta}$ — (3×1)-вектор, $\boldsymbol{\varepsilon}$ — (25×1)-вектор.

Используя результаты из гл. 2, получим МНК-оценки для β_0, β_8 и β_6:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y},$$

где $\mathbf{b}$ — вектор оценок элементов $\boldsymbol{\beta}$ при условии, что $\mathbf{X}'\mathbf{X}$ не вырождена. Тогда

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix} = \left\{ \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ 35{,}3 & 29{,}7 & 30{,}8 & \dots & 28{,}6 \\ 20 & 20 & 23 & \dots & 22 \end{bmatrix} \times \right.$$

$$\times\left\{\begin{bmatrix} 1 & 35{,}3 & 20 \\ 1 & 29{,}7 & 20 \\ 1 & 30{,}8 & 23 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 28{,}6 & 22 \end{bmatrix}\right\}^{-1}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 35{,}3 & 29{,}7 & 30{,}8 & \cdots & 28{,}6 \\ 20 & 20 & 23 & \cdots & 22 \end{bmatrix}\times$$

$$\times\begin{bmatrix} 10{,}98 \\ 11{,}13 \\ \cdot \\ 12{,}51 \\ \vdots \\ 11{,}08 \end{bmatrix}$$

Приведем размеры матриц, построенных выше:

$$[3\times 1]=\{[3\times 25]\,[25\times 3]\}^{-1}\,[3\times 25]\,[25\times 1].$$

Перемножая матрицы в больших фигурных скобках, получим

$$\underset{[3\times 1]}{\begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix}}=\underset{[3\times 3]^{-1}}{\begin{bmatrix} 25{,}00 & 1315{,}00 & 506{,}00 \\ 1315{,}00 & 76323{,}42 & 26353{,}30 \\ 506{,}00 & 26353{,}30 & 10460{,}00 \end{bmatrix}^{-1}}\times$$

$$\times\underset{[3\times 25]}{\begin{bmatrix} 1 & 1 & \cdots & 1 \\ 35{,}3 & 29{,}7 & \cdots & 28{,}6 \\ 20 & 20 & \cdots & 22 \end{bmatrix}}\underset{[25\times 1]}{\begin{bmatrix} 10{,}98 \\ 11{,}13 \\ 12{,}51 \\ \cdot \\ \cdot \\ \cdot \\ 11{,}08 \end{bmatrix}}.$$

Тогда

$$\underset{[3\times 1]}{\begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix}}=\underset{[3\times 3]^{-1}}{\begin{bmatrix} 25{,}00 & 1315{,}00 & 506{,}00 \\ 1315{,}00 & 76323{,}42 & 26353{,}30 \\ 506{,}00 & 26353{,}30 & 10460{,}00 \end{bmatrix}^{-1}}\underset{[3\times 1]}{\begin{bmatrix} 235{,}6000 \\ 11821{,}4320 \\ 4831{,}8600 \end{bmatrix}}.$$

Затем, обращая матрицу $[3 \times 3]$, найдем:

$$\underset{[3\times 1]}{\begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix}} = \underset{[3\times 3]}{\begin{bmatrix} 2{,}778747 & -0{,}011242 & -0{,}106098 \\ & 0{,}146207\times 10^{-3} & 0{,}175467\times 10^{-3} \\ \text{(симметрично)} & & 0{,}478599\times 10^{-2} \end{bmatrix}} \underset{[3\times 1]}{\begin{bmatrix} 235{,}6000 \\ 11821{,}4320 \\ 4831{,}8600 \end{bmatrix}}.$$

Вычисление обратной матрицы можно проверить, перемножая $(\mathbf{X}'\mathbf{X})^{-1}$ на исходную матрицу $(\mathbf{X}'\mathbf{X})$, что должно дать единичную матрицу (3×3). Заметим, что так как обратная матрица (подобно прямой) симметрична, то приводится только ее верхняя треугольная часть. Выполняя перемножение матриц, наконец, получим

$$\underset{[3\times 1]}{\begin{bmatrix} b_0 \\ b_8 \\ b_6 \end{bmatrix}} = \underset{[3\times 1]}{\begin{bmatrix} 9{,}1266 \\ -0{,}0724 \\ 0{,}2029 \end{bmatrix}}.$$

Таким образом, имеем подобранное с помощью метода наименьших квадратов уравнение:

$$\hat{Y} = 9{,}1266 - 0{,}0724X_8 + 0{,}2029X_6.$$

Фактически при проведении этих матричных операций на машине мы не получаем большей точности, чем следуя нашим путем. Основная причина здесь заключается в том, что могут быть большие ошибки округления, если придерживаться такой последовательности операций. Этот вопрос обсуждается в параграфе 5.5.

Для полноты картины приведем алгебраическую форму нормальных уравнений в случае двух независимых переменных:

$$b_0 n + b_1 \sum_{i=1}^{n} X_{1i} + b_2 \sum_{i=1}^{n} X_{2i} = \sum_{i=1}^{n} Y_i,$$

$$b_0 \sum_{i=1}^{n} X_{1i} + b_1 \sum_{i=1}^{n} X_{1i}^2 + b_2 \sum_{i=1}^{n} X_{1i}X_{2i} = \sum_{i=1}^{n} X_{1i}Y_i,$$

$$b_0 \sum_{i=1}^{n} X_{2i} + b_1 \sum_{i=1}^{n} X_{2i}X_{1i} + b_2 \sum_{i=1}^{n} X_{2i}^2 = \sum_{i=1}^{n} X_{2i}Y_i.$$

Мы получили подобранное выше уравнение с помощью простых регрессионных вычислений. То же уравнение можно получить еще и с помощью последовательности простых линейных регрессий. Хотя практически это не лучший путь, рассмотрим в учебных целях, как этого добиться. Прежде чем исследовать подбор уравнения в параграфе 4.2, обсудим альтернативную процедуру.

4.1. СВЕДЕНИЕ МНОЖЕСТВЕННОЙ РЕГРЕССИИ С ДВУМЯ ПРЕДИКТОРНЫМИ ПЕРЕМЕННЫМИ К ПОСЛЕДОВАТЕЛЬНОСТИ ПРОСТЫХ ЛИНЕЙНЫХ РЕГРЕССИЙ

В параграфе 4.0 мы применили метод наименьших квадратов и получили уравнение

$$\hat{Y} = 9{,}1266 - 0{,}0724X_8 + 0{,}2029X_6.$$

Другой путь получения того же решения следующий.

1. Строится регрессия Y на X_8. Эта линейная регрессия уже была получена в гл. 1, и окончательное уравнение имело вид

$$\hat{Y} = 13{,}6215 - 0{,}0798X_8.$$

Оно не предсказывает Y точно (см. табл. 1.2). Включение новой переменной, скажем X_6 (число рабочих дней), в предсказывающее уравнение может значительно улучшить предсказание. Чтобы достичь этого, мы хотим соотнести число рабочих дней с необъясненным разбросом данных после того, как исключен («снят») эффект температуры воздуха. Однако если вариации температуры воздуха так или иначе связаны с вариацией, обусловленной числом рабочих дней, то следует прежде всего внести поправку на это. Таким образом, нам предстоит определить зависимость между необъясненным разбросом в количестве используемого пара (после того, как исключено влияние температуры воздуха) и остаточным разбросом в числе рабочих дней (после исключения из него эффекта температуры воздуха).

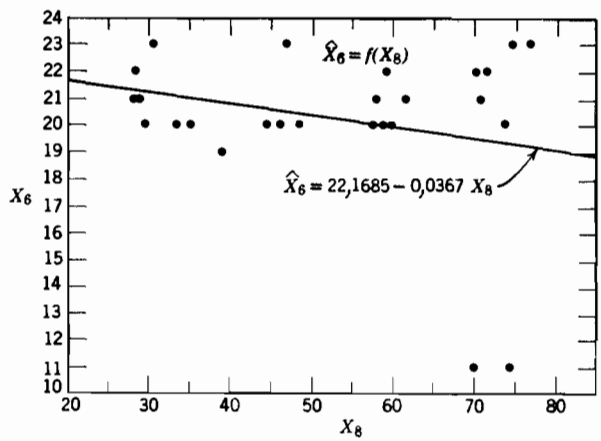


Рис. 4.1. Прямая метода наименьших квадратов для регрессии X_6 на X_8

2. Строится регрессия X_6 на X_8; вычисляются остатки $X_{6i} - \hat{X}_{6i}$, $i = 1, 2, \ldots, n$. График зависимости X_6 от X_8 показан на рис. 4.1. Используя обозначения и методы из гл. 2, получим оценки коэффи-

циентов регрессии:

$$\begin{bmatrix} b_0 \\ b_8 \end{bmatrix} = \left\{ \begin{bmatrix} 1 & 1 & \dots & 1 \\ 35{,}3 & 29{,}7 & \dots & 28{,}6 \end{bmatrix} \begin{bmatrix} 1 & 35{,}3 \\ 1 & 29{,}7 \\ \cdot & \cdot \\ \cdot & \cdot \\ \cdot & \cdot \\ 1 & 28{,}6 \end{bmatrix} \right\}^{-} \times$$

$$\times \begin{bmatrix} 1 & 1 & \dots & 1 \\ 35{,}3 & 29{,}7 & \dots & 28{,}6 \end{bmatrix} \begin{bmatrix} 20 \\ 20 \\ \cdot \\ \cdot \\ \cdot \\ 22 \end{bmatrix}.$$

Отсюда $\hat{X}_6 = 22{,}1685 - 0{,}0367 X_8$, а остатки показаны в табл. 4.1.

Таблица 4.1. **Остатки:** $X_{6i} - \hat{X}_{6i}$

Наблюдение с номером i	X_{6i}	$\hat{X}_{6i}$	$X_{6i} - \hat{X}_{6i}$	Наблюдение с номером i	X_{6i}	$\hat{X}_{6i}$	$X_{6i} - \hat{X}_{6i}$
1	20	20,87	—0,87	14	19	20,73	—1,73
2	20	21,08	—1,08	15	23	20,45	2,55
3	23	21,04	1,96	16	20	20,39	—0,39
4	20	20,01	—0,01	17	22	19,99	2,01
5	21	19,92	1,08	18	22	19,60	2,40
6	22	19,55	2,45	19	11	19,60	—8,60
7	11	19,44	—8,44	20	23	19,44	3,56
8	23	19,36	3,64	21	20	19,53	0,47
9	21	19,58	1,42	22	21	20,04	0,96
10	20	20,06	—0,06	23	20	20,53	—0,53
11	20	20,47	—0,47	24	20	20,94	—0,94
12	21	21,11	—0,11	25	22	21,12	0,88
13	21	21,14	—0,14				

Отметим, что два остатка (— 8,44) и (— 8,60) имеют абсолютные значения, существенно большие, чем остальные. Они приходятся на те месяцы, когда число рабочих дней было необычно мало,— одиннадцать в каждом случае. Мы можем, конечно, сделать заключение, что это «выбросы» и что месяцы со столь малым числом рабочих дней не должны даже рассматриваться при анализе. Однако если мы хотим получить удовлетворительное уравнение для предсказания, пригодное для *всех* месяцев, независимо от числа рабочих дней, то важно учесть эти частные результаты и построить уравнение, позволяющее включить эту информацию. Как можно видеть из данных рис.4.1 и табл. 4.2,

Таблица 4.2. **Отклонения $\hat{Y}_i = f(X_8)$ и $\hat{X}_{6i} = f(X_8)$ от Y_i и X_{6i} соответственно**

Наблюдение с номером i	$Y_i - \hat{Y}_i$	$X_{6i} - \hat{X}_{6i}$	Наблюдение с номером i	$Y_i - \hat{Y}_i$	$X_{6i} - \hat{X}_{6i}$
1	0,17	—0,87	14	—0,93	—1,73
2	—0,12	—1,08	15	1,05	2,55
3	1,34	1,96	16	—0,17	—0,39
4	—0,53	—0,01	17	1,20	2,01
5	0,55	1,08	18	0,08	2,40
6	0,80	2,45	19	—1,20	—8,60
7	—1,32	—8,44	20	1,20	3,56
8	1,00	3,64	21	—0,19	0,47
9	—0,16	1,42	22	—0,51	0,96
10	0,11	—0,06	23	—1,20	—0,53
11	—1,68	—0,47	24	—0,60	—0,94
12	0,87	—0,11	25	—0,26	0,88
13	0,50	—0,14			

если игнорировать эти месяцы, то кажущееся влияние числа рабочих дней на отклик будет слабым. Это может быть *не* следствием несущественности фактора, а следствием того, что фактически наблюдаемая вариация его значений была слишком мала, чтобы фактор смог проявить сколько-нибудь ощутимое влияние на отклик. Если некий фактор значимо влияет на отклик в одном исследовании и незначимо в другом, то вполне возможно, что варьирование в первом множестве данных было в более широком диапазоне, чем во втором. В этом, между прочим, один из недостатков использования производственных данных в том виде, «как они поступают» [1]. Часто размах варьирования фактора так мал, что влияние на отклик не обнаруживается, даже когда фактор в более широких интервалах имеет отчетливый эффект. Поэтому планируемый эксперимент, в котором уровни назначаются шире, чем при нормальной работе объекта, часто обнаруживает эффекты, не замеченные ранее.

3. Теперь строим регрессию $Y - \hat{Y}$ на $X_6 - \hat{X}_6$, подбирая модель

$$(Y_i - \hat{Y}_i) = \beta (X_{6i} - \hat{X}_{6i}) + \varepsilon_i.$$

Заметим, что член β_0 отсутствует в этой модели первого порядка, так как мы используем два множества остатков, суммы которых равны нулю, и, таким образом, линия должна пройти через начало координат. (Если включить член β_0, то мы найдем, что $b_0 = 0$ в любом случае.) Для удобства оба множества остатков, используемых как данные, извлечены из табл. 1.2 и 4.1 и представлены в табл. 4.2. График этих остатков показан на рис. 4.2.

Используя формулы из гл. 1, найдем

$$b = \frac{\Sigma (Y_i - \hat{Y}_i)(X_{6i} - \hat{X}_{6i})}{\Sigma (X_{6i} - \hat{X}_{6i})^2} = \frac{42{,}0821}{208{,}8523} = 0{,}2015.$$

[1] Речь идет о данных, получаемых по итогам пассивных наблюдений за ходом производственного процесса.— *Примеч. пер.*

Тогда уравнение прямой будет иметь вид

$$\widehat{(Y-\hat{Y})} = 0{,}2015\,(X_6 - \hat{X}_{6i}).$$

В скобки можно подставить $\hat{Y}$ и $\hat{X}_6$ как функции $\hat{X}_8$ и, перенеся члены, содержащие $\hat{Y}$, в левую часть, получить полную зависимость в виде $\hat{Y} = \hat{Y}\,(X_6,\ X_8)$:

$$[\hat{Y} - (13{,}6215 - 0{,}0798X_8)] = 0{,}2015\,[X_6 - (22{,}1685 - 0{,}0367X_8)]$$

или

$$\hat{Y} = 9{,}1545 - 0{,}0724X_8 + 0{,}2015X_6.$$

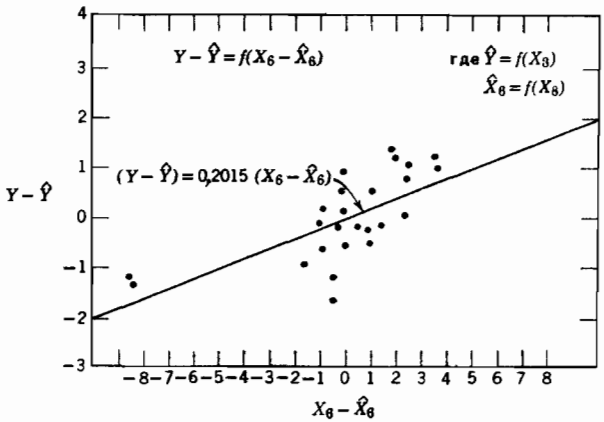


Рис. 4.2. График зависимости для остатков из табл. 4.2

Прежний результат был

$$\hat{Y} = 9{,}1266 - 0{,}0724X_8 + 0{,}2029X_6.$$

Теоретически эти результаты должны быть идентичными; практически, как мы можем видеть, они немного расходятся вследствие ошибок округления. Игнорируя пока ошибки округления, покажем геометрически на простом примере, что оба метода должны давать одинаковые результаты. (Конец этого параграфа при первом чтении можно пропустить.)

Геометрическая интерпретация

Рассмотрим пример. Предположим, мы имеем $n = 3$ наблюдения отклика Y, а именно Y_1, Y_2 и Y_3, которые получены в трех множествах условий (X_1, Z_1), (X_2, Z_2), (X_3, Z_3). Тогда, взяв трехмерное пространство с осями координат, обозначенными 1, 2 и 3 и с началом координат в нуле, можно построить точки:

$$\mathrm{Y} \equiv (Y_1,\ Y_2,\ Y_3), \quad \mathrm{X} \equiv (X_1,\ X_2,\ X_3) \text{ и } \mathrm{Z} \equiv (Z_1,\ Z_2,\ Z_3).$$

Геометрическая интерпретация регрессии выглядит следующим образом. Чтобы получить регрессию Y на X, опускаем перпендикуляр YP на OX. Координаты точки P — это предсказанные значения $\hat{Y}_1$, $\hat{Y}_2$, $\hat{Y}_3$. Квадрат длины отрезка OP, OP^2,— это сумма квадратов, обусловленная регрессией, OY^2 — полная сумма квадратов и YP^2 — остаточная сумма квадратов. По теореме Пифагора $OP^2 + YP^2 = OY^2$, что соответствует разложению суммы квадратов в дисперсионном анализе (см. рис. 4.3).

Если мы построим параллелограмм, который имеет диагональ OY, а стороны OP и PY, то получится параллелограмм $OP'YP$. Тогда координаты P' будут значениями остатков для регрессии переменной Y на переменную X. В векторной форме можно записать:

$$\overrightarrow{OP} + \overrightarrow{OP'} = \overrightarrow{OY},$$

или в «статистической» векторной записи

$$\hat{\mathbf{Y}} + (\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}.$$

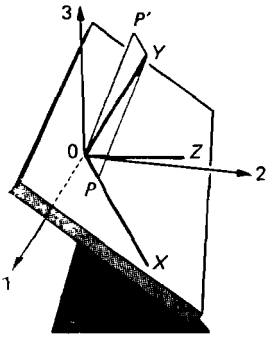


Рис. 4.3. Геометрическая интерпретация регрессии Y на X

Этот результат справедлив и для общего n-мерного случая. (Мы рассматриваем $n = 3$ всего лишь из-за возможности построить картинку.)

Положим, что мы хотим построить регрессию для переменной Y на переменные X и Z одновременно. Прямые OX и OZ определяют плоскость в трехмерном пространстве. Мы опускаем перпендикуляр YT на эту плоскость. Тогда координаты точки T есть значения $\hat{Y}_1$, $\hat{Y}_2$, $\hat{Y}_3$ для *этой* регрессии. OT^2 — сумма квадратов, обусловленная регрессией, YT^2 — остаточная сумма квадратов и OY^2 — полная сумма квадратов. Снова по теореме Пифагора $OY^2 = OT^2 + YT^2$, что дает разложение суммы квадратов, которое мы видим в таблице дисперсионного анализа. Построение параллелограмма $OT'YT$ с диагональю OY и сторонами OT, TY дает OT' — вектор остатков этой регрессии, а координаты T' дают остатки $\{(Y_1 - \hat{Y}_1), (Y_2 - \hat{Y}_2), (Y_3 - \hat{Y}_3)\}$ регрессии Y на X и Z одновременно. Снова в векторной записи

$$\overrightarrow{OT} + \overrightarrow{OT'} = \overrightarrow{OY},$$

или в «статистической» векторной записи

$$\hat{\mathbf{Y}} + (\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{Y}$$

для этой регрессии (см. рис. 4.4).

Как мы видели выше в численном примере, те же самые окончательные остатки получаются (если игнорировать ошибку округления), когда мы строим регрессии: (1) Y на X и (2) Z на X, а затем регрес-

сию остатков (1) на остатки (2). Справедливость этого можно показать геометрически. На рис. 4.5 построены три параллелограмма в трехмерном пространстве:

1) $OP'YP$ для регрессии Y на X;
2) $OQ'ZQ$ для регрессии Z на X;
3) $OT'YT$ для регрессии Y одновременно на X и Z.

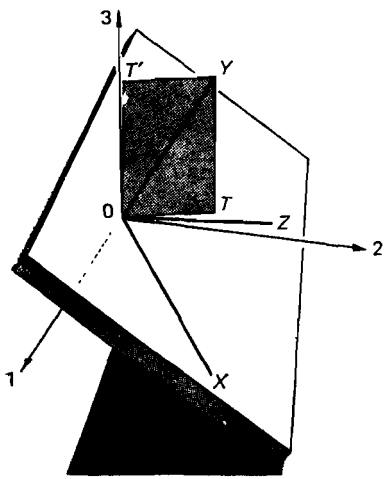


Рис. 4.4. Геометрическая интерпретация регрессии Y на X и Z

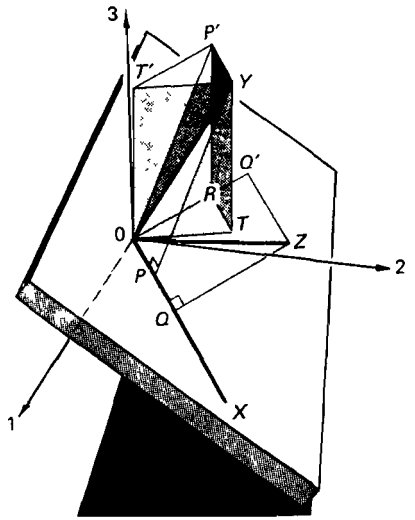


Рис. 4.5. Построение регрессии Y на X и Z можно рассматривать и как двухступенчатую процедуру, описанную в тексте

Теперь регрессия остатков (1) на остатки (2) достигается с помощью перпенднкуляра из P' на OQ'. Положим, точка встречи есть точка R. Тогда прямая из O, параллельная RP', длиною RP' будет остаточным вектором двухступенчатой регрессии Y на X и на Z. Однако точки O, Q', Z, P, Q, X и T лежат в плоскости π, определяемой прямыми OZ и OX. Так получается точка R. Поскольку $OP'YP$— параллелограмм, а отрезки $P'R$ и YT — перпендикуляры к плоскости π, то $P'R = YT$ по длине. Из того, что $TY = OT'$, следует, что $OT' = RP'$. Однако OT', RP' и TY все параллельны друг другу и перпендикулярны к плоскости π. Следовательно, $OT'P'R$— параллелограмм, откуда вытекает, что $\overrightarrow{OT'}$ — вектор остатков для двухступенчатой регрессии. Поэтому результаты для регрессии Y на X и Z, получаемые независимо двумя методами, должны быть эквивалентны. Таким образом, мы можем видеть, что «плоскостная» регрессия Y на X и Z одновременно может рассматриваться как совокупность последовательных линейных регрессий:

1) Y на X;
2) Z на X;

3) остатков (1) на остатки (2).

Если поменять X и Z ролями, то получится то же самое. Все линейные регрессии могут быть разложены, таким образом, на серии простых регрессий.

4.2. ИССЛЕДОВАНИЕ УРАВНЕНИЯ РЕГРЕССИИ

Чем полезно уравнение $\hat{Y}=f(X_8,X_6)$?

Взяв данные из гл. 1 и 2, будем рассматривать уравнение, найденное для $\hat{Y}$, как функцию от X_8 и X_6. Мы можем вычислить остатки,

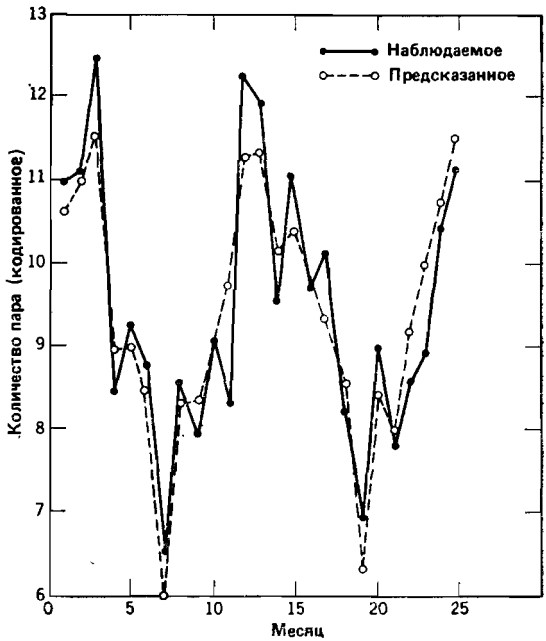


Рис. 4.6. Количество пара, используемого на заводе в месяц

используя уравнение и результаты опытов. Эти остатки приведены в табл. 4.3. Имеем следующий дисперсионный анализ для регрессии:

ANOVA

Источник вариации	Число степеней свободы	SS	MS	F
Общий (нескорректированный)	25	2284,1102		
Среднее (b_0)	1	2220,2944		
Общий (скорректированный)	24	63,8158		
Регрессия | b_0	2	54,1871	27,0936	61,8999
Остаток	22	9,6287	0,4377	

Таблица 4.3. **Остатки из модели $\hat{Y} = f(X_8, X_6)$ для расхода пара**

Номер наблюдения	X_8	X_6	Y	$\hat{Y}$	Остатки
1	35,3	20	10,98	10,63	0,35
2	29,7	20	11,13	11,03	0,10
3	30,8	23	12,51	11,56	0,95
4	58,8	20	8,40	8,93	—0,53
5	61,4	21	9,27	8,94	0,33
6	71,3	22	8,73	8,43	0,30
7	74,4	11	6,36	5,97	0,39
8	76,7	23	8,50	8,24	0,26
9	70,7	21	7,82	8,27	—0,45
10	57,5	20	9,14	9,02	0,12
11	46,4	20	8,24	9,82	—1,58
12	28,9	21	12,19	11,29	0,90
13	28,1	21	11,88	11,35	0,53
14	39,1	19	9,57	10,15	—0,58
15	46,8	23	10,94	10,40	0,54
16	48,5	20	9,58	9,67	—0,09
17	59,3	22	10,09	9,30	0,79
18	70,0	22	8,11	8,52	—0,41
19	70,0	11	6,83	6,29	0,54
20	74,5	23	8,88	8,40	0,48
21	72,1	20	7,68	7,96	—0,28
22	58,1	21	8,47	9,18	—0,71
23	44,6	20	8,86	9,96	—1,10
24	33,4	20	10,36	10,77	—0,41
25	28,6	22	11,08	11,52	—0,44
			235,60		$\Sigma (Y_i - \hat{Y}_i) = 0$
			$\bar{Y} = 9{,}424$		$\Sigma (Y_i - \hat{Y}_i)^2 = 9{,}6432$

При α-риске, равном 0,05, МНК-уравнение служит хорошим «предсказателем», расчетное значение $F = 61{,}8999$ для регрессии больше, чем табличное $F\ (2;\ 22;\ 0{,}95) = 3{,}44$.

График наблюденных значений Y и предсказанных $\hat{Y}$ изображен на рис. 4.6, где показано, что подобранная модель хорошо предсказывает месячное использование пара. Однако было ли полезным включение в модель X_6?

Что происходит при добавлении второго предиктора X_6?

Существует несколько полезных критериев, которые можно применять для ответа на этот вопрос и теперь мы их обсудим.

Квадрат множественного коэффициента корреляции R^2. Квадрат множественного коэффициента корреляции R^2 определяется (см. уравнение (2.6.11)) как

$$R^2 = \frac{\text{сумма квадратов, обусловленная регрессией}}{\text{полная (скорректированная) сумма квадратов}}.$$

Его часто представляют в процентах, $100\,R^2$; чем он больше, тем лучше

подобранное уравнение объясняет вариацию в данных. Мы можем сравнить величины R^2 на каждой стадии регрессионной задачи.

Стадия 1. $\hat{Y} = f(X_8)$.

Уравнение регрессии	$100R^2$
$\hat{Y} = 13{,}6230 - 0{,}0798X_8$	71,44 % (см. параграф 1.4)

Стадия 2. $\hat{Y} = (X_8, X_6)$.

Уравнение регрессии	$100R^2$
$\hat{Y} = 9{,}1266 - 0{,}0724X_8 + 0{,}2029X_6$	84,89 %

Таким образом, мы видим значительный рост R^2. Однако эту статистику следует применять с осторожностью, поскольку можно всегда сделать $R^2 = 1$, как показано в параграфе 2.6.

Если число наблюдений значительно больше, чем число X-переменных, которые потенциально могут быть рассмотрены, то добавление новой переменной всегда будет повышать R^2, но не обязательно будет повышаться точность оценки отклика. Это происходит потому, что уменьшение остаточной суммы квадратов может быть меньшим, чем уменьшение величины первоначального остаточного среднего квадрата. Кроме того, так как из остаточных степеней свободы исключается одна, средний квадрат может оказаться даже больше. Подобный пример содержится в приложении Б (см. кн. 2), которое мы еще не обсуждали. Можно провести следующее сравнение:

R^2	Номера факторов в регрессионной модели	Остаток		
		сумма квадратов	число степеней свободы	средний квадрат
98,23	1, 2, 3	48,11	9	5,35
98,24	1, 2, 3, 4	47,86	8	5,98

Мы видим, что, хотя в регрессионную модель был включен дополнительный фактор, остаточный средний квадрат увеличился, так как дополнительный фактор уменьшил остаточную сумму квадратов на $48{,}11 - 47{,}86 = 0{,}25 < 5{,}35$ при потере одной степени свободы. Величина R^2 в результате повысилась, правда, незначительно.

Оценка стандартной ошибки, s. Остаточный средний квадрат s^2 есть оценка для $\sigma^2_{Y \cdot X}$ — дисперсии относительно регрессии. До и после включения фактора в модель мы можем проверить, что

$$s = \sqrt{\text{остаточный средний квадрат}}.$$

Исследование этой статистики показывает, что чем она меньше, тем лучше, тем более точными будут предсказания. Но поскольку s можно сделать равным нулю, включая в модель достаточно параметров

(точно так же, как можно R^2 сделать равным единице), этот критерий тоже следует использовать с осторожностью. Уменьшение s желательно, если только почти нет повторений и остается много степеней свободы для ошибки. В нашем примере на стадии 1

$$s = \sqrt{0{,}7923} = 0{,}89,$$

на стадии 2

$$s = \sqrt{0{,}4377} = 0{,}66.$$

Отсюда вывод, что включение X_6 уменьшило s и увеличило точность оценивания.

Оценка стандартной ошибки s в процентах от среднего отклика [2]. Другой подход к оценке уменьшения s — это рассмотрение ее относительно отклика. В нашем примере на стадии 1 s в процентах от среднего $\bar{Y}$ составляет

$$0{,}89/9{,}424 = 9{,}44\ \%.$$

На стадии 2 s в процентах от среднего $\bar{Y}$ есть

$$0{,}66/9{,}424 = 7{,}00\ \%.$$

Следовательно, включение X_6 уменьшает стандартную ошибку оценки до величины порядка 7 % среднего отклика. Удовлетворителен ли такой уровень точности — это вопрос для экспериментатора, и его он должен решать на основе априорных знаний и личного опыта.

Последовательный F-критерий (показывающий влияние X_6, когда X_8 уже включен в уравнение). Этот метод, оценивающий значение X_6 как дополнительного фактора в модели $\hat{Y} = f(X_8)$, состоит в разложении суммы квадратов, обусловленной регрессией, на следующие части:

Дисперсионный анализ

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий (нескорректированный)	25	2284,1102		
Среднее (b_0)	1	2220,2944		
Общий (скорректированный)	24	63,8158		
Регрессия $\mid b_0$	2	54,1871	27,0936	61,8999
обусловленный $b_8 \mid b_0$	1	45,5924	45,5924	104,1636
обусловленный $b_6 \mid b_8, b_0$	1	8,5947	8,5947	19,6361
Остаток	22	9,6287	0,4377	

Поскольку 19,6361 превышает $F(1; 22; 0{,}95) = 4{,}30$, включение X_6 имело смысл. Этот F-критерий обычно называют «последовательным F-критерием» (см. параграф 2.9).

Проверка по частному F-критерию (см. параграф 2.9). Еще один путь оценки значения X_8 — это рассмотрение порядка включения

[2] Эту величину называют еще коэффициентом вариации.— *Примеч. пер.*

двух факторов в процедуре метода наименьших квадратов. Например, можно поставить следующие вопросы:

1. Если мы введем фактор X_6 в уравнение первым, то как он будет влиять?

2. Если X_6 был использован первым, то как будет влиять X_8 при включении его в регрессию?

Ответы на эти вопросы дают вычисления, приведенные выше, но выполненные в обратном порядке. Результаты таковы:

Дисперсионный анализ

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий (нескорректированный)	25	2284,1102		
Среднее (b_0)	1	2220,2944		
Общий (скорректированный)	24	63,8158		
Регрессия $\mid b_0$	2	54,1871	27,0936	61,8999
обусловленный $b_6 \mid b_0$	1	18,3424	18,3424	41,9063
обусловленный $b_8 \mid b_6, b_0$	1	35,8447	35,8447	81,8933
Остаток	22	9,6287	0,4377	

Заметим, что влияние X_6 более велико в данном случае, чем после включения X_8. Заметим также, что это отражается в наблюдаемых значениях F для X_8 в двух вариантах, т. е.:

вариант 1: 104,1636;

вариант 2: 81,8933.

Однако в обоих случаях X_8 — все же более важная переменная, так как ее влияние на уменьшение остаточной суммы квадратов наибольшее независимо от порядка включения переменных.

Стандартная ошибка b_i

Используя результат, полученный в параграфе 2.6, найдем матрицу дисперсий-ковариаций для $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$. Отсюда дисперсия $b_i = V(b_i) = c_{ii}\sigma^2$, где c_{ii} — диагональный элемент $(\mathbf{X}'\mathbf{X})^{-1}$, соответствующий переменной с номером i.

Ковариация b_i и b_j равна $c_{ij}\sigma^2$, где c_{ij} — недиагональный элемент $(\mathbf{X}'\mathbf{X})^{-1}$, соответствующий пересечению i-й строки и j-го столбца или j-й строки и i-го столбца, так как $(\mathbf{X}'\mathbf{X})^{-1}$ симметрична. Поэтому стандартная ошибка b_i есть $\sigma\sqrt{c_{ii}}$. Например, используя данные со с. 247 и 258, получим следующую оценку стандартной ошибки b_8:

$$\text{оценка дисперсии } b_8 = s^2 c_{88} = (0{,}4377)\,(0{,}146207 \times 10^{-3}) =$$

$$= 0{,}639948 \times 10^{-4}.$$

Отсюда[3] оценка стандартной ошибки $b_8 = \sqrt{\text{оценка дисперсии } b_8} =$

$$= \sqrt{0{,}639948 \times 10^{-4}} = 0{,}008000.$$

[3] В оригинале отсутствует слово «оценка», что искажает смысл.— *Примеч. пер.*

Доверительные пределы для «истинного» среднего значения Y при заданных значениях X-ов

Предсказанное значение $\hat{Y} = b_0 + b_1X_1 + \ldots + b_pX_p$ есть оценка для

$$E(Y) = \beta_0 + \beta_1X_1 + \ldots + \beta_pX_p.$$

Дисперсия величины $\hat{Y}$, т. е. $V[b_0 + b_1X_1 + \ldots + b_pX_p]$, есть

$$V(b_0) + X_1^2V(b_1) + \ldots + X_p^2V(b_p) + 2X_1\operatorname{cov}(b_0, b_1) + \ldots$$
$$\ldots + 2X_{p-1}X_p\operatorname{cov}(b_{p-1}, b_p).$$

Это выражение можно переписать совсем компактно в матричных обозначениях, полагая $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$:

$$V(\hat{Y}) = \sigma^2(\mathbf{X}_0'\mathbf{C}\mathbf{X}_0) =$$
$$= \sigma^2 \{1\ X_1\ \ldots\ X_p\} \begin{bmatrix} c_{00} & c_{01} & \ldots & c_{0p} \\ c_{10} & c_{11} & \ldots & c_{1p} \\ \vdots & & & \vdots \\ c_{p1} & & & c_{pp} \end{bmatrix} \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ X_p \end{bmatrix}.$$

Отсюда $(1-\alpha)$-доверительные пределы для «истинного» среднего значения Y при X_0 получаются из выражения

$$\hat{Y} \pm t\left\{(n-p-1),\quad 1-\frac{1}{2}\alpha\right\} \cdot s\sqrt{\mathbf{X}_0'\mathbf{C}\mathbf{X}_0}.$$

Например, в точке X-пространства с координатами ($X_8 = 32$; $X_6 = 22$) получается следующая дисперсия $\hat{Y}$:

$$\operatorname{var}(\hat{Y}) = s^2(\mathbf{X}_0'\mathbf{C}\mathbf{X}_0) = (0{,}4377)\,(1,\ 32,\ 22) \times$$
$$\times \begin{bmatrix} 2{,}778747 & -0{,}011242 & -0{,}106098 \\ -0{,}011242 & 0{,}146207 \times 10^{-3} & 0{,}175467 \times 10^{-3} \\ -0{,}106098 & 0{,}175467 \times 10^{-3} & 0{,}478599 \times 10^{-2} \end{bmatrix} \times$$
$$\times \begin{bmatrix} 1 \\ 32 \\ 22 \end{bmatrix} = (0{,}4377)\,(0{,}104140) = 0{,}045582.$$

95 %-ные доверительные пределы «истинного» среднего значения Y при $X_8 = 32$, $X_6 = 22$ будут:

$$\hat{Y} \pm t(22;\ 0{,}975)\, s\sqrt{\mathbf{X}_0'\mathbf{C}\mathbf{X}_0} = 11{,}2736 \pm (2{,}074)\ (\mathbf{0{,}213499}) =$$
$$= 11{,}2736 \pm 0{,}4418 = \{10{,}8318;\ \mathbf{11{,}7154}\}.$$

Эти пределы интерпретируются следующим образом. Пусть несколько

раз повторяются выборки Y того же объема, что и исходная, и при тех же фиксированных значениях (X_8, X_6), которые использовались при определении коэффициентов подобранного выше уравнения. Если всякий раз строить 95 %-ные доверительные интервалы для среднего значения Y при $X_8 = 32$, $X_6 = 22$, то 95 % из этих интервалов будут содержать «истинное» среднее значение Y в данных условиях. С практической точки зрения с вероятностью 0,95 справедливо утверждение, что «истинное» среднее значение Y при $X_8 = 32$, $X_6 = 22$ лежит между 10,8318 и 11,7154.

Доверительные пределы для среднего из g наблюдений при заданных значениях X

Эти пределы вычисляются из соотношения

$$\hat{Y} \pm t\left(v,\ 1-\frac{1}{2}\alpha\right) s \sqrt{1/g + \mathbf{X}_0' \mathbf{C} \mathbf{X}_0}.$$

Например, 95 %-ные доверительные пределы для единичного наблюдения в точке ($X_8 = 32$, $X_6 = 22$) есть

$$\hat{Y} \pm t\,(22;\ 0{,}975)\, s \sqrt{1 + \mathbf{X}_0' \mathbf{C} \mathbf{X}_0} =$$

$$= 11{,}2736 \pm (2{,}074)\,(0{,}661589) \sqrt{1 + 0{,}10413981} =$$

$$= 11{,}2736 \pm (2{,}074)\,(0{,}661589)\,(1{,}050781) =$$

$$= 11{,}2736 \pm 1{,}4418 = \{9{,}8318,\ 12{,}7154\}.$$

(Примечание. Для получения совместных доверительных поверхностей во всем диапазоне изменения регрессионной функции надо было бы подставить вместо t (v, 1—1/2α) выражение $\{qF(q, n-q, 1-\alpha)\}^{1/2}$, где q — общее число параметров модели с учетом и β_0. Стало быть, $v = n-q = n-p-1$. В нашем примере $n = 25$, $p = 2$, $q = 3$. См., например[4]: Miller R. G. Simultaneous Statistical Inference.— New York: McGraw-Hill, p. 110—116.)

Исследование остатков

Остатки, показанные в табл. 4.3, могут быть, как видно, исследованы, если имеются какие-либо определенные признаки неадекватности. Мы предоставим это читателю в качестве упражнения, ограничившись следующими замечаниями:

1) остатки в зависимости от $\hat{Y}$ (см. рис. 4.7). Из этого графика не видно какого-либо необычного поведения;

[4] К сожалению, нам не известна какая-либо книга на русском языке, аналогичная работе Р. Миллера. Но применительно к ситуации регрессионного анализа дополнительную информацию можно найти в кн.: Себер Дж. Линейный регрессионный анализ/Пер. с англ. Под ред. М. Б. Малютова.— М.: Мир, 1980, гл. 5, с. 122—137.— *Примеч. пер.*

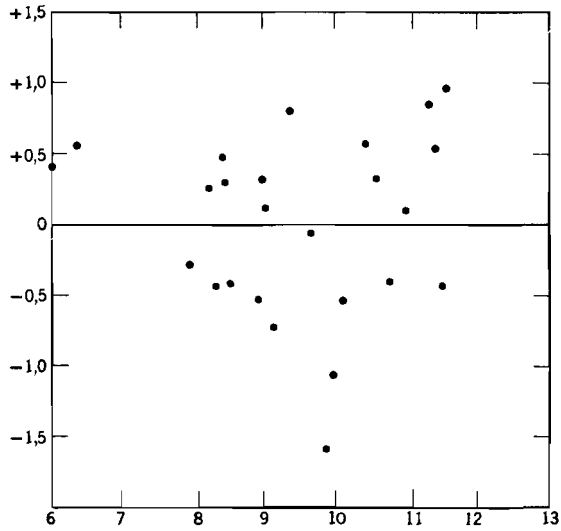


Рис. 4.7. Остаткн в зависимости от $\hat{Y}$

2) критерий серий и критерий Дарбина—Уотсона не обнаруживают каких-либо отклонений от случайного характера временной последовательности (см. упражнение 7 из гл. 3).

Упражнения

(Примечание. Если вы считаете на компьютере, то берите данные так, как они есть. Если же вы работаете с карманным калькулятором, то можете искать модели в форме:

$$Y - \overline{Y} = \beta_1 (X_1 - \overline{X}_1) + \beta_2 (X_2 - \overline{X}_2) + \varepsilon.$$

Это должно приводить в точности к тем же ответам, но вам понадобится обращать матрицы размера 2×2, а не 3×3. Почему это так, вы можете узнать, прочитав с. 318—320.)

1. Задача множественной регрессии.

Данные

X_0	X_1	X_2	Y	X_0	X_1	X_2	Y
1	1	8	6	1	5	0	2
1	4	2	8	1	10	—12	—4
1	9	—8	1	1	2	4	10
1	11	—10	0	1	7	—2	—3
1	3	6	5	1	6	—4	5
1	8	—6	3				

1) Применяя метод наименьших квадратов, оцените β-коэффициенты в модели:

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon.$$

2) Составьте таблицу дисперсионного анализа.

3) Используя $\alpha = 0{,}05$, проверьте, является ли вся регрессия статистически значимой.

4) Вычислите квадрат множественного коэффициента корреляции R^2. Какая часть общего разброса объясняется двумя факторами?

5) Обратная матрица $(\mathbf{X}'\mathbf{X})^{-1}$ для этой задачи такова:

$$\begin{bmatrix} 4{,}3705 & -0{,}8495 & -0{,}4086 \\ -0{,}8495 & 0{,}1690 & 0{,}0822 \\ -0{,}4086 & 0{,}0822 & 0{,}0422 \end{bmatrix}.$$

На основании результатов из таблицы дисперсионного анализа и с помощью этой матрицы вычислите:

а) дисперсию b_1;

б) дисперсию b_2;

в) дисперсию предсказанного значения Y в точке $X_1 = 3$, $X_2 = 5$.

6) Насколько полезна регрессия только по X_1? Что дает добавление X_2, когда X_1 уже включен в модель?

7) Насколько полезна регрессия только по X_2? Что дает добавление X_1, когда X_2 уже включен в регрессию?

8) Каковы ваши выводы?

2. В приведенной ниже таблице содержится двенадцать наборов наблюдений трех переменных X, Y, Z. Найдите регрессионную «плоскость» X по Y и Z, т. е. линейную комбинацию Y и Z, которая предсказывает значение X наилучшим образом, когда даны только Y и Z. С помощью таблицы дисперсионного анализа для X или иным способом проверьте, имеется ли преимущество от включения в формулу предсказания обоих факторов Y и Z.

X	Y	Z	X	Y	Z
1,52	98	77	1,63	97	82
1,41	76	139	1,38	91	100
1,16	58	179	1,37	79	125
1,45	94	95	1,36	92	96
1,24	73	142	1,40	92	99
1,21	57	186	1,03	54	190

Источник: Дипломная работа, Кембридж, 1949.

3. Приведенные ниже данные выбраны из обширного материала, содержащего сведения о кандидатах на получение «Аттестата зрелости», специально для критического анализа. Здесь Y обозначает суммарную оценку (отметку) кандидата, набранную из 1000 возможных баллов на экзамене для получения «Аттестата зрелости». Эта оценка складывается из оценок по предметам, выбранным самим кандидатом — максимум 800 баллов, и оценок по двум обязательным письменным работам — основы естествознания и по английскому языку,— за которые дается максимум 200 баллов [5]. Эта оценка обозначена X_1.

[5] В 1951 г. система школьного образования в Великобритании подверглась очередной реформе. В этом и предыдущем упражнениях приведены фрагменты данных, собранных в связи с подготовкой реформы и оценкой ее целесообразности. Статистический анализ этих данных показал, что нет смысла обременять выпускников обязательным экзаменом по английскому языку непосредственно в школе. Достаточно ограничиться лишь государственным экзаменом. Для государственного экзамена предусмотрены три варианта: после 5-го класса (обычный уровень), после 6-го класса (повышенный уровень) и то же после 6-го класса, но на стипендию. Обработка данных такого рода ставит вопрос о выполнении предпосылок регрессионного анализа. Но в приведенных упражнениях речь идет просто о тренировке в вычислениях, так что эта проблема не возникает.— *Примеч. пер.*

Обозначение X_2 принято для оценок (из 100 баллов) за обязательное школьное выпускное сочинение по английскому языку («Школьный аттестат»), служащих предварительным ориентиром.

Вычислите множественную регрессию Y на X_1 и X_2 и сделайте необходимые проверки, позволяющие обобщить ваши заключения об интеллекте кандидатов. К ним можно добавить текущую характеристику по обязательным письменным контрольным работам для прогноза суммарной оценки экзамена на «Аттестат зрелости». Имеют ли предварительные характеристики из «Школьного аттестата» по английскому языку хоть какое-нибудь значение для предсказания итога независимо от того, что уже было выяснено из текущих результатов обязательных контрольных работ?

Кандидат	Y	X_1	X_2	Кандидат	Y	X_1	X_2
1	476	111	68	9	645	117	59
2	457	92	46	10	556	94	97
3	540	90	50	11	634	130	57
4	551	107	59	12	637	118	51
5	575	98	50	13	390	91	44
6	698	150	66	14	562	118	61
7	545	118	54	15	560	109	66
8	574	110	51				

Источник. Дипломная работа, Кембридж, 1953 Упражнения 2 и 3 публикуются с разрешения издательства Кембриджского университета (Cambridge University Press)

4. Восемь опытов содержат различные условия насыщения (X_1) и количества трансизомеров (X_2). Отклики Y и соответствующие уровни X_1 и X_2 приведены в таблице [6]:

Y	X_1	X_2	Y	X_1	X_2
66,0	38	47,5	22,0	31	29,5
43,0	41	21,3	14,0	34	14,2
36,0	34	36,5	12,0	29	21,0
23,0	35	18,0	7,6	32	10,0

Суммы, нескорректированные суммы квадратов и смешанные произведения следующие:

$$\Sigma Y = 223{,}6, \qquad \Sigma X_1 = 274, \qquad \Sigma X_2 = 198,$$

[6] Трансизомерами (или геометрическими изомерами) называются молекулы с одинаковой последовательностью и типом химической связи, но различным пространственным строением. Атомы углерода в этих молекулах располагаются по разные стороны плоскости двойной связи. Они обычно обладают различными оптическими свойствами. В этом примере величина отклика обозначена аббревиатурой *SCI*, которую не удалось расшифровать. Можно предположить, что это сокращение от слова *sciagraphy* — рентгенография. Оно, видимо, обозначает какую-то характеристику поведения смеси в рентгеновской части спектра. Отсутствие полной ясности не мешает в данном случае решению задачи. — *Примеч. пер.*

$$\Sigma Y^2 = 8911{,}76, \qquad \Sigma X_1^2 = 9488, \qquad \Sigma X_2^2 = 5979{,}08,$$

$$\Sigma X_1 Y = 8049{,}2, \qquad \Sigma X_2 Y = 6954{,}7, \qquad \Sigma X_1 X_2 = 6875{,}6$$

1) Подберите модель $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$.
2) Значима ли регрессия в целом? (Используйте $\alpha = 0{,}05$.)
3) Какая часть разброса в Y объясняется с помощью X_1 и X_2?

5. Величина зазора затворной пластинки и ее температура влияют на процент заготовок, которые благополучно проходят контрольную проверку на машинах по упаковке мыла. Были собраны некоторые данные об этих факторах, они приведены в таблице:

Зазор затворной пластинки, X_1	Температура затворной пластинки, X_2	Процент заготовок, проходящих зазор, Y	Зазор затворной пластинки, X_1	Температура затворной пластинки, X_2	Процент заготовок, проходящих зазор, Y
130	190	35,0	139	240	56,7
174	176	81,7	188	230	84,4
134	205	42,5	175	200	94,3
191	210	98,3	156	218	44,3
165	230	52,7	190	220	83,3
194	192	82,0	178	210	91,4
143	220	34,5	132	208	43,5
186	235	95,4	148	225	51,7

1) Предположите, что модель линейна: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ и определите МНК-оценки параметров β_0, β_1 и β_2.
2) Значима ли регрессия в целом? (Примите $\alpha = 0{,}05$.)
3) Можно ли сказать, что одна из двух переменных более полезна, чем другая, для предсказания отклика?
4) Какие рекомендации вы можете дать относительно работы упаковочной машины?

6. Используя 17 наблюдений, которые приведены ниже:
1) подберите модель $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$;
2) проверьте неадекватность с помощью «чистой» ошибки;
3) исследуйте остатки;
4) оцените величину вклада каждого из факторов X_1 и X_2 в регрессионную модель.

X_1	X_2	Y	X_1	X_2	Y
17	42	90	25	34	75, 82
19	45	71, 76	27	98	99
20	29	63, 63, 80, 80	28	9	73
21	93	80, 64, 82, 66	30	73	67, 74

7. По приведенным в таблице данным постройте модель $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$. Проверьте неадекватность как по «чистой» ошибке, так и по анализу остатков. Оцените вклад каждого из предикторов X_1 и X_2 в регрессионную модель.

X_1	X_2	Y	X_1	X_2	Y
2,6	3,9	83	3,0	9,0	73
2,8	4,2	64	3,0	9,0	75
2,8	4,2	69	3,4	3,1	68
2,9	2,6	56	3,4	3,1	75
2,9	2,6	56	3,6	9,5	92
2,9	2,6	73	3,7	0,6	66
2,9	2,6	73	3,9	7,0	60
3,0	9,0	57	3,9	7,0	67
3,0	9,0	59			

8. (Сначала обратитесь, пожалуйста, к упражнению 8 из гл. 3.) После того, как вы проанализировали исходные данные, управляющий вернулся к своим записям, пытаясь найти какую-нибудь дополнительную информацию, которую можно было бы включить в его модель. Ему удалось обнаружить сведения о числе мужчин, которые нанимались на работу только на один день. В том порядке, в каком были записаны исходные данные, эти числа оказались следующими:

$$Z = 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 1,\ 3,\ 6,\ 6.$$

По имеющимся данным постройте с помощью метода наименьших квадратов модель плоскости $Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon$. Проверьте неадекватиость и (если неадекватность отсутствует) всю регрессию. Проверьте гипотезу H_0: $\beta_2 = 0$ против альтернативы H_1: $\beta_2 \neq 0$ с помощью принципа дополнительной суммы квадратов. Какие выводы вы можете извлечь из проделаииого анализа?

Вот результаты, которые вам пригодятся:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 13 & 65 & 17 \\ 65 & 437 & 155 \\ 17 & 155 & 83 \end{bmatrix}^{-1} =$$

$$= \frac{1}{24\,780} \begin{bmatrix} 12\,246 & -2\,760 & 2\,646 \\ -2\,760 & 790 & -910 \\ 2\,646 & -910 & 1\,456 \end{bmatrix},$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 2990 \\ 19\,120 \\ 6\,050 \end{bmatrix}, \qquad \mathbf{b} = \frac{1}{24\,780} \begin{bmatrix} -147\,360 \\ 1\,346\,900 \\ -678\,860 \end{bmatrix} = \begin{bmatrix} -5{,}947 \\ 54{,}354 \\ -27{,}396 \end{bmatrix}.$$

9. Подберите модель $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ по приведенным ниже данным. Постройте таблицу дисперсионного анализа и примените частиый F-критерий для проверки гипотезы H_0: $\beta_i = 0$ против альтернативы H_1: $\beta_i \neq 0$ при $i = 1, 2$, полагая, что вторая переменная включена в модель. Обсудите относительные вклады факторов X_1 и X_2 в зависимости от того, какой из иих включается в модель первым, а какой — вторым.

X_1	X_2	Y	X_1	X_2	Y
—5	5	11	2	—2	5
—4	4	11	3	—2	5
—1	1	8	3	—3	4
2	—3	2			

10. Дж. Джон (J. A. John) однажды рассказал нам о таком случае. Экспериментатор говорит Вам, что он хочет методом наименьших квадратов подобрать модель $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$ при условии, что $\beta_1 = 1$. В частности, он интересуется тем, что надо ему делать, если он захочет получить методом наименьших квадратов, скажем, модель вида $Y - X_1 = \beta_0 + \beta_2 X_2 + \varepsilon$. Сможет ли он сделать это? (Да.)

11. Т. Митчел (T. J. Mitchell) рассказал о другом случае. Экспериментатор хочет построить квадратичную модель $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$. Он «знает» (по его утверждению), что при $X = 1$ отклик равен 10, т. е. если пренебречь членом, содержащим ошибку, справедливо выражение: $10 = \beta_0 + \beta_1 + \beta_{11}$. Исходя из этого он подставляет выражение $\beta_0 = 10 - \beta_1 - \beta_{11}$ в первую модель и получает $Y - 10 = \beta_1 Z_1 + \beta_{11} Z_2 + \varepsilon$, где $Z_1 = X - 1$ и $Z_2 = X^2 - 1$. Теперь он находит методом наименьших квадратов коэффициенты второй модели b_1 и b_{11}, определяет b_0 из $b_0 = 10 - b_1 - b_{11}$ и заявляет, что он получил методом наименьших квадратов решение первой задачи при ограничении, что значению $X = 1$ соответствует отклик, равный 10. Прав ли он? (Да.)

Ответы к упражнениям

1.1) $b_0 = 14, \quad b_1 = -2, \quad b_2 = -\frac{1}{2}$.

2) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий (скорректированный)	10	190		
Обусловленный регрессией	2	122	61,0	7,17*
Остаток	8	68	8,5	

3) Проверка значимости:

$$\text{Сравним} \quad F = \frac{\text{MS регрессии}}{\text{MS остаточная}} \quad \text{с} \quad F(2;8;\ 0{,}95) = 4{,}46.$$

Так как 7,17 больше, чем критическое F, отвергаем гипотезу о нелинейности модели и используем подобранное уравнение

$$\hat{Y} = 14 - 2X_1 - \frac{1}{2} X_2.$$

4) $R^2 = \frac{122}{190} = 64{,}21\%$.

5а) Оценка дисперсии $b_1 = 1{,}4365$.
5б) Оценка дисперсии $b_2 = 0{,}3587$.
5в) Оценка дисперсии $\hat{Y} = 1{,}95075$.

6) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий (скорректированный)	10	190,00		
Регрессия	2	122,00	61,00	7,17
обусловленный b_1	1	116,08	116,08	13,64*
b_2 при данном b_1	1	5,92	5,92	< 1
Остаток	8	68,00	8,50	

7) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий (скорректированный)	10	190,00		
Регрессия	2	122,00	61,00	7,17
обусловленный b_2	1	98,33	98,33	11,57*
b_1 при данном b_2	1	23,67	23,67	2,78
Остаток	8	68,00	8,50	

8) *Выводы*:

а) Хотя уравнение регрессии $\hat{Y} = 14 - 2X_1 - \frac{1}{2} X_2$ статистически значимо, средний квадрат ошибки для него больше, чем для уравнения $\hat{Y} = 9{,}162 - 1{,}027X_1$.

б) Из этих данных нельзя получить независимые оценки для β_1 и β_2. А если их получение все же желательно, то надо провести для X_1 и X_2 сбалансированный спланированный эксперимент.

в) Когда ставится задача выбора модели, обычно требуется большой объем экспериментальной работы, причем эксперименты следует планировать.

2 $$\hat{X} = 1{,}0607 + 0{,}0056Y - 0{,}0013Z$$

Дисперсионный анализ

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий (скорректированный)	11	0,294867		
Регрессия	2	0,236409	0,118204	18,20
обусловленный Y	1	0,236275	0,236275	36,38*
обусловленный $Z \mid Y$	1	0,000134	0,000134	<1
обусловленный Z	1	0,236006	0,236006	36,34*
обусловленный $Y \mid Z$	1	0,000403	0,000403	<1
Остаток	9	0,058458	0,006495	

Вывод: Одновременное введение в модель Y и Z нецелесообразно. Дальнейшим подтверждением этому служит коэффициент корреляции $r_{yz} = -0{,}9978$

3. **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий (скорректированный)	14	85 386,0000		
Регрессия	2	49 791,1751	24 895,58755	8,39*
обусловленный одним X_1	1	48 186,1482	48 186,1482	16,24*
обусловленный $X_2 \mid X_1$	1	1 605,0249	1 605,0269	<1
Остаток	12	35 594,8249	2 966,2354	

Выводы:
1) Предсказывающая модель

$$\hat{Y} = 124{,}063977 + 3{,}512038X_1 + 0{,}834632X_2$$

объясняет только 58,31 % общего разброса экзаменационных баллов для «Аттестата зрелости». Несмотря на то что уравнение оказывается статистически значимым для уровня значимости $\alpha = 0{,}011$, стандартное отклонение остатков равно 54,46, а будучи выражено в процентах от среднего экзаменационного балла оно составляет всего 9,725 %. Это указывает на то, что существует значительная часть необъясненного разброса, и, следовательно, уравнение будет мало применимым для предсказания.

2) Дополнительное введение в модель X_2, предварительной информации об успеваемости по английскому языку из «Школьного аттестата», не добавляет ничего нового к методике предсказания итоговой оценки кандидата на экзамене для получения «Аттестата зрелости». С таким же успехом можно пользоваться простой моделью $Y = \beta_0 + \beta_1 X_1 + \varepsilon$.

4. 1) $\hat{Y} = -94{,}552026 + 2{,}801551X_1 + 1{,}072683X_2$.

2) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий (скорректированный)	7	2662,14		
Регрессия	2	2618,98	1309,49	151,74*
Остаток	5	43,16	8,63	

Так как F (2; 5; 0,95) = 5,79, регрессия в целом статистически значима, т. е. $151{,}74 > 5{,}79$.

3) $$R^2 = \frac{\text{SS регрессии}}{\text{общая скор. SS}} = \frac{2618{,}98}{2662{,}14} = 98{,}38\ \%.$$

5. 1) $\hat{Y} = 67{,}234527 + 0{,}906089\ (X_1 - 164) - 0{,}064122\ (X_2 - 213)$.

2) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий (скорректированный)	15	8429,14444		
Регрессия	2	6796,77105	3398,385525	26,90*
Остаток	13	1632,37339	126,336415	

Квадрат коэффициента множественной корреляции $R^2 = 80{,}5\ \%$.

Стандартное отклонение остатков равно 11,239947.

Подобранная модель статистически значима. Однако 20 % разброса остаются необъясненными. Нужна дальнейшая работа, чтобы справиться с этой трудностью.

3) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	*F*
Общий (скорректированный)	15	8429,14444		
Регрессия	2	6796,77105	3398,385525	26,90*
X_1	1	6777,72877	6777,72877	53,65*
$X_2 \mid X_1$	1	19,04228	19,04228	Не значимо
X_2	1	25,10057	25,10057	Не значимо
$X_1 \mid X_2$	1	6771,67048	6771,67048	53,60*
Остаток	13	1632,37339	126,33645	

X_1 — более важная переменная.

4) *Выводы*:

в) в рабочем диапазоне этой машины, как показывают уровни «зазора» и «температуры», «зазор» имеет резко выраженное влияние на отклик;

б) существует некоторое указание на то, что «температура» тоже оказывает аддитивное влияние на отклик.

5) Данные до некоторой степени указывают на то, что может быть пригодна другая модель. Полезно рассмотреть наблюдения, расположенные по приведенной ниже форме:

Температура пластин, °С

		176—208	210—220	225—240
Зазор пластин	130—148	35 42,5 43,5	34,5	56,7 51,7
	156—178	81,7 94,3	44,3 91,4	52,7
	186—194	82,0	98,3 83,3	95,4 84,4

Можно видеть определенное взаимодействие между зазором и температурой. Следовательно, более приемлемой будет модель второго порядка.

6. 1) $\hat{Y} = 72{,}25 + 0{,}0286X_1 + 0{,}0487X_2$.

2)

Источник	Число степеней свободы	SS	MS	F
$b_1, b_2 \mid b_0$	2	35,0	17,5	<1
$b_1 \mid b_0$	1	1,7	1,7	<1
$b_2 \mid b_0, b_1$	1	33,3	33,3	<1
$b_2 \mid b_0$	1	34,8	34,8	<1
$b_1 \mid b_0, b_2$	1	0,6	0,6	<1
Остаток	14	1509,1	$s^2 = 107,8$	
неадекватность	5	898,6	179,7	2,65
«чистая» ошибка	9	610,5	67.8	
Общий, скорректированный	16			

Неадекватность не проявилась.

3) Остатки таковы: 15, — 4, 1, — 11, — 11, 6, 6, 3, — 13, 5, — 11, 0, 7, 21, 0, — 9, — 2. (Исследование остатков оставляем читателю в качестве упражнения.)

4) Ни X_1, ни X_2 не имеют значений, объясняющих Y на основе рассматриваемого набора данных. Оба вместе они объясняют всего лишь 100 $R^2 = 2,3$ % разброса относительно среднего.

7. Для модели, включающей обе переменные, X_1 и X_2, мы получаем $\hat{Y} = 65,13 + 0,286X_1 + 0,487X_2$ и следующую таблицу дисперсионного анализа:

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
b_0	1	79 973,88		
$b_1, b_2 \mid b_0$	2	35,01	17,51	0,16
Неадекватность	5	898,61	179,72	2,65
«Чистая» ошибка	9	610,50	67,83	
Общий	17	81 518,00		

Поскольку F (5; 9; 0,95) = 3,48, нет никакого смысла подозревать неадекватность. Значит, $s^2 = (898,61 + 610,50)/(5 + 9) = 107,79$. Критерий для всей регрессии $F = 17,51/107,79 = 0,16$ не значим. Остатки не демонстрируют ничего примечательного.

$$\mathrm{SS}(b_1 \mid b_0,\ b_2) = 0,21;\ \mathrm{SS}(b_2 \mid b_0 \mid b_1) = 33,32.$$

Оба частных F-критерия не значимы. Отсюда мы заключаем, что модель $\hat{Y} = \bar{Y} = 68,588$ не хуже любой другой.

(Примечание. Вполне возможно, что общий F-критерий для гипотезы $H_0: \beta_1 = \beta_2 = 0$ окажется незначимым, но частный F-критерий для одной из гипотез $H_0: \beta_j = 0$, $j = 1, 2$ должен быть значимым. Это указывало бы на слабую связь с соответствующим X_j.)

8. Вот подобранное уравнение:

$$\hat{Y} = -5,95 + 54,35X - 27,40Z.$$

Тогда $\boldsymbol{b}'\mathbf{X}'\boldsymbol{Y} = 21205018600/24780 = 855731$.

Это последнее число есть не что иное, как SS (b_0, b_1, b_2). Поскольку значения переменной Z были найдены позднее, сумма квадратов «чистой» ошибки будет в точности той же, что и раньше, а именно 1467 с пятью степенями свободы. (В общем случае это не верно. Обычно введение нового фактора, такого, как в нашем случае, ведет к уменьшению числа степеней свободы для «чистой» ошибки, поскольку отклики точно с теми же значениями X-ов могут, вообще говоря, иметь различные значения. Однако в наших данных такие отклики всегда имеют одни и те же значения, что не типично.) Вот общая таблица дисперсионного анализа.

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
b_0	1	687 700		
b_1, $b_2 \mid b_0$	2	168 031	84 015	
Неадекватность	5	302	60	<1 не значимо*
«Чистая» ошибка	5	1 467	293	
Общий	13	857 500		

* Следовательно, $s^2 = (302 + 1,467)\,(5 + 5) = 177$.

Проверка для всей регрессии приводит к F (2; 10) = 84015/177 = 474,7, что в высшей степени значимо.

А вот что дает критерий дополнительной суммы квадратов для гипотезы $H_0 : \beta_2 = 0$ против альтернативы $H_1 : \beta_1 \neq 0$:

$$\mathrm{SS}\,(b_2 \mid b_1,\ b_0) = \mathrm{SS}\,(b_2,\ b_1 \mid b_0) - \mathrm{SS}\,(b_1 \mid b_0) = 168031 - 155258 = 12773.$$

Критерий F (1; 10) = $\mathrm{MS}\,(b_2 \mid b_1,\ b_0)/s^2 = 12773/177 = 72,2$, что в высшей степени значимо. Таким образом, введение в модель Z в качестве предиктора чрезвычайно полезно.

Выводы: «Число работающих» мужчин — важная переменная. Поскольку коэффициент $b_2 = -27,40$, она имеет отрицательный эффект. Это, конечно, *не* означает, что мужчины выполняют «отрицательную» работу, но отсюда следует, что их присутствие служит причиной уменьшения объема выполненной работы по сравнению с тем, что могло бы быть без них. Подбор уравнения в форме

$$Y = \beta_0 + \beta_1 (X - Z) + \beta_2 Z + \varepsilon \qquad (1)$$

был бы, видимо, несколько более информативным. Здесь два предиктора (X—Z) и Z представляют собой число работающих женщин и мужчин соответственно. Тогда наши подобранные ранее уравнения для X и Z можно было бы переписать так:

$$\hat{Y} = -5,95 + 54,35\,(X - Z) + 26,95Z,$$

что дает как раз то уравнение, которое получилось бы, если бы мы находили уравнение (1) «в лоб» методом наименьших квадратов. Теперь мы видим, для данных такого типа, что мы сейчас имеем, женщины выполняют примерно вдвое больше работы, чем мужчины. Это приводит к практическому выводу, что в будущем стоит нанимать на работу женщин вместо мужчин (или по крайней мере совсем других мужчин) и посмотреть, что из этого получится.

9. 1) Модель, подобранная для двух переменных X_1 и X_2:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 7 & 0 & 0 \\ 0 & 68 & -67 \\ 0 & -67 & 68 \end{bmatrix}^{-1} \begin{bmatrix} 46 \\ -66 \\ 69 \end{bmatrix} =$$

$$= \begin{bmatrix} \frac{46}{7} \\ 1 \\ 2 \end{bmatrix}.$$

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
b_0	1	302,29		
$b_1,\ b_2 \mid b_0$	2	72,00	36,00	83,72
Остаток	4	1,71	$s^2 = 0{,}43$	
Общий	7	376,00		

Критерий F значим на 1 %-ном уровне, поэтому мы отвергаем нуль-гипотезу $H_0 : \beta_1 = \beta_2 = 0$.

2) Модель для одного X_1:

$$\hat{Y} = 46/7 - (66/68)\,X_1,$$

$$\mathrm{SS}\,(b_1|b_0) = 64{,}06.$$

Следовательно, SS $(b_2|b_1,\ b_0) = 72 - 64{,}06 = 7{,}94$.

Критерий для $\beta_2 = 0$ (при β_1, включенном в модель): $F = 7{,}94/0{,}43 = 18{,}53$ значим на 5 %-ном уровне, но не значим на 1 %-ном уровне, так как $F\ (1;\ 4;\ 0{,}99) = 21{,}20$.

3) Модель для одного X_2:

$$\hat{Y} = 46/7 + (69/68)\,X_2,$$

$$\mathrm{SS}\,(b_2|b_0) = 70{,}01.$$

Следовательно, SS $(b_1|b_2,\ b_0) = 72 - 70{,}01 = 1{,}99$.

Критерий для $\beta_1 = 0$ (при β_2, включенном в модель) : $F = 1{,}99/0{,}43 = 4{,}64$ не значим на 5 %-ном уровне, поскольку $F\ (1;\ 4;\ 0{,}95) = 7{,}71$.

4) *Следствия*. Если X_2 включен в модель, то X_1 нам не нужен. Если же включен в модель X_1, то X_2 способствует повышению значимости. Отсюда ясно, что X_2 — более полезный фактор, который сам по себе объясняет $R^2 = 70{,}01/73{,}71 = 0{,}9498$ разброса относительно среднего. Фактор X_1, взятый сам по себе, объясняет только 0,8691, а факторы X_1 и X_2 вместе — 0,9768 разброса. Отметим, что в нашем наборе данных факторы X_1 и X_2 сильно закоррелированы.

10—11. Решение не приводится.

5.0. Введение

В гл. 1 довольно подробно обсуждались линейные модели первого порядка с одной предикторной переменной. Были также рассмотрены концепция адекватности различных моделей и статистический критерий проверки адекватности. Математический анализ из гл. 1 был представлен в гл. 2 в матричной форме, так что переход от модели первого порядка с одним предиктором к общей модели, линейной по оцениваемым параметрам и содержащей несколько предикторов, можно сделать наиболее эффективно. Построение регрессии в общем случае было приведено в матричной форме в последних параграфах гл. 2. Результаты этих параграфов используются далее повсюду в книге. В гл. 4 модель первого порядка с двумя предикторами обсуждалась как в алгебраическом, так и в геометрическом аспектах. Были введены некоторые критерии для проверки множественного уравнения регрессии и формулы доверительных интервалов для β-коэффициентов и для предсказанных значений Y.

До сих пор основное внимание уделялось моделям первого порядка, линейным, с одним или двумя предикторами. В эту главу включены примеры более сложных моделей. Некоторые из них требуют преобразования переменных и используют «фиктивные» переменные. В этой главе мы будем также обсуждать отдельные аспекты подготовки данных для регрессионного анализа.

Можно записать наиболее общий тип линейной модели с переменными $X_1, X_2, \ldots, X_k$ в виде

$$Y = \beta_0 Z_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p + \varepsilon. \qquad (5.0.1)$$

Переменная $Z_0 = 1$ — это фиктивная переменная, которая всегда равна единице и обычно не записывается. Однако иногда с математической точки зрения в модель удобно включать Z_0. Например, если

$$(Z_{1i}, Z_{2i}, \ldots, Z_{pi}), \quad i = 1, 2, \ldots, n$$

есть n наборов переменных Z_j, $j = 1, 2, \ldots, p$, соответствующих Y_i, $i = 1, 2, \ldots, n$, то при $j \neq 0$ и $Z_{0i} = 1$

$$\sum_{i=1}^{n} Z_{ji} = \sum_{n=1}^{n} Z_{ji} Z_{0i},$$

и поэтому при составлении нормальных уравнений можно воспользо-

ваться общим выражением вида $\sum_{i=1}^{n} Z_{ji}Z_{li}$. Заметим, что $\sum_{i=1}^{n} Z_{0i}^2 = n$.

Каждая Z_j, $j = 1, 2, \ldots, p$, есть известная функция от $X_1, X_2, \ldots, X_k$:

$$Z_j = Z_j(X_1, X_2, \ldots, X_k),$$

которая может иметь любую форму. Иногда каждая функция Z_j включает только одну X-переменную.

Любую такую модель можно записать после перестройки или преобразования в виде уравнения (5.0.1) и анализировать ее общими методами, изложенными в параграфах 2.6—2.15. Теперь мы приведем частные примеры моделей, относящиеся к общей форме уравнения (5.0.1) и допускающие обработку этими методами.

5.1. ПОЛИНОМИАЛЬНЫЕ МОДЕЛИ РАЗЛИЧНЫХ ПОРЯДКОВ ПО X_j

Модели первого порядка

1. Если $p = 1$ и $Z_1 = X$ в уравнении (5.0.1), то мы имеем простую модель первого порядка с одним предиктором:

$$Y = \beta_0 + \beta_1 X + \varepsilon. \tag{5.1.1}$$

2. Если $p = k$, $Z_j = X_j$, то мы имеем модель первого порядка с k предикторами:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon. \tag{5.1.2}$$

Модели второго порядка

1. Если $p = 2$, $Z_1 = X$, $Z_2 = X^2$ и $\beta_2 = \beta_{11}$, то мы имеем модель второго порядка с одним предиктором:

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon. \tag{5.1.3}$$

2. Если $p=5$, $Z_1=X_1$, $Z_2=X_2$, $Z_3=X_1^2$, $Z_4=X_2^2$, $Z_5 = X_1X_2$, $\beta_3 = \beta_{11}$, $\beta_4=\beta_{22}$, $\beta_5=\beta_{12}$, то мы имеем модель второго порядка с двумя предикторами:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{12} X_1 X_2 + \varepsilon. \tag{5.1.4}$$

Полную модель второго порядка для k переменных можно получить аналогичным путем, когда $p = k + k + \frac{1}{2}k(k-1) = \frac{1}{2}(k^2 + 3k)$.

Модели второго порядка применяются, в частности, при исследовании поверхностей отклика методами планирования эксперимента. Здесь целью служит определение или аппроксимация характеристик некоторой неизвестной поверхности отклика полиномом низкой степени. Заметим, что в модель входят все возможные члены второго порядка. Это целесообразно, потому что пропуск членов предполагает наличие информации о том, что не могут встретиться определенные типы поверхностей, которые немыслимы *без* пропущенных членов. Такого рода случаи не часты. Когда же подобная информация есть, обычно можно провести исследование на более строгой теоретической основе.

Пример анализа поверхности отклика второго порядка приведен в гл. 7.

Модели третьего порядка

1. Если $p = 3$, $Z_1 = X$, $Z_2 = X^2$, $Z_3 = X^3$, $\beta_2 = \beta_{11}$ и $\beta_3 = \beta_{111}$, то мы имеем модель третьего порядка с одним предиктором:

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \beta_{111} X^3 + \varepsilon. \qquad (5.1.5)$$

2. Если $p = 9$ и установлено подходящее соответствие для β_i и Z_l (мы опускаем детали, так как разобранные выше примеры делают идею ясной), то модель (5.0.1) можно представить как модель третьего порядка с двумя предикторами:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \\ + \beta_{22} X_2^2 + \beta_{111} X_1^3 + \beta_{112} X_1^2 X_2 + \beta_{122} X_1 X_2^2 + \beta_{222} X_2^3 + \varepsilon. \qquad (5.1.6)$$

Общую модель третьего порядка для k факторов $X_1, X_2, \ldots, X_k$ можно получить аналогично. Модели третьего порядка тоже применяются в работах по исследованию поверхностей отклика методами планирования эксперимента, хотя и значительно реже, чем модели второго порядка [1]. Отметим метод индексации β-коэффициентов. На

---

[1] Поскольку читателя может заинтересовать существо упомянутых здесь задач, сообщим некоторые ссылки на работы по планированию эксперимента при построении поверхностей отклика второго, третьего и более высоких порядков, тем более, что в нашей стране получен ряд важных результатов в этой области. Общее начальное представление о планах второго порядка можно получить, например, из работы Адлер Ю. П. Введение в планирование эксперимента.— М.: Металлургия, 1969.— 157 с. (особо с. 96—115). Более формальное описание см. в работах: Налимов В. В., Чернова Н. А. Планирование эксперимента при поиске оптимальных условий.— М.: Наука, 1965.— 340 с. (особо с. 135—144); Федоров В. В. Теория оптимального эксперимента.— М.: Наука, 1971.— 312 с. Методологические аспекты см. в работах: Налимов В. В., Голикова Т. И. Логические основания планирования эксперимента. — 2-е изд.— М.: Металлургия, 1981.— 151 с. (особо гл. 2, 3, с. 34—66); Вознесенский В. А., Ковальчук А. Ф. Принятие решений по статистическим моделям.— М.: Статистика, 1978.— 192 с. Существуют и обширные таблицы экспериментальных планов второго порядка, см.: Голикова Т. И., Панченко Л. А., Фридман М. З. Каталог планов второго порядка. Препринт МГУ.— М.: Изд-во МГУ, 1977, вып. 1— 387 с.; вып. 2.—384 с.; Таблицы планов эксперимента для факторных и полиномиальных моделей (Справочное издание)/Под ред. В. В. Налимова.— М.: Металлургия, 1982.— 752 с. (особо с. 299—533).

Планы третьего порядка описаны в работе: Никитина Е. П. Планирование и анализ эксперимента (Модели третьего порядка).— М.: Изд-во МГУ, 1976.— 119 с. И здесь тоже есть каталог: Мержанова Р. Ф., Никитина Е. П. Каталог планов третьего порядка.— М.: Изд-во МГУ, 1979.— 171 с., да и в упомянутых выше таблицах под ред. В. В. Налимова они приводятся на с. 534—686. Теоретические вопросы обсуждаются на с. 257—258 монографии В. В. Налимова и Н. А. Черновой, а в более прикладном плане см., например, Тихомиров В. Б. Планирование и анализ эксперимента (при проведении исследований в легкой и текстильной промышленности.)— М.: Легкая индустрия, 1974.— 263 с. (особо с. 122—131). Между прочим, здесь же обсуждаются и модели от первого до десятого (!) порядка, правда, только для одного фактора, см. с. 131—134.

первый взгляд он может показаться странным, однако он удобен, поскольку позволяет легко установить, с какими переменными X и в каких степенях связан данный коэффициент. Например, при $X_1X_2^2 = X_1X_2X_2$ будет коэффициент β_{122} и т. д. Аналогичные обозначения используются выше для моделей второго порядка, они являются стандартными в работах по анализу поверхностей отклика методами планирования эксперимента.

Продолжая процесс, проиллюстрированный выше, к уравнению (5.0.1) можно привести модели любых возможных порядков.

Преобразования

Если модель второго порядка не адекватна, то, может быть, подойдет модель третьего порядка. Однако вряд ли стоит механически добавлять в модель члены более высоких порядков. Часто оказывается продуктивным исследование возможностей каких-то иных преобразований предикторов, откликов или и тех, и других одновременно. То же замечание относится и к решению о переходе от первого порядка ко второму [2]. Так, например, прямая, подобранная в координатах

Известны работы, в которых строятся и исследуются многофакторные планы более высоких порядков—четвертого, пятого, даже шестого. См., например: Баданов А. Г. Построение математических моделей высших порядков сложных физических объектов на основе синтеза частных математических моделей.— В кн.: Тезисы докладов IV Всесоюзной конференции по планированию и автоматизации эксперимента в научных исследованиях.— М.: Изд-во МЭИ, 1973; Чалый В. Д., Яценко Ю. И., Грицаненко А. В. Ортогональное планирование эксперимента четвертого порядка для построения математических моделей функционирования сложных систем.— В кн.: Применение планирования эксперимента в радиоэлектронике и смежных областях техники. Материалы семинара.— М.: МГДНТП, 1975, с. 24—31; Еханин М. В., Слободчикова Р. И. Обобщенный алгоритм построения полиномиальных моделей высоких порядков по многоуровневым сверхнасыщенным планам.— В кн.: Научные труды Гиредмета.— М.: Металлургия, 1982, т. 113, с. 3—7.

Кроме полных полиномов соответствующих степеней, можно строить еще и неполные (например, квадратные без полных квадратов) или полиномы специального вида. Конечно, для такого выбора требуется, как правило, априорная информация. Здесь часто полезны планы: Бродский В. З. Введение в факторное планирование эксперимента.— М.: Наука, 1974.— 223 с., Raktoe B. L., Hedayat A., Federer W. T. Factorial Designs.— New York: J. Wiley, 1981.— 209 p.

Еще по затронутым вопросам см.: Cochren W., Cox G. M. Experimental designs.— 2-ed.— New York: J. Wiley, 1957.— 617 p. (особо p. 335—375); Anderson V. L., McLean R, Design of experiments: A realistic approach.— New York: Marcel Dekker, 1974.— 462 p.

Проблема аппроксимации поверхностей отклика полиномами заданных степеней — одна из центральных в планировании эксперимента.— *Примеч. пер.*

[2] В рамках теории планирования эксперимента, при непосредственном участии первого автора этой книги, разрабатывалась еще одна возможность. Она сводится к ответу на вопрос типа: «Как построить полином первого порядка, если «истинная» поверхность отклика представляет собой полином второго порядка, причем так, чтобы минимизировалось смещение, обусловленное этим несоответствием порядков? » Аналогично, можно искать полином второго порядка для поверхности, описываемой полиномом третьей степени и т. д. См.: Box G. E. P., Draper N. R. A Basic for the selection of a response sur-

log Y от X, если она возможна, нередко предпочтительнее, чем квадратичная модель зависимости Y от X, если, конечно, поведение остатков делает оба эти выбора работоспособными.

5.2. МОДЕЛИ, ВКЛЮЧАЮЩИЕ ПРЕОБРАЗОВАНИЯ, ОТЛИЧНЫЕ ОТ ЦЕЛЫХ СТЕПЕНЕЙ

Полиномиальные модели из параграфа 5.1 включали степени и смешанные произведения степеней предикторов $X_1, X_2, \ldots, X_k$. Здесь мы приведем примеры других типов преобразований, часто полезных при построении регрессионных моделей.

Модели, получаемые при преобразовании только X_j

«Обратное» преобразование. Если в уравнении (5.0.1) мы положим $p = 2$, $Z_1 = 1/X_1$, $Z_2 = 1/X_2$, то получим модель

$$Y = \beta_0 + \beta_1 (1/X_1) + \beta_2 (1/X_2) + \varepsilon. \qquad (5.2.1)$$

Логарифмическое преобразование. При выборе $p = 2$, $Z_1 = \ln X_1$, $Z_2 = \ln X_2$ уравнение (5.0.1) можно записать так:

$$Y = \beta_0 + \beta_1 \ln X_1 + \beta_2 \ln X_2 + \varepsilon. \qquad (5.2.2)$$

Преобразование типа квадратного корня. Например,

$$Y = \beta_0 + \beta_1 X_1^{1/2} + \beta_2 X_2^{1/2} + \varepsilon. \qquad (5.2.3)$$

Ясно, что существует много преобразований и можно постулировать модели, содержащие меньше или больше таких членов. В одной и той же модели может содержаться, конечно, несколько различных преобразований. Нередко трудно решить, что предпочесть, если можно сделать любое преобразование. Выбор часто осуществляется на основе предыдущих знаний о факторах данной задачи. Цель преобразований такого рода состоит в том, чтобы получить для преобразованных переменных более простую регрессионную модель, чем для исходных.

В статье Дж. Бокса и П. Тайдуэлла о преобразованиях независимых переменных (см.: Box G. E. P., Tidwell P. W. Transformation of independent variables.— Technometrics, 1962, **4**, p. 531—550) предложена итеративная процедура поиска преобразований отдельных переменных на основе имеющихся данных. Преобразования могут также включать одновременно несколько переменных X_j, на-

face design.— J. Amer. Statist, Assoc., 1959, **54**, № 257, p. 622—630; Box G. E. P., Draper N. R. The choise of a second order rotatable design.— Biometrika, 1963, **50**, № 3/4, p. 335—343; Налимов В. В., Чернова Н. А. Планирование эксперимента при поиске оптимальных условий— М.: Наука, 1965, с. 258—275; Седунов Е. В. Оптимальное планирование и анализ регрессионных экспериментов с учетом систематической ошибки. Препринт.— М.: Научн. совет «Кибернетика» АН СССР, 1978,— 60. с.; Седунов Е. В. Планирование и анализ регрессионных экспериментов с учетом систематической ошибки (обзор).— Заводская лаборатория, 1979, № 1, с. 55—62.— *Примеч. пер.*

пример $Z_1 = X_1^{1,2} \ln X_2$. Преобразования такого типа иногда предлагаются исходя из формы искомого уравнения в преобразованных переменных. Простой пример приведен в работе Дж. Бокса и Д. Кокса, посвященной анализу преобразований (см.: Box G. E. P., Cox D. R. An analysis of transformations.— Journal of the Royal Statistical Society. Series B, 1964, **26**, p. 211—243, discussion — p. 244—252, см. p. 222—223), где рассматриваются, однако, в основном преобразования зависимой переменной. Поскольку модель такого типа еще не упоминалась, мы ниже (см. с. 279) обсудим пример.

Хорошие преобразования предикторов иногда предлагаются также на основе построения различных диаграмм. См., например, работы Э. Хёрла о подгонке кривых к данным (Hoerl A. E. Fitting curves to data.— Chemical Business Handbook.— New York: McGrow-Hill, 1954) и Дж. Долби о быстром методе выбора преобразования (Dolby J. L. A quick method for choosing a transformation.— Technometrics, 1963, **5**, p. 317—325), а также работу Дж. Тьюки о сравнительной анатомии преобразований (Tukey J. W. On the comparative anatomy of transformations.— Annals of Mathematical Statistics, 1957, **28**, p. 602—632).

Основная информация, необходимая для обоснованного выбора ряда преобразований, которые можно использовать для зависимой переменной, отражена в работе: Bartlett M. S. The use of transformations.— Biometrics, 1947, **3**, p. 39—52.

Нелинейные модели, которые «внутренне» линейны

Нелинейные модели (т. е. модели, нелинейные по оцениваемым параметрам) можно подразделить на два класса, которые удобно назвать *внутренне линейными и внутренне нелинейными*. Если модель внутренне линейна, то ее с помощью подходящего преобразования можно привести к стандартной форме линейной модели в виде уравнения (5.0.1).

Если же нелинейную модель нельзя представить в такой форме, то она внутренне нелинейна [3] (т. е. действительно нелинейна, периодична). В этом параграфе мы сосредоточим внимание на моделях, которые внутренне линейны и могут обрабатываться с помощью матричных методов, описанных в гл. 2. Необходимые преобразования обычно охватывают как отклики, так и предикторы. Рассмотрим примеры.

Мультипликативная модель.

$$Y = \alpha X_1^{\beta} X_2^{\gamma} X_3^{\delta} \varepsilon, \tag{5.2.4}$$

где α, β, γ и δ — неизвестные параметры, ε — мультипликативная случайная ошибка.

Логарифмирование уравнения (5.2.4) по основанию e переводит модель в линейную форму:

$$\ln Y = \ln \alpha + \beta \ln X_1 + \gamma \ln X_2 + \delta \ln X_3 + \ln \varepsilon. \tag{5.2.5}$$

Преобразованная модель (5.2.5) имеет форму уравнения (5.0.1), и

[3] Такие модели описываются в гл. 10 (см. кн. 2).— *Примеч. пер.*

поэтому здесь можно применять стандартные методы исследования линейной регрессии, описанные в гл. 2. Однако следует подчеркнуть, что для того, чтобы критерии значимости и оценки доверительных интервалов были обоснованными, необходимо соблюдение условий $\ln \varepsilon \sim N(0, 1\sigma^2)$ (а вовсе не для ε). Поэтому экспериментатор должен быть готов после построения модели проверить справедливость этого предположения с помощью исследования остатков, как описано в гл. 3.

Альтернативная модель, которую часто рассматривают применительно к этому случаю, есть

$$Y = \alpha X_1^{\beta} X_2^{\gamma} X_3^{\delta} + \varepsilon. \qquad (5.2.4A)$$

Общие методы линейной регрессии, приведенные в гл. 2, непригодны для этой модели, так как она внутренне нелинейна. При применении метода наименьших квадратов приходится обращаться к итеративным процедурам нахождения оценок α, β, γ, δ. Эти процедуры кратко обсуждаются в гл. 10.

Теперь мы перескажем пример преобразования предикторов, приведенный Дж. Боксом и Д. Коксом. Он содержит модель типа уравнения (5.2.5), для которой на основании имеющихся данных были получены следующие оценки b, c и d параметров β, γ и δ:

$$b = 4{,}96 \pm 0{,}20; \quad c = -5{,}27 \pm 0{,}30;$$
$$d = -3{,}15 \pm 0{,}30.$$

Числа, следующие за знаками плюс и минус, представляют собой стандартные отклонения[4]. Эти данные показывают, например, что предположение $\beta = 5 = -\gamma$, $\delta = -3$ не бессмысленно. Если мы положим $Z = X_1/X_2$, так что $\ln Z = \ln X_1 - \ln X_2$, и подставим в уравнение (5.2.4), то получим модель, предполагавшуюся при анализе:

$$Y \sim Z^{+5} X_3^{-3} \varepsilon.$$

Дж. Бокс и Д. Кокс замечают, что эта модель «соответствует данным удивительно хорошо».

Экспоненциальная модель.

$$Y = e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2} \cdot \varepsilon. \qquad (5.2.6)$$

Логарифмируя обе части по натуральному основанию, получим

$$\ln Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ln \varepsilon. \qquad (5.2.7)$$

«Обратная» модель.

$$Y = \frac{1}{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}. \qquad (5.2.8)$$

---

[4] В оригинале здесь уже «привычная» опечатка: вместо «отклонений» стоят «ошибки». Мы, следуя рекомендации авторов, везде правим на «отклонения». — *Примеч. пер.*

Обращая обе части, получим

$$\frac{1}{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon. \qquad (5.2.9)$$

В этом случае экспериментатор должен использовать в качестве отклика обратную величину зависимой переменной.

Более сложная экспоненциальная модель.

$$Y = \frac{1}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon}}. \qquad (5.2.10)$$

Применяя обращение, вычитая единицу и затем логарифмируя по натуральному основанию обе части, получим

$$\ln\left(\frac{1}{Y} - 1\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon. \qquad (5.2.11)$$

Это пример последовательного преобразования зависимой переменной для сведения сложной нелинейной модели к линейному виду.

Способ, каким ошибка входит в первые из каждой пары моделей, приведенных выше (см. уравнение (5.2.10)), может показаться несколько странным. Причина этого обстоятельства состоит в том, что если мы преобразуем и подбираем именно ту модель, которая соответствует вторым уравнениям каждой пары (см. уравнение (5.2.11)), то первые уравнения должны соответствовать нашим действительным гипотезам относительно Y.

Во всех случаях, когда модели преобразуются, как в этих примерах, метод наименьших квадратов применяется к преобразованной модели вида (5.0.1) и, конечно, оцениваемые коэффициенты будут «МНК-оценками» только относительно преобразованной модели. И остатки надо исследовать методами из гл. 3 именно для этих *преобразованных* откликов, а вовсе не для исходных.

Вот два других примера внутренне нелинейных моделей:

$$Y = \beta_0 + \beta_1 e^{-\beta_2 X} + \varepsilon$$

и

$$Y = \beta_0 + \beta_1 X + \beta_2 (\beta_3)^X + \varepsilon.$$

Последняя модель обсуждается в работе: Shah B. K., Khatri C. G. A method of fitting the regression curve $E(y) = \alpha + \delta x + \beta p^x$.— Technometrics, 1965, **7**, p. 59—65.

Заключение

Рассмотренные в этом параграфе преобразования для сведения сложных моделей к линейным в настоящее время начали более или менее широко применяться. Когда, как мы здесь предполагаем, предикторы не содержат ошибок, в решении таких задач нет трудностей.

Однако, когда преобразования включают зависимую переменную Y, следует особенно тщательно следить, чтобы предпосылки метода регрессионного анализа (независимость ошибок, $N(0, \sigma^2)$) при преобразовании не нарушились. Часто можно избежать преобразования зависимой переменной путем отыскания подходящего преобразования X-переменных (как, скажем, в ранее упомянутой работе Дж. Бокса и П. Тайдуэлла).

5.3. СЕМЕЙСТВА ПРЕОБРАЗОВАНИЙ

Преобразование отклика

Одно из полезных семейств преобразований для (обязательно положительного) отклика Y — степенные преобразования:

$$W = \begin{cases} (Y^{\lambda} - 1)/\lambda & \text{при } \lambda \neq 0, \\ \ln Y & \text{при } \lambda = 0'^{*}. \end{cases} \tag{5.3.1}$$

Это непрерывное семейство, зависящее от единственного параметра λ. Мы можем воспользоваться для оценки этого параметра имеющимися данными точно так же, как мы оцениваем вектор параметров $\boldsymbol{\beta}$ в модели, которую строим, скажем

$$\mathbf{W} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{5.3.2}$$

где $\mathbf{W} = (W_1, W_2, \ldots, W_n)'$. Существуют два главных способа оценки. Один из них основан на методе максимума правдоподобия в предположении, что остатки распределены нормально ($\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$) для подходящего выбора λ. Этот подход (так же, как и его байесовский эквивалент) хорош для любого семейства преобразований, в том числе и для упомянутого выше, что отмечали Дж. Бокс и Д. Кокс еще в своей работе, вышедшей в 1964 г. Ниже перечислены необходимые шаги этого метода.

Оценка λ методом максимума правдоподобия

1. Берем какое-нибудь значение λ из заданного диапазона. (Обычно мы начинаем подбор λ с просмотра диапазона (— 2, 2) или даже (— 1, 1), постепенно расширяя диапазон настолько, насколько это окажется необходимым. Как правило, в заданном диапазоне выбирается от одиннадцати до двадцати одного значения λ. В дальней-

---

* Мы пишем здесь $\ln Y$, поскольку это предел для выражения $(Y^{\lambda}-1)/\lambda$ при λ, стремящемся к нулю, и, следовательно, наше семейство преобразований непрерывно по λ. Правда, если $\lambda = 0$ — это значение λ, которое выбрано фактически для преобразования наших данных, то мы можем еще выбрать либо натуральный логарифм (ln), либо логарифм по любому другому основанию, скажем 10 (log). Этот выбор повлияет только на постоянный множитель и, значит, только на масштаб чисел, используемых в анализе, но не окажет никакого влияния на природу дальнейшего исследовании. Точно таким же образом, когда для преобразования данных выбрано ненулевое значение λ, мы можем, если угодно, вести наш анализ для Y^{λ}, а вовсе не для $(Y^{\lambda}-1)/\lambda$. Это, как и прежде, ничуть не повлияет на осно у основ дальнейшего анализа, а скажется только в различии масштабов и точки начала координат.

шем мы можем дробить интервал и на более мелкие части, если потребуются дополнительные подробности, но необходимость в этом возникает нечасто, см. пункт 3 ниже.)

2. Для выбранного значения λ вычисляем

$$L_{\max}(\lambda) = -\frac{1}{2} n \ln \hat{\sigma}^2(\lambda) + \ln J(\lambda, \mathbf{Y}), \tag{5.3.3}$$

где n — общее число наблюдений,

$$\hat{\sigma}^2(\lambda) = \mathbf{W}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{W}/n \tag{5.3.4}$$

(т. е. остаточная сумма квадратов от уравнения регрессии, подобранного по модели (5.3.2) при выбранном значении λ, взятая $1/n$ раз), а

$$J(\lambda, \mathbf{Y}) = \prod_{i=1}^{n} \frac{\partial W_i}{\partial Y_i} = \prod_{i=1}^{n} Y_i^{\lambda-1} \text{ для всех } \lambda, \tag{5.3.5}$$

т. е.

$$\ln J(\lambda, \mathbf{Y}) = (\lambda - 1) \sum_{i=1}^{n} \ln Y_i.$$

Подставляя все это в уравнение (5.3.3), получим

$$L_{\max}(\lambda) = -1/2n \ln(\text{остаточная SS}/n) + (\lambda - 1) \sum_{i=1}^{n} \ln Y_i \tag{5.3.6}$$

для каждого значения λ, которое мы выбрали. (Не забывайте использовать $W = \ln Y$, когда $\lambda = 0$. Или же вообще избегайте применения значения λ, *точно* равного нулю, когда покрываете заданный диапазон значений λ.)

3. После того как уравнение (5.3.6) вычислено для нескольких значений λ в заданном диапазоне, постройте график значений $L_{\max}(\lambda)$ в зависимости от λ и соедините точки гладкой кривой. Отыщите то значение λ, которое максимизирует величину $L_{\max}(\lambda)$. Это и будет $\hat{\lambda}$, оценка метода максимума правдоподобия для λ. Чаще всего мы предпочитаем не использовать такого точного значения λ в дальнейших вычислениях. Вместо этого мы берем ближайшее удобное значение в последовательности . . . ; — 2; — 1,5; — 1; — 0,5; 0; 0,5; 1; 1,5; 2; . . . после первоначальной прикидки, так что искомое значение попадает в заданный доверительный интервал (см. ниже). Если, например, вычисленное значение $\hat{\lambda}$ окажется порядка 0,11, то мы, по-видимому, сможем воспользоваться величиной $\lambda = 0$. А если бы $\hat{\lambda}$ оказалась примерно 0,94, то мы могли бы взять $\lambda = 1$ и т. д. (Правда, существует масса вариантов персональных решений при выборе λ после того, как вычисления уже закончены. Так в некоторых случаях могут пригодиться значения вроде $^1/_3$, $^2/_3$. Некоторые специалисты предпочитают округлять до ближайшей четверти, а не до половины. Другие чувствуют себя неуютно при любом округлении и продолжают пользоваться теми значениями, какие они получили без всяких

округлений.) Итак, мы анализируем преобразованные данные, т. е. данные, преобразованные * с помощью выбора какого бы то ни было конечного значения λ, и описываем результаты.

Альтернативный подход

Альтернативный, но эквивалентный способ вычислений, который предпочитают отдельные специалисты, сводится к преобразованию наблюдений к виду

$$V_i = W_i / \{J(\lambda, \mathbf{Y})\}^{1/n} \tag{5.3.7}$$

и максимизации выражения

$$L_{\max}(\lambda) = -1/2n \ln \hat{\sigma}^2(\lambda, \mathbf{V}), \tag{5.3.8}$$

где $\mathbf{V} \quad (V_1\ V_2, \ldots, V_n)'$,

$$\hat{\sigma}^2(\lambda, \mathbf{V}) = S(\lambda, \mathbf{V})/n, \tag{5.3.9}$$

а величина $S(\lambda, \mathbf{V})$ — остаточная сумма квадратов, полученная для модели $\mathbf{V} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$. Проще говоря, это означает, что мы можем минимизировать функцию $S(\lambda, \mathbf{V})$. Заметим, что с учетом уравнения (5.3.5) величина V_i в уравнении (5.3.7) может быть переписана:

$$V_i = W_i / \dot{Y}^{\lambda-1}, \tag{5.3.7а}$$

где

$$\dot{Y} = (Y_1 Y_2 \ldots Y_n)^{1/n} \tag{5.3.7б}$$

— среднее геометрическое значений Y-ов. Форма, в которой представлено уравнение (5.3.7а),— одна из тех, что обычно приводится. У нее несколько преимуществ. Она гораздо проще (см. указанную выше статью Дж. Бокса и Д. Кокса 1964 г., с. 216) благодаря изменению масштаба обеспечивает большую точность вычислений, особенно для больших значений λ, и позволяет выполнять все вычисления с помощью любой стандартной программы регрессионного анализа. А еще она допускает прямое сравнение остаточных сумм квадратов, поскольку масштабный делитель $\dot{Y}^{\lambda-1}$ фактически возвращает значения W_i назад к исходным величинам в выражении (5.3.7а). Пример, который мы приводим ниже, рассчитан в терминах W_i, но если читатель пожелает, он может для сравнения вести параллельные вычисления через V_i.

---

* См. примечание на с. 281. Некоторые специалисты считают нужным еще уменьшать соответствующим образом число степеней свободы с учетом числа оцениваемых параметров преобразования (т. е. на 1 для уравнения (5.3.1), на 2 для уравнения (5.3.22)). Это должно быть особенно важно при малых n. В наших примерах мы не делаем этих поправок, хотя, конечно, их можно было бы сделать.

Приближенный доверительный интервал для λ

Приближенный * 100 $(1-\alpha)$ %-ный доверительный интервал для λ включает те значения λ, которые удовлетворяют неравенству

$$L_{max}(\hat{\lambda}) - L_{max}(\lambda) \leqslant \chi_1^2(1-\alpha), \qquad (5.3.10)$$

где $\chi_1^2(1-\alpha)$ — процентная точка распределения хи-квадрат с одной степенью свободы, которая отсекает площадь, равную α, от верхнего хвоста этого распределения. Вот некоторые из этих значений:

α	0,10	0,05	0,025	0,01	0,001
$\chi_1^2(1-\alpha)$	2,71	3,84	5,02	6,63	10,83

Чтобы удовлетворить неравенству (5.3.10), мы просто нанесем на график зависимости $L_{max}(\lambda)$ от λ некоторую горизонтальную линию на уровне

$$L_{max}(\hat{\lambda}) - \frac{1}{2}\chi_1^2(1-\alpha) \qquad (5.3.11)$$

вертикальной шкалы. Эта линия пересечет кривую в двух точках (при двух значениях λ). Они и будут крайними точками приближенного доверительного интервала.

Пример 1. Данные в табл. 5.1 представляют собой фрагмент более обширного набора, приведенного в работе Дж. Дерринджера, где обсуждается эмпирическая модель вязкости сложного эластомера [5]

---

* А. Эткинсон (A. C. Atkinson), исследовав один частный факторный план 3 × 4 с четырьмя параллельными наблюдениями в ячейке, показал, что в случае нормально распределенных ошибок 95 %-ная доверительная полоса имеет более или менее обычные размеры (см.: Atkinson A. C. Testing transformations to normality.—Journal of the Royal Statistical Society, 1973, B—35, p. 473—479, особо p. 478). Типично это или нет для нормально распределенных ошибок — не известно.

[5] Эластомеры — это синтетические смеси, обладающие высокой эластичностью в сочетании со многими другими свойствами, что позволяет заменять ими естественные каучуки в огромном числе самых разнообразных вещей, но прежде всего, конечно, в автомобильных, самолетных и т. п. покрышках и в резино-технических изделиях. Обычно эластомеры — многокомпонентные смеси, но главных компонентов, как правило, три. Это полимер (смола, основа смеси), наполнитель (придающий ей жесткость) и пластификатор (или мягчитель, обеспечивающий гибкость, пластичность). В качестве основы часто используется бутадиен-стирольный сополимер, который рассматривается и в данном примере, где он обозначен SBR-1500 (последнее число, видимо, обозначает давление). Между прочим, каучуки такого рода были впервые получены промышленно по способу академика С. В. Лебедева в 1932 г. в СССР. Вот брутто-формула вулканизированного бутадиен-стирольного каучука: $[-CH_2-CH=CH-CH_2-]_n - [-CH_2-CH(C_6H_5)-]_m$. Как наполнители часто берутся сажа, силикагель и другие вещества, а пластификатором обычно служит какая-нибудь маслянистая жидкость, нередко получаемая из нефти. Поскольку смола — основной компонент, принято описывать состав, указывая число грамм наполнителя или пластификатора, приходящееся на 100 г смолы. В работе для обозначения этого соотношения принята аббревиатура phr = г/100 г смолы. Ясно, что (динамическая) вязкость (или, как еще говорят, внутреннее трение) служит одной из главных характеристик таких смесей. Она характеризует пластичность смесей, т. е. говорит о связи растяжения с деформацией. Теоретически эта связь может описываться так называемым двупараметрическим урав-

с наполнителем и пластификатором (см.: Derringer G. C. An empirical model for viscosity of filled and plasticized elastomer compounds.— Journal of Applied Polymer Science, 1974, **18**, p. 1083—1101). (Эти данные воспроизводятся с разрешения издателя John Wiley & Sons, Inc.) Мы хотим найти некоторое преобразование вида $(Y^{\lambda}-1)$ для $\lambda \neq 0$ или $\ln Y$ для $\lambda = 0$, которое обеспечило бы хорошее соответствие данным модели первого порядка. Наша модель для уравнения (5.3.2) имеет вид

$$W = \beta_0 + \beta_1 f + \beta_2 p + \varepsilon, \qquad (5.3.12)$$

где f — уровень наполнителя, а p — уровень пластификатора. (Название пластификатора приведено в первом столбце табл. 5.1.)

Таблица 5.1. **Вязкость по Муни MS_4 при 100 °С в зависимости от уровней наполнителя и масла в SBR-1500***

Нафтеновое масло**, phr, p	Наполнитель, phr, f					
	0	12	24	36	48	60
0	26	38	50	76	108	157
10	17	26	37	53	83	124
20	13	20	27	37	57	87
30	—	15	22	27	41	63

* «Филлипс Петролеум Ко».
** «Секолит Процесс Ойл», «Сан Ойл Ко».

нением Муни. На его основе разработан метод измерения «вязкости по Муни», широко применяемый для невулканизированных каучуков и резиновых смесей. Это промышленно-технологический метод, обеспечивающий измерение некоторого произвольного значения эффективной вязкости при заданных геометрических размерах и конструкции вискозиметра, а также скорости сдвига и напряжении сдвига. В СССР действует ГОСТ 10722—64, по которому определяется зависимость вязкости (M) от времени при различных температурах и скоростях вращения ротора на сдвиговом дисковом ротационном вискозиметре. Причем, за единицу «вязкости по Муни» принят момент сопротивления сдвигу M, равный 0,083 н. м. Судя по обсуждаемой работе, в американской практике принято проводить измерения при двух различных конструкциях вискозиметра, а именно с длинным цилиндром (ротором)—ML и с коротким цилиндром — MS. Цифры в подстрочных индексах обозначают время. Так, MS_4 означает, что измерение велось с коротким цилиндром в течение четырех минут.

Пластификатором в нашей работе служит нафтеновое масло. Это продукт переработки нефти нафтенового основания. Причем, сополимер получен на известной фирме «Филлипс Петролеум Ко.», а нафтеновое масло — с помощью варианта крекинг-процесса, запатентованного фирмой «Сан Ойл Ко.».

В исследованиях процессов производства эластомеров широко используются методы математического моделирования, особенно планирование эксперимента. См., например: Маркова Е. В., Путилина С. Н. Планирование эксперимента при получении и переработке полимерных материалов.— Журнал ВХО им. Д. И. Менделеева, 1980, № 1, с. 72—80 или сборник: Планирование эксперимента и применение вычислительной техники в процессе синтеза резины/Под ред В. Ф. Евстратова, А. Г. Шварца.— М.: Химия, 1970.— 255 с.— *Примеч. пер.*

Обратите внимание, что значения отклика разбросаны в диапазоне от 157 до 13, отношение границ которого дает 157/13 = 12,1. Когда отношение наибольшего значения отклика к наименьшему составляет примерно порядок (т. е. около 10) или более того, есть надежда, что преобразование Y окажется эффективным.

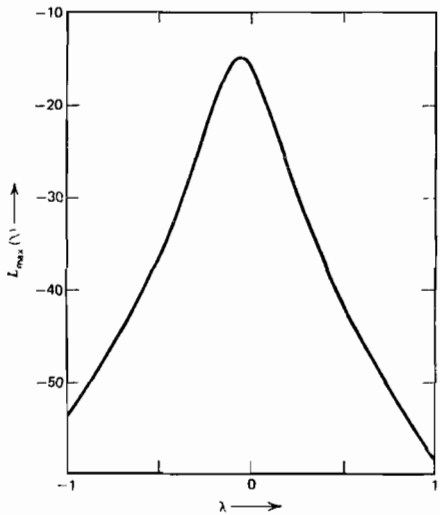


Рис. 5.1. Графнк зависимости L_{max} (λ) от λ для данных о вязкости эластомера

В табл. 5.2 приведены значения L_{max} (λ) для различных λ. (Начальный набор значений для $\lambda = -2\,(0,1)\,2$ был затем уточнен по более густой сетке $\lambda = -0,2\,(0,01)\,0,1$ в окрестности пика кривой.) Сглаженная кривая, проведеиная по этим точкам, представлена на рис. 5.1. Мы видим, что максимум L_{max} (λ) приходится на значение, близкое к $\lambda = -0,05$. Это совсем рядом с нулем, значит, преобразование

$$W = \ln Y \qquad (5.3.13)$$

может оказаться подходящим для этого набора данных.

Приближенный 95%-ный доверительный интервал, полученный по уравнению (5.3.11), оказался равным $-0,135 \leqslant \lambda \leqslant 0,03$. Чтобы показать более подробно, как это вычислено, на рис. 5.2 пик кривой L_{max} (λ) построен в крупном масштабе. Мы видим, что использование $\lambda = 0$ состоятельно и согласуется с этими вычислениями и что наше преобразование служит хорошей оценкой. Такие значения λ, как $\lambda = 1$ (отказ от всяких преобразований), $\lambda = \frac{1}{2}$ (преобразование квадратного корня), $\lambda = -1$ (обратное преобразование), да и многие другие, полностью исключаются теперь из числа преобразований, возможных при наших данных. (Плохое оценивание могло бы проявиться в относительно более широком приближенном доверительном

Таблица 5.2. **Значении L_{max} (λ), соответствующие выбраниым значениям λ для данных о вязкости**

λ	L_{max} (λ)	λ	L_{max} (λ)	λ	L_{max} (λ)	λ	L_{max} (λ)
—1,0	—53,70	—0,15	—17,40	—0,04	—14,82	0,2	—26,53
—0,8	—47,68	—0,10	—15,47	—0,02	—15,09	0,4	—37,27
—0,6	—40,52	—0,08	—15,02	0,00	—15,60	0,6	—45,69
—0.4	—31,46	—0,06	—14,80	0,05	—17,65	0,8	—52,67
—0,2	—20,07	—0,05	—14,78	0,10	—20,43	1,0	—58,80

интервале для λ, который указывал бы, что в широком диапазоне значений λ различие в их использовании крайне мало.) Воспользовавшись преобразованием натурального логарифма к исходным данным, получим преобразованные данные, представленные в табл. 5.3. Вот какова самая лучшая плоскость, которую можно подобрать методом наименьших квадратов по этим преобразованным данным:

$$\widehat{\ln Y} = 3{,}212 + 0{,}03088f - 0{,}03152p. \qquad (5.3.14)$$

Соответствующая таблица дисперсионного анализа представлена табл. 5.4. Из вариации относительно среднего $100\,R^2 = 99{,}51$ % объясняет эта модель с тремя параметрами, а величина статистики F-критерия для всей регрессии равна 2045, что действительно весьма значимо. Ясно, что получено превосходное согласие между данными и моделью.

Если бы мы построили модель первого порядка по *непреобразованным* данным, то мы получили бы

$$\hat{Y} = 28{,}184 + 1{,}55f - 1{,}717p \qquad (5.3.15)$$

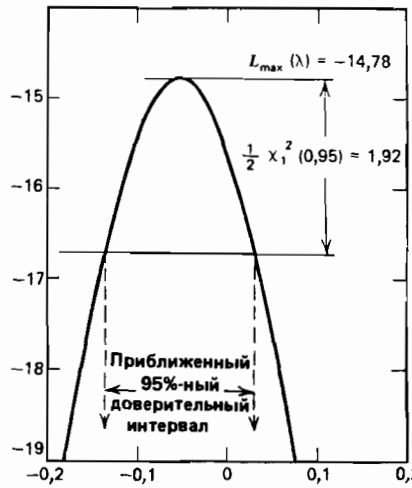


Рис. 5.2. Получение приближенных 95 %-ных доверительных интервалов для λ по данным о вязкости эластомера

с величиной $100\,R^2$, равной 87,93, и с общим $F = 72{,}9$ (см. табл. 5.5). Это само по себе отличное приближение, но улучшение при переходе к $\ln Y$ весьма впечатляюще. (В иных случаях начальное приближение могло бы оказаться довольно плохим, а подходящее преобразование благополучно обеспечило бы значимость подобранной модели. Иногда преобразования позволяют ограничиться полиномом более низкой степени, чем это было бы нужно в противном случае. Ниже мы увидим, что это верно и в данном примере тоже.)

Таблица 5.3. **Преобразованные значения $W = \ln Y$ для данных из табл. 5.1**

Нафтеновое масло, phr, p	Наполнитель, phr, f					
	0	12	24	36	48	60
0	3,258	3,638	3,912	4,331	4,682	5,056
10	2,833	3,258	3,611	3,970	4,419	4,820
20	2,565	2,996	3,296	3,611	4,043	4,466
30	—	2,708	3,091	3,296	3,714	4,143

Таблица 5.4. Дисперсионный анализ модели первого порядка для f и p, подобранной по логарифмированным данным о вязкости

Источник	Число степеней свободы	SS	MS	F
b_0	1	319,44855	—	
$b_1, b_2 \mid b_0$	2	10,55167	5,27583	2045
Остаток	20	0,05171	0,00258	
Общий	23	330,05193		

Таблица 5.5. Дисперсионный анализ модели первого порядка для f и p, подобранной для непреобразованных данных о вязкости

Источник	Число степеней свободы	SS	MS	F
$b_1, b_2 \mid b_0$	2	27 842,62	13 921,31	72,9
Остаток	20	3 820,60	191,03	
Общий, скорректированный	22	31 663,22		

Кодирование предикторов. Для преодоления трудностей с нашим примером на следующих этапах его обсуждения мы воспользуемся двумя предикторами f и p в тех единицах, в которых они и были заданы. В случаях, подобных нашему, когда уровни f и p выбраны с равным шагом, кодирование

$$x_1 = (f - 30)/6 \quad \text{и} \quad x_2 = (p - 15)/5 \tag{5.3.16}$$

приводит к уровням (кодированным) $x_1 = -5, -3, -1, 1, 3, 5$ и $x_2 = -3, -1, 1, 3$, несколько упрощая вычисления. Заметим, что простое кодирование предикторов таким способом не оказывает какого бы то ни было влияния на оценивание λ. Зато иногда подходящее кодирование будет упрощать регрессионные вычисления. Так, например, если бы при $f = 0$ и $p = 30$ наблюдение в табл. 5.1 не было бы потеряно, то кодирование такого рода, как показано в (5.3.16), сделало бы столбцы x_1 и x_2 взаимно ортогональными и ортогональными к единичному столбцу в матрице $\mathbf{X}$. (Заметим, однако, что *преобразование* предикторов, скажем, $x_1 = f^{\alpha_1}, x_2 = p^{\alpha_2}$, изменило бы сложность задачи и *повлияло бы* на оценивание λ.)

Важность проверки остатков

Преобразования отклика влияют на распределение ошибок. Наше предположение состоит в том, что *после* преобразования ошибки преобразованного отклика должны быть нормальными $N(\mathbf{0}, \mathbf{I}\sigma^2)$. Значит, важно провести анализ остатков для модели, которая в конце концов

подобрана, чтобы посмотреть, не проявится ли что либо, нарушающее эти предположения. Остатки для модели первого порядка, заданной уравнением (5.3.14), представлены в табл. 5.6. Мы оставляем их анализ в качестве упражнения для читателей.

Таблица 5.6. **Остатки, умноженные на 1000, для модели первого порядка, подобранной по логарифмированным данным о вязкости**

Нафтеновое масло, phr, p	Наполнитель, phr, f					
	0	12	24	36	48	60
0	46	55	—41	7	—13	—9
10	—64	—10	—27	—39	39	70
20	—17	43	—27	—83	—21	31
30	—	71	83	—83	—36	23

Второй метод оценки λ

Во втором методе оценивания мы выбираем λ так, чтобы минимизировать определенную величину, которую мы хотим уменьшить, и/или максимизировать определенную величину, которую мы хотим увеличить. Пусть, например, исходный отклик Y допускает разумное описание моделью второго порядка относительно X_1 и X_2:

$$\beta_0+\beta_1X_1+\beta_2X_2+\beta_{11}X_1^2+\beta_{22}X_2^2+\beta_{12}X_1X_2+\varepsilon, \qquad (5.3.17)$$

а мысль о преобразовании Y в W возникла в связи с тем, что для преобразованного отклика может подойти модель первого порядка $\beta_0+\beta_1X_1+\beta_2X_2+\varepsilon$. Мы могли бы подобрать уравнение (5.3.17) для W методом наименьших квадратов для некоторого множества значений λ и выбрать в качестве наилучшего для наших целей то из них, что минимизирует какую-либо подходящую статистику. Вполне возможно выбрать, скажем, значение F-критерия, связанного с дополнительной суммой квадратов SS (b_{11}, b_{22}, $b_{12}|b_0$, b_1, b_2) или отношение средних квадратов, получающихся из моделей второго и первого порядка. Для осуществления нашей идеи было бы хорошо, чтобы члены второго порядка оказались незначимыми при том значении λ, на котором мы в конце концов остановимся.

Пример 2. Возьмем снова данные о вязкости из табл. 5.1. Мы хотим найти преобразование в виде $W=(Y^\lambda-1)/\lambda$ при $\lambda\neq 0$ или $W=\ln Y$ при $\lambda=0$, которое допускало бы хорошую подгонку моделью первого порядка без необходимости в членах второго порядка. Сначала строим модель

$$W=\beta_0+\beta_1f+\beta_2p+\beta_{11}f^2+\beta_{22}p^2+\beta_{12}fp+\varepsilon, \qquad (5.3.18)$$

где, как и раньше, f — уровень наполнителя, а p — уровень пластификатора для наборов выбранных значений λ. (Понятно, можно было бы строить уравнение в кодированных переменных аналогично тому, как это сделано в уравнении (5.3.16), что, впрочем не оказало бы влияния на основные результаты.) Для каждого λ вычисляем:

MS_1 — средний квадрат, получающийся как $SS(b_1,\ b_2 \mid b_0)/2$,

MS_2 — средний квадрат, получающийся как $SS(b_{11}, b_{22}, b_{12} \mid b_0,\ b_1, b_2)3$,

$$\gamma = MS_2/MS_1. \tag{5.3.19}$$

Наносим на график значения γ в зависимости от значений λ, как показано на рис. 5.3. Числа, требующиеся для получения этого графика, приводятся в табл. 5.7. Отметим, что использование отношения γ позволяет обойти проблему масштаба, связанную с применением W вместо Y. Мы видим, что минимум γ приходится на значение порядка $\lambda = -0{,}05$, указывая на то, что выбор $\lambda = 0$ и логарифмического преобразования вполне разумен. Это точно тот же результат, к которому мы пришли и предыдущим методом. (Правда, у данной процедуры есть неудобство *, заключающееся в том, что мы не можем легко найти доверительный интервал для λ.) Такое преобразование отклика ведет к следующей модели второго порядка:

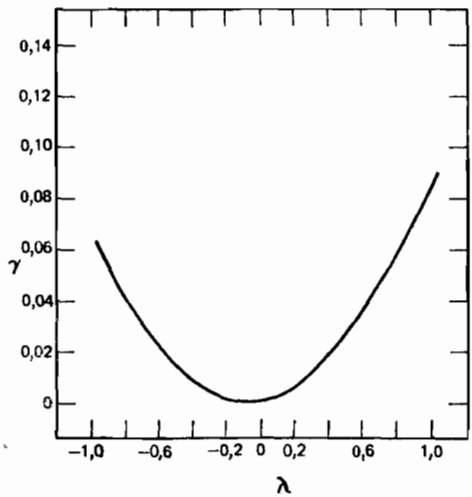


Рис. 5.3. Подбор модели второго порядка для преобразованных данных о вязкости. График зависимости $\gamma = MS_2/MS_1$ от λ

$$\widehat{\ln Y} = 3{,}231 + 0{,}02861f - 0{,}03346p + 0{,}00004416f^2 + \\ + 0{,}00011207p^2 - 0{,}00003718fp. \tag{5.3.20}$$

Соответствующая таблица дисперсионного анализа представлена табл. 5.8. Ясно, что выбранное преобразование вполне удачно, что в полной модели второго порядка нет никакой необходимости и что модель первого порядка из уравнения (5.3.14) безусловно адекватна. Для сравнения приведем уравнение модели второго порядка для непреобразованных данных:

$$\hat{Y} = 24{,}067 + 0{,}57387f - 0{,}82628p + 0{,}02639f^2 + 0{,}02752p^2 - 0{,}04930fp, \tag{5.3.21}$$

а также таблицу дисперсионного анализа (табл. 5.9). Выходит, что,

---

* Это неудобство проявляется и в этом примере. Правда, если в данном случае воспользоваться F-статистикой, то приближенный доверительный интервал все-таки можно получить с помощью $F(\upsilon_1, \upsilon_2, 1-\alpha)$, как описано в работе: Draper N. R., Hunter W. G. Ttansformations: some examples revisited.— Technometrics, 1969, **11**, p. 23—40.

Таблица 5.7. **Значении MS_1, MS_2 и $\gamma = MS_2/MS_1$ для выбранных λ по данным о вязкости**

λ	MS_1	MS_2	$\gamma = MS_2/MS_1$	λ	MS_1	MS_2	$\gamma = MS_2/MS_1$
—1,0	0,0037	0,00024	0,0649	0,025	6,375	0,00504	0,0008
—0,8	0,0152	0,00063	0,0415	0,05	7,705	0,00911	0,0012
—0,6	0,0633	0,00144	0,0228	0,1	11,272	0,02576	0,0023
—0,4	0,2698	0,00251	0,0093	0,2	24,236	0,1382	0,0057
—0,2	1,1782	0,00207	0,0018	0,4	114,23	1,980	0,0173
—0,1	2,485	0,00079	0,0003	0,6	552,36	19,379	0,0351
—0,05	3,618	0,00073	0,0002	0,8	2 739,1	160,05	0,0584
—0,025	4,368	0,00130	0,0003	1,0	13 921,3	1206,66	0,0867
0,0	5,276	0,00259	0,0005				

когда преобразование не делается, значимая кривизна поверхности второго порядка проявляется в данных.

Таблица 5.8. **Дисперсионный анализ модели второго порядка для f и p, построенной по логарифмированным данным о вязкости**

Источник	Число степеней свободы	SS	MS	F
b_1, $b_2 \mid b_0$	2	10,55167	5,27583	2037,0
b_{11}, b_{22}, $b_{12} \mid b_0$, b_1, b_2	3	0,00776	0,00259	1,0
Остаток	17	0,04395	0,00259	
Общий, скорректированный	22	10,60338		

Таблица 5.9. **Дисперсионный анализ модели второго порядка для f и p, построенной по исходным данным о вязкости**

Источник	Число степеней свободы	SS	MS	F
b_1, $b_2 \mid b_0$	2	27 842,616	13 921,308	1179,7
b_{11}, b_{22}, $b_{12} \mid b_0$, b_1, b_2	3	3 619,987	1 206,662	102,3
Остаток	17	200,615	11,801	
Общий, скорректированный	22	31 663,217*		

* Округлено с точностью до 0,001.

Преимущества метода максимума правдоподобия

Из двух методов, предназначенных для оценивания параметров преобразования, в большинстве практических ситуаций мы предпочли бы метод максимума правдоподобия. С его помощью мы всегда можем получить приближенный доверительный интервал или область, а, кроме того, здесь нужно только подобрать ту самую модель, которой мы интересуемся, без всяких дополнительных сложностей, обычно возникающих при втором методе. (Действительно, в некоторых случаях данные могут оказаться неадекватными альтернативной модели более высокого порядка.) Правда, второй метод может оказаться полезным, когда нам понадобится исследовать ряд критериев. Тогда можно одновременно построить графики зависимостей каждого критерия от λ и сравнить значения λ, получающиеся на каждом из графиков.

Приближенный метод оценивания

Относительно простой приближенный метод оценки λ в больших массивах данных, когда множество остаточных средних квадратов само собой возникает в ходе анализа, описан в работе: Hinz P. N., Eagles H. A. Estimation of a transformation for the analysis of some agronomic and genetic experiments.— Crop Science, 1976, **16**, p. 280—283.

Семейство преобразований отклика с большими возможностями

Вводя новый параметр, мы можем расширить диапазон допустимых преобразований сверх того, что обсуждалось выше. Рассмотрим двухпараметрическое семейство:

$$W = \begin{cases} \dfrac{(Y+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \text{при } \lambda_1 \neq 0, \\ \ln(Y+\lambda_2) & \text{при } \lambda_1 = 0, \end{cases} \tag{5.3.22}$$

где обязательно $Y > -\lambda_2$. Методы, применяемые для однопараметрического семейства (когда $\lambda_2 = 0$), можно распространить и на этот случай. Теперь вместо Y мы имеем $Y + \lambda_2$ и вынуждены искать оценки метода максимума правдоподобия $(\hat{\lambda}_1, \hat{\lambda}_2)$ на двумерной сетке. Точно так же и при вычислениях, связанных с получением доверительной области для (λ_1, λ_2), мы берем теперь хи-квадрат с *двумя* степенями свободы, вместо одной поскольку имеются два параметра преобразования. Работают в точности те же идеи, только вычисления становятся более сложными. Пример такого рода кратко обсуждается в упомянутой выше работе Дж. Бокса и Д. Кокса 1964 г. на с. 225—226.

Нормирующее преобразование (аналогичное уравнению (5.3.7а), когда $\lambda_2 = 0$) в этом случае таково:

$$V_l = W_l / \dot{Y}^{\lambda_1 - 1}, \tag{5.3.22а}$$

где

$$\dot{Y} = \{(Y_1 + \lambda_2)(Y_2 + \lambda_2) \ . \ . \ . \ (Y_n + \lambda_2)\}^{1/n} \qquad (5.3.22б)$$

есть среднее геометрическое значений $(Y_i + \lambda_2)$.

Заметим, что поскольку $Y > -\lambda_2$, мы должны избегать выбора значений λ, удовлетворяющих неравенству $\lambda_2 \leqslant -Y_{\min}$. Хотя и проще всего отбрасывать те значения λ_2, которые нарушают ограничение,— это не годится. Требуется локальный максимум, удовлетворяющий данному ограничению.

Альтернативное семейство преобразований откликов

Когда графики остатков явно свидетельствуют о симметричном, но не нормальном распределении ошибок, может оказаться полезным однопараметрическое семейство степенных функций от *модулей*:

$$W = \begin{cases} (\text{знак } Y)\,[\{|Y|+1\}^{\lambda} - 1]/\lambda, & \lambda \neq 0, \\ (\text{знак } Y)\ln\{|Y|+1\}, & \lambda = 0. \end{cases}$$

За примером обратитесь к работе: J o h n J. A., D r a p e r N. R. An alternative family of transformations.— Applied Statistics, 1980, 29, p. 190—197.

Семейство степенных функций для долей

Все то же самое можно использовать и для оценки параметра λ в семействе

$$P = \{p^{\lambda} - (1-p)^{\lambda}\}/\lambda, \qquad (5.3.23)$$

где p — наблюдаемая доля случаев, когда некоторое событие имеет место. Вообще говоря, наблюдаемые значения p могут зависеть от массы предикторов X_1, X_2, . . . , и следовало бы постулировать модель в общем виде:

$$P = f(X_1,\ X_2,\ \ldots,\ \lambda,\ \boldsymbol{\beta}) + \varepsilon,$$

где $\boldsymbol{\beta}$ — вектор параметров. Значение λ следовало бы выбирать так, чтобы получилась наилучшая подгонка к имеющимся данным в предположении, что $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\ \mathbf{I}\sigma^2)$.

Степенное преобразование (5.3.23) было предложено Дж. Тьюки. Для работы со статистическим распределением величины W, когда p имеет равномерное распределение, хорошей предварительной подготовкой будет чтение статьи Б. Джойнера и Дж. Розенблатта о некоторых свойствах размаха в выборках из симметричного λ-распределения Тьюки (см.: J o i n e r B. L., R o s e n b l a t t J. R. Some properties of the range in samples from Turkey's symmetric λ distributions.— J. Amer. Statist. Assoc., 1971, 66, p. 394—399; см. также ссылки на литературу в этой статье).

Два примера преобразования долей приведены ниже на с. 294—296. Одно из них — частный случай обсуждавшегося выше семейства, а второе — приближение к некоторому частному случаю.

Преобразования для стабилизации дисперсии

Если преобразованные данные анализируются методом наименьших квадратов, то важно, чтобы дисперсия отклика оказалась независимой от его среднего значения. Там, где известно заранее или где можно установить эмпирически, что стандартное отклонение непреобразованного отклика Y, скажем σ_Y, связано некоторой определенной функцией $f(\eta)$ со своим средним значением, $\eta = E(Y)$, мы можем получить подходящее преобразование непосредственно, воспользовавшись преобразованным значением $h(Y)$ из выражения:

$$\frac{\partial h(Y)}{\partial Y} \propto \frac{1}{f(Y)}. \qquad (5.3.24)$$

Иными словами, мы получим $h(Y)$, интегрируя $1/f(Y)$ по Y. Несколько хорошо известных преобразований, получающихся таким образом, приведено в табл. 5.10. Отметим, что некоторые из членов этого семейства определяются по уравнению (5.3.1).

Таблица 5.10. **Преобразования, подходящие для стабилизации дисперсий, когда $\sigma_Y = f(\eta)$**

Природа зависимости $\sigma_Y = f(\eta)$	Преобразование, стабилизирующее дисперсию
$\sigma_Y^* \propto \eta^k$ $(Y \geqslant 0)$	Y^{1-k}
и, в частности,	
$\sigma_Y \propto \eta^{1/2}$ (распределение Пуассона) $(Y \geqslant 0)$	$Y^{1/2}$
$\sigma_Y \propto \eta$ $(Y \geqslant 0)$	$\ln Y$
$\sigma_Y \propto \eta^2$ $(Y \geqslant 0)$	Y^{-1}
$\sigma_Y \propto \eta^{1/2}(1-\eta)^{1/2}$ (биномиальное распределение) $(0 \leqslant Y \leqslant 1)$	$\sin^{-1}(Y^{1/2})$**
$\sigma_Y \propto (1-\eta)^{1/2}/\eta$ (отрицательное биномиальное распределение) $(0 \leqslant Y \leqslant 1)$	$(1-Y)^{1/2}-(1-Y)^{3/2}/3$
$\sigma_Y \propto (1-\eta^2)^{-2}$ $(-1 \leqslant Y \leqslant 1)$	$\ln\{(1+Y)/(1-Y)\}$

* Символ $\propto$ означает «пропорционально». — *Примеч. пер.*

** В отечественной литературе вместо $\sin^{-1}$ часто употребляют обозначение arcsin. — *Примеч. пер.*

Преобразования откликов, задаваемых долями

Многие виды данных об откликах представляют собой доли, $0 \leqslant Y_i \leqslant 1$, получаемые как числа «успехов» (они могут быть и неудачами), появляющихся при большом числе «опытов». Так, например, шесть крыс из десяти, справившихся с поставленной задачей, дают $Y_i = 0{,}60$. Данные, представленные в виде долей, как правило, не имеют одинаковых дисперсий, так как $V(Y_i) = \pi_i(1-\pi_i)/m_i$, где $E(Y_i) = \pi_i$, а m_i — число опытов. Вот два наиболее распространенных преобразования для данных такого рода.

1. *Преобразование логарифма преобладания*[6]. Положим

$$W_i = \ln\{Y_i/(1-Y_i)\}.$$

Тогда W будет натуральным логарифмом «отношения преобладания» $Y_i/(1-Y_i)$, отношения доли успехов к доле неудач. При подборе модели

$$W_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi} + \varepsilon_i$$

по нашим данным $(W_i, X_{1i}, \ldots, X_{pi})$, $i = 1, 2, \ldots, n$, мы возьмем взвешенный метод наименьших квадратов, так как приближенно $V(W_i) = 1/\{\pi_i(1-\pi_i)m_i\}$, и это — отнюдь не константа. Для демонстрации такого утверждения воспользуемся тем фактом, что при малых x $\ln(1+x) = x - x^2/2 + x^3/3 - \ldots \simeq x$. Тогда, опуская на минуту подстрочный индекс i, мы увидим, что (все результаты приближенные)

$$\ln Y = \ln\pi + (Y-\pi)/\pi,$$

т. е.

$$E(\ln Y) = \ln\pi \quad \text{и} \quad V(\ln Y) = (1-\pi)/(\pi m).$$

Аналогично

$$E\{\ln(1-Y)\} = \ln(1-\pi) \quad \text{и} \quad V\{\ln(1-Y)\} = \pi/\{(1-\pi)m\},$$

а

$$\operatorname{covar}\{\ln Y,\ \ln(1-Y)\} = -1/m.$$

Отсюда следует, что $V(W) = \{(1-\pi)/\pi + \pi/(1-\pi) - 2(-1)\}/m = 1/\{\pi(1-\pi)m\}$. Понятно, что эти n дисперсий $V(W_i) = 1/\{\pi_i(1-\pi_i)m_i\}$ не известны, но их оценками служат соответствующие значения $s_i^2 = 1/\{Y_i(1-Y_i)m_i\}$ для $i = 1, 2, \ldots, n$, а само оценивание происходит так же, как в параграфе 2.11. В этом случае матрица $\mathbf{V}$ диагональна с оцененными диагональными элементами $s_1^2, s_2^2, \ldots, s_n^2$.

2. *Преобразование арксинуса.* Как показано в табл. 5.10, преобразование $U = \sin^{-1}Y^{1/2}$ будет стабилизировать дисперсию, если все те выборки, по которым определяются наблюдаемые значения Y, будут иметь одинаковый объем, допустим m. На самом деле несколько лучше преобразование $W = 2\sin^{-1}Y^{1/2}$, поскольку оно дает постоянную теоретическую дисперсию, равную $1/m$. Заметим, что если только m не постоянно для наших данных, это преобразование не будет стабилизировать дисперсию. В таком случае нужно преобразование $Z_i = 2\,m_i^{1/2}\sin^{-1}Y^{1/2}$, где Y_i определяется по m_i опытам. Еще отметим, что данные в середине диапазона долей (скажем, между 0,30 и 0,70) не будут слишком меняться под воздействием такого преобразования,

---

[6] Термин «логарифм преобладания» («log odds») введен в связи с анализом таблиц сопряженности признаков (см., например: Аптон Г. Анализ таблиц сопряженности/Пер. с англ.— М.: Статистика, 1982.— 139 с. (особо с. 19)). Он представляет собой логарифм отношения числа благоприятных исходов к числу неблагоприятных при схеме испытаний с двумя исходами (схеме Бернулли). Это отношение в свою очередь называется «отношением преобладания».— *Примеч. пер.*

так как само преобразование, в этом интервале значений приблизительно линейно. Табл. 5.11 представляет собой краткое извлечение из таблиц преобразований $U = \sin^{-1} Y^{1/2}$ и $W = 2 \sin^{-1} Y^{1/2}$. Обратите внимание, что и U и W приведены не в градусах, а в радианах. Если же использовались градусы, то следует умножить значение U на $360/(2\pi)$, а W — на $360/\pi$. Кроме того, для перевода в градусы дисперсии W, равной $1/m$ в радианах, следует умножить ее на $\{360/(\pi)\}^2$.

В общем, всегда надо помнить, что не существует гарантии того, что использование этих преобразований всегда лучше, чем прямой анализ непреобразованных долей. Многое зависит от данных. Эффективность преобразования лучше всего оценить, осуществив это преобразование, затем проверив адекватность модели и исследовав структуру остатков, которые получатся в результате.

Преобразование предикторов

Известно очень много возможных преобразований предикторов. Один из полезных типов преобразований, применяемых во множестве случаев, это степенные преобразования:

$$Z_i = \begin{cases} X^{\alpha_i} & \text{при} \quad \alpha_i \neq 0, \\ \ln X_i & \text{при} \quad \alpha_i = 0 \end{cases} \tag{5.3.25}$$

для $i = 1, 2, \ldots, k$, где $X_1, X_2, \ldots, X_k$ — исходные (непреобразованные) предикторы, а α_i — параметры, подлежащие оценке. Самый лучший способ оценки α_i — это оценка их одновременно с параметрами постулируемой модели с помощью методов нелинейного оценивания (см. гл. 10). Или же можно воспользоваться итеративным методом, описанным в работе: Box G. E., Tidwell P. W. Transformation of the independent variables.— Technometrics, 1962, 4, p. 531—550.

Важно иметь в виду, что преобразования предикторов не влияют на распределения ошибок отклика. Но, конечно, они воздействуют на исследование остатков после подбора заданной модели, что, впрочем, верно для построения любой регрессионной модели.

Комбинируя оба множества методов, описанных выше, мы можем одновременно преобразовывать и отклик, и предикторы, или же делать это в последовательных итерациях. В общем вычисления получаются более сложными.

Комментарии

Когда мы делаем преобразование, невозможно соотносить параметры модели, построенной по преобразованным данным, с параметрами первоначальной модели для непреобразованных данных. Обычно здесь нет математической эквивалентности, если не принимать во внимание приближения такого типа, как разложение в ряд Тейлора. Так, например, если вместо модели $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$ мы построим модель $Y^\lambda = \alpha_0 + \alpha_1 X + \varepsilon$, то соотношение между β_0, β_1,

β_{11} и λ, α_0, α_1 совсем не ясно. Попытки отыскать такую связь обычно не продуктивны.

Когда несколько наборов данных получаются в аналогичных экспериментальных ситуациях, вовсе не обязательно придется испытывать полным анализом все эти наборы, чтобы отыскать подходящие преобразования. Довольно часто находится одно такое преобразование, которое срабатывает во всех случаях.

Таблица 5.11. **Краткая таблица преобразований $U=\sin^{-1}Y^{1/2}$ и $W = 2U$, используемых для преобразования данных типа долей. Дисперсия $V(U) = 1/4m$, а $V(W) = 1/m$, где m — число наблюдений, по которым вычислялся Y**

Y	U	W	Y	U	W	Y	U	W	Y	U	W
0,01	0,10	0,20	0,26	0,54	1,07	0,51	0,80	1,59	0,76	1,06	2,12
0,02	0,14	0,28	0,27	0,55	1,09	0,52	0,81	1,61	0,77	1,07	2,14
0,03	0,17	0,35	0,28	0,56	1,12	0,53	0,82	1,63	0,78	1,08	2,17
0,04	0,20	0,40	0,29	0,57	1,14	0,54	0,83	1,65	0,79	1,09	2,19
0,05	0,23	0,45	0,30	0,58	1,16	0,55	0,84	1,67	0,80	1,11	2,21
0,06	0,25	0,49	0,31	0,59	1,18	0,56	0,85	1,69	0,81	1,12	2,24
0,07	0,27	0,54	0,32	0,60	1,20	0,57	0,86	1,71	0,82	1,13	2,27
0,08	0,29	0,57	0,33	0,61	1,22	0,58	0,87	1,73	0,83	1,15	2,29
0,09	0,30	0,61	0,34	0,62	1,25	0,59	0,88	1,75	0,84	1,16	2,32
0,10	0,32	0,64	0,35	0,63	1,27	0,60	0,89	1,77	0,85	1,17	2,35
0,11	0,34	0,68	0,36	0,64	1,29	0,61	0,90	1,79	0,86	1,19	2,37
0,12	0,35	0,71	0,37	0,65	1,31	0,62	0,91	1,81	0,87	1,20	2,40
0,13	0,37	0,74	0,38	0,66	1,33	0,63	0,92	1,83	0,88	1,22	2,43
0,14	0,38	0,77	0,39	0,67	1,35	0,64	0,93	1,85	0,89	1,23	2,47
0,15	0,40	0,80	0,40	0,68	1,37	0,65	0,94	1,88	0,90	1,25	2,50
0,16	0,41	0,82	0,41	0,69	1,39	0,66	0,95	1,90	0,91	1,27	2,53
0,17	0,42	0,85	0,42	0,71	1,41	0,67	0,96	1,92	0,92	1,28	2,57
0,18	0,44	0,88	0,43	0,72	1,43	0,68	0,97	1,94	0,93	1,30	2,61
0,19	0,45	0,90	0,44	0,73	1,45	0,69	0,98	1,96	0,94	1,32	2,65
0,20	0,46	0,93	0,45	0,74	1,47	0,70	0,99	1,98	0,95	1,35	2,69
0,21	0,48	0,95	0,46	0,75	1,49	0,71	1,00	2,00	0,96	1,37	2,74
0,22	0,49	0,98	0,47	0,76	1,51	0,72	1,01	2,03	0,97	1,40	2,79
0,23	0,50	1,00	0,48	0,77	1,53	0,73	1,02	2,05	0,98	1,43	2,86
0,24	0,51	1,02	0,49	0,78	1,55	0,74	1,04	2,07	0,99	1,47	2,94
0,25	0,52	1,05	0,50	0,79	1,57	0,75	1,05	2,09	1,00	1,57	3,14

Из того факта, что существует общий анализ для поиска преобразований, отнюдь не следует, что его надо всегда использовать. Нередко неформальные графики данных ясно показывают, что нужно преобразование вполне определенного вида (такое, как $\ln Y$ или $1/Y$). В таких случаях более формальный анализ можно рассматривать как полезный метод проверки, остающийся в резерве [7].

[7] Проблема преобразований — одна из важнейших в статистике и особенно в математическом моделировании. Она имеет множество аспектов. Оригинальный взгляд на преобразования, частично нашедший отражение и в настоящей работе, принадлежит американскому статистику Дж. Тьюки. См., например: Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ/Пер. с англ. Под ред. В. Ф. Писаренко.—М.: Мир, 1981.— 693 с.; Мостеллер Ф.,

5.4. ИСПОЛЬЗОВАНИЕ «ФИКТИВНЫХ» ПЕРЕМЕННЫХ В МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Общая концепция «фиктивных» переменных и пример их использования

Факторы, применяемые в регрессионных задачах, обычно могут принимать значения из какого-либо непрерывного интервала. Иногда мы можем вводить фактор, который имеет два или более различных уровня. Например, данные можно получать на трех машинах, или на двух фабриках, или с помощью шести операторов. В таком случае мы не можем построить непрерывную шкалу для факторов «машина», или «фабрика», или «оператор». Мы можем приписать этим факторам некоторые уровни по порядку, учитывая тот факт, что различные машины, фабрики или операторы могут иметь независимые детерминированные эффекты в отклике. Переменные такого типа обычно называют фиктивными переменными. Обычно (но не всегда) они не связаны с физическими уровнями, которые могут существовать у факторов сами по себе.

Первый пример фиктивной переменной — это дополнительная переменная X_0 при члене β_0 в регрессионной модели (она всегда равна единице). Переменную X_0 совсем не обязательно вводить в модель, но ее использование иногда обеспечивает удобство в обозначениях. Другие фиктивные переменные вводятся, как мы увидим, из соображений, более важных, чем просто удобство обозначений, что имеет место, скажем, при применении регрессионных методов к задачам дисперсионного анализа, как это показано в гл. 9.

Фиктивные переменные для разбиения данных на блоки

Допустим, мы хотим отразить в модели представление о том, что два типа машин (скажем, тип A и тип B) дают различные уровни отклика в дополнение к вариации, обусловленной другими факторами. Один путь состоит в том, чтобы включить в модель фиктивную переменную Z и коэффициент регрессии, скажем α, так что в модели появится дополнительный член αZ. Коэффициент α можно оценить од-

---

Тьюки Дж. Анализ данных и регрессия/Пер. с англ. Под ред. Ю. П. Адлера.— М.: Финансы и статистика, 1982, вып. 1.— 317 с. (особо гл. 4—6, с. 92—127, 252—265).

С одной стороны, преобразования тесно связаны со стремлением удовлетворить предпосылкам регрессионной модели, а с другой — с априорными представлениями и возможностями теоретического анализа. Вопрос с преобразованиями далеко не исчерпан. См.: Кабанова О. В. Критерии и методы преобразования переменных при построении статистических моделей.— Заводская лаборатория, 1979, 45, № 3, с. 245—248; Кабанова О. В. Методы преобразования переменных.— В кн.: Рузинов Л. П., Слободчикова Р. И. Планирование эксперимента в химии и химнческой технологии.— М.: Химня, 1980.— 280 с.; Маркова Е. В., Шляпников Д. А. Преобразование данных. — В кн.: Статистические методы в теории обеспечения эксплуатации. — Вопросы кибернетики-94.— М.: ВИНИТИ, 1982, с. 88—104.— *Примеч. пер.*

новременно с оцениванием β-коэффициентов. Фактору Z можно приписать следующие значения:

$Z = 0$, если наблюдение получено с машины A;

$Z = 1$, если наблюдение получено с машины B.

Фактически годятся любые два различных значения Z, хотя приведенные выше обычно оказываются наилучшими. Однако иногда удобнее другие обозначения. Пусть, например, из общего числа наблюдений n часть n_1 получена на машине типа A, а $n_2 = n - n_1$ — на машине типа B. Тогда если мы выберем уровни

$$Z = \frac{-n_2}{\sqrt{n_1 n_2 (n_1 + n_2)}} \quad \text{для машины } A$$

и

$$Z = \frac{n_1}{\sqrt{n_1 n_2 (n_1 + n_2)}} \quad \text{для машины } B,$$

то соответствующие столбцы матрицы **X** будут ортогональны к «столбцу β_0», и сумма их квадратов будет равна единице, что может быть также удобно.

(Примечание. Если желательно рассматривать три разные машины, то потребуются две фиктивные переменные: Z_1 и Z_2. Тогда мы получим

$$\begin{aligned}(Z_1, Z_2) &= (1, 0) \quad \text{для машины } A,\\ &= (0, 1) \quad \text{для машины } B,\\ &= (0, 0) \quad \text{для машины } C,\end{aligned}$$

и модель будет включать дополнительные члены $\alpha_1 Z_1 + \alpha_2 Z_2$ с коэффициентами α_1 и α_2, требующими оценивания. Снова возможно много различных вариантов уровней. Если желательны столбцы, которые ортогональны к «столбцу β_0» и имеют единичную сумму квадратов, то можно достигнуть этого, положив

$$\begin{aligned}(Z_1, Z_2) &= \left(\frac{-n}{\sqrt{n_1 n_3 (n_1 + n_3)}}, 0\right) \quad \text{для машины } A,\\ &= \left(0, \frac{-n_3}{\sqrt{n_2 n_3 (n_2 + n_3)}}\right) \quad \text{для машины } B,\\ &= \left(\frac{n_1}{\sqrt{n_1 n_3 (n_1 + n_3)}}, \frac{n_2}{\sqrt{n_2 n_3 (n_2 + n_3)}}\right) \quad \text{для машины } C,\end{aligned}$$

где n_1, n_2 и n_3 — соответственно числа наблюдений на машинах A, B и C.)

В общем, при продолжении такой процедуры мы можем прийти к r уровням для $(r-1)$ фиктивных переменных. Структура такой системы фиктивных переменных получится, если выписать единичную матрицу **I** размером $(r-1) \times (r-1)$ и приписать к ней строку, состоящую из $(r-1)$ нулей. Для случая $r = 6$ это показано на столбцах $X_1, X_2, \ldots, X_5$ во второй таблице из примера 2.

Теперь приведем пример такого использования фиктивных переменных.

Пример 1. Данные в табл. 5.12 представляют собой вес (Y) в фунтах и возраст (X) в неделях для тринадцати индеек, выращенных ко Дню Благодарения [8]. Четыре из этих индеек были выращены в штате Джорджия (Д), четыре — в Виргинии (В) и пять — в Висконсине (Ви). Нам хотелось бы связать Y и X простой линейной моделью, но разное происхождение индеек может стать камнем преткновения. Если это так, то как нам его обойти?

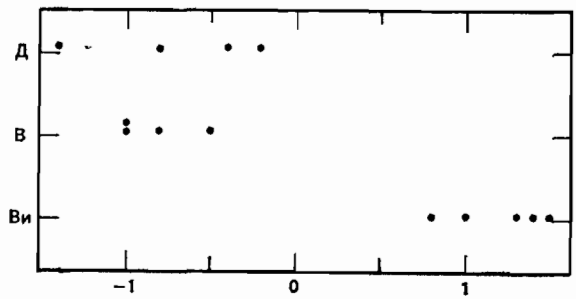


Рис. 5.4. Данные об индейках. Остатки для модели $\hat{Y} = 1{,}98 + 0{,}4167X$ в зависимости от происхождения индеек

Если бы мы построили регрессию Y на X, то получили бы такое уравнение: $\hat{Y} = 1{,}98 + 0{,}4167X$. Вот его остатки, выписанные по порядку: — 0,4; — 1,4; — 0,2; — 0,8; — 1,0; — 0,8; — 0,5; — 1,0; 0,8; 1,0; 1,3; 1,5; 1,4. Когда их нанесут на график в зависимости от происхождения индеек, они будут выглядеть так, как на рис. 5.4, где совершенно ясно видно, что надо брать различные уровни отклика. Для достижения этого введем фиктивные переменные Z_1 и Z_2, показанные в табл. 5.12, а затем методом наименьших квадратов подберем модель

$$Y = \beta_0 + \beta_1 X + \alpha_1 Z_1 + \alpha_2 Z_2 + \varepsilon. \tag{5.4.1}$$

Тогда получится уравнение

$$\hat{Y} = 1{,}43 + 0{,}4868X - 1{,}92Z_1 - 2{,}19Z_2. \tag{5.4.2}$$

[8] День Благодарения (Thanksgiving day) — национальный праздник США, введенный решением президента Авраама Линкольна в 1863 г. в честь события, имевшего место в 1621 г., через год после создания первыми колонистами Массачусетса, прибывшими из Англии на знаменитом корабле «Мейфлауэр», колонии Новый Плимут. Тогда им удалось собрать урожай маиса, обеспечивший выживание колонии. Празднуется каждый последний четверг ноября. См., например: Слезкин Л. Ю. Легенда, утопия, быль в ранней американской истории.— М.: Наука, 1981.— 160 с. (особо с. 102). Сложилась традиция подавать в этот день на праздничный стол жареную индейку. С этим обстоятельством и связан данный пример.— *Примеч. пер.*

Таблица 5.12. **Данные об индейках (X, Y, происхождение) и фиктивные переменные (Z_1, Z_2)**

X	Y	Происхождение	Z_1	Z_2	X	Y	Происхождение	Z_1	Z_2
28	13,3	Д	1	0	26	11,8	В	0	1
20	8,9	Д	1	0	21	11,5	Ви	0	0
32	15,1	Д	1	0	27	14,2	Ви	0	0
22	10,4	Д	1	0	29	15,4	Ви	0	0
29	13,1	В	0	1	23	13,1	Ви	0	0
27	12,4	В	0	1	25	13,8	Ви	0	0
28	13,2	В	0	1					

Оценки $a_1 = -1{,}92$ и $a_2 = -2{,}19$ указывают на различия в индейках, первая — из Джорджии и Висконсина, а вторая — из Виргинии и Висконсина соответственно. Подставляя три различных набора значений (Z_1, Z_2), мы получим уравнения для трех различных штатов:

$$\hat{Y} = -0{,}49 + 0{,}4868X \quad \text{для Д},$$

$$\hat{Y} = -0{,}76 + 0{,}4868X \quad \text{для В},$$

$$\hat{Y} = 1{,}43 + 0{,}4868X \quad \text{для Ви}. \qquad (5.4.3)$$

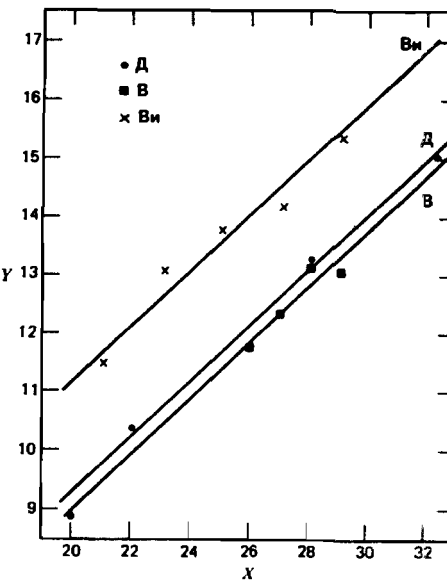


Рис. 5.5. График для данных об индейках и три прямые, подобранные методом наименьших квадратов

Экспериментальные данные и три прямые, подобранные методом наименьших квадратов, приведены на рис. 5.5. Все три линии параллельны, но имеют разные свободные члены. Дисперсионный анализ модели можно представить так, как показано в табл. 5.13. Оба значения F-критерия весьма значимы, и это указывает, что введение фиктивных переменных явно оправданно и что линии имеют определенно не нулевой наклон. Это уравнение объясняет 97,94 % вариации относительно среднего. (А без фиктивных переменных удается объяснить только 66,47 %.)

Если угодно, можно построить t-критерий для проверки значимости различий между свободными членами этих уравнений. Так, например, истинное различие между Ви и Д оценивает коэффициент $-a_1 = 1{,}92$, значит, разделив его на оценку стандартного отклонения, т. е. на корень квадратный из соответствующего диагонального элемента матрицы $(\mathbf{X}'\mathbf{X})^{-1}s^2$, мы получим значение t-критерия, модуль

(абсолютное значение) которого надо сравнить с процентной точкой t (9,1—1/2 α) двустороннего критерия, проверяющего нуль-гипотезу $H_0 : \alpha_1 = 0$ против альтернативы $H_1 : \alpha_1 \neq 0$. Пользуясь нашими данными, получим: $t = 1{,}92/0{,}201 = 9{,}55$, что значимо на уровне 0,1 %. Другой, но равноценный критерий получится, если воспользоваться соотношением

$$F = \{\mathrm{SS}(a_1 | b_0,\ b_1,\ a_2)/1\}/s^2 = 8{,}145/0{,}090 = 90{,}50.$$

Этот результат сравнивается с F (1, 9, 1—α) для критерия с тем же уровнем значимости. Результат идентичен предыдущему, поскольку теоретически величина F-критерия должна равняться квадрату величины t-критерия, полученного выше. В данном случае $t^2 = 91{,}20$, что должно было бы быть равно: $F = 90{,}50$. То, что нет полного совпадения, объясняется ошибками округления. Критерий для нуль-гипотезы $H_0 : \alpha_2 = 0$, где α_2 представляет собой истинную разность между В и Ви, можно построить точно так же. Величина t-критерия окажется равной — 2,19/0,21 = — 10,43, что также значимо на уровне 0,1 %. Оценкой различия между Д и В служит разность $a_1 - a_2 = 0{,}27$, оценкой дисперсии которой в свою очередь служит выражение: оц. $V(a_1)$+оц. $V(a_2)$ — 2 оц. cov (a_1, a_2), все три члена которого можно извлечь из матрицы $(\mathbf{X}'\mathbf{X})^{-1}s^2$. В результате получим: оц. $V(a_1 - a_2) = 0{,}040369 + 0{,}044310 - 2\,(-0{,}018690) = 0{,}122059 = (0{,}349)^2$. Тогда величина t-критерия, равная 0,27/0,349 = 0,77, окажется незначимой. Таким образом, фактические различия существуют между Д и Ви и между В и Ви, но не проявляются между Д и В.

Таблица 5.13. **Дисперсионный анализ для примера с индейками**

Источник	Число степеней свободы	SS	MS	f
b_0	1	2124,803		
$a_1, a_2 \mid b_0$	2	6,382	3,191	35,46
$b_1 \mid b_0, a_1, a_2$	1	32,224	32,224	358,04
Остаток	9	0,811	$s^2 = 0{,}090$	
Общий	13	2164,220		

Представление фиктивных переменных не единственно

Как можно понять из того, что сказано выше, для данной регрессионной задачи существует не единственный способ выбора фиктивных переменных, а в большинстве случаев путей их представления превеликое множество. Это обстоятельство может оказаться выгодным, если мы сумеем использовать его для объяснения некоторых особенностей, проявляющихся в наших данных. Правда, должна быть уверенность в том, что выбранное представление действительно сработает, т. е. даст возможность сосчитать результат для всех уровней (категорий) фиктивного фактора, не приводя к вырожденности мат-

рицы $\mathbf{X}'\mathbf{X}$. Основные представления мы советуем брать среди простейших. Все другие представления должны обязательно обеспечить такое же число линейно независимых столбцов матрицы $\mathbf{X}$, причем так, чтобы они были линейными комбинациями исходных столбцов.

Пример 2. Ниже приведена схема фиктивного фактора. Пригодна ли она, если иметь в виду возможные различия в уровнях для шести групп?

Группа	Z_1	Z_2	Z_3	Z_4	Z_5	Группа	Z_1	Z_2	Z_3	Z_4	Z_5
1	1	1	1	1	1	4	0	0	0	1	1
2	0	1	1	1	1	5	0	0	0	0	1
3	0	0	1	1	1	6	0	0	0	0	0

Ответ утвердительный. Вспомните, что наша схема базисных векторов для описываемого случая, записанная ниже со столбцом X_0, была вот какой:

X_0	X_1	X_2	X_3	X_4	X_5	X_0	X_1	X_2	X_3	X_4	X_5
1	1	0	0	0	0	1	0	0	0	1	0
1	0	1	0	0	0	1	0	0	0	0	1
1	0	0	1	0	0	1	0	0	0	0	0

Сразу видно, что:

$$Z_0 = X_0, \qquad Z_3 = X_1 + X_2 + X_3,$$

$$Z_1 = X_1, \qquad Z_4 = X_1 + X_2 + X_3 + X_4,$$

$$Z_2 = X_1 + X_2, \qquad Z_5 = X_1 + X_2 + X_3 + X_4 + X_5.$$

Таким образом, система столбцов Z представляет собой независимые линейные комбинации столбцов системы X.

Пример 3. Другая вполне пригодная схема в том же контексте, что и в примере 2, могла бы содержать столбцы $Z_0 = X_0$, $Z_i = X_0 + X_i$, $i = 1, 2, \ldots, 5$. Это привело бы к схеме:

Группа	Z_1	Z_2	Z_3	Z_4	Z_5	Группа	Z_1	Z_2	Z_3	Z_4	Z_5
1	2	1	1	1	1	4	1	1	1	2	1
2	1	2	1	1	1	5	1	1	1	1	2
3	1	1	2	1	1	6	1	1	1	1	1

Члены с взаимодействиями, включающие фиктивные факторы

Положим, для определенности, что мы имеем два аналогичных набора данных об отклике Y и предикторе X и что для каждого из этих наборов мы имеем в виду модель в форме

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon.$$

Нужно выяснить, можно ли использовать для обоих множеств данных одну и ту же модель и если можно, то как подобрать ее коэффициенты? Один из путей подхода к этой задаче заключается в том, чтобы одновременно подбирать модель для обоих наборов данных в виде

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \alpha_0 Z + \alpha_1 XZ + \alpha_{11} X^2 Z + \varepsilon,$$

где Z — фиктивный фактор, принимающий уровень 0 для одного набора даниых и 1 — для другого. Тогда критерий дополнительной суммы квадратов позволяет нам проверять различные варианты гипотез, такие, например, как:

1. Гипотеза $H_0 : \alpha_0 = \alpha_1 = \alpha_{11} = 0$ против альтернативы H_1: что это не так. Если эта гипотеза будет отвергнута, то мы придем к выводу, что модели не одинаковы, ну а если нет, то мы будем пользоваться одной моделью.

2. Если гипотеза H_0 в п. 1 окажется отвергнутой, то мы можем рассмотреть подмножества значений α. Так, например, мы могли бы проверить гипотезу $H_0 : \alpha_1 = \alpha_{11} = 0$ против альтернативы H_1: что это не так. Если бы H_0 не была отвергнута, то мы могли бы заключить, что имеющиеся два набора данных отличаются только уровнем отклика, но имеют одинаковые углы наклона и кривизну.

3. Если гипотеза H_0 в п. 2 окажется отвергнутой, то мы сможем проверить новую гипотезу $H_0 : \alpha_{11} = 0$ против альтернативы H_1: $\alpha_{11} \neq 0$, чтобы увидеть, не отличаются ли модели только членами нулевого и первого порядка, на что указывало бы не отбрасывание H_0.

Могли бы быть выбраны и другие последовательности проверок, если бы это было разумно в контексте решаемой задачи. Выбраниая последовательность представляет естественный порядок различий, который часто разумен.

В принципе нет никаких проблем, препятствующих распространению такого подхода на ситуации с большим числом наборов данных и с другими моделями, включающими больше предикторов, X_1, X_2, . . . , X_k. Если бы было r наборов данных, нам пришлось бы образовать $(r-1)$ фиктивных факторов Z_1, Z_2, . . . , Z_{r-1} с уровнями, задаваемыми элементами матрицы $\mathbf{I}_{r-1}$, к которой снизу приписана строка с $(r-1)$ нулями. Тогда строки будут соответствовать группам, а столбцы — фиктивным факторам.

Если бы для одного набора данных основная модель была бы

$$\mathbf{Y} = f(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon},$$

то по всем данным мы могли бы построить модель[9]

$$Y = f(\mathbf{X},\ \beta) + \sum_{j=1}^{r-1} Z_j f(\mathbf{X},\ \alpha_j) + \varepsilon,$$

где α_j — вектор параметров такого же размера, что и β, как в приведенном выше примере для $r = 2$.

Мы должны получить такие же ответы, как если бы мы обрабатывали каждый набор даиных отдельно. Так, если $\mathbf{X}_i$ — матрица $\mathbf{X}$ для i-го набора данных, а всего имеются два набора, то модель имеет вид:

$$E(\mathbf{Y}) = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{X}_2 & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix},$$

и мы можем ее представить, скажем, так:

$$E(\mathbf{Y}) = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \beta \\ \alpha + \beta \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{0} \end{bmatrix}.$$

Преимущество использования взаимодействий с фиктивными факторами заключается в том, что появляется возможность простой формализации и естественный способ применения критериев дополнительной суммы квадратов.

Пример 4. Проиллюстрируем сказанное на примере с индейками. Для построения трех отдельных прямых (см. табл. 5.12) мы возьмем модель

$$Y = \beta_0 + \beta_1 X + Z_1(\gamma_0 + \gamma_1 X) + Z_2(\delta_0 + \delta_1 X) + \varepsilon,$$

т. е.

$$Y = \beta_0 + \beta_1 X + \gamma_0 Z_1 + \gamma_1 (Z_1 X) + \delta_0 Z_2 + \delta_1 (Z_2 X) + \varepsilon.$$

Тогда получится следующее уравнение МНК:

$$\hat{Y} = 2{,}475 + 0{,}4450X - 3{,}454Z_1 - 2{,}775Z_2 + 0{,}06104(Z_1 X) + \\ + 0{,}02500(Z_2 X).$$

А вот три отдельных уравнения прямых линий:

$$\hat{Y} = -0{,}979 + 0{,}5060X \quad (\text{при подстановке } Z_1 = 1, \quad Z_2 = 0),$$

$$\hat{Y} = -0{,}300 + 0{,}4700X \quad (\text{при подстановке } Z_1 = 0, \quad Z_2 = 1),$$

$$\hat{Y} = 2{,}475 + 0{,}4450X \quad (\text{при подстановке } Z_1 = 0, \quad Z_2 = 0).$$

Эти линии, которые в точности те же, что получились бы при подборе уравнений для каждого набора данных в отдельности, несколько отличаются от тех линий, что приведены на рис. 5.5, в чем читатель может убедиться, если построит их на графике или просто сравнит

[9] В приведенном ниже уравнении в оригинале содержится опечатка. Чтобы ее устранить, надо либо указать, что значение отклика, предсказанное, $\hat{Y}$, либо добавить член, содержащий случайную ошибку. Мы предпочли последнее.— *Примеч. пер.*

с уравнениями (5.4.3). Таблица дисперсионного анализа для этих данных имеет вид:

ANOVA

Источник	Число степеней свободы	SS	MS	F
b_0	1	2124,803		
$b_1, c_0, c_1, d_0, d_1 \mid b_0$	5	38,711	7,742	76,6
Остаток	7	0,706	0,101	
Общий	13	2164,220		

Эти три подобранные прямые были бы идентичны, если бы была верна нуль-гипотеза $H_0 : \gamma_0 = \gamma_1 = \delta_0 = \delta_1 = 0$. Проверка этой гипотезы против альтернативы H_1: что H_0 не верна, требует дополнительной суммы квадратов для c_0, c_1, d_0 и d_1 с четырьмя степенями свободы:

$$\mathrm{SS}(b_1, c_0, c_1, d_0, d_1 | b_0) - \mathrm{SS}(b_1 | b_0) = 38{,}71 - 26{,}20 = 12{,}51.$$

(Величина 26,20 представляет собой сумму квадратов относительно регрессии для общего уравнения; она ранее не использовалась.) Соответствующее значение F-критерия равно:

$$F = (12{,}51/4)/(0{,}101) = 30{,}97,$$

что превышает табличное значение F (4, 7, 0,99) = 7,85, так что гипотеза H_0 отвергается. Это, конечно, отнюдь не удивительно, как мы уже видели, когда рассматривали исходные данные.

Можно проверить гипотезу о существовании трех параллельных линий, т. е. $H_0 : \gamma_1 = \delta_1 = 0$ против альтернативы H_1: что H_0 не верна. Для этого находим дополнительную сумму квадратов для c_1 и d_1 в виде

$$\mathrm{SS}(b_1, c_0, c_1, d_0, d_1 | b_0) - \mathrm{SS}(b_1, c_0, d_0 | b_0) = 38{,}71 - 38{,}61 = 0{,}10,$$

где величина 38,61 представляет собой сумму второй и третьей строк из табл. 5.13. Это дает две степени свободы и приводит к незначимому F-отношению (0,10/2)/0,101 = 0,50. Мы не отбрасываем гипотезу H_0, а это значит, что модель, показанная на рис. 5.5, вполне удовлетворяет ей.

Как показывает наш пример, использование взаимодействия с фиктивными факторами упрощает построение подходящих критериев и получение правильных статистик для проверки гипотез. Быть может, это наиглавнейшее достоинство данного метода [10].

---

[10] Иногда предпочитают называть регрессионной моделью такую зависимость, в которой все предикторы имеют непрерывные шкалы. Тогда введение фиктивного предиктора означает переход к иной модели — модели ковариационного анализа. См., например: Адлер Ю. П. Предпланирование эксперимента.— М.: Знание, 1978.— 63 с.; Крастынь О. П. Изучение статистических зависимостей по многолетним данным.— М.: Финансы и статистика,

Временны́е тренды в данных

Во многих практических случаях временны́е тренды проявляются в откликах. Иногда тренд представляет собой единственный фактор, влияющий на отклик, а иногда он налагается на эффекты других предикторов. Вообще говоря, мы можем описать временно́й тренд с помощью одного или нескольких подходящим образом определенных фиктивных факторов. Соответствующие члены модели, отражающие эти фиктивные переменные, просто приписываются к модели для всех остальных предикторов, а затем оценивается вся модель аналогично тому, как показано в примере с блоковым фактором. Хотя наше обсуждение ниже сфокусировано непосредственно на временных трендах, следовало бы помнить, что любые другие параметры, относящиеся к решаемой задаче, как правило, можно оценивать точно так же.

Единственный временно́й тренд. Когда в данных представлен простой линейный тренд, для его учета достаточно ввести одну фиктивную переменную. Продемонстрируем это на примере.

Пример 5. Данные табл. 5.14 показывают паритетную цену[11] в центах за фунт живого веса цыплят через равные промежутки времени. Две альтернативные фиктивные переменные, используемые для элиминирования линейного временного тренда, приведены в столбцах X и X'. Годится любая из них, однако центрированный столбец X' лучше, так как он ортогонален к единичному столбцу **X**-матрицы. Подходящие модели в обоих случаях выражаются уравнениями:

$$Y = \beta_0 + \beta_1 X + (\text{другие члены с предикторными переменными}) + \varepsilon \tag{5.4.4}$$

и (так как $X_i' = X_i - \bar{X}$, $i = 1, 2, \ldots, n$)

$$\begin{aligned} Y &= (\beta_0 + \beta_1 \bar{X}) + \beta_1 X' + (\text{другие члены}) + \varepsilon = \\ &= \beta_0' + \beta_1 X' + (\text{другие члены}) + \varepsilon. \end{aligned} \tag{5.4.5}$$

1981.— 136 с.; Маркова Е. В., Денисов В. И., Полетаева И. А, Пономарев В. В. Дисперсионный анализ и синтез планов на ЭВМ.— М.: Наука, 1982.— 196 с.; Себер Дж. Линейный регрессионный анализ/Пер. с англ. Под ред. М. Б. Малютова.— М.: Мир, 1980.— 456 с. (особо с. 273—282).

Разбиение на блоки — прием, характерный для планирования эксперимента. Материал об этом можно найти во многих работах. Например, в упомянутой выше монографии В. В. Налимова и Н. А. Чериовой на с. 101—105 или в кн.: Маркова Е. В., Лисенков А. Н. Комбинаторные планы в задачах многофакторного эксперимента.— М.: Наука, 1979.— 348 с.— *Примеч. пер.*

[11] Паритетными называют цены, допускающие непосредственное сравнение. Поскольку курс доллара во времени постоянно колеблется, приходится соотносить все текущие цены с некоторой постоянной базой. Именно такие относительные цены и становятся паритетными, сравнимыми. В США решением Конгресса от 1933 г. существуют паритетные цены на сельскохозяйственные продукты.— См.: Миллс Ф. Статистические методы/Пер. с англ. Под ред. П. П. Маслова.— М.: Госстатиздат, 1958.— 799 с. (особо с. 478—482).— *Примеч. пер.*

(Примечание. Так как здесь $n = 21$, т. е. нечетно, величины $X_i' = X_i - \overline{X}$ — все целые числа. Когда n четно, мы можем применить замену $X_i' = 2(X_i - \overline{X})$,

чтобы избежать появления дробей. Например,

$$X_i = 1\ 2\ 3\ 4\ \left(\overline{X} = 2\frac{1}{2}\right),$$

$$X_i - \overline{X} = -1\frac{1}{2} \quad -\frac{1}{2} \quad \frac{1}{2} \quad 1\frac{1}{2},$$

$$2(X - \overline{X}) = -3 \quad -1 \quad 1 \quad 3.)$$

В случае квадратичного временного тренда придется добавлять к модели члены $\beta_0 + \beta_1 X + \beta_{11} X^2$ (или $\beta_0' + \beta_1 X' + \beta_{11} X_1'^2$), либо выражать такой тренд через ортогональные полиномы, описанные в параграфе 5.6. С трендами более высоких порядков надо обходиться таким же образом, используя члены более высоких порядков.

Таблица 5.14. **Паритетная цена (центы) за фунт живого веса цыплят**

Период	Паритетная цена, Y	X или X'	Период	Паритетная цена, Y	X или X'
Январь 1955 г.	29,1	1 —10	Сентябрь 1958 г.	28,6	12 1
Май 1955 г.	29,0	2 —9	Январь 1959 г.	26,9	13 2
Сентябрь 1955 г.	28,6	3 —8	Май 1959 г.	27,0	14 3
Январь 1956 г.	28,1	4 —7	Сентябрь 1959 г.	26,8	15 4
Май 1956 г.	28,6	5 —6	Январь 1960 г.	25,7	16 5
Сентябрь 1956 г.	28,7	6 —5	Май 1960 г.	25,9	17 6
Январь 1957 г.	28,2	7 —4	Сентябрь 1960 г.	25,6	18 7
Май 1957 г.	28,6	8 —3	Январь 1961 г.	25,1	19 8
Сентябрь 1957 г.	28,6	9 —2	Май 1961 г.	25,2	20 9
Январь 1958 г.	28,1	10 —1	Сентябрь 1961 г.	25,1	21 10
Май 1958 г.	28,7	11 0			

В табл. 5.14 приведены данные, полученные через равные промежутки времени, поэтому и значения X-ов здесь тоже выбраны с равным шагом. А если бы данные оказались не равномерными во времени, то и X-ы пришлось бы выбирать соответственно. Так, например, если бы данные относились к январю 1955 г., февралю 1955 г., апрелю 1955 г., июню 1955 г., . . . , то для X-ов пришлось бы взять значения 1, 2, 4, 6, . . . и т. д. Использование столбца $X - \overline{X}$ в таком случае могло бы оказаться неудобным из-за возможного появления дробных значений. В таком случае пришлось бы либо воспользоваться множителем, превращающим все числа в целые, либо, если бы это оказалось невозможным, перейти к величинам $X_i - A$, где A было бы некоторым произвольным целым числом, близким к $\overline{X}$. Правда, использование преобразования $X_i - A$ вряд ли даст какие-либо преимущества, если только не считать того, что числа, с которыми предстоит работать,

станут более удобными. Поскольку шаг не равный, приходится вычислять специальные ортогональные полиномы, если, конечно, решено, что ими стоит воспользоваться. Именно в силу этого ортогональные полиномы очень редко применяются при неравномерных данных.

Два временны́х тренда. Когда представлены два временны́х тренда, то фиктивная переменная должна быть выбрана для каждого из них. Эта задача может иметь два уровня сложности в зависимости от того, известно ли, какому тренду принадлежат данные, или же это не известно.

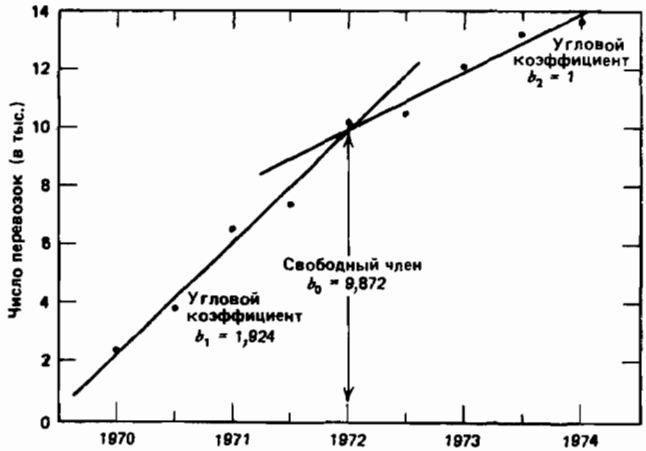


Рис. 5.6. Использование фиктивных перемеиных. Две линии, абсцисса точки пересечения известна

1. *Когда известно, какие точки принадлежат каким трендам.* Пусть, например, есть два временны́х тренда, причем оба линейные. Тогда мы можем выделить в этой ситуации еще два подкласса: (1а) когда об абсциссе точки пересечения двух линий можно предположить, что она соответствует определенному значению, в котором есть одно или несколько наблюдений, и (1б) когда абсцисса точки пересечения этих двух линий не известна.

Пример 6. Равномерно расположенные данные, представленные на рис. 5.6, относятся к варианту (1а). Известно, что первые пять точек лежат (если пренебречь случайной ошибкой) на первой прямой, а последние пять точек (опять же если пренебречь случайной ошибкой) — на второй. Значит, пятая точка в этом случае оказывается общей для обеих прямых. Мы можем ввести две фиктивные переменные X_1 и X_2 для этих двух прямых следующим образом. Положим обе фиктивные переменные равными нулю в известной точке пересечения, а именно в точке пятого наблюдения, из которой X_1 для первой прямой пойдет назад, а X_2 для второй прямой — вперед, причем каждая переменная будет обращаться в нуль там, где действует дру-

гая. (Шаги здесь равные, поскольку и данные собраны с равным шагом. Если бы это было не так, то пришлось бы выбрать иные уровни для соответствующих переменных.) Итоговая матрица данных в предположении, что нет никаких других предикторных переменных, приведена в табл. 5.15. Если теперь мы подберем модель

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \tag{5.4.6}$$

то полученные оценки будут играть такие роли:

b_0 — значение $\hat{Y}$ в точке пересечения, $X_1 = X_2 = 0$,
b_1 — угловой коэффициент прямой первого тренда,
b_2 — угловой коэффициент прямой второго тренда.

Нормальные уравнения для приведенных данных имеют вид:

$$\begin{aligned} 9b_0 - 10b_1 + 10b_2 &= 79{,}6, \\ -10b_0 + 30b_1 &= -41{,}0, \\ 10b_0 + 30b_2 &= 128{,}7, \end{aligned} \tag{5.4.7}$$

а их решение таково: $b_0 = 9{,}871$, $b_1 = 1{,}924$, $b_2 = 1{,}000$, как показано на рис. 5.6. Если бы в задаче играли роль и другие предикторы, то соответствующие члены надо было бы дописать в правой части уравнения (5.4.6).

Таблица 5.15. **Фиктивные переменные для примера с двумя прямыми, абсцисса точки пересечения которых известна**

Номер наблюдения	Дата	X_0	X_1	X_2	Y
1	1970 г.	1	—4	0	2,3
2		1	—3	0	3,8
3	1971 г.	1	—2	0	6,5
4		1	—1	0	7,4
5	1972 г.	1	0	0	10,2
6		1	0	1	10,5
7	1973 г.	1	0	2	12,1
8		1	0	3	13,2
9	1974 г.	1	0	4	13,6

Как бывает всегда с фиктивными переменными, такое их представление не единственно. Пример альтернативного представления показан в табл. 5.16, где (X_1 новая) = (X_1 старая) + 5. Это представление даст точно такие же оценки угловых коэффициентов, что и предыдущее, но зато постоянный член b_0, соответствующий значению $\hat{Y}$, когда $X_1 = X_2 = 0$, стал бы теперь свободным членом первого уравнения с абсциссой в точке $1969\frac{1}{2}$.

Наличие временных трендов более высоких порядков привело бы к добавлению к модели членов более высоких порядков. За прямой линией следует квадратичная кривая, для которой, например, мы могли бы найти

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{22} X_2^2 + \varepsilon. \tag{5.4.8}$$

Если сверх того мы пожелали бы, чтобы модель имела в общей точке непрерывную производную, то нам пришлось бы потребовать выполнения в точке пересечения равенства $\partial Y/\partial X_1 = \partial Y/\partial X_2$. А это означает, что в пятой точке, где $X_2 = 0$, должно соблюдаться условие

$$\beta_1 = \beta_2 + 2\beta_{22}X_2. \tag{5.4.9}$$

Тогда мы должны положить $\beta_1 = \beta_2 = \beta$, скажем, в уравнении (5.4.8), а это сводит модель к виду

$$Y = \beta_0 + \beta(X_1 + X_2) + \\ + \beta_{22}X_2^2 + \varepsilon.$$

Рис. 5.7. Использование фиктивных переменных. Две линии, абсцисса точки пересечения не известна

Пример 7. Данные с равным шагом на рис. 5.7 относятся к варианту (1б). Здесь известно, что первые четыре точки лежат (если пренебречь случайной ошибкой) на первой прямой, а последние пять (снова без учета случайной ошибки) — на второй. Однако точка их пересечения не известна. Чтобы обнаружить эту неизвестную точку, понадобится третья фиктивная переменная X_3. Ее естественно положить равной нулю для всех точек первой прямой и соответственно единице для точек второй прямой, чтобы отразить скачок (положительный или отрицательный) от первой прямой ко второй. Фиктивные переменные X_1 и X_2 выбираются точно так же, как это сделано в примере 6. В табл. 5.17 приведена соответствующая матрица данных. Если не включать в нее никаких дополнительных предикторов, то можно получить такую модель:

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon. \tag{5.4.10}$$

Параметр β_3 представляет собой шаг изменения, приводящий к эффекту наблюдения в пятой точке. Фактически это вертикальное расстояние, на котором в этой точке вторая прямая проходит *выше* первой. (Если вторая прямая лежит ниже первой, то коэффициеит β_3 будет отрицательным.)

Таблица 5.16. **Альтернативный вариант фиктивной переменной для примера с двумя прямыми, точка пересечения которых известна**

Номер наблюдения	Дата	X_0	X_1	X_2	Y
1	1970, г.	1	1	0	2,3
2		1	2	0	3,8
3	1971 г.	1	3	0	6,5
4		1	4	0	7,4
5	1972 г.	1	5	0	10,2
6		1	5	1	10,5
7	1973 г.	1	5	2	12,1
8		1	5	3	13,2
9	1974 г.	1	5	4	13,6

Таблица 5.17. **Фиктивные переменные для примера с двумя прямыми, точка пересечения которых не известна**

Номер наблюдения	Дата	X_0	X_1	X_2	X_3	Y
1	1970 г.	1	1	0	0	1,8
2		1	2	0	0	4,3
3	1971 г.	1	3	0	0	5,6
4		1	4	0	0	8,2
5	1972 г.	1	5	0	1	9,1
6		1	5	1	1	10,7
7	1973 г.	1	5	2	1	11,5
8		1	5	3	1	12,5
9	1974 г.	1	5	4	1	14,0

Для данных из табл. 5.17 получаются следующие нормальные уравнения:

$$\begin{aligned} 9b_0 + 35b_1 + 10b_2 + 5b_3 &= 77{,}4, \\ 35b_0 + 155b_1 + 50b_2 + 25b_3 &= 347{,}5, \\ 10b_0 + 50b_1 + 30b_2 + 10b_3 &= 126{,}3, \\ 5b_0 + 25b_1 + 10b_2 + 5b_3 &= 57{,}5 \end{aligned} \qquad (5.4.11)$$

с таким решением:

$b_0 = -0{,}15$ (свободный член прямой 1 при $X_1 = 0$),

$b_1 = 2{,}05$ (угловой коэффициент прямой 1),

$b_2 = 1{,}13$ (угловой коэффициент прямой 2),

$b_3 = -0{,}86$ (расстояние по вертикали между прямыми 2 и 1 в пятой точке).

Вся ситуация графически представлена на рис. 5.7. Отрицательный знак коэффициента b_3 и тот факт, что $b_1 > b_2$ указывают на смещение точки пересечения двух прямых влево от пятой точки. Практи-

чески это произойдет тогда, когда X_1 станет равным 4,065. Эту точку можно обнаружить, если выписать обе прямые в шкале X_1. Первая прямая имеет вид

$$\hat{Y} = -0{,}15 + 2{,}05X_1, \tag{5.4.12}$$

а вторая прямая запишется так:

$$\hat{Y} = -0{,}15 + 2{,}05\,(5) + 1{,}13X_2 - 0{,}86, \tag{5.4.13}$$

т. е.

$$\hat{Y} = 9{,}24 + 1{,}13X_2. \tag{5.4.14}$$

Взглянув на шкалы X_1 и X_2 отиосительно шкалы на рис. 5.7, мы обнаружим, что $X_2 = 0$ при $X_1 = 5$, так что можно подставить $X_2 = X_1 - 5$ в уравнение второй прямой, что сведет его к виду

$$\hat{Y} = 3{,}59 + 1{,}13X_1. \tag{5.4.15}$$

Приравнивая правые части уравнений (5.4.12) и (5.4.15), получим, что $X_1 = 4{,}065$ соответствует точке пересечения.

2. *Когда не известно, какие точки относятся к какому тренду.* В предыдущем случае мы оценивали параметры составной модели с помощью линейного метода наименьших квадратов. Здесь же решение следовало бы получать, просматривая все возможные варианты разбиения точек между двумя прямыми, оценивая в каждом таком разбиении параметры линейным методом наименьших квадратов и вычисляя остаточные суммы квадратов. А затем можно выбрать такое разбиение вместе с набором оценок параметров, которое порождает наименьшее из всех значение остаточной суммы квадратов [12]. (На практике обычно нет никакой необходимости просматривать каждое возможное разбиение точек, поскольку даже малые вычисления обычно показывают ту «танцплощадку», где и находится наилучшее

---

[12] Проблема трендов особенно тщательно разработана в эконометрии, поскольку рассмотрение любых экономических данных во времени сразу приводит к мысли о временном тренде. Среди многих работ отметим, например: Драймз Ф. Распределенные лаги. Проблемы выбора и оценивания модели/Пер. с англ. Под ред. Э. Б. Ершова.— М.: Финансы и статистика, 1982.— 383 с.; Песаран М., Слейтер Л. Динамическая регрессия: теория и алгоритмы/Пер. с англ. Под ред. Э. Б. Ершова. — М.: Финансы и статистика, 1984.— 310 с.; Вайну Я. Я.-Ф. Корреляция рядов динамики.— М.: Статистика, 1977.— 119 с.

Другая область, где к трендам, называемым также дрейфами, проявляют традиционное внимание — это планирование эксперимента. Здесь разработана и используется идея о том, что можно построить план для всех представляющих интерес предикторов так, чтобы он по возможности не зависел от временного дрейфа (т. е. был ортогонален к нему), задаваемого полиномом известного порядка (но с неизвестными коэффициентами). О дрейфах см. упомянутую выше монографию В. В. Налимова и Н. А. Чериовой, с. 281—285, а также кн.: Маркова Е. В., Лисеиков А. Н. Планирование эксперимента в условиях неоднородностей.— М.: Наука, 1973.— 219 с., гл. 7—9, с. 124—174; Планирование эксперимента в исследовании технологических процессов/Пер. с нем. Под ред. Э. К. Лецкого.— М.: Мир, 1977.— 547 с. (особо с. 278—316). Отметим, что при планировании эксперимента в условиях дрейфа применяются ортогональные полиномы Чебышева (см. параграф 5.6).— *Примеч. пер.*

разбиение. Только этот ограниченный набор разбиений и надо сосчитать.) С другой стороны, эту задачу можно представить как задачу нелинейного оценивания и решать ее методами, обсуждаемыми в гл. 10. (При этом иногда надо проявлять бдительность, поскольку могут встретиться локальные минимумы.) В поисках дополнительной информации читатель может обратиться к литературе, где обсуждаются отрезки прямых и сплайны, которая приведена в конце кн. 2 настоящей монографии [13].

5.5. ЦЕНТРИРОВАНИЕ И МАСШТАБИРОВАНИЕ. ПРЕДСТАВЛЕНИЕ РЕГРЕССИИ В КОРРЕЛЯЦИОННОЙ ФОРМЕ

Если в регрессионную модель включены только одна или две предикторные переменные, то непосредственное вычисление по формуле $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, как показано в примере для двух переменных в гл. 4, обычно не вызывает затруднений при условии, что вычисления проводятся с достаточным числом значащих цифр. В задачах с несколькими предикторами и с большим объемом данных результаты могут оказаться совершенно не верными вследствие ошибок округления. Вот типичный пример, когда возникают ошибки округления: допустим, требуется вычислить $(a/b) - (c/d) = e$. Если (a/b) и (c/d) велики, а e мало, то слишком большое округление чисел (a/b) и (c/d) может привести к тому, что все значащие цифры в числе e будут потеряны.

Пример. $a = 100$, $b = 3$, $c = 166{,}663$, $d = 5$. Допустим, что вычисления ведутся вручную и мы округляем числа до трех значащих цифр после запятой. Тогда $a/b = 33{,}333$, $c/d = 33{,}333$, так что $e = 0$. (Более точно $e = (ad - bc)/bd = (500 - 499{,}989)/15 = 0{,}011/15$.) Поэтому если умножить e на последней стадии, скажем, на 1 000 000, то в результате будет нуль (вместо правильного $11000/15 = 733{,}3$ — огромная разница.)

---

[13] В добавление к тому, что содержится в списке литературы, укажем еще из многочисленных работ по сплайнам две общие: Альберг Дж., Нильсон Э., Уолш Дж. Теория сплайнов и ее приложения./Пер. с англ.— М.: Мир, 1972.— 316 с.; Завьялов Ю. С., Квасов Б. И., Мирошниченко В. Л. Методы сплайн-функций. — М.: Наука, 1980.— 352 с., а также несколько специальных, с приложениями в разных областях, в том числе и в регрессии: Себер Дж. Линейный регрессионный анализ/Пер. с англ. Под ред. М. Б. Малютова.— М.: Мир, 1980.— 456 с. (особо с. 222—226); Пуарье Д. Эконометрия структурных изменений (с применением сплайн-функций)/Пер с англ. Под ред. Г. Г. Пирогова.— М.: Финансы и статистика, 1981.— 183 с. (особо с. 136—150); Котюков В. И. Многофакторные кусочно-линейные модели.— М.: Финансы и статистика, 1984.— 216 с.; Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Исследование зависимостей. Справочное издание.— М.: Финансы и статистика, 1985.—487 с. (особо с. 328—335); Стечкин С. Б., Субботин Ю. Н. Сплайны в вычислительной математике. — М.: Наука, 1976.— 248 с.; Maindonald J. H. Statistical computation.— New York: J. Wiley, 1984.— 370 p. (особо p. 255—259); Де Бор К. Практическое руководство по сплайнам/Пер. с англ. Под ред. В. И. Скурихина.— М.: Радио и связь, 1985.— 304 с.— *Примеч. пер.*

Хотя цифровая машина перерабатывает значительно больше цифр, чем человеческая «вычислительная машина», ошибки такого типа нередко происходят и часто приводят к совершенно не верным результатам или к увеличению времени счета. Заключения, которые, по общему мнению, логичны, иногда базируются целиком на капризах ошибок округления.

В работе Р. Фройнда о предупреждении ошибок округления в регрессионном анализе (Freund R. J. A warning of round-off errors in regression.— American Statistician, December, 1963, **17**, p. 13—15) приведен пример, в котором пять различных регрессионных вычислений с использованием четырех разных регрессионных программ привели к значительным различиям в оцениваемых коэффициентах, обусловленным ошибками округления. Для преодоления этого некоторые программы позволяют получать результаты с *удвоенной точностью*. Это означает, что машина (по требованию) работает с числами, вдвое более длинными, чем обычно. Применение такого приема как стандартного способа приводит к излишнему расходованию машинного времени и часто оказывается неоправданной предосторожностью. Гораздо лучше сначала выяснить, что ошибки округления могут иметь место, и лишь тогда предпринимать шаги, позволяющие уменьшить ошибки, а возможно, и исключить их полностью.

Две основные причины ошибок округления таковы:

1. Числа, включенные в регрессионные вычисления, могут резко различаться по порядку, как, например, если включить в расчет числа вроде 52793,— 943 и 6.

2. Матрица, которую надо обращать, может оказаться очень близкой к вырожденной. Из уравнений (2.1.10) и (2.1.11) мы можем видеть, что определитель матрицы входит в каждый элемент обратной матрицы. Если определитель матрицы мал по сравнению с остальными числами в расчете, то помехи от округления, вероятно, будут иметь место, и это справедливо не только для матриц 2×2 и 3×3, но и в общем случае. Когда $\det(\mathbf{X}'\mathbf{X})$ очень мал по сравнению с другими числами в расчете, о матрице $\mathbf{X}'\mathbf{X}$ говорят, что она *плохо (или слабо) обусловлена*. Когда же $\det(\mathbf{X}'\mathbf{X}) = 0$, о матрице $\mathbf{X}'\mathbf{X}$ говорят, что она *сингулярная (вырожденная)*; если это случается при машинном счете, то возникает переполнение и машина останавливается. (Точнее говоря, мы обычно имеем дело с величиной определителя *корреляционной матрицы*, которая обсуждается ниже в этом параграфе.)

Плохая обусловленность

Когда существуют строгие зависимости между столбцами матрицы $\mathbf{X}$, т. е. когда один (или более) столбцов можно строго выразить как линейную комбинацию (с различными численными коэффициентами) других столбцов, определитель $\det(\mathbf{X}'\mathbf{X}) = 0$. Мы можем рассуждать об этом двумя способами. Либо модель переопределена, т. е. постулировалось больше параметров, чем действительно нужно для описания данных, либо наши данные приводят к неадекватной оценке выбранной модели. (Конечно, у всякой монеты есть две стороны, но и «вина»

ложится либо на «модельера», либо на «сборщика данных»!) Дело, по существу, сводится к выбору между простотой модели и охватом, если, конечно, можно собрать данные, позволяющие оценить такую модель.

Когда же зависимости проявляются лишь приближенно, в матрице $\mathbf{X}'\mathbf{X}$ может встретиться плохая обусловленность, и тогда потребуются те же самые выборы или, быть может, использование риджрегрессии, которая описана в параграфе 6.7. Полезное обсуждение того, что же такое плохая обусловленность (она обычно называется *мультиколлинеарностью предикторов*), можно найти в работе: Willan A. W., Watts D. G. Meaningful multicollinearity measures.— Technometrics, 1978, **20**, p. 407—412 [14]. Мы обсудим теперь шаги, которые можно сделать для улучшения метода вычислений. Это — *центрирование данных* и использование корреляционной матрицы вместо матрицы $\mathbf{X}'\mathbf{X}$. Ортогонализация столбцов $\mathbf{X}$-матрицы методом Грама—Шмидта будет обсуждаться в параграфе 5.7 после краткого обсуждения ортогональных полиномов в параграфе 5.6. Центрирование и использование корреляционной матрицы стандартны для большинства программ линейной регрессии. Ортогонализация — полезная процедура, которая может применяться для проверки матрицы $\mathbf{X}'\mathbf{X}$ на вырожденность.

Допустим, что мы хотим подобрать общую линейную модель методом наименьших квадратов в виде

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p + \varepsilon, \qquad (5.5.1)$$

где $Z_i = Z_i(X_1, X_2, \ldots, X_k)$ — некоторые (определенные) функции предикторов $X_1, X_2, \ldots, X_k$. Запишем вектор параметров:

$$\beta' = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$$

и вектор n наблюдений

$$\mathbf{Y}' = (Y_1, Y_2, \ldots, Y_i, \ldots, Y_n).$$

---

[14] Проблема обусловленности имеет логический и вычислительный аспект. Ее логический аспект связан с априорной информацией об объекте и способом формирования его математической модели. Что же касается вычислительного аспекта, то он интенсивно обсуждается в литературе. См., например: Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры.— 2-е изд. — М.— Л.: Физматгиз, 1963.— 734 с.; Уилкинсон Дж. Х. Алгебраическая проблема собственных значений. Пер. с англ.— М.: Наука, 1970.— 564 с.; Райс Дж. Матричные вычисления и математическое обеспечение/Пер. с англ. Под ред. В. В. Воеводина.—М.: Мир, 1984.—264 с.; Маркова Е. В., Деннсов В. И., Полетаева И. А., Пономарев В. В. Дисперсионный анализ и синтез планов на ЭВМ.— М.: Наука, 1982.—196 с.; Лоусон Ч., Хенсон Р. Численное решение задач метода наименьших квадратов. Пер. с англ.— М.: Наука, 1986. — *Примеч. пер.*

Тогда оценка вектора $\boldsymbol{\beta}$, а именно

$$\mathbf{b} = (b_0,\ b_1,\ b_2,\ \ldots,\ b_p)'$$

дается выражением $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ с применением формул из гл. 2, где

$$\mathbf{X} = \begin{bmatrix} 1 & Z_{11} & Z_{21} & \ldots & Z_{p1} \\ 1 & Z_{12} & Z_{22} & \ldots & Z_{p2} \\ & \vdots & \vdots & & \\ 1 & Z_{1i} & Z_{2i} & \ldots & Z_{pi} \\ & \vdots & \vdots & & \\ 1 & Z_{1n} & Z_{2n} & \ldots & Z_{pn} \end{bmatrix}, \qquad (5.5.2)$$

а Z_{ji} — наблюденное значение Z_j, соответствующее наблюдению Y_i. Чтобы показать это на простом примере, допустим, что мы используем модель

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon, \qquad (5.5.3)$$

так что $Z_1 = X$, $Z_2 = X^2$, $\beta_2 = \beta_{11}$ в общей форме, указанной выше. Если имеющиеся данные выражаются так:

$$X = 1 \quad 2 \quad 3 \quad 4 \quad 5,$$

$$Y = Y_1\ Y_2\ Y_3\ Y_4\ Y_5,$$

то

$$\begin{matrix} & 1 & Z_1 = X & Z_2 = X^2 \end{matrix} \qquad \begin{matrix} Y \end{matrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{bmatrix}, \qquad \mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix}.$$

(Наоборот, если $Z_2 = \sqrt{X}$, то элементами столбца были бы $1, \sqrt{2}$, $\sqrt{3,2}$, $\sqrt{5}$ и так далее для более общих случаев.) Когда данные подготавливают для вычислений на машинах, матрицу $\mathbf{X}$ и вектор $\mathbf{Y}$ обычно записывают подряд без разделения и называют это *матрицей данных* или *матрицей исходных данных*. Например, матрица данных

для приведенного выше простого примера имеет вид:

	X_0	X	X^2	Y
	1	1	1	Y_1
	1	2	4	Y_2
	1	3	9	Y_3
	1	4	16	Y_4
	1	5	25	Y_5
Итого по столбцу	—	15	55	$\sum_{i=1}^{n} Y_i$
Среднее по столбцу	—	3	11	$\overline{Y}$
Сумма квадратов для столбца	—	55	979	

(5.5.4)

«Центрирование» данных

Пусть мы имеем следующую матрицу исходных данных вместе со средними по столбцам:

Z_0	Z_1	Z_2	. . .	Z_p	Y
1	Z_{11}	Z_{21}	. . .	Z_{p1}	Y_1
1	Z_{12}	Z_{22}	. . .	Z_{p2}	Y_2
. . .					
1	Z_{1n}	Z_{2n}	. . .	Z_{pn}	Y_n
Сумма по столбцу —	ΣZ_{1i}	ΣZ_{2i}	. . .	ΣZ_{pi}	ΣY_i
Среднее по столбцу —	$\overline{Z}_1$	$\overline{Z}_2$	. . .	$\overline{Z}_p$	$\overline{Y}$

(5.5.5)

(суммирование по i, $i=1, 2, \ldots, n$)

Наша модель такова:

$$Y=\beta_0+\beta_1 Z_1+\beta_2 Z_2+\ldots+\beta_p Z_p+\varepsilon. \qquad (5.5.6)$$

Мы можем переписать ее в виде

$$Y=\{\beta_0+\beta_1\bar{Z}_1+\beta_2\bar{Z}_2+\ \ldots\ +\beta_p\bar{Z}_p\}+\beta_1(Z_1-\bar{Z}_1)+\beta_2(Z_2-\bar{Z}_2)+$$
$$+\ \ldots\ +\beta_p(Z_p-\bar{Z}_p)+\varepsilon,$$

где $\bar{Y}$, $\bar{Z}_1$, $\bar{Z}_2$, . . . , $\bar{Z}_p$—фактические численные значения, полученные на основании данных. Если обозначить

$$\bar{z}_j=Z_j-\bar{Z}_j,\quad j=1,\ 2,\ \ldots,\ p,\quad \beta_0'=\beta_0+\beta_1\bar{Z}_1+\beta_2\bar{Z}_2+$$
$$+\ \ldots\ +\beta_p\bar{Z}_p,$$

то модель можно будет выразить так:

$$Y=\beta_0'+\beta_1z_1+\beta_2z_2+\ \ldots\ +\beta_pz_p+\varepsilon. \tag{5.5.7}$$

Теперь можно преобразовать данные, как мы сделали раньше с переменными, так что $z_{ji}=Z_{ji}-\bar{Z}_j$, $j=1,\ 2,\ \ldots,\ p$ при $i=1,\ 2,\ \ldots,\ n$. Отсюда $\bar{z}_j=0$, $j=1,\ 2,\ \ldots,\ n$. И таким образом, первое нормальное уравнение, получаемое путем дифференцирования остаточной суммы квадратов по β_0', сводится к

$$b_0'+b_1\bar{z}_1+b_2\bar{z}_2+\ \ldots\ +b_p\bar{z}_p=\bar{Y}$$

или

$$b_0'=\bar{Y}$$

независимо от того, какими могут быть значения $b_1,\ b_2,\ \ldots,\ b_p$.

Поскольку это будет верно всегда, мы можем исключить β_0' из модели и применять ее в виде

$$Y-\bar{Y}=\beta_1(Z_1-\bar{Z}_1)+\beta_2(Z_2-\bar{Z}_2)+\ \ldots\ +\beta_p(Z_p-\bar{Z}_p)+\varepsilon' \tag{5.5.8}$$

для оценивания методом наименьших квадратов. И она будет давать точно те же оценки параметров и предсказанные значения, какие мы получили бы, если бы воспользовались МНК для уравнения (5.5.6). (При вычислениях на карманном калькуляторе это иногда полезно, так как уменьшает на единицу размер матрицы, которую требуется обратить.) Кажущийся выигрыш, состоящий в том, что теперь надо оценивать на один параметр меньше, компенсируется тем, что n (новых) значений зависимой переменной, а именно $(Y_1-\bar{Y})$, $(Y_2-\bar{Y})$, . . . , $(Y_n-\bar{Y})$, теперь связаны ограничением

$$\sum_{i=1}^{n}(Y_i-\bar{Y})=0,$$

и поэтому одна степень свободы исключается из общего числа.

Матрица данных для уравнения (5.5.8) будет иметь следующий вид:

z_1	z_2	. . .	z_p	y
$Z_{11}-\overline{Z}_1$	$Z_{21}-\overline{Z}_2$		$Z_{p1}-\overline{Z}_p$	$Y_1-\overline{Y}$
$Z_{12}-\overline{Z}_1$	$Z_{22}-\overline{Z}_2$		$Z_{p2}-\overline{Z}_p$	$Y_2-\overline{Y}$
. . .	. . .		. . .	. . .
$Z_{1n}-\overline{Z}_1$	$Z_{2n}-\overline{Z}_2$		$Z_{pn}-\overline{Z}_p$	$Y_n-\overline{Y}$

Для иллюстрации ситуации в очень простом случае мы воспользуемся примером, построенным выше, с матрицей исходных данных (5.5.4). Положим $y_i=Y_i-\overline{Y}$, $z_{1i}=X_i-3$, $z_{2i}=X_i^2-11$, что приведет к новой матрице исходных данных:

	z_1	z_2	y
	-2	-10	$y_1=Y_1-\overline{Y}$
	-1	-7	$y_2=Y_2-\overline{Y}$
	0	-2	$y_3=Y_3-\overline{Y}$
	1	5	$y_4=Y_4-\overline{Y}$
	2	14	$y_5=Y_5-\overline{Y}$
Итого по столбцу	0	0	0
Сумма квадратов для столбца	10	374	

Заметим, что благодаря центрированию столбцов X и X^2 уменьшаются абсолютные значения чисел, участвующих в вычислениях, и подчеркиваются не столько абсолютные значения, сколько разброс и распределение элементов X-столбца относительно их среднего. Центрирование также необходимо для получения корреляционной матрицы переменных, которая очень важна далее в процедурах отбора, обсуждаемых в гл. 6.

Корреляционная матрица

Пусть мы хотим построить модель

$$Y=\beta_0+\beta_1 Z_1+\beta_2 Z_2+\varepsilon.$$

Центрируя данные, как показано выше, можно привести модель к виду

$$Y-\overline{Y}=\beta_1(Z_1-\overline{Z}_1)+\beta_2(Z_2-\overline{Z}_2)+\varepsilon$$

аналогично уравнению (5.5.8) при $p=2$.

Когда модель записана в такой форме, «матрица $\mathbf{X}'\mathbf{X}$» имеет вид

$$\mathbf{S}=\begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix},$$

где $S_{jl}=\sum_{i=1}(Z_{ji}-\overline{Z}_j)(Z_{li}-\overline{Z}_l)$, $j, l=1, 2$, и где Z_{ji}, $i=1, 2, \ldots, n$, есть n наблюдений, получающихся при Z_j. Числа S_{jl} нередко могут сильно разниться. В больших $\mathbf{S}$-матрицах, скажем 5×5 и более, это может часто приводить к ошибкам округления при обращении матрицы, даже когда работа ведется на вычислительной машине. Пусть производится следующее преобразование центрированных данных:

$$x_{ji}=\frac{(Z_{ji}-\overline{Z}_j)}{S_{jj}^{1/2}}, \qquad j=1, 2, \qquad y_i=\frac{(Y_i-\overline{Y})}{S_{yy}^{1/2}},$$

где $S_{yy}=\sum_{i=1}^{n}(Y_i-\overline{Y})^2$. Мы можем сделать подобные преобразования переменных Z_1, Z_2 и Y, отбрасывая всюду индекс i. Это приведет к новой форме «центрированной модели»:

$$yS_{yy}^{1/2}=\beta_1 S_{11}^{1/2}x_1+\beta_2 S_{22}^{1/2}x_2+\varepsilon$$

или

$$y=\alpha_1 x_1+\alpha_2 x_2+\varepsilon',$$

где $\alpha_1=\beta_1(S_{11}/S_{yy})^{1/2}$ и $\alpha_2=\beta_2(S_{22}/S_{yy})^{1/2}$ — новые коэффициенты, которые подлежат оцениванию по преобразованным данным (y_i, x_{1i}, x_{2i}), $i=1, 2, \ldots, n$, и представляют собой *масштабированные* исходные коэффициенты β_1 и β_2. Когда модель записывается в такой форме, «матрица $\mathbf{X}'\mathbf{X}$» приобретает вид

$$\begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix}$$

и называется корреляционной матрицей Z-переменных, где

$$r_{12}=\frac{S_{12}}{(S_{11}S_{22})^{1/2}}=r_{21}$$

есть коэффициент корреляции между Z_1 и Z_2, как сказано в параграфе 1.6. Так как мы можем теперь записать

$$r_{jy} = \frac{S_{jy}}{(S_{jj}S_{yy})^{1/2}},$$

где $S_{jy} = \sum_{i=1}^{n} (Z_{ji} - \bar{Z}_j)(Y_i - \bar{Y})$ — корреляция между Z_j ($j = 1$ или 2) и Y (снова см. параграф 1.6), нормальные уравнения для новой модели упростятся и примут вид

$$\begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix},$$

где a_1 и a_2 — МНК-оценки для α_1 и α_2. Легко найти (используя тот факт, что $r_{21} = r_{12}$) решение этой системы уравнений:

$$a_1 = \frac{(r_{1y} - r_{12}r_{2y})}{D},$$

$$a_2 = \frac{(r_{2y} - r_{12}r_{1y})}{D},$$

где D — определитель матрицы корреляций, $D = 1 - r_{12}^2$.

Отметим здесь следующие моменты. В общем, преобразования регрессионной задачи к виду, в котором используются корреляции, удобны, так как они делают все числа, участвующие в вычислениях, лежащими между — 1 и 1. Когда все числа одного порядка, неблагоприятные эффекты ошибок округления минимизируются. Хотя при рассмотрении всего двух факторов опасность и невелика, она резко возрастает, когда задачи со многими независимыми переменными решаются на машине. (Общее правило таково: применение корреляции не необходимо, если задача достаточно проста и можно считать вручную. Однако оно входит существенной частью в хорошую машинную программу). При определении значений обеих оценок a_1 и a_2, приведенных выше, мы должны делить на $D = 1 - r_{12}^2$ определитель корреляционной матрицы. Отсюда, если величина r_{12}^2 очень близка к единице, D очень близко к нулю, и формальный подход к вычислению a_1 и a_2 приводит к неопределенности, делая их значения очень большими. Далее, r_{12} — это корреляция между Z_1 и Z_2. Следовательно, если Z_1 и Z_2 полностью или почти полностью зависят друг от друга или просто меняются одновременно или почти одновременно, то r_{12} будет равен единице или почти единице, а D будет нулем или почти нулем. Фактически, конечно, если D равен нулю, то мы не можем произвести вычисления a_1 и a_2, как раньше, поскольку вместо двух нормальных уравнений в действительности имеется только одно. Общее следствие отсюда: важно вычислить значение определителя $D = 1 - r_{12}^2$ с тем, чтобы установить, имеется ли зависимость между нормальными

уравнениями. В более общих задачах корреляционная матрица имеет вид

$$\begin{bmatrix} 1 & r_{12} & r_{13} & \cdots & r_{1p} \\ r_{21} & 1 & r_{23} & \cdots & r_{2p} \\ r_{31} & r_{32} & 1 & \cdots & r_{3p} \\ \vdots & & & & \\ r_{p1} & r_{p2} & r_{p3} & \cdots & 1 \end{bmatrix},$$

но сделанные замечания справедливы, и вычисление определителя такой матрицы — важная часть любой хорошей вычислительной регрессионной программы. Конечно, когда рассматривается задача только с одной независимой переменной, то матрица станет просто числом — единицей [15].

Преобразования, сделанные выше для получения корреляционной матрицы 2×2 из «матрицы $\mathbf{X}'\mathbf{X}$», включали репараметризацию $\alpha_j = \beta_j (S_{jj}/S_{yy})^{1/2}$, $j = 1, 2$. Напомним также (см. с. 319), что $b_0' = \overline{Y}$ есть оценка для величины $\beta_0 + \beta_1\overline{Z}_1 + \beta_2\overline{Z}_2$. Следовательно, оценки b_0, b_1, b_2 исходных коэффициентов β_0, β_1, β_2 получаются так

$$b_1 = a_1\left(\frac{S_{yy}}{S_{11}}\right)^{1/2},$$

$$b_2 = a_2\left(\frac{S_{yy}}{S_{22}}\right)^{1/2},$$

$$b_0 = \overline{Y} - b_1\overline{Z}_1 - b_2\overline{Z}_2.$$

Некоторые машинные программы выводят на печать оба множества коэффициентов.

Частные корреляции

Корреляции играют важную роль в различных методах выбора предикторов, обсуждаемых в гл. 6. В некоторых из них мы хотим добавлять независимые переменные одну за другой в выбранную модель. Первый предиктор, включаемый в выбранную модель,— это переменная, наиболее сильно коррелирующая с Y, т. е. переменная Z_j,

---

[15] Здесь авторы, видимо, оговорились и имеют в виду не «одну независимую переменную», а модель с одним определяемым коэффициентом. В конце концов нам безразлично, какова природа столбцов матрицы $\mathbf{X}$: порождены ли они различными функциями одного предиктора или множеством разных предикторов, или же и тем и другим одновременно. Лишь бы между ними не было «корреляции» (точнее было бы сказать «сопряженности» («коллинеарности»)).— *Примеч. пер.*

для которой r_{jy} — наибольший из всех r_{ly}, $l = 1, 2, \ldots, p$. Пусть для простоты это будет Z_1. Построим модель

$$Y = \beta_0 + \beta_1 Z_1 + \varepsilon.$$

Затем построим новые переменные $Z_2^*, Z_3^*, \ldots, Z_p^*$, используя остатки от регрессии Z_2 на Z_1, т. е. остатки от подобранной модели $Z_2 = \alpha_0 + \alpha_1 Z_1 + \varepsilon'$, остатки от регрессии Z_3 на $Z_1, \ldots,$ остатки от регрессии Z_p на Z_1 соответственно. Новая зависимая переменная Y^* имеет значения, содержащие остатки от регрессии Y на Z_1 (т. е. используют приведенную выше модель $Y = \beta_0 + \beta_1 Z_1 + \varepsilon$). Аналогичная работа была проделана в параграфе 4.1 при $p = 2$.

Значения новой зависимой переменной Y^* и новых независимых переменных $Z_2^*, Z_3^*, \ldots, Z_p^*$ представляют собой доли соответствующих векторов исходных данных, которые не зависят от значений переменной Z_1. Теперь мы можем получить новое множество корреляций, которые используют переменные со звездочками. Они называются *частными* (*парциальными*) *корреляциями*, могут записываться, например, как $r_{2y \cdot 1}$, имея смысл корреляции переменных Z_2^* и Y^*, а читаются так: «частная корреляция переменных Z_2 и Y после исключения влияния (снятия) переменной Z_1». На второй стадии процедуры отбора мы будем включать в модель переменную Z_j, для которой коэффициент частной корреляции с Y без учета Z_1 (т. е. $r_{jy \cdot 1}$) оказался наибольшим; другими словами, будем выбирать переменную Z_j, наиболее коррелированную с Y после того, как эффект Z_1 исключен из обеих переменных Y и Z_j. Если отобранная таким образом вторая Z переменная есть, скажем, Z_2, то на третьей стадии процедуры отбора рассматривается частная корреляция вида $r_{jy \cdot 12}$, т. е. корреляция между: 1) остатками регрессии Z_j на Z_1 и Z_2 и 2) остатками регрессии Y на Z_1 и Z_2. Этот процесс можно продолжать сколько угодно [16]. Квадраты значений таких частных корреляций можно увидеть в распечатках из приложения Б. (Квадраты используются, когда в знаках нет нужды.)

Если потребуется, то частные корреляции можно выразить через обычные. Например,

$$r_{2y \cdot 1} = (r_{2y} - r_{12} r_{1y}) / \{(1 - r_{1y}^2)(1 - r_{12}^2)\}^{1/2}.$$

Правда, такие формулы нужны редко.

5.6. ОРТОГОНАЛЬНЫЕ ПОЛИНОМЫ

Ортогональные полиномы применяются для построения полиномиальных моделей любого порядка от одной переменной. Идея такова. Пусть имеется n наблюдений (X_i, Y_i), $i = 1, 2, \ldots, n$, где X — пре-

[16] Более подробно вопрос о частных корреляциях освещен, например, в кн. Езекиэл М., Фокс К. А. Методы анализа корреляций и регрессий, линейных и криволинейных/Пер. с англ. Под ред. Н. К. Дружинина.— М.: Статистика, 1966.— 557 с.— *Примеч. пер.*

диктор, а Y — отклик, и мы хотим подобрать модель

$$Y=\beta_0+\beta_1X+\beta_2X^2+\ .\ .\ .\ +\beta_pX^p+\varepsilon. \tag{5.6.1}$$

Как правило, столбцы, образующие **X**-матрицу, не ортогональны. Если мы захотим в дальнейшем добавить в модель новый член $\beta_{p+1}X^{p+1}$, то в общем случае произойдут изменения оценок всех остальных коэффициентов. Однако мы можем построить полиномы вида

$$\psi_0(X_i)=1 \text{ — полином нулевого порядка,}$$

$$\psi_1(X_i)=P_1X_i+Q_1 \text{ — полином первого порядка,}$$

$$\psi_2(X_i)=P_2X_i^2+Q_2X_i+R_2 \text{ — полином второго порядка,}$$

. . . . . .

$$\psi_r(X_i)=P_rX_i^r+Q_rX_i^{r-1}+\ .\ .\ .\ +T_r \text{ — полином } r\text{-го порядка,}$$

обладающие тем свойством, благодаря которому их называют *ортогональными полиномами*, т. е.

$$\sum_{i=1}^{n}\psi_j(X_i)\psi_l(X_i)=0 \quad (j\neq l) \tag{5.6.2}$$

для всех j, $l<n-1$. Теперь можно переписать модель так:

$$Y=\alpha_0\psi_0(X)+\alpha_1\psi_1(X)+\ .\ .\ .\ +\alpha_p\psi_p(X)+\varepsilon. \tag{5.6.3}$$

В этом случае получим

$$\mathbf{X}=\begin{bmatrix} 1 & \psi_1(X_1) & \psi_2(X_1) & \ .\ .\ . & \psi_p(X_1) \\ 1 & \psi_1(X_2) & \psi_2(X_2) & \ .\ .\ . & \psi_p(X_2) \\ \multicolumn{5}{c}{.\ .\ .\ .\ .\ .\ .\ .\ .\ .\ .\ .\ .\ .\ .\ .\ .\ .\ .} \\ 1 & \psi_1(X_n) & \psi_2(X_n) & \ .\ .\ . & \psi_p(X_n) \end{bmatrix}, \tag{5.6.4}$$

так что

$$\mathbf{X}'\mathbf{X}=\begin{bmatrix} A_{00} & & & & \\ & A_{11} & & & 0 \\ & & A_{22} & & \\ & & & \ddots & \\ 0 & & & & A_{pp} \end{bmatrix}, \tag{5.6.5}$$

где $A_{jj}=\sum_{i=1}^{n}\{\psi_j(X_i)\}^2$, а все недиагональные элементы обращаются в нуль в соответствии с уравнением (5.6.2). Так как обратная матрица $(\mathbf{X}'\mathbf{X})^{-1}$ также диагональна и получается обращением каждого эле-

мента отдельно (см. уравнение (2.1.13)), то метод наименьших квадратов дает в качестве оценок коэффициентов α_j величины:

$$a_j = \frac{\sum_{i=1}^{n} Y_i \psi_j(X_i)}{\sum_{i=1}^{n} [\psi_j(X_i)]^2}, \qquad j = 0, 1, 2, \ldots, p, \tag{5.6.6}$$

$$a_j = \frac{A_{jY}}{A_{jj}}$$

с очевидными обозначениями. Так как $V(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\sigma^2$ в общей модели регрессии, то дисперсия a_j есть

$$V(a_j) = \frac{\sigma^2}{A_{jj}}, \tag{5.6.7}$$

где σ^2 обычно оценивается на основе таблицы дисперсионного анализа. Для получения данных такой таблицы мы вычислим сумму квадратов, обусловленную a_j:

$$\mathrm{SS}(a_j) = \frac{A_{jY}^2}{A_{jj}}, \tag{5.6.8}$$

и тогда сможем получить следующую таблицу дисперсионного анализа:

ANOVA

Источник	Число степеней свободы	SS	MS
a_0 (среднее)	1	SS (a_0)	—
a_1	1	SS (a_1)	SS (a_1)
a_2	1	SS (a_2)	SS (a_2)
. . .	. . .	. . .	. . .
a_p	1	SS (a_p)	SS (a_p)
Остаток	$n-p-1$	По разности	s^2
Общий	n	$\sum_{i=1}^{n} Y_i^2$	

Если модель корректна, то s^2 есть оценка для σ^2. Обычно, когда средний квадрат незначимо больше s^2, мы объединяем суммы квадратов с остатком и получаем оценку величины σ^2, основанную на большем числе степеней свободы. Заметим, что если ожидается добавление в уравнение (5.6.3) нового члена $\alpha_{p+1}\psi_{p+1}(X)$, то в новых перерасчетах нет необходимости вследствие ортогональности полиномов. Так можно легко подбирать полиномы все более и более высокого порядка, прекращая этот процесс, когда подобранное уравнение станет удовлетворительным.

Можно построить $\psi_j(X_i)$ и реализовать рассмотренную выше процедуру для любых значений X_i. Однако, когда X_i меняется с неравным шагом, полиномы приходится строить специально. (См., например, работы: Wishart J., Metakides T. Orthogonal polynomial fitting.— Biometrika, 1953, **40**, p. 361—369 и Robson D. S. A simple method for constructing orthogonal polynomials when the independent variable is unequally spaced.—Biometrics, 1959, **15**, p. 187—291.) Если X_i меняется с равным шагом, можно воспользоваться таблицами. Фактические численные значения $\psi_j(X_i)$ и A_{jj}, так же как и общие функциональные формы для $\psi_j(X_i)$, при $j = 1, 2, \ldots, 6$ и $n \leqslant 52$ даны в таблицах Е. Пирсона и Х. Хартли (см.: Pearson E. S., Hartley H. O. Biometrika Tables for Statisticians. Cambridge University Press, 1958, 1). Это значит, что мы обеспечены при $p \leqslant 6$ элементами матриц $\mathbf{X}$ и $\mathbf{X}'\mathbf{X}$. Для $j \leqslant 5$ и $n \leqslant 75$ аналогичные данные приводятся в статистических таблицах Р. Фишера и Ф. Йейтса для биологических, сельскохозяйственных и медицинских исследований (см.: Fisher R. A., Yates F. Statistical Tables for Biological, Agricultural and Medical Research.— New York: Hafner Publishing Co., 1964 (6-е изд.)). Когда $p > 6$, нужны более обширные таблицы, например книга Де Лури, включающая значения и интегралы ортогональных полиномов до $n = 26$ (см.: De Lury D. B. Values and Integrals of the Orthogonal Polynomials up to n = 26, University of Toronto Press, 1960). Еще см. работу Д. Уилки о полном множестве приведенных коэффициентов $\lambda(r, n)$ для ортогональных полиномов до $n = 26$ (Wilkie D. Complete set of leading coefficients, $\lambda(r, n)$, for orthogonal polynomials up to $n = 26$.— Technometrics, 1965, **7**, p. 644—648). Краткая таблица ортогональных полиномов для равномерно расположенных данных при n не больше 12 приведена [17] на с. 333 — 334 .

---

[17] В связи с важностью вопроса приведем еще ряд источников, где можно найти ортогональные полиномы в виде таблиц, графиков или формул, кроме тех, что названы в тексте. На русский язык переведены подробные таблицы, составленные К. Ланцошем, см.: Таблицы полиномов Чебышева $S_n(x)$ и $C_n(x)$.— М.: ВЦ АН СССР, 1963.— (Библиотека математических таблиц, вып. 19). Наиболее доступные обширные таблицы: Большев Л. Н., Смирнов Н. В. Таблицы математической статистики.— 3-е изд.— М.: Наука, 1983.— 416 с. (особо с. 100—102, 376—385 ($n = 3$ (1) 52, $i = 6$)); Хотимский В. Выравнивание статистических рядов по методу наименьших квадратов (способ Чебышева).— М.: Госстатиздат, 1959.— 85 с. Оригинальные таблицы с удобными вспомогательными коэффициентами можно найти в кн.: Литтл Т., Хиллз Ф. Сельскохозяйственное опытное дело. Планирование и анализ/Пер. с англ. Под ред. Д. В. Васильевой.— М.: Колос, 1981.— 319 с. (особо с. 210—219 ($n = 3$ (1) 25), с. 299—306); Янке Е., Эмде Ф., Лёш Ф. Специальные функции. Формулы, графики, таблицы.—3-е изд./Пер. с нем. Под ред. Л. И. Седова.— М.: Наука, 1977.— 342 с. (особо с. 142—144); Оуэн Д. Б. Сборник статистических таблиц.— 2-е изд./Пер. с англ.— М.: ВЦ АН СССР, 1973.— 586 с. (особо с. 515—516; таблицы практически совпадают с теми, что приведены в тексте); Таблицы специальных функций/Под ред. Я. Н. Шпильрейна.— М.: ГТТИ, 1934; Справочник по специальным функциям с формулами, графиками и таблицами/Под ред. М. А. Абрамовица и И. Стиган/Пер. с англ. Под ред. В. А. Дидкина и Л. М. Кармазиной.—М.: Наука, 1979.— 830 с.; Anderson R. L., Houseman E. E. Tables of Orthogonal Polinomial Values

Хотя ортогональные полиномы рекомендуют, как правило, только тогда, когда применяются карманные компьютеры, Дж. Брайт и Дж. Доукинз в работе о некоторых аспектах подбора кривой с помощью ортогональных полиномов (см.: Bright J. W., Dawkins G. S. Some aspects of curve fitting using orthogonal polynomials.— Industrial and Engineering Chemistry Fundamentals, February, 1965, 4, p. 93—97) пришли к заключению, что даже если есть вычислительная машина, особенно в случае полиномов высокого порядка, ортогональные полиномы имеют смысл [18]. Их использование обеспечивает бо́льшую машинную точность и сокращает время счета. Мы проиллюстрируем применение ортогональных полиномов на примере.

---

extended to N = 104.— Research Bulletin № 297.— Ames, Iowa: Iowa Agricultural Experiment Station, 1942; National Bureau of Standards, Applied Mathematics, Series 9: Tables of Chebyshev polinomials $S_n(x)$ and $C_n(x)$.— Washington: 1952.— *Примеч. пер.*

[18] Ортогональные полиномы описанного вида связаны с именем замечательного русского математика П. Л. Чебышева (1821—1894). О нем см., например: Прудников В. Е. Пафнутий Львович Чебышев.— Л.: Наука, 1976.— 282 с. Первую работу «О непрерывных дробях», где вводились ортогоиальные полиномы, П. Л. Чебышев опубликовал в 1855 г., а дальнейшее развитие эта идея получила в 1858 г. в работе «Об интерполировании в случае большого числа данных, полученных из наблюдений» (см.: Чебышев П. Л. Полное собрание сочинений.— М.: Изд-во АН СССР, 1944—1951, т. 2, с. 103—126 и 245—314). Полиномы Чебышева быстро нашли применение в различных областях. В связи с МНК существенное продвижение в 1878 г. осуществил американский вычислитель-геодезист М. Дулиттл в работе: Doolittle M. H. Adjustment of the primary triangulation between Kent Island and Atlanta base lines. (Paper N 3. Method employed in solution of normal equations and the adjustment of a triangulation.) Report of Superintendant, Coast and Geodetic Survey. 1878, p. 115—120. Он построил вычислительную схему МНК, основанную на ортогональных полиномах. В «домашинную эру» она обеспечивала практически единственную реальную возможность решения задач относительно большой размерности. Среди тех, кто усовершенствовал схему Дулиттла, был академик В. С. Немчинов, много сделавший для внедрения МНК в статистические задачи, особенно в задачи сельскохозяйственной статистики. Общую идею применения полиномов Чебышева он изложил в работе: Немчинов В. С. Сельскохозяйственная статистика с основами общей теории.— М.: Сельхозгиз, 1945. А на следующий год было опубликовано специальное исследование: Немчинов В. С. Полиномы Чебышева и математическая статистика.— М.: СХА им. К. А. Тимирязева, 1946. И еще раз В. С. Немчинов вернулся к этому вопросу уже в конце жизни, см.: Немчинов В. С. Экономико-математические методы и модели.— М.: Соцэкгиз, 1962.—410 с. (особо с. 96—154). Описание схемы Дулиттла можно найти еще, например, в: Езекиэл М., Фокс К. А. Методы анализа корреляций и регрессий, линейных и криволинейных/Пер. с англ. Под ред. Н. К. Дружинина. — М.: Статистика, 1966.— 557 с. (особо с. 503—523). Важный шаг в переосмысливании возможностей применения полиномов Чебышева в задачах обработки данных совершил недавно умерший американский статистик Корелиус Ланцош. Его подход, оказавший и продолжающий оказывать значительное влияние, блестяще изложен в работе: Ланцош К. Практические методы прикладного анализа. Справочное руководство/Пер. с англ. Под ред. А. М. Лопшица.— М.: Физматгиз, 1961.— 524 с. Более современная сводка основных результатов в области теории и приложений полиномов Чебышева содержится в монографии: Карлин С., Стаддеи В. Чебышевские системы и их применение в анализе и статистике/Пер. с англ. Под ред. С. М. Ермакова.— М.: Наука, 1976.— 568 с. Материал о применении ортогонализации в регрессионном анализе можно найти, например,

Пример. Фирма «Жиллетт» (Gillette) [19] опубликовала чистый доход на акцию по годам за 1957—1964 гг.:

Год (z_i)	1957	1958	1959	1960	1961	1962	1963	1964
Чистый доход на акцию в дол. (Y_i)	0,93	0,99	1,11	1,33	1,52	1,60	1,47	1,33

Требуется подобрать полином подходящего порядка, который удовлетворительно аппроксимирует эти данные.

Решение. Мы игнорируем в годовых данных все характеристики, кроме того, что они меняются с равным шагом. Из таблицы ортогональных полиномов мы найдем значения $\psi_j(X_i)$, соответствующие $n = 8$ наблюдениям, как показано в табл. 5.18.

Таблица 5.18. **Расчетная таблица для примера**

$Y - 0{,}93$	Ψ_0	Ψ_1	Ψ_2	Ψ_3	Ψ_4	Ψ_5	Ψ_6
0	1	—7	7	—7	7	—7	1
0,06	1	—5	1	5	—13	23	—5
0,18	1	—3	—3	7	—3	—17	9
0,40	1	—1	—5	3	9	—15	—5
0,59	1	1	—5	—3	9	15	—5
0,67	1	3	—3	—7	—3	17	9
0,54	1	5	1	—5	—13	—23	—5
0,40	1	7	7	7	7	7	1
A_{jj}	8	168	168	264	616	2184	264

Для упрощения вычислений будем использовать $(Y—0{,}93)$ вместо Y. Заметим, что $A_{00} = n = 8$. Мы рассматриваем модель

$$Y - 0{,}93 = \sum_{j=0}^{6} \alpha_j \psi_j (X).$$

в работах: Линник Ю. В. Метод наименьших квадратов и основы математико-статистической теории обработки наблюдений.— 2-е изд.— М.: Физматгиз, 1962.— 352 с.; Себер Дж. Линейный регрессионный анализ/Пер. с англ. Под ред. М. Б. Малютова.— М.: Мир, 1980.— 456 с. (особо с. 213—221). Лоусон Ч., Хенсон Р. Численное решение задач метода наименьших квадратов. Пер. с англ. — М.: Наука, 1986. О применении ортогональных полиномов в планировании эксперимента см., например, в уже упоминавшейся монографии В. В. Налимова и Н. А. Черновой, с. 277—281, или в работе: Маркова Е. В., Лисенков А. Н. Планирование эксперимента в условиях неоднородностей.— М.: Наука, 1973.— 219 с. Связи ортогонализации с обработкой данных в более широком аспекте посвящены, например, работы: Мили Э. Численный анализ/Пер. с англ.— М.: ИЛ, 1951; Хемминг Р. В. Численные методы для научных работников и инженеров/Пер. с англ. Под ред. Р. С. Гутера.— М.: Наука, 1968.— 400 с. (особо с. 257—272); Гутер Р. С., Овчинский Б. В. Элементы численного анализа и математической обработки результатов опыта.— 2-е изд.— М.: Наука, 1970.— 432 с. (особо с. 382—392). См. также примечание [17] на с. 327. Отметим еще, что существуют и иные способы ортогонализации, использующие другие системы функций. См., например: Суетин П. К. Классические ортогональные многочлены.—М.: Наука, 1976.—328 с. Таким образом, ортогонализация имеет множество точек соприкосновения с регрессионным анализом и планированием эксперимента. Последнее слово здесь еще не сказано.— *Примеч. пер.*

[19] Фирма «Жиллетт» (Gillette) — старая швейцарская фирма (ныне США), славившаяся лезвиями для бритья.— *Примеч. пер.*

На основании уравнения (5.6.6) получим

$$a_0 = \frac{(0\,(1) + 0{,}06\,(1) + \ldots + 0{,}40\,(1))}{8} = \frac{2{,}84}{8} = 0{,}355,$$

$$a_1 = \frac{(0\,(-7) + 0{,}06\,(-5) + \ldots + 0{,}40\,(7))}{168} = \frac{6{,}86}{168} = 0{,}040833,$$

$$a_2 = \frac{-4{,}10}{168} = -0{,}024405,$$

$$a_3 = \frac{-3{,}60}{264} = -0{,}013636,$$

$$a_4 = \frac{1{,}36}{616} = 0{,}002208,$$

$$a_5 = \frac{2{,}94}{2184} = 0{,}001346,$$

$$a_6 = \frac{(0\,(1) + 0{,}06\,(-5) + \ldots + 0{,}40\,(1))}{264} = \frac{0{,}10}{264} = 0{,}000379.$$

Подобранная модель имеет вид

$$\hat{Y} - 0{,}93 = \sum_{j=0}^{6} a_j \psi_j\,(X),$$

где a_j уже найдены, а $\psi_j\,(X)$ находятся из таблиц (мы найдем их после преобразования уравнения). Для вычисления элементов таблицы дисперсионного анализа сначала найдем следующие значения:

$j = 0$	1	2	3	4	5	6
$A_{jY} = 2{,}84$	6,86	—4,10	—3,60	1,36	2,94	0,10
$A_{jj} = 8$	168	168	264	616	2184	264

Используя уравнение (5.6.8), получим таблицу дисперсионного анализа:

ANOVA

Источник	Число степеней свободы	SS	MS
a_0 (среднее)	1	1,008	(равные SS)
a_1	1	0,280	
a_2	1	0,100	
a_3	1	0,049	
a_4	1	0,003	
a_5	1	0,004	
a_6	1	0,000	
Остаток	1	0,001	
Общий	8	1,445	

(Примечание. Если порядок подбираемого полинома максимален и равен $p = N-1$, то модель предсказывает точно, без остатков. Здесь же $p = N-2$ и таким образом остаток фактически равен $SS(a_7)$.)

Если оценка σ^2 была получена независимо, то можно сравнить с ней средние квадраты. Проделав это, мы увидим, что члены третьей и более низких степеней вносят в данные наибольшую вариацию. Таким образом, можно принять в качестве модели выражение

$$\hat{Y} - 0{,}93 = 0{,}355 + 0{,}041\psi_1(X) - 0{,}024\psi_2(X) - 0{,}014\psi_3(X)$$

и перестроить таблицу дисперсионного анализа:

ANOVA

Источник	Число степеней свободы	SS	MS
a_0 (среднее)	1	1,008	1,008
a_1	1	0,280	0,280
a_2	1	0,100	0,100
a_3	1	0,049	0,049
Остаток	4	0,008	$s^2 = 0{,}002$
Общий	8	1,445	

Чтобы получить уравнение относительно исходных переменных, мы должны прежде всего выразить функции ψ_j через Z-переменные. Из таблицы ортогональных полиномов [20] (см., например, Biometrika Tables, p. 212, $n = N = 8$ или информацию, приводимую ниже в этом параграфе) находим

$$\psi_0(X) = 1, \quad \psi_1(X) = 2X, \quad \psi_2(X) = X^2 - \frac{21}{4},$$

$$\psi_3(X) = \frac{2}{3}\left[X^3 - \frac{37}{4}\right].$$

Между Z и X имеется следующее соответствие:

$\psi_1(X_i) = 2X_i$:	-7	-5	-3	-1	1	3	5	7
X_i:	$-\frac{7}{2}$	$-\frac{5}{2}$	$-\frac{3}{2}$	$-\frac{1}{2}$	$\frac{1}{2}$	$\frac{3}{2}$	$\frac{5}{2}$	$\frac{7}{2}$
Z_i:	1957	1958	1959	1960	1961	1962	1963	1964

Ясно, что требуемое кодирование задается формулой

$$X = Z - 1960\frac{1}{2}.$$

[20] Более доступную таблицу см. в примечании 17.— *Примеч. пер.*

Итоговый полином таков:

$$\hat{Y} = 1{,}285 + 0{,}082\left(Z - 1960\frac{1}{2}\right) - 0{,}024\left[\left(Z - 1960\frac{1}{2}\right)^2 - \frac{21}{4}\right] - \\ - \frac{0{,}028}{3}\left[\left(Z - 1960\frac{1}{2}\right)^3 - \frac{37}{4}\left(Z - 1960\frac{1}{2}\right)\right].$$

Его можно преобразовать в кубический полином от Z, но приведенная выше форма удобнее для получения предсказанных значений и остатков. Читателю рекомендуется нанести данные на график, вычислить предсказанные значения, а также исследовать остатки, как описано в гл. 3.

Ортогональные полиномы для $n = 3, \ldots, 12$

Вот формулы ортогональных полиномов $\psi_j(X)$ до шестого порядка при любых значениях n и X-ах, заданных с равным шагом:

$$\psi_0(X) = 1,$$

$$\psi_1(X) = \lambda_1 X,$$

$$\psi_2(X) = \lambda_2\left\{X^2 - \frac{1}{12}(n^2 - 1)\right\},$$

$$\psi_3(X) = \lambda_3\left\{X^3 - \frac{1}{20}(3n^2 - 7)\,X\right\},$$

$$\psi_4(X) = \lambda_4\left\{X^4 - \frac{1}{14}(3n^2 - 13)\,X^2 + \frac{3}{560}(n^2 - 1)(n^2 - 9)\right\},$$

$$\psi_5(X) = \lambda_5\left\{X^5 - \frac{5}{18}(n^2 - 7)\,X^3 + \frac{1}{1008}(15n^4 - 230n^2 + 407)\,X\right\},$$

$$\psi_6(X) = \lambda_6\left\{X^6 - \frac{5}{44}(3n^2 - 31)\,X^4 + \frac{1}{176}(5n^4 - 110n^2 + 329)\,X^2 - \right. \\ \left. - \frac{5}{14784}(n^2 - 1)(n^2 - 9)(n^2 - 25)\right\}.$$

(Если $n \leqslant 6$, то мы действуем точно так же, только полагая, что $j = n - 1$.) Соответствующие значения λ приводятся под каждым столбцом. Они выбраны так, чтобы гарантировать в таблицах только целые значения. Приведенные выше значения λ представляют собой суммы квадратов элементов, образующих соответствующий столбец, т. е. то, что обозначено A_{jj} в уравнении (5.6.5). Отметим, что хотя X-ы сами по себе и не фигурируют в таблице, всегда справедливо условие $\psi_1 = \lambda_1 X$, где λ_1 равна 1 для нечетных n и 2 — для четных, так что если мы захотим, то всегда можем считать X как отношение ψ_1/λ_1. Все столбцы ψ_j для четных j симметричны, а для нечетных — антисимметричны.

(В работе Б. Купера об использовании ортогональных полиномов

Таблица коэффициентов ортогональных полиномов

$n=3$		$n=4$			$n=5$				$n=6$				
ψ_1	ψ_2	ψ_1	ψ_2	ψ_3	ψ_1	ψ_2	ψ_3	ψ_4	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5
−1	1	−3	1	−1	−2	2	−1	1	−5	5	−5	1	−1
0	−2	−1	−1	3	−1	−1	2	−4	−3	−1	7	−3	5
1	1	1	−1	−3	0	−2	0	6	−1	−4	4	2	−10
		3	1	1	1	−1	−2	−4	1	−4	−4	2	10
					2	2	1	1	3	−1	−7	−3	−5
									5	5	5	1	1
2	6	20	4	20	10	14	10	70	70	84	180	28	252
1	3	2	1	$\frac{10}{3}$	1	1	$\frac{5}{6}$	$\frac{35}{12}$	2	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{7}{12}$	$\frac{21}{10}$

Таблица коэффициентов ортогональных полиномов

$n=7$						$n=8$					
ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
−3	5	−1	3	−1	1	−7	7	−7	7	−7	1
−2	0	1	−7	4	−6	−5	1	5	−13	23	−5
−1	−3	1	1	−5	15	−3	−3	7	−3	−17	9
0	−4	0	6	0	−20	−1	−5	3	9	−15	−5
1	−3	−1	1	5	15	1	−5	−3	9	15	−5
2	0	−1	−7	−4	−6	3	−3	−7	−3	17	9
3	5	1	3	1	1	5	1	−5	−13	−23	−5
						7	7	7	7	7	1
28	84	6	154	84	924	168	168	264	616	2184	264
1	1	$\frac{1}{6}$	$\frac{7}{12}$	$\frac{7}{20}$	$\frac{77}{60}$	2	1	$\frac{2}{3}$	$\frac{7}{12}$	$\frac{7}{10}$	$\frac{11}{60}$

Таблица коэффициентов ортогональных полиномов

$n=9$						$n=10$					
ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
−4	28	−14	14	−4	4	−9	6	−42	18	−6	3
−3	7	7	−21	11	−17	−7	2	14	−22	14	−11
−2	−8	13	−11	−4	22	−5	−1	35	−17	−1	10
−1	−17	9	9	−9	1	−3	−3	31	3	−11	6
0	−20	0	18	0	−20	−1	−4	12	18	−6	−8
1	−17	−9	9	9	1	1	−4	−12	18	6	−8
2	−8	−13	−11	4	22	3	−3	−31	3	11	6
3	7	−7	−21	−11	−17	5	−1	−35	−17	1	10
4	28	14	14	4	4	7	2	−14	−22	−14	−11
						9	6	42	18	6	3
60	2772	990	2002	468	1980	330	132	8580	2860	780	660
1	3	$-\frac{5}{6}$	$\frac{7}{12}$	$\frac{3}{20}$	$\frac{11}{60}$	2	$\frac{1}{2}$	$\frac{5}{3}$	$\frac{5}{12}$	$\frac{1}{10}$	$\frac{11}{240}$

при равных значениях x-ов (см.: Cooper B. E. The use of orthogonal polynomials with equal x values.— Applied Statistics, 1971, **20**, p. 209—213) приведен алгоритм AS 42, который реализует на языке Фортран генерирование ортогональных полиномов.)

Таблица коэффициентов ортогональных полиномов

$n=11$						$n=12$					
ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6	ψ_1	ψ_2	ψ_3	ψ_4	ψ_5	ψ_6
−5	15	−30	6	−3	15	−11	55	−33	33	−33	11
−4	6	6	−6	6	−48	−9	25	3	−27	57	−31
−3	−1	22	−6	1	29	−7	1	21	−33	21	11
−2	−6	23	−1	−4	36	−5	−17	25	−13	−29	25
−1	−9	14	4	−4	−12	−3	−29	19	12	−44	4
0	−10	0	6	0	−40	−1	−35	7	28	−20	−20
1	−9	−14	4	4	−12	1	−35	−7	28	20	−20
2	−6	−23	−1	4	36	3	−29	−19	12	44	4
3	−1	−22	−6	−1	29	5	−17	−25	−13	29	25
4	6	6	−6	−6	−48	7	1	−21	−33	−21	11
5	15	30	6	3	15	9	25	−3	−27	−57	−31
						11	55	33	33	33	11
110	858	4290	286	156	11 220	572	12 012	5148	8008	15 912	4488
1	1	5/6	1/12	1/40	11/120	2	3	2/3	7/24	3/20	11/360

5.7. ПРЕОБРАЗОВАНИЕ МАТРИЦЫ X ДЛЯ ПОЛУЧЕНИЯ ОРТОГОНАЛЬНЫХ СТОЛБЦОВ

В регрессионной задаче матрица **X** должна быть такой, чтобы ни один столбец нельзя было представить как линейную комбинацию других столбцов. Из этого вытекает также, что должно быть по крайней мере столько же независимых строк, сколько и оцениваемых параметров, иначе зависимость будет проявляться также и в столбцах. Пусть, например, наблюдения Y зарегистрированы только при трех уровнях X, а именно при $X = a$, b и c, причем постулируется модель

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon.$$

Матрица **X** имеет вид

$$\begin{bmatrix} 1 & a & a^2 & a^3 \\ 1 & b & b^2 & b^3 \\ 1 & c & c^2 & c^3 \end{bmatrix}$$

и столбцы зависимы, так как (столбец 4) $-$ $(a + b + c)$ $\times$ (столбец 3) $+$ $(ab + bc + ca)$ $\times$ (столбец 2) $-$ abc $\times$ (столбец 1) $= 0$. Распознать такую зависимость в регрессионных задачах часто бывает очень трудно. Если она существует, то матрица $\mathbf{X'X}$ будет всегда особенной и ее нельзя будет обратить. Если столбцы **X**-матрицы

почти зависимы, то матрица $\mathbf{X}'\mathbf{X}$ будет почти особенной и трудно обращаемой, с возможными большими ошибками округления.

Одна из процедур, которую можно запрограммировать и использовать как программу проверки X-матрицы (либо во всех, либо в подозрительных случаях), содержит последовательные преобразования столбцов, так что каждый новый столбец ортогонален ко всем предыдущим преобразованным столбцам. Если среди столбцов есть зависимые, то в итоге мы будем получать новые столбцы, состоящие целиком из нулей. Если столбцы почти зависимы, то новые столбцы будут состоять из очень малых чисел, возможно, с некоторыми нулями. Метод весьма общий. Преобразования столбцов имеют следующий вид:

$$\mathbf{Z}_{iT} = (\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\mathbf{Z}_i = \mathbf{Z}_i - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}_i, \qquad (5.7.1)$$

где $\mathbf{Z}$ — матрица из уже преобразованных вектор-столбцов; $\mathbf{Z}_i$ — следующий вектор-столбец $\mathbf{X}$, подвергаемый преобразованию; $\mathbf{Z}_{iT}$ — вектор, оказывающийся ортогональным к вектор-столбцам, уже содержащимся в $\mathbf{Z}$.

Заметим, что $\mathbf{Z}_{iT}$ — фактически остаточный вектор (вектор остатков) от $\mathbf{Z}_i$, после того как была построена регрессия $\mathbf{Z}_i$ на столбцы $\mathbf{Z}$.

Для иллюстрации этого процесса мы возьмем частный случай, он приводит к получению ортогональных полиномов при $n = 5$. Пусть значения Y зафиксированы при $X = 1, 2, 3, 4, 5$ и постулируется модель

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon. \qquad (5.7.2)$$

Исходная $\mathbf{X}$-матрица есть

$$\begin{array}{c} \begin{array}{cccc} 1 & X & X^2 & X^3 \end{array} \\ \left[\begin{array}{cccc} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \\ 1 & 5 & 25 & 125 \end{array}\right] \end{array}.$$

Положим на первом этапе

$$\mathbf{Z}_{iT} = \left[\begin{array}{c} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{array}\right] = \mathbf{Z}.$$

(Для начала процесса преобразования надо выбрать какой-нибудь вектор-столбец.)

Выберем

$$\mathbf{Z}_2 = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}.$$

Тогда по уравнению (5.7.1)

$$\mathbf{Z}_{2T} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} - \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} (5)^{-1}(15) = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} - \begin{bmatrix} 3 \\ 3 \\ 3 \\ 3 \\ 3 \end{bmatrix} = \begin{bmatrix} -2 \\ -1 \\ 0 \\ 1 \\ 2 \end{bmatrix}.$$

На данном этапе

$$\mathbf{Z} = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{bmatrix} = [\mathbf{Z}_{1T},\ \mathbf{Z}_{2T}]$$

и третий столбец $\mathbf{X}$ используем как $\mathbf{Z}_3$. Мы получим

$$(\mathbf{Z}'\mathbf{Z})^{-1} = \begin{bmatrix} 5 & 0 \\ 0 & 10 \end{bmatrix}^{-1} = \begin{bmatrix} 1/5 & 0 \\ 0 & 1/10 \end{bmatrix},$$

$$\mathbf{Z}'\mathbf{Z}_i = \begin{bmatrix} 55 \\ 60 \end{bmatrix},$$

тогда

$$(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}_i = \begin{bmatrix} 11 \\ 6 \end{bmatrix}.$$

Отсюда

$$\mathbf{Z}_{3T} = \begin{bmatrix} 1 \\ 4 \\ 9 \\ 16 \\ 25 \end{bmatrix} - \begin{bmatrix} 11-12 \\ 11-6 \\ 11-0 \\ 11+6 \\ 11+12 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ -2 \\ -1 \\ 2 \end{bmatrix}.$$

Вычисление $\mathbf{Z}_{4T}$ оставляем читателю для упражнения. Окончательная матрица с ортогональными столбцами имеет вид

$$\begin{bmatrix} 1 & -2 & 2 & -1{,}2 \\ 1 & -1 & -1 & 2{,}4 \\ 1 & 0 & -2 & 0 \\ 1 & 1 & -1 & -2{,}4 \\ 1 & 2 & 2 & 1{,}2 \end{bmatrix}.$$

Заметим, что первые три столбца есть ψ_0, ψ_1 и ψ_2 — ортогональные полиномы нулевого, первого и второго порядка для $n = 5$. Четвертый столбец — ортогональный полином ψ_3 третьего порядка, умноженный на 1, 2 для $n = 5$.

Как отмечалось, этот процесс совершенно общий. Зависимость между столбцами можно также обнаружить, зная, что в этом случае определитель матрицы $\mathbf{X'X}$ (или корреляционной матрицы) равен нулю. Однако процедура преобразования имеет дополнительные преимущества: она выявляет зависимые столбцы.

Описанная выше процедура хорошо известна как метод ортогонализации столбцов Грама—Шмидта [21] (Gram—Schmidt). Другой подход к этой задаче заключается в том, чтобы перевести матрицу $\mathbf{X'X}$ в корреляционную форму, как показано в той части параграфа 5.5, что начинается на с. 320, а затем найти собственные значения (характеристические значения, или латентные (скрытые) корни — все эти термины обозначают одно и то же) этой корреляционной матрицы. Если между **Z**-столбцами существуют линейные зависимости, то появятся нулевые собственные значения, а малые собственные значения (малые относительно диапазона от 0 до 1) будут служить указателями на возможные тесные зависимости. (См., например, с. 78 в работе Р. Сни о некоторых аспектах анализа неортогональных данных (Snee R. D. Some aspects of non-orthogonal data analysis. Part I. Developing prediction equations.— Journal of Quality Technology, 1973, **5**, April, p. 67—79).)

Соответствующие алгоритмы опубликованы в работах: Clayton D. G. Gram—Schmidt orthogonalization.— Applied Statistics, 1971, **20**, p. 335—338 (Fortran); Farebrother R. W. Gram—Schmidt regression. — Applied Statistics, 1974, **23**, p. 470—476 (Algol 60).

5.8. РЕГРЕССИОННЫЙ АНАЛИЗ УСРЕДНЕННЫХ ДАННЫХ

Пусть мы имеем k наборов параллельных наблюдений $\{Y_{iu}, u = 1, 2, \ldots, n_i\}$, $i = 1, 2, \ldots, k$, но поскольку данные свернуты, результаты отдельных параллельных опытов не приведены, а известны только k средних $\overline{Y}_i$ и k выборочных оценок дисперсий (σ^2), равных

$$s_i^2 = \sum_{u=1}^{n_i} (Y_{iu} - \overline{Y}_i)^2/(n_i - 1).$$

---

[21] О процессе ортогонализации, использующем процедуру Грама—Шмидта, еще см.: Уилкинсон Дж Х. Алгебраическая проблема собственных значений. Пер. с англ.— М.: Наука, 1970.— 564 с. (особо с. 224—226); Райс Дж. Матричные вычисления и математическое обеспечение/Пер. с англ. Под ред. В. В. Воеводина.— М.: Мир, 1984.— 264 с. (особо с. 166—169); Maindonald J. H. Statistical computation.— New York: J. Wiley, 1984.— 370 p. (особо p. 129—145). Если чебышевские ортогональные полиномы первоначально предназначались для однофакторных моделей, то процедуры ортогонализации сделали их пригодными для построения многомерных систем взаимно ортогональных векторов, что весьма важно в регрессии.— *Примеч. пер.*

Как же нам поступить? *Для целей получения регрессионных коэффициентов* мы можем просто действовать так, как если бы $Y_{iu} = \overline{Y}_i$ т. е. как будто каждый набор параллельных наблюдений состоит из равного числа наблюдений, причем все они равны среднему значению. Это станет сразу ясно, если мы представим себе, что случится с i-м набором параллельных в произведении матриц $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. Для части $\mathbf{X}'\mathbf{Y}$ мы имеем

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \cdots & 1 & 1 & \cdots & 1 & \cdots \\ \cdots & a & a & \cdots & a & \cdots \\ \cdots & b & b & \cdots & b & \cdots \\ \cdots & & \cdots & \cdots & \cdots & \cdots \\ \cdots & z & z & \cdots & z & \cdots \end{bmatrix} \times$$

$$\times \begin{bmatrix} \cdots \\ \cdots \\ \cdots \\ Y_{i1} \\ Y_{i2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{in_i} \\ \cdot \\ \cdot \\ \cdot \\ \cdots \end{bmatrix}.$$

Все значения a, b и т. п. равны между собой, поскольку представляют параллельные опыты. Значит, все вклады элементов i-го набора параллельных наблюдений в произведение $\mathbf{X}'\mathbf{Y}$ будут иметь приблизительно такой вид:

$$\begin{aligned} aY_{i1} + aY_{i2} + \cdots + aY_{in_i} &= a \sum_{u=1}^{n_i} Y_{iu}, \\ &= an_i\overline{Y}_i, \\ &= a\overline{Y}_i + a\overline{Y}_i + \cdots + \\ &+ a\overline{Y}_i, \end{aligned}$$

так что замена всех Y_{iu} на $\overline{Y}_i$ не сможет повлиять на вычисление вектора $\mathbf{b}$.

Правда, использование $\overline{Y}_i$ вместо индивидуальных значений Y_{iu} приведет к ошибочной общей скорректированной сумме квадратов. Например, если $\overline{Y}$ — общее среднее всех наблюдений, то вклад (ошибочный) скорректированной суммы квадратов от i-го набора наблюдений должен был бы получиться таким:

$$\sum_{u=1}^{n_i} (\overline{Y}_i - \overline{Y})^2 = n_i (\overline{Y}_i - \overline{Y})^2,$$

в то время как на самом деле должно быть

$$\sum_{u=1}^{n_i} (Y_{iu} - \overline{Y})^2.$$

Мы могли бы, конечно, легко показать, что

$$\sum_{u=1}^{n_i} (Y_{iu} - \overline{Y})^2 = \sum_{u=1}^{n_i} [(Y_{iu} - \overline{Y}_i) + (\overline{Y}_i - \overline{Y})]^2 =$$

$$= \sum_{u=1}^{n_i} (Y_{iu} - \overline{Y}_i)^2 + n_i (\overline{Y}_i - \overline{Y})^2,$$

так как член, содержащий парное произведение, сокращается при суммировании. Значит, для получения *правильного* вклада в скорректированную сумму квадратов от i-го набора параллельных опытов нам надо добавить к $n_i (\overline{Y}_i - \overline{Y})^2$ еще величину

$$\sum_{u=1}^{n_i} (Y_{iu} - \overline{Y}_i)^2 = (n_i - 1)\, s_i^2.$$

(В некоторых случаях приводится только величина s_i, тогда важно помнить, что ее надо сначала возвести в квадрат, а лишь затем умножить на $n_i - 1$.)

Величины $(n_i - 1)\, s^2$, $i = 1, 2, \ldots, k$, играют еще и иную роль. Они образуют k вкладов с $(n_i - 1)$ степенями свободы, $i = 1, 2, \ldots, k$ соответственно, которые должны при объединении дать сумму квадратов «чистой» ошибки.

Таким образом, мы можем все-таки выполнить основные регрессионные вычисления, если только известны выборочные средние и оценки дисперсий для всех наборов параллельных опытов. И нет необходимости знать индивидуальные значения наблюдений [22].

(Примечание. Приведенные выше рассуждения должны помочь при решении упражнения 14.)

---

[22] Основные результаты относительно обработки усредненных данных при различных вариантах дублирования опытов можно найти, например, в статье: Горский В. Г., Адлер Ю. П. О методологии регрессионного и дисперсионного анализа при планировании эксперимента с неравномерным дублированием опытов.— Заводская лаборатория, 1971, 37, № 3, с. 319—325.— *Примеч. пер.*

Упражнения

1. Через равные промежутки времени (с равным шагом) получены следующие данные наблюдений:

Номер опыта	Отклик, Y	Номер опыта	Отклик, Y
1	1	6	11
2	4	7	11,5
3	6	8	13
4	7	9	13,5
5	9,5		

Предположите, что один временно́й тренд распространяется на первые четыре наблюдения, а другой временно́й треид распространяется на остальные пять наблюдений.

1) Используя фиктивные переменные, как в примере 7 из параграфа 5.4, определите наклоны двух линий. Интерпретируйте β-коэффициенты.

2) Какова наилучшая оценка для β_3? Что отражает этот коэффициент? Что вы можете сказать относительно точки пересечения двух линий?

3) Определите, будет ли подобранная модель статистически значимой?

2. Используя обобщенную процедуру преобразования, приведенную в параграфе 5.7 для получения матрицы с ортогональными столбцами, покажите, что для приведенных ниже данных процесс подбора модели методом наименьших квадратов в виде $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ приводит к особенной матрице $\mathbf{X'X}$.

X_1	X_2	X_3	Y
1	—2	4	81
2	—7	11	88
4	3	5	94
7	1	13	95
8	—1	17	123

3. В некотором процессе было получено восемнадцать наблюдений над четырьмя предикторами и одним откликом. Предполагается, что имеет смысл модель

$$Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_{11} X_1^2 + \beta_{12} X_1 X_2 + \beta_{22} X_2^2 + \\ + \beta_{13} X_1 X_3 + \beta_{14} X_1 X_4 + \varepsilon.$$

Данные приведены в следующей таблице:

X_1	X_2	X_3	X_4	Y	X_1	X_2	X_3	X_4	Y
20	50	75	15	27	27	55	60	20	24
27	55	60	20	23	40	90	78	32	16
22	62	68	16	18	32	79	71	11	28
27	55	60	20	26	50	84	72	12	31
24	75	72	8	23	40	90	78	32	22
30	62	73	18	27	20	50	75	15	24
32	79	71	11	30	50	84	72	12	31
24	75	72	8	23	30	62	73	18	29
22	62	68	16	22	27	55	60	20	22

1) Исследуйте данные и модель. Можно ли построить предложенную модель по этим данным? Да или нет?

2) Определите оценку дисперсии случайной ошибки σ^2.

4. Для двух факторов на четырех уровнях (закодированных как —3, — 1, 1 и 3) каждый были реализованы шестнадцать опытов, образующих все возможные комбинации X_1 и X_2. Было решено использовать результаты шестнадцати опытов для построения уравнения регрессии, включающего свободный член, а также все возможные члены первого, второго, третьего и четвертого порядка для X_1 и X_2. Данные были введены в машину вместе с программой, которая обычно выдает вектор оценок

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Машина отказалась выдать оценки. Почему?

Экспериментатор, который тем временем исследовал данные, решил на этом этапе игнорировать уровни фактора X_2 и построить модель четвертого порядка для X_1 по *тем же самым* наблюдениям. Машина снова отказалась выдать оценки. Почему?

5. В течение 15 лет, с 1950 по 1964 г., фирма «Жиллетт» публиковала чистый доход на акцию: 0,64; 0,60; 0,56; 0,73; 0,92; 1,04; 1,13; 0,93; 0,99; 1,11; 1,33; 1,52; 1,60; 1,47; 1,33 (часть этих данных приведена в параграфе 5.6 в качестве примера). Постройте полином подходящего порядка по этим данным, используя ортогональные полиномы, и нанесите на график экспериментальные и расчетные значения. Вычислите остатки (наблюденное значение минус предсказанное значение) и исследуйте их с помощью методов из гл. 3.

6. Новорожденные дети еженедельно взвешивались, цифры отбирались в каждом случае как средние веса за три последовательных дня. Двадцать таких весов приведены ниже. Постройте, используя ортогональные полиномы, по этим данным полиномиальную модель той степени, которая оправдана при определенной точности данных, т. е. проверьте значимость линейных, квадратичных и других членов.

Номер недели	1	2	3	4	5	6	7	8
Вес	141	144	148	150	158	161	166	170
Номер недели	9	10	11	12	13	14	15	16
Вес	175	181	189	194	196	206	218	229
Номер недели			17	18	19	20		
Вес			234	242	247	257		

(Источник. Дипломная работа, Кембридж, 1950. Публикуется с разрешения Cambridge University Press.)

7. В процессе производства куски мыла метятся для их разбраковки по внешнему виду. Эти отметки делаются по шкале 1—10, и чем выше отметка, тем лучше. Различия между характеристиками операторов и скоростью конвейера, возможно, существенно влияют на качество изделий. В этой задаче были собраны следующие данные:

Оператор	Скорость	Внешний вид (сумма по 30 кускам)	Оператор	Скорость	Внешний вид (сумма по 30 кускам)
1	150	255	2	200	231
1	175	246	3	150	265
1	200	249	3	175	247
2	150	260	3	200	256
2	175	223			

1) Используя фиктивные факторы, постройте по этим данным модель множественной регрессии.

2) Приняв $\alpha = 0{,}05$, определите, важно ли для качества изделий различие в операторах. С помощью регрессионной модели покажите, что среднее значение качества для оператора № 1 равно 250, для оператора № 2—238 и для оператора № 3—256.

3) Влияет ли скорость на качество? (Примите $\alpha = 0{,}05$).

4) Пригодна ли ваша модель для предсказания качества куска мыла?

8. Известно, что конечный продукт теряет в весе с течением времени. Следующие данные демонстрируют это снижение веса.

Время после производства, t	Разница в весе (в 1/16 унции), Y	Время после производства, t	Разница в весе (в 1/16 унции), Y
0	0,21	2,5	—5,37
0,5	—1,46	3,0	—6,03
1,0	—3,04	3,5	—7,21
1,5	—3,21	4,0	—7,46
2,0	—5,04	4,5	—7,96

1) Используя ортогональные полиномы, постройте модель второго порядка, которая представляет потерю веса как функцию времени, прошедшего после приготовления продукта.

2) Проанализируйте остатки от этой модели и дайте заключение относительно ее адекватности.

9. Точка (или температура) помутнения жидкости служит мерой кристаллизации исходного материала и ее можно измерить с помощью коэффициента преломления (показателя рефракции) [23]. Предполагается, что процент фракции 1—8 в исходном сырье прекрасно предсказывает точку помутнения с помощью модели второго порядка:

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon.$$

Были собраны следующие данные:

%, 1—8, X	Точка помутнения, Y	%, 1—8, X	Точка помутнения, Y	%, 1—8, X	Точка помутнения, Y	%, 1—8, X	Точка помутнения, Y
0	22,1	2	26,1	5	28,9	0	22,8
1	24,5	4	28,5	6	30,0	3	27,3
2	26,0	6	30,3	7	30,4	6	29,8
3	26,8	8	31,5	8	31,4	9	31,8
4	28,2	10	33,1	0	21,9		

[23] Речь идет об экспрессном методе определения одного из показателей качества сырья — степени его кристаллизации. После растворения об этом однозначно свидетельствует показатель преломления, но если бы о степени кристаллизации знать заранее, можно было бы лучше управлять технологическим процессом. Поэтому была высказана гипотеза о том, что доля фракции 1—8 (это, видимо, номер сита, используемого в ситовом анализе; какая именно фракция имеется в виду — сказать трудно, возможно, что мелкая) тоже несет в себе нужную информацию. Если бы это оказалось так, то показатель преломления можно было бы предсказывать с требуемой точностью, пользуясь регрессионной моделью, которая играла бы в этом случае роль градуировочного графика.— *Примеч. пер.*

1) Определите наилучшую модель второго порядка.
2) Приняв $\alpha = 0{,}05$, проверьте полиую регрессию.
3) Проверьте адекватность.
4) Будет ли достаточной модель первого порядка $Y = \beta_0 + \beta_1 X + \varepsilon$? Используйте остатки от этой простой модели как основу для ваших заключений.
5) Объясните применение подобранной модели второго порядка как предсказывающего уравнения.

10. Рассмотрите следующие показательные данные:

Год, X	Скорость, мили в час, Y	Средства достижения скорости
1830	30	Железная дорога
1905	130	» »
1930	400	Самолет
1947	760	»
1952	1500	»
1969	25000	Космический корабль

* Одна миля равна 1609 м. — *Примеч. пер.*

1) Нанесите значения X и Y на график. Как вы считаете, информативен этот график или нет? Почему?
2) Преобразуйте данные по формуле $Z = \log Y$ и снова нанесите их на график. Стал ли график для X и Z более удобным, чем предыдущий, или нет? Почему?
3) Можете ли вы найти достаточно простое преобразование $u = f(Y)$, которое привело бы к (более или менее) прямой линии на графике в координатах X—U?
4) Если вы еще ие сделали этого в пункте 3, то постройте график X—V, где $V = \log \log Y$. Постройте по этим точкам методом наименьших квадратов прямую $V = \beta_0 + \beta_1 X + \varepsilon$. Нанесите полученную таким образом прямую на ваш график. Найдите остатки и обсудите их.
5) Постройте для даиных из пункта 4 подходящую таблицу дисперсионного анализа, проверьте всю регрессию и найдите R^2. Прокомментируйте свои результаты.
6) Используя полученное в пункте 4 уравнение прямой, предскажите, когда же человек достигнет, наконец, скорости света (186 000 миль в с, обратите внимание — именно в *секунду!*).
7) Подумайте, разумно ли ваше предсказание? Каких предположений оно требует? Считаете ли вы свое предсказание реалистическим или нереалистическим? Изложите свои соображения точно, но кратко.

11. По приведенным ниже данным постройте модель

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon \qquad (1)$$

и найдите суммы SS $(b_1 | b_0)$, SS $(b_2 | b_0, b_1)$, SS $(b_3 | b_0, b_1, b_2)$ и SS $(b_2, b_3 | b_0, b_1)$. Затем проверьте гипотезу $H_0 : \beta_2 = \beta_3$. Какую модель вы должны были бы использовать для представления этих данных? Найдите остатки e_i, соответствующие полученной модели, и убедитесь, что если пренебречь ошибками округления, то справедливо тождество $\Sigma e_i = \Sigma e_i \hat{Y}_i = 0$.

X:	−5	−3	−1	1	3	5
Y:	13	4	3	4	10	22

Найдите также смещения оценок параметров *выбранной* вами модели относительно коэффициентов модели (1), приведенной выше, от которой вы отказались.

12. Отклик измерялся в следующие даты: в ноябре 17, 19, 20, 22, 26, 29, 30, в декабре 1, 2, 3 и 5. Предполагалось, что он зависит от двух факторов X_1 и X_2, значения которых также записывались, но здесь не приведены. И кроме того, в данных предполагался квадратичный временной тренд. Как вам следует поступить, чтобы учесть это обстоятельство? (Если вы имеете в виду какие-либо новые переменные, то укажите их фактические уровни.)

13. (Источник. Работа Дж. Дерринджера об изменчивости сдвигового коэффициента (динамической) вязкости систем SBR-наполнитель-пластификатор [24] (Derringer G. C. Variable shear rate viscosity of SBR-filler-plasticizer systems.— Rubber Chemistry and Technology, September 1974, 47, p. 825—836.)

Подберите модель [25]

$$Y = (\alpha_0 + \alpha_1 Z + \alpha_2 Z^2) + (\beta_0 + \beta_1 Z + \beta_{11} Z^2) X_1 + (\gamma_0 + \gamma_1 Z + \gamma_{11} Z^2) X_2 + \varepsilon,$$

где $Z = ln\ (X_3 + 1)$. Воспользуйтесь приведенными ниже данными и проведите полный анализ. Обратите внимание, что есть шесть параллельных опытов.

X_1	X_2	X_3	Y
47,1	33,9	7,5	11,97
72,9	33,9	750	8,63
47,1	8,1	750	8,80
60	21	75	10,73
60	21	75	10,69
72,9	8,1	7,5	13,12
47,1	8,1	7,5	12,58
72,9	33,9	7,5	12,24
60	21	75	10,64
72,9	8,1	750	9,09

X_1	X_2	X_3	Y
47,1	33,9	750	8,46
60	21	75	10,65
60	21	3000	7,60
60	21	3	13,06
39	21	75	10,51
60	0	75	11,22
60	21	75	10,67
60	42	75	10,24
81	21	75	10,74
60	21	75	10,69

14. Приведенные ниже данные извлечены из работы Дж. Бханота и Шри Патела, посвященной ретроспективному исследованию за пятилетие (1961—1962 по 1965—1966) здоровья выпускников университета Махараджа Сайяджирао (см.: Bhanot J. V., Patel Shri R. G. Health and Ailments of M. S. U. Alumni: A Study in Retrospect Over a Five Year Period (1961—1962 to 1965—1966)). Она была опубликована департаментом статистики в издательстве этого университета в г. Барóда [26] в октябре 1968 г. (Department of Statistics at the

[24] См. примечание 5 на с. 284. SBR — здесь, как и прежде, бутадиен-стирольный сополимер, а сдвиговый коэффициент динамической вязкости (внутреннего трения) — «вязкость по Муни».— *Примеч. пер.*

[25] Модели такого типа, когда коэффициенты полинома относительно некоторого набора предикторов сами рассматриваются как функции какого-то фактора (часто времени), характерны для исследований кинетики различных процессов с помощью экспериментальных планов, предложенных Дж. Боксом и У. Хантером. См., например: Box G. E. P., Hunter W. G. A useful method for model building.— Technometrics, 1962, 4, № 3, p. 301—312; Налимов В. В., Чернова Н. А. Цит. соч., с. 203—212; Адлер Ю. П., Рохваргер А. Е. Изучение кинетики термической поризации перлита с помощью планирования эксперимента.— Заводская лаборатория, 1966, 32, № 1, с. 59—65. См. также параграф 10.8.— *Примеч. пер.*

[26] Существует распространенное мнение, что вес человека существенно зависит от диеты. Ясно, конечно, что при этом надо считаться еще с полом и возрастом. В Индии, где широко распространено вегетарианство, представляется благоприятная возможность количественного исследования такой зависимости на большом числе студентов. Эта возможность и была реализована

Maharaja Sayajirao University of Baroda). В этих данных представлены средние арифметические веса в фунтах и их стандартные отклонения для различных категорий студентов в 1965—1966 гг. Здесь воспроизведена лишь часть этих данных, причем нет никакой увереиности в том, что они полно представляют всю совокупность. Поскольку данные по индивидуальным точкам не приводятся, обрабатывайте каждый набор параллельных опытов так, как если бы он представлял собой иабор того же числа одииаковых опытов, чего вполне достаточно для целей регрессионного анализа. Причем воспользуйтесь соответствующими значениями стаидартных отклонений, возведя их предварительно в квадрат и умножив на соответствующее число степеней свободы (а именно на число студентов в группе минус 1), чтобы получить разумные вклады в сумму квадратов «чистой» ошибки. Будьте внимательны, поскольку вам придется делать соответствующие поправки при использовании действительных значений сумм квадратов где-нибудь в другом месте.

Постройте по данным модель вида $Y = \beta_0 + \beta_1$ (возраст) $+ \varepsilon$, где Y — вес студента. Добавьте подходящую фиктивную переменную, чтобы учесть различные категории студентов. Проведите подробный анализ. Каковы ваши выводы?

Если вы чувствуете, что подобрали неудачную модель, то какую модель вы могли бы предложить взамен и почему? Подберите любую альтернативную модель, какую вы считаете нужной, проведите ее анализ и сделайте выводы.

Возраст, лет	Число студентов	Средний вес, фунт	Оценка стандартного отклонения	Категория *	Возраст, лет	Число студентов	Средний вес, фунт	Оценка стандартного отклонения	Категория *
16	19	103,82	12,70	*A*	23	7	113,21	7,28	*B*
17	19	105,39	12,50	*A*	16	18	100,83	16,75	*C*
18	16	107,50	15,30	*A*	17	33	99,32	14,50	*C*
19	8	103,12	7,68	*A*	18	24	100,83	19,67	*C*
20	6	105,83	16,75	*A*	19	18	96,39	13,29	*C*
21	6	107,50	5,77	*A*	20	6	101,67	18,12	*C*
22	1	102,50	—	*A*	21	4	100,00	15,21	*C*
23	1	117,50	—	*A*	22	0	—	—	*C*
16	12	99,17	9,20	*B*	23	1	112,50	—	*C*
17	29	107,67	19,55	*B*					
18	28	103,57	13,05	*B*					
19	32	112,19	17,60	*B*	Всего	322			
20	22	110,00	12,04	*B*					
21	8	113,13	14,87	*B*					
22	4	112,00	11,18	*B*					

* *A* — студенты мужского пола — индусы-вегетарианцы; *B* — то же, не индусы, со смешанным питанием; *C* — студентки-индуски со смешанным питанием

15. (Источник. Eppright E. S., Fox H. M., Fryer B. A., Lamkin G. H., Vivian V. M., Fuller E. S. Nutrition of infants and preschool children in the north central region of the United States of America.— World Review of Nutrition and Dietetics, 1972, **14**, p. 269—332.) Ниже при-

в университете города Барóда (бывшей столице одноименного княжества на северо-востоке Декана, которое входит теперь административно в штат Бомбей) в рамках широкой программы изучения здоровья и болезней поколений студентов. Трудно ожидать, что представленный здесь фрагмент может быть интересным вне контекста всего образа жизни обследованных студеитов. Слова «eggitarian» в расшифровке категории *A* не удалось обнаружить в словарях, поэтому приходится считать, что здесь допущена опечатка и надо читать: «Vegitarian»— «вегетарианец».— *Примеч. пер.*

водятся семьдесят два наблюдения над откликом Y = отношение рост/вес для мальчиков (Р/В) в зависимости от расположенного с равным шагом предиктора X = возраст в месяцах [27]. Предполагается, что эти данные распадаются на две группы: (1) первые семь наблюдений, (2) остальные шестьдесят пять наблюдений. Еще предполагается, что такое разбиение на группы можно объяснить наличием двух линейных временных трендов. Воспользовавшись методами из параграфа 5.4, найдите угловые коэффициенты обоих трендов и точку их пересечения. Нанесите на график экспериментальные точки и полученные прямые. Затем постройте таблицу дисперсионного анализа и выполните анализ остатков.

Р/В	Возраст	Р/В	Возраст	Р/В	Возраст	Р/В	Возраст	Р/В	Возраст	Р/В	Возраст
0,46	0,5	0,75	12,5	0,88	24,5	0,87	36,5	0,92	48,5	0,97	60,5
0,47	1,5	0,80	13,5	0,81	25,5	0,87	37,5	0,96	49,5	0,94	61,5
0,56	2,5	0,78	14,5	0,83	26,5	0,85	38,5	0,92	50,5	0,96	62,5
0,61	3,5	0,82	15,5	0,82	27,5	0,90	39,5	0,91	51,5	1,03	63,5
0,61	4,5	0,77	16,5	0,82	28,5	0,87	40,5	0,95	52,5	0,99	64,5
0,67	5,5	0,80	17,5	0,86	29,5	0,91	41,5	0,93	53,5	1,01	65,5
0,68	6,5	0,81	18,5	0,82	30,5	0,90	42,5	0,93	54,5	0,99	66,5
0,78	7,5	0,78	19,5	0,85	31,5	0,93	43,5	0,98	55,5	0,99	67,5
0,69	8,5	0,87	20,5	0,88	32,5	0,89	44,5	0,95	56,5	0,97	68,5
0,74	9,5	0,80	21,5	0,86	33,5	0,89	45,5	0,97	57,5	1,01	69,5
0,77	10,5	0,83	22,5	0,91	34,5	0,92	46,5	0,97	58,5	0,99	70,5
0,78	11,5	0,81	23,5	0,87	35,5	0,89	47,5	0,96	59,5	1,04	71,5

16. (Сокращенный вариант предыдущего упражнения.) Воспользуйтесь только первыми тридцатью двумя наблюдениями. Предполагается, что эти тридцать два наблюдения распадаются на две группы: (1) первые семь наблюдений, (2) остальные двадцать пять наблюдений. Еще предполагается, что такое разбиение на группы можно объяснить наличием двух линейных временных трендов. Воспользовавшись методами из параграфа 5.4, найдите угловые коэффициенты обоих трендов и точку их пересечения. Нанесите на график экспериментальные точки и полученные прямые. Затем постройте таблицу дисперсионного анализа и выполните анализ остатков.

17. С помощью метода ортогонализации из параграфа 5.7 проверьте, нет ли у матрицы, приведенной ниже, зависимых столбцов. Каковы ваши выводы?

$$\begin{bmatrix} 1 & -4 & 1 & 3 \\ 1 & 3 & 2 & -5 \\ 1 & 1 & 3 & -4 \\ 1 & 4 & 4 & -8 \\ 1 & -3 & 5 & -2 \\ 1 & -1 & 6 & -5 \end{bmatrix}.$$

18. Девять вариантов красителя с равномерно расположенными уровнями использовались для окраски внешне одинаковых кусков ткани. Экспертная

[27] После работ Ф. Гальтона, которые мы упоминали в примечаниях к гл. 1, в широком масштабе начались антропометрические исследования. Одно из их направлений — выбор наиболее информативных показателей физического развития людей, а один из таких показателей — отношение рост/вес.— *Примеч. пер.*

оценка цвета в порядке возрастания уровня красителя получилась такой:

$$Y = 11,\ 12,\ 10,\ 12,\ 11,\ 14,\ 16,\ 22,\ 28.$$

Найдите с помощью ортогональных полиномов подходящую полиномиальную зависимость между Y и уровнем красителя.

19. (Источник[28]. Derringer G. C. An empirical model for viscosity of filled and plasticized elastomer compounds.—Journal of Applied Polymer science, 1974, **18**, p. 1083—1101.) В таблице к этому упражнению приведены два набора значений откликов (обозначенных *** и ****). Для каждого из этих наборов найдите преобразование вида $W = (Y^{\lambda}-1)/\lambda$ при $\lambda \neq 0$ и $W = \ln Y$ при $\lambda = 0$, чтобы можно было построить методом наименьших квадратов модель вида $W = \beta_0 + \beta_1 f + \beta_2 p + \varepsilon$, где f — уровень наполнителя, а p — уровень нафтенового масла. Проделайте полностью обычный регрессионный анализ для наилучшего, по вашему мнению, значения λ, включая и анализ остатков. (Заметьте, что кодирование, которое обсуждалось в параграфе 5.3, было бы чрезвычайно полезно для второго набора данных.)

Таблица к упражнению 19. **Вязкость по Муни MS_4 при 100° С в зависимости от уровней наполнителя и пластификатора в SBR-1500***

Нафтеновое масло **, *phr*, p	Наполнитель, *phr*, f					
	0	12	24	36	48	60
0	26***	28	30	32	34	37
	25****	30	35	40	50	60
10	18	19	20	21	24	24
	18	21	24	28	33	41
20	12	14	14	16	17	17
	13	15	17	20	24	29
30	—	12	12	13	14	14
	11	14	15	17	18	25

* «Филлипс Петролеум Ко.».
** «Секолайт процесс Ойл», «Сан Ойл Ко.».
*** № 990, «Кейбот Корп.».
**** Силикагель В, «Пи-Пи-Джи Индастриз».

20. (Источник.[29] Тот же, что и в упражнении 19.)

[28] Этот пример — продолжение работы, уже упомиинавшейся в данной главе. См. примечание 5 на с. 284. Здесь сравниваются два наполнителя — сажа марки № 990, производимая «Кейбот Корпорейшн», и силикагель марки В (2-й сорт), производимый «Пи-Пи-Джи Индастриз». — *Примеч. пер.*

[29] Продолжение предыдущего упражнения. Снова см. примечание 5 на с. 284. Напомним, что ML_4 — измерение вязкости по Муни с длинным цилиндром в течение четырех минут, а MS_4 — то же, но с коротким цилиндром. Новые ингредиенты: Silica A — силикагель марки А (1-й сорт), № 330 — обозначение марки сажи, кумарон-инденовая смола, выработанная с помощью процесса, запатентованного «Оллайд Кемикал Корпорейшн» (это смола, получаемая полимеризацией ненасыщенных соединений, содержащихся в продуктах коксования каменных углей и пиролиза нефти; в данном случае используется как пластификатор).— *Примеч. пер.*

Таблица к упражнению 20. **Вязкость по Муки Y_1 (ML_4) и Y_2 (MS_4) при 100 °С для различных комбинаций кодированных факторов* x_1, x_2, x_3 и x_4**

x_1	x_2	x_3	x_4	Y_1	Y_2	x_1	x_2	x_3	x_4	Y_1	Y_2
—3	—3	—3	—3	51	29	1	—3	1	1	50	27
—1	—1	—1	—1	61	34	—1	1	—3	1	88	49
—1	—1	—1	—1	64	35	1	—1	—3	—1	124	68
—3	—1	—1	1	36	20	—1	—1	1	—3	54	30
—3	1	1	—1	39	21	1	1	—1	—3	133	74
—1	—3	—1	—1	55	30						

* $x_i = (X_i - 22\frac{1}{2})7\frac{1}{2}$, $i=1,2$, $x_j = (X_j - 15)/5$, $j=3,4$, где X_1 — уровень силикагеля марки А (1-й сорт); X_2 — уровень сажи марки № 330, «Кейбот Корп.»; X_3 — нафтеновое масло; X_4 — кумарон-инденовая смола, Camar MH 2 $\frac{1}{2}$, «Оллайд Кемикал Корп».

Выберите для каждого отклика в отдельности наилучшее значение λ в преобразовании $W = (Y^\lambda - 1)/\lambda$ при $\lambda \neq 0$, $W = \ln Y$ при $\lambda = 0$, чтобы можно было подобрать по имеющимся данным методом наименьших квадратов хорошую модель в виде $W = \beta_0 + \Sigma\beta_i x_i + \varepsilon$. В каждом случае после выбора наилучшего параметра преобразования и его оценки $\hat{\lambda}$ проведите обычный полный регрессионный анализ, включая и анализ остатков.

21. (Источник. Draper N.R. Third order rotatable design in three factors: analysis.—Technometrics, 1962,4, p. 219—234.) Приведенные ниже данные специально построены для иллюстрации использования последовательного плана третьего порядка для генерирования полиномиальной модели третьей степени. Найдите наилучшее значение λ в преобразовании $W = (Y^\lambda - 1)$ при $\lambda \neq 0$ и $W = \ln Y$ при $\lambda = 0$, которое позволило бы оценить десять параметров полной модели второго порядка для трех факторов x_1, x_2 и x_3. После оценивания выбранного параметра $\hat{\lambda}$ проведите полный регрессионный анализ и сделайте выводы.

x_1	x_2	x_3	Y	x_1	x_2	x_3	Y	
—1	—1	—1	34,727	$2^1/_2$	0	0	33,414	34,453
1	—1	—1	38,917	$-2^1/_2$	0	0	38,540	39,201
—1	1	—1	44,907	0	$2^1/_2$	0	40,393	38,335
1	1	—1	24,641	0	$-2^1/_2$	0	40,687	40,092
—1	—1	1	24,658	0	0	$2^1/_2$	23,869	25,823
1	—1	1	45,636	0	0	$-2^1/_2$	33,727	33,068
—1	1	1	33,702	0	0	0	43,832	44,502
1	1	1	5,374	0	0	0	42,165	41,187

22. (Источник [30]. Ryan B. F., Wishart E. R., Show D. E. Commonwealth Scientific and Industrial Research Organisation (C. S. I. R. O.), Australia.) Соответствующая журнальная ссылка: The growth rates and densities of ice crystals between — 3 °C and — 21 °C.— Journal of the Atmospheric Sciences, 1976, 33, p. 842—850.) Кристаллы льда помещались в камеру, внутри

[30] Это упражнение — продолжение упражнения 18 из гл. 1, только там изучалась длина кристаллов льда, а здесь — их масса. См. наше примечание 41 в гл. 1.— *Примеч. пер.*

[31] Нанограмм — миллиардная доля грамма = $1 \cdot 10^{-9}$ г.— *Примеч. пер.*

которой поддерживалась постоянная температура (— 5 °С) и постоянный уровень насыщения воздуха водяным паром. Наблюдался рост кристаллов во времени. Представленные здесь сорок три набора измерений — это масса кристаллов (M) в нанограммах [31] за время от $T = 50$ с до 180 с от начала кристаллизации. Каждое измерение получено в полном отдельном эксперименте; эксперименты продолжались несколько дней и были рандомизированы во времени. (Фактический порядок, в котором они выполнялись, здесь не показан). Хотелось бы связать отклик M с предиктором T какой-нибудь простой зависимостью. (Вполне возможно, что подошла бы модель $E(M) = \alpha T^{\beta}$.) Проделайте следующие операции.

1) Подберите модель $W = \gamma + \beta \ln T$, где $W = (M^{\lambda} - 1)/\lambda$ при $\lambda \neq 0$ и $W = \ln M$ при $\lambda = 0$ из подходящего множества значений λ, а затем отыщите самое лучшее преобразование для M, пользуясь методами, описаннымн в параграфе 5.3.

2) Воспользовавшись отобранными значениямн λ подробно проделайте обычные вычисления метода наименьших квадратов для ваших данных и сделайте соответствующие выводы. В частности, не указывает ли график остатков на то, что структура дисперсии ошибок не остается постоянной, как предполагалось?

T	M	T	M
50	11,5	125	47,7
60	8,2, 11,5	130	92,0, 87,2
70	14,1, 17,2	135	58,0, 47,7
80	33,5, 28,8	140	73,2, 58,0
90	15,6, 24,4, 33,5	145	47,7
95	38,8	150	118,9, 58,0
100	47,7, 58,0, 36,1	155	143,9, 87,2
105	47,7, 65,5	160	143,9, 73,2, 73,7
110	58,0, 47,7, 33,5	165	97,0
115	69,5, 69,5, 47,7	170	112,3
120	87,2, 51,0, 33,5	180	113,2

23. (Источник. Wayne Nelson. A short life test for comparing a sample with previous accelerated test results.— Technometrics, 1972, **14**, p. 175—185.) Данные в таблице представляют собой результаты ускоренных испытаний двадцати пяти образцов некоторого типа солнечных батарей [32]. T — температура в °F, а Y — долговечность (время жизни) в часах при условии, что температура фиксируется для каждого образца в отдельности. При каждой температуре испытывалось по шесть образцов. Нанесите данные на график и изучите его. Постройте модель

$$\log_{10} Y = \beta_0 + \beta_1 \{1000/(T + 460)\} + \varepsilon$$

и проделайте весь обычный анализ.

(Примечание. $T + 460$ — это абсолютная температура в °F.)

T	Y					
1520	1953	2135	2471	4727	6143	6314
1620	1190	1286	1550	2125	2557	2845
1660	651	837	848	1038	1361	1543
1708	511	651	651	652	688	729

[32] Солнечные батареи представляют собой полупроводниковые преобразователи, превращающие тепловую энергию солнечных лучей в электрическую. Нашли широкое применение иа космических кораблях и искусственных спутниках, где их надежиость и долговечность имеют огромное значение. В последнее время начали применяться в вычислительной технике.— *Примеч. пер.*

24. Хозяйка дома записывала ежемесячно число порций мороженого, которое она готовила для своей семьи в течение года. Ее записи воспроизведены ниже. Она удивилась бы, если бы полученную картину удалось объяснить с помощью модели полиномиальной регрессии. Она просит вас проделать такой анализ с использованием ортогональных полиномов шестого и более низких порядков. Проделайте это, проведите полный стандартный анализ, как указано ниже, и сделайте выводы, вытекающие из вашего решения.

Месяц	Число порций мороженого	Месяц	Число порций мороженого	Месяц	Число порций мороженого
Январь	2	Май	103	Сентябрь	14
Февраль	72	Июнь	82	Октябрь	23
Март	106	Июль	54	Ноябрь	64
Апрель	116	Август	30	Декабрь	129

Основные моменты задания

Подбор модели.

Построение таблицы дисперсионного анализа.

Принятие решения о порядке модели, который целесообразно использовать, и сжатие таблицы дисперсионного анализа и т. д.

Вычисление R^2, F-критерия и т. д.

Вычисление матрицы дисперсий-ковариаций.

Вычисление остатков *для выбранной модели*.

Оценка матрицы дисперсий-ковариаций остатков, только диагональные элементы и построение графиков остатков.

Исследования остатков, включая критерий Дарбина—Уотсона и временные последовательности опытов.

Результаты, как полагается.

Выводы.

25. Ниже приведена прибыль на акцию в долларах Y для некоторых американских компаний за несколько последовательных лет [33]. Для каждой из приведенных компаний постройте полиномиальную модель $Y_i = f$ (год) $+ \varepsilon$ подходящего порядка, чтобы она могла удовлетворительно объяснять данные. (Для получения дополнительной информации об этих компаниях обратитесь, например, к текущим номерам издания Moody's Handbook of Common Stocks, которое можно найти во многих библиотеках.)

---

[33] Выбор наиболее выгодного способа инвестиций представляет для западного читателя практический интерес. А тот факт, что регрессионный анализ может в этом помочь — лучшая реклама метода. Представленные здесь компании принадлежат к числу крупнейших. Так, например, «Доу Кемикал» (Dow Chemical) имеет активы 1271 млн. дол., продажи — 1077,4 млн. дол., число работающих — 33 тыс. чел. Производит химические товары: пластмассы, смолы, удобрения, химикаты, промышленные газы, редкие металлы. «Дженерал Моторс» (General Motors) соответственно имеет активы 11245, продажи — 17000, работающих—661 тыс., работа идет на 124 заводах в разных странах мира. Производит легковые и грузовые автомобили и запчасти к ним, двигатели, электрооборудование, строительные машины, военную технику, тепловозы, подшипники. Все данные относятся к 1965 г. Moody's Handbook of Common Stocks.— Известное ежегодное справочное издание, содержащее разнообразную информацию о деятельности многих компаний.— *Примеч. пер.*

Примените метод ортогональных полиномов.

Год	«Доу Кемикал»	«Дженерал Моторс»	«Пфицер»	«Вулворт»	Год	«Доу Кемикал»	«Дженерал Моторс»	«Пфицер»	«Вулворт»
1964	0,52	6,04	0,76	1,99	1974	2,89	3,27	1,93	2,14
1965	0,60	7,41	0,90	2,41	1975	3,41	4,32	2,00	3,33
1966	0,68	6,24	1,02	2,34	1976	3,30	10,08	2,28	3,62
1967	0,73	5,65	0,96	2,34	1977	3,01	11,62	2,50	3,03
1968	0,75	6,01	1,03	2,29	1978	3,16	12,24	2,93	4,34
1969	0,79	5,94	1,13	2,32	1979	4,33	10,04	3,26	6,02
1970	0,73	2,08	1,27	2,52	1980	4,42	—2,65	3,48	5,30
1971	0,85	6,72	1,38	2,48	1981	3,00	1,07	3,63	2,64
1972	1,04	7,51	1,50	2,60	1982	1,77	3,09	4,26	2,67
1973	1,49	8,34	1,73	3,15					

Держатели акций любят находить такие компании, которые распределяют свои прибыли в виде некоторого приемлемого *процента на капитал*, измеряемого в процентах (или пропорционально им) годового дохода, найденного *по результатам деятельности за предыдущие годы*. С этой точки зрения иногда имеет смысл изучать не только доходы на акцию, но и логарифмы этих доходов. Если, например, доходы растут с постоянной скоростью, то логарифмы доходов должны будут иметь линейный треид. Возьмите логарифмы по основанию e или по основанию 10 (результаты будут отличаться только постоянным множителем) от данных из этого упражнения и, используя ортогональные полиномы, подберите полином такого порядка, который соответствует результатам. Для каждой из изучаемых компаний сравните результаты для прологарифмированных и для непрологарифмированных данных. Исследуйте остатки, дабы увидеть, не следует ли отбросить предположения об ошибках. Прокомментируйте ваши результаты.

26. Есть надежда, что для приведенных ниже данных разумна функция отклика вида $\eta = \alpha X_1^{\beta} X_2^{\gamma}$, причем не ясно, как в этой ситуации учитывать ошибки эксперимента. Подберите по данным методом наименьших квадратов модель вида $\log Y = \log \alpha + \beta \log X_1 + \gamma \log X_2 + \varepsilon$, исследуйте полученную модель любым подходящим способом и сделайте выводы, которые могли бы пролить свет на сложившуюся ситуацию. (Возьмите логарифмы по основанию 10.)

X_1	X_2	Y	X_1	X_2	Y
10	10	2 040	10	1 000	10 370
100	10	7 350	100	1 000	1 150
1 000	10	12 210	1 000	1 000	23 580
10 000	10	23 580	10 000	1 000	296 120
10	100	18 200	10	10 000	9 040
100	100	10	100	10 000	1 960
1 000	100	2 960	1 000	10 000	96 980
10 000	100	108 040	10 000	10 000	1 004 020

27. (Источник. Power J. F., Alessi J. Tiller development and yield of standard and semidwarf spring wheat varieties as affected by nitrogen fertilizer. —Journal of Agricultural Science, Cambridge, 1978, 90, p. 97—108. Воспроизведено с разрешения издательства Кембриджского университета.)

В таблице к этому упражнению столбец Y представляет собой урожай зерна в кг/га, который получился в результате выращивания двух сортов твердой краснозерной яровой пшеницы, «Уолдрон» (Waldron) и «Киано» (Ciano), при трех различных количествах азота $N = 0, 50, 270$, что соответствует недостаточному, нормальному и избыточному содержанию азота [34]. Побег нумеруется в зависимости от того, развился ли он первым, вторым или третьим (последним), считая от главного стебля, вырастающего из пазухи исходного листа (см. с. 98 в источнике). В последнем столбце приведено содержание азота в миллиграммах на побег от точки ветвления (кущения), X.

Постройте для отклика Y квадратичную модель по X с присоединением дополнительных членов, отражающих в модели номер побега, сорт пшеницы и количество азота. Исследуйте полученные результаты и сделайте выводы.

Таблица к упражнению 27. **Урожай с трех типов побегов от двух сортов пшеницы, обработаниых различным количеством азотных удобрений**

Y	Номер побега	Сорт (У/К)	Количество азота	N на побег, мг (X)	Y	Номер побега	Сорт (У/К)	Количество азота	N на побег, мг (X)
370	1	У	0	4,2	188	2	К	0	2,8
659	1	У	50	7,2	632	2	К	50	6,0
935	1	У	270	9,8	538	2	К	270	7,7
390	1	К	0	3,6	27	3	У	0	2,0
753	1	К	50	7,6	141	3	У	50	2,8
733	1	К	270	10,3	262	3	У	270	3,6
182	2	У	0	3,1	34	3	К	0	2,7
417	2	У	50	5,2	222	3	К	50	3,1
686	2	У	270	7,8	242	3	К	270	4,4

28. Данные об отклике Y и k предикторах $X_1, X_2, \ldots, X_k$ собраны на двух заводах A и B. Хотелось бы построить модель в виде

$$\hat{Y}_Q = b_{0Q} + b_{1Q}X_1 + b_2X_2 + \ldots + b_kX_k,$$

где $Q = A$ или B обозначает тот завод, для которого делается предсказание. Иначе говоря, эффекты переменных $X_2, \ldots, X_k$ на обоих заводах одинаковы, а свободный член и угловой коэффициент переменной X_1 для каждого завода свои.

Покажите, что эту задачу можно решить с помощью дополнительного фиктивного фактора Z, который принимает значение 1 для завода A, и 0 — для B, и что искомая модель имеет вид

$$Y = \beta_0 + \beta_Z Z + \beta_{1Z}X_1Z + \beta_1X_1 + \ldots + \beta_kX_k + \varepsilon,$$

[34] Сравнение урожайности сортов различных сельскохозяйственных культур — исходная и традиционная задача планирования эксперимента. Решающий вклад в развитие этого направления внес выдающийся английский ученый сэр Р. Фишер (см., например: Fisher R. A. The Design of Experiments. 8th ed.— New York: Hafner Publising Company, 1966). В данном примере реализован полный факторный план 2^13^2. В СССР тоже ведутся работы по созданию высокоурожайных сортов яровой пшеницы.— *Примеч. пер.*

откуда следует, что в данном случае

$$b_{0A} = b_0 + b_Z, \qquad b_{1A} = b_1 + b_{1Z},$$
$$b_{0B} = b_0, \qquad b_{1B} = b_1.$$

Такая же идея может использоваться при введении в модель еще эффектов взаимодействия $\beta_{jZ}ZX_j$, по отношению к остальным факторам. Покажите, что если эту идею использовать по отношению ко *всем* X-ам, то фактически мы подберем отдельные модели для данных с заводов A и B.

29. Переработайте пример (см. уравнение (5.3.7а)), заменив W_i на V_i в данных, которые представлены в табл. 6.2 и 5.7. Сравните ваши новые таблицы с теми, что были в тексте, и обсудите проблемы масштабирования, возникающие при использовании W. Вместо (5.3.10) возьмите выражение

$$S(\lambda, \mathbf{V}) = \mathbf{S}(\hat{\lambda}, \mathbf{V}) \exp\{n^{-1}\chi_1^2(1-\alpha)\}.$$

Ответы к упражнениям

1. 1) $b_0 = -\frac{1}{2}$, отрезок, отсекаемый на оси ординат линией № 1.

$b_1 = 2$, наклон линии № 1.
$b_2 = 1$, наклон линии № 2.

2) $b_3 = 0{,}2$ указывает на то, что $\hat{Y}$, предсказанное при пятом наблюдении линией № 2, на $+0{,}2$ единицы выше, чем $\hat{Y}$, предсказанное при пятом наблюдении линией № 1. Следовательно, точка пересечения двух линий находится справа от пятого наблюдения.

3) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий (скорректированный)	8	146,50		
Регрессия	3	145,20	48,40	186,15*
$b_1 \mid b_0$	1	132,22	132,22	508,54*
$b_2 \mid b_0,\ b_1$	1	12,83	12,83	49,35*
$b_3 \mid b_0,\ b_1,\ b_2$	1	0,15	0,15	< 1 не значимо
Остаток	5	1,30	0,26	

2. Заключительная матрица следующая:

$$\mathbf{Z} = \begin{bmatrix} 1 & -\frac{17}{5} & \frac{215}{186} & 0 \\ 1 & -\frac{12}{5} & -\frac{822}{186} & 0 \\ 1 & -\frac{2}{5} & \frac{824}{186} & 0 \\ 1 & \frac{13}{5} & \frac{131}{186} & 0 \\ 1 & \frac{18}{5} & -\frac{348}{186} & 0 \end{bmatrix}$$

3. 1) Модель содержит десять параметров. Рассмотрение данных выявляет восемь различных точек. Следовательно, установлено, что подобрать модель невозможно.

2) $s^2 = 4{,}325$ с 10 степенями свободы.

4. Основная причина, по которой вычислительная машина не может давать оценки, состоит в том, что используемый экспериментальный план и модель дают вырожденную матрицу $(\mathbf{X}'\mathbf{X})$, которую нельзя обратить. Здесь в каждой строке матрицы $\mathbf{X}$ справедливо

$$X_1^4 = 10X_1^2 - 9 \quad \text{и} \quad X_2^4 = 10X_2^3 - 9.$$

Обе модели приводят к одной и той же вырожденной задаче.

5. Проще работать со значениями преобразованных наблюдений $Y_i - 0{,}56$ вместо первоначальных Y_i. Если модель первого порядка подобрана, графики остатков указывают на необходимость введения членов более высокого порядка. Таблица дисперсионного анализа, приведенная ниже, показывает, что статистически значимы только коэффициенты первого и пятого порядков. Если мы оставим все члены до пятого порядка включительно, то получим довольно громоздкое уравнение, приведенное ниже, которое можно было бы преобразовать так, чтобы объединились члены, содержащие одинаковые степени выражения $(Z - 1957)$.

$$\hat{Y} = 1{,}060 + 0{,}070143\,(Z - 1957) - 0{,}000386\,[3\,(Z - 1957)^2 - 56] - \frac{0{,}000824}{6} \times$$

$$\times\,[5\,(Z - 1957)^3 - 167\,(Z - 1957)] - \frac{0{,}000058}{12}\,[35\,(Z - 1957)^4 - 1655\,(Z - 1957)^2 +$$

$$+ 9072] - 0{,}000102\left(\frac{21}{22}\right)\left[(Z - 1957)^5 - \frac{545}{9}\,(Z - 1957)^3 + \frac{708032}{1008}\,(Z - 1957)\right].$$

Источник рассеяния	Число степеней свободы	SS	MS	Источник рассеяния	Число степеней свободы	SS	MS
a_0	1	3,750000	3,750000	a_4	1	0,022108	0,022108
a_1	1	1,377606	1,377606*	a_5	1	0,109029	0,109029*
a_2	1	0,005539	0,005539	Остаток	9	0,071906	0,007990
a_3	1	0,027012	0,027012	Общий	15	5,36200	

6. Пусть Z равно номеру недели. Подобранное уравнение:

$$\hat{Y} = 136{,}227 + 2{,}687Z + 0{,}167Z^2.$$

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий	20	74 628,00		
a_0	1	48 609,80		
a_1	1	25 438,75	25 438,75	4558,92*
a_2	1	489,00	489,00	87,63*
a_3	1	1,15	1,15	0,21
Остаток	16	89,30	5,58	

7.1)

$$\mathbf{X} = \begin{array}{c} \begin{matrix} X_0 & X_1 & X_2 & X_3 \end{matrix} \\ \left[\begin{matrix} 1 & 1 & 0 & -1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 0 \\ 1 & -1 & -1 & 1 \end{matrix}\right] \end{array}.$$

$$\hat{Y} = 248 + 2X_1 - 10X_2 - 7{,}33X_3.$$

2) **Дисперсионный анализ**

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий (скорректированный)	8	1466,0		
Регрессия	3	826,7	257,57	не значимо
$b_1 \mid b_0$	1	54,0 } 504,0	252,00	не значимо
$b_2 \mid b_0, b_1$	1	450,0 }		
$b_3 \mid b_0, b_1, b_2$	1	322,7	322,7	не значимо
Остаток	5	639,3	127,9	

Различия между операторами статистически незначимы, т. е. $\dfrac{252{,}0}{127{,}9} = 1{,}87$ меньше, чем F (2; 5; 0,95) = 5,79.

Оператор № 1: $\hat{Y} = 248 + 2\,(1) = 250.$

Оператор № 2: $\hat{Y} = 248 - 10\,(1) = 238.$

Оператор № 3: $\hat{Y} = 248 + 2\,(-1) - 10\,(-1) = 256.$

3) Здесь нет достаточных оснований, чтобы сказать, что линейный эффект линейной скорости значим с α-риском 0,05.

4) График остатков указывает, что модель второго порядка для линейной скорости — это наилучший выбор.

8. 1) $\hat{Y} = -0{,}0037 - 2{,}8008t + 0{,}2314t^2.$

2) Анализ остатков: нет оснований считать, что надо увеличивать степень полинома по t.

9. 1) $\hat{Y} = 22{,}561235 + 1{,}668017X - 0{,}067958X^2.$

2) Дисперсионный анализ

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий (скорректированный)	18	204,481053		
Регрессия	2	201,994394	100,997197	649,8*
Остаток	16	2,486659	0,155416	

Регрессия статистически значима при $\alpha = 0{,}05$.

3) Критерий неадекватности — разложение остатка из приведенной выше таблицы дисперсионного анализа.

Источник	Число степеней свободы	SS	MS	F
Остаток	16	2,486659		
неадекватность	8	1,733325	0,216666	2,30
«чистая» ошибка	8	0,753334	0,094168	

Неадекватность незначима, так как $2{,}30 \leqslant F\ (8;\ 8;\ 0{,}95) = 3{,}44$. Следовательно, квадратичная модель достаточна для предсказания.

4) $\hat{Y} = 23{,}346374 + 1{,}045463X$.

Дисперсионный анализ

Источник рассеяния	Число степеней свободы	SS	MS	F
Общий	18	204,481053		
Регрессия	1	195,242967	195,242967	359,29*
Остаток	17	9,238086	0,543417	
неадекватность	9	8,484752	0,942750	10,01*
«чистая» ошибка	8	0,753334	0,094168	

Неадекватность статистически значима, так как $10{,}01 > F\ (9;\ 8;\ 0{,}95) = 3{,}39$.

Неадекватность модели подтверждается также остатками для полученного уравнения.

При вычерчивании графика остатков в зависимости от значений независимой переменной можно увидеть определенную кривизну, которая указывает на необходимость введения члена второго порядка по X.

5) *Выводы.*

Точку помутнения можно предсказать функцией второго порядка по X:

$$\hat{Y} = 22{,}561235 + 1{,}668017X - 0{,}067958X^2,$$

Причем нет никаких указаний на то, что нужна какая-то иная, более сложная модель.

10.

X	Y	$Z = \log_1 Y$	$V = \log_{10}\log_{10} Y$
1830	30	1,4771	0,1694
1905	130	2,1139	0,3251
1930	400	2,6021	0,4153
1947	760	2,8808	0,4595
1952	1 500	3,1761	0,5019
1969	25 000	4,3979	0,6432

1) График зависимости Y от X показывает все «события» как могущие произойти в более поздние годы, он не слишком информативен. Потребность в преобразовании очевидна. В такой ситуации, когда наблюдается пропорциональный рост, определение Z часто более информативно.

2) График зависимости Z от X оценить гораздо легче, он, несомненно, лучше, чем график зависимости Y от X.

3) $U \equiv Y$ — тоже не плохо, хотя мы можем найти и кое-что получше.

4) $\widehat{V} = -5{,}4874 + 0{,}00307284X$.

Графики остатков как общий, так и в зависимости от времени мало удовлетворительны, но они содержат всего шесть остатков при четырех степенях свободы, так что у нас нет особой надежды на нечто большее, чем то, что мы видим.

5) **Дисперсионный анализ**

Источник	Число степеней свободы	SS	MS	F
$b_1 \mid b_0$	1	0,1185	0,11850	41,58
Остаток	4	0,0114	0,00285	
Общий (скорректированный)	5	0,1299		

$41{,}58 > F\ (1;\ 4;\ 0{,}99)$, что значимо на 1 %-ном уровне. $R^2 = 0{,}9122$. Бо́льшая часть вариации в данных объясняется с помощью этой модели.

Отметим, что ковариация $(b_0,\ b_1) = -0{,}999717$, что очень много. Это вызвано удаленностью начала координат и репараметризация, переносящая начало координат в точку, близкую к $\overline{X} = 1922{,}167$, должна помочь (см. параграф 10.4).

6) Используя обратную интерполяцию, получим

$$\log\{\log(186000 \times 3600)\} = -5{,}4874 + 0{,}00307284\widehat{X},$$

$$\widehat{X} = 2094.$$

7) Это рискованная экстраполяция по малому набору не очень представительных данных, которые к тому же подверглись произвольному преобразованию, так что мы не можем рассчитывать на доказательные утверждения относительно новых данных.

Прогноз зависит от того, сохранится ли существенный тренд на следующие 100 лет или около того. Можно ли этому поверить!

11. $\mathrm{SS}\,(b_1|b_0) = 58{,}51$,

$\mathrm{SS}\,(b_2|b_0, b_1) = 210{,}58$,

$$SS(b_3|b_0,\ b_1,\ b_2) = 0{,}005,$$
$$SS(b_2,\ b_3|b_0,\ b_1) = 210{,}59.$$

Для проверки гипотезы $H_0 : \beta_2 = \beta_3$ нам надо найти модель

$$Y = \beta_0 + \beta_1 X + \beta (X^2 + X^3) + \varepsilon.$$

$$\frac{(SSW - SSR)/q}{SSR/(n-p)} = \frac{(187{,}02 - 2{,}23)/1}{2{,}33/2} = 165{,}73.$$

Гипотеза H_0 отвергается.

Вот наилучшая модель:

$$\hat{Y} = 2{,}40625 + 0{,}914286X + 0{,}59375X^2.$$

$$\mathbf{A} = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{X}_2 = \begin{bmatrix} 0 \\ 20{,}2 \\ 0 \end{bmatrix}.$$

Таким образом, оценки b_0 и b_2 — несмещенные, а $E(b_1) = \beta_1 + 20{,}2\,\beta_3$.

12. Возьмем фиктивный фактор Z с уровнями 1, 3, 4, 6, 10, 13, 14, 15, 16, 17, 19 соответственно для чисел от 17, 19, . . . ноября до 5 декабря. Тогда надо подбирать модель $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \alpha_1 Z + \alpha_2 Z^2 + \varepsilon$. (Заметим, что член Z, как, впрочем, и Z^2, нужен для представления квадратичного тренда. Нам бы понадобился еще и член α_0, если бы в модели не было уже члена β_0.) Альтернативными вариантами фиктивной переменной могли бы быть $(Z-1)$ или $(Z-\overline{Z})$; это повлекло бы за собой только изменения в интерпретации постоянного члена.

13.
$a_0 = 12{,}565663$, $b_0 = 0{,}038437$, $c_1 = -0{,}032454$,
$a_1 = -0{,}006327$, $b_1 = -0{,}013571$, $c_2 = 0{,}001248$,
$a_2 = -0{,}090698$, $b_2 = 0{,}001376$, $c_3 = 0{,}000198$.

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
Регрессия $\|\, b_0$	8	44,40797	5,55100	
Остаток	11	0,03989	0,00363	
Неадекватность	6	0,03461	0,00577	5,46 значимо на уровне $\alpha = 0{,}05$
«Чистая» ошибка	5	0,00528	0,001056	
Общий (скорректированный)	19	44,44785		

$R^2 = 44{,}40797/44{,}44785 = 0{,}9991$. Это очень интересный набор данных. Практически вся вариация объясняется моделью, а неадекватность тем не менее значима. В этой связи прежде всего возникает вопрос, а *действительно* ли параллельные опыты в точности параллельны? Если дело обстоит именно так, то анализ показывает, что существует вариация, выходящая за границы естественного разброса, который удается объяснить. (С практической точки зрения, однако, львиная доля вариации *уже* объясиена. Так почему бы не воспользоваться полученным уравнением? Автор так и поступил.) Графики остатков в пространстве X указывают на возможность эффекта взаимодействия X_1X_2 и наводят на мысль

о включении в модель дополнительного члена вида $(\delta_0 + \delta_1 Z + \delta_{11} Z^2)\ X_1 X_2$. Правда, если это проделать, то коэффициент δ_{11} не удастся оценить независимо из-за коррелированности столбцов матрицы **X**, поэтому метод наименьших квадратов даст лишь следующие оценки:

$a_0 = 11{,}918982$, $b_0 = 0{,}049215$, $c_0 = 0{,}001660$, $d_0 = 0{,}000513$,
$a_1 = 0{,}057033$, $b_1 = -0{,}014627$, $c_1 = -0{,}001769$, $d_1 = -0{,}000050$.
$a_2 = -0{,}090698$, $b_{11} = 0{,}001376$, $c_{11} = 0{,}000198$,

Дополнительная сумма квадратов SS для d_0 и d_1 при заданных остальных оценках равна 0,02183, что ведет к F-критерию, равному $5{,}43 > F(1; 9; 0{,}95) = 5{,}12$, а это едва значимо. Новый критерий неадекватности F равен $3{,}03 < F(4; 5; 0{,}95) = 5{,}19$, что не значимо, а R^2 увеличился совсем чуть-чуть до 0,9996. Значение F-критерия для регрессии (при заданном a_0) равно 2214, что в высшей степени значимо.

14. Сначала прочтите параграф 5.8. Всего, как описано в данной задаче, 322 наблюдения. Пусть Z_1, Z_2 — фиктивные переменные, такие, что $Z_1 = 1$ для категории A и нулю для прочих категорий, $Z_2 = 1$ для категории B и нулю для остальных. Построим модель $Y = \beta_0 + \alpha_1 Z_1 + \alpha_2 Z_2 + \beta_1$ (возраст) $+ \varepsilon$. Имеем: $\hat{Y} = 80{,}58 + 5{,}73 Z_1 + 7{,}54 Z_2 + 1{,}08$ (возраст). Скорректированная таблица дисперсионного анализа такова:

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
$a_1, a_2 \mid b_0$	2	4 398,97	2199,49	
$b_1 \mid b_0, a_1, a_2$	1	1 000,18	1000,18	4,52
Остаток	318	70 441,50	$s^2 = 221{,}51$	
неадекватность	19	2 379,08	125,21	0,55
«чистая» ошибка	299	68 062,42	227,63	
Общий, скорректированный	321	75 840,64		

Неадекватность не значима, и мы отбрасываем гипотезу $\beta_1 = 0$ на уровне значимости $\alpha = 0{,}05$, поскольку $4{,}52 >$ (значения, интерполированного между 3,92 и 3,84 в таблице F-критерия). $R^2 = (4398{,}97 + 1000{,}18)/75840{,}64 = 0{,}0712$. Это значение (объясняющее всего лишь 7 % общей вариации относительно среднего) и тот факт, что F-критерий, равный 4,52, лишь незначительно больше, чем процентная точка, с которой он сравнивался, говорят о том, что полученное уравнение не очень-то полезно. Критерии для гипотез $\alpha_1 = 0$, $\alpha_2 = 0$ и $\alpha_1 - \alpha_2 = 0$ подтверждают тот факт, что две группы мужчин весят примерно одинаково и что они тяжелее, чем группа женщин. (Соответствующие значения t-критерия равны 2,55; 3,82 и 0,84). Отсюда мы заключаем, что хотя, как видно из данных, вес и увеличивается с возрастом, из него нельзя получить хороший предиктор и что хотя *существуют* различия между диетами, они оказываются смешанными с полом и могут объясняться скорее полом, чем диетой.

15. Мы можем воспользоваться фиктивными переменными со следующими значениями:

$$X_1 = 1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 8,\ 8,\ 8,\ 8,\ \ldots,\ 8,$$
$$X_2 = 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 2,\ 3,\ \ldots,\ 64,$$
$$X_3 = 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 1,\ 1,\ 1,\ \ldots,\ 1,$$

для получения модели

$$\hat{Y} = 0{,}421429 + 0{,}039643X_1 + 0{,}004062X_2 + 0{,}016366X_3.$$

$$\text{SS}(b_1,\ b_2,\ b_3|b_0) = 1{,}00910 \text{ (3 степени свободы).}$$

$$\text{Остаточная SS} = 0{,}03829 \text{ (68 степеней свободы).}$$

$F = 597{,}3 > F\ (3;\ 68;\ 0{,}99) = 4{,}10$ (приближенно). Регрессия значима.

$R^2 = 0{,}9634$. Вообще, с точки зрения данного критерия кажется, что получилась отличная модель (однако см. ниже).

Точка пересечения линий соответствует возрасту 7,96 ($X_1 = 8{,}46$).

Статистика критерия Дарбина—Уотсона $d = 2{,}514$; $4 - d = 1{,}486$, что значимо на уровне 5 %. Это указывает на отрицательную сериальную корреляцию остатков, которая нуждается в дальнейшем исследовании. Соответствующий критерий серий в верхнем хвосте распределения тоже значим ($u = 46$, $n_1 = 32$, $n_2 = 40$, $Z = 2{,}27$, вероятность верхнего хвоста равна 0,0116). Наличие отрицательной сериальной корреляции в известной степени воздействует на состоятельность приведенных выше регрессионных критериев. Возможный следующий шаг мог бы состоять в оценке значения ρ_s и в использовании взвешенного метода наименьших квадратов.

(Примечание. В исходной работе строилась квадратичная модель, а вовсе не две прямых, как здесь. Поэтому вообще не делалось предположений о том, какие части модели в каких точках работают. Описание такой более общей модели можно найти в приведенном источнике.)

16. Мы можем воспользоваться здесь той же структурой фиктивных переменных, что и в предыдущем упражнении, однако в данном случае число наблюдений ограничено только 32. Тогда получим

$$\hat{Y} = 0{,}421429 + 0{,}039643X_1 + 0{,}004277X_2 + 0{,}012905X_3.$$

$$\text{SS}(b_1,\ b_2,\ b_3|b_0) = 0{,}33925 \text{ (3 степени свободы).}$$

Остаточная SS = 0,02152 (28 степеней свободы).

$F = 147{,}13 > F\ (3;\ 28;\ 0{,}99) = 4{,}57$. Регрессия значима.

$R^2 = 0{,}9403$. Это снова отличное приближение для такого критерия (однако см. ниже).

Точка пересечения линий соответствует возрасту 7,86 ($X_1 = 8{,}36$).

Статистика критерия Дарбина—Уотсона $d = 2{,}722$; $4 - \text{d} = 1{,}278$, что почти значимо на 5 %-ном уровне (где $d_L = 1{,}24$). Соответствующий критерий серий для верхнего хвоста распределения значим, правда, на 2,5 %-ном (одностороннем) уровне. (Действительно, $u = 23$, $n_1 = 18$, $n_2 = 14$, $z = 2{,}10$, вероятность верхнего хвоста равна 0,0179.) Наличие отрицательной сериальной корреляции некоторым образом влияет на состоятельность приведенных выше регрессионных критериев. Возможный следующий шаг мог бы заключаться в оценке величины ρ_s и использовании взвешенного метода наименьших квадратов (см. примечание к предыдущему решению).

17. Второй столбец уже ортогонален к первому. Мы получим:

$$Z_{3\text{т}} = \frac{1}{2}(-5,\ -3,\ -1,\ 1,\ 3,\ 5)',$$

$Z_{4\text{т}} = (0,\ 0,\ 0,\ 0,\ 0,\ 0)'$, значит, зависимость между столбцами существует. В самом деле сумма второго, третьего и четвертого столбцов матрицы **X** равна нулю.

18. Дисперсионный анализ приведен в следующей таблице:

Значение b_j	Источник	Число степеней свободы	SS
15,111	b_0	1	2055,11
1,866667	b_1	1	209,07
0,165945	b_2	1	70,33
0,072727	b_3	1	5,24
—	Остаток	5	4,36
	Общий	9	2350,00

Кубический член не значим.
Вот подходящая модель.

$$\hat{Y} = b_0 + b_1X + b_2(3X^2 - 20) = 11{,}792 + 1{,}8667X + 0{,}4978X^2,$$

где уровни красителя закодированы числами $X = -4, -3, -2, -1, 0, 1, 2, 3, 4$. Эта модель охватывает $R^2 = 285{,}4/294{,}89 = 0{,}9678$ разброса относительно среднего; $s^2 = 9{,}6/6 = 1{,}6$.

19. Решение не приводится.

20. (Y_1 — данные.)

λ	$L_{max}(\lambda)$	λ	$L_{max}(\lambda)$
—1,0	—22,9	0,1	—5,3
—0,5	—11,0	0,2	—6,5
—0,4	—9,5	0,3	—8,1
—0,3	—8,0	0,4	—9,8
—0,2	—6,5	0,5	—11,5
—0,1	—5,3	1,0	—24,8
0,0	—4,8		

Соответствующий 95 %-ный доверительный интервал равен $-0{,}35 \leqslant \lambda \leqslant 0{,}32$.
Мы выбираем $\lambda = 0$, т. е. используем преобразование $W = \ln Y_1$.
Имеем следующее уравнение:

$$\hat{W} = 4{,}234 + 0{,}204X_1 + 0{,}098X_2 - 0{,}139X_3 - 0{,}070X_4.$$

$R^2 = 0{,}9963$. Все частные F-критерии для отдельных коэффициентов весьма значимы. При исследовании остатков в количестве одиннадцати штук всего с семью степенями свободы не приходится ожидать, что они выявят что-нибудь загадочное. И этого действительно не случилось.
(Для данных Y_2 решение не приводится.)

21. Решение не приводится.

22. $\hat{\lambda} = 0{,}11$. Доверительная полоса при 95 % лежит около $-0{,}16 \leqslant \lambda \leqslant 0{,}39$, значит, мы можем взять $\lambda = 0$, т. е. использовать преобразование $W = \ln M$. Тогда

$$\hat{W} = -5{,}728 + 2{,}031 \ln T,$$

$$R^2 = 0{,}8056.$$

$$\mathrm{SS}(b_1|b_0) = 16{,}240 \text{ (1 степень свободы)}.$$

Остаточная SS = 3,919 (41 степень свободы).
Общая, скорректированная SS = 20,159 (42 степени свободы).
$F = 169{,}9 > F(1; 41; 0{,}99) = 4{,}08$, так что регрессия весьма значима.

На графиках остатков в зависимости от T и от $\hat{Y}$ не видно ничего необычного. Критерий Дарбина—Уотсона равен $d = 1{,}94$, получен в обычном порядке, не значим.

23. $\widehat{\log Y} = -4{,}927 + 16{,}85\,\{1000/(T + 460)\}$.

Дисперсионный анализ

Источник	Число степеней свободы	SS	MS	F
b_0	1	238,731		
$b_1 \mid b_0$	1	1,785	1,785	
Неадекватность	2	0,064	0,032	1,24
«Чистая» ошибка	20	0,515	0,026	не значимо
Общий	24	241,095		

Проверка для регрессии $F = 1{,}785/(0{,}579/22) = 67{,}87 > F\ (1;\ 22;\ 0{,}95) = 4{,}30$ говорит о явной значимости.

$R^2 = 0{,}7552$.

Остатки демонстрируют снижение точек по мере роста T, что указывает на неадекватность преобразования $\log Y$ в силу неоднородности дисперсий в исходном массиве данных. Теперь есть две возможности: выбор нового преобразования и использование взвешенного метода наименьших квадратов.

Возрастание наклона остатков, когда они наносятся на график в зависимости от предсказанных значений отдельно для каждой температуры,— это ложный «эффект», обусловленный группированием. Так как фактически (для групп равного объема) попытки подобрать модель для групповых средних сводятся к подбору прямой, само собой получается, что самые малые наблюдения внутри группы будут иметь малые остатки, и т. д. Общий (несгруппированный) график остатков показывает только эффект «воронки», отмеченный выше.

24. Частичное решение:

i	Оцененный коэффициент a_i	Перед сжатием		После сжатия	
		число степеней свободы	SS (a_i)	число степеней свободы	SS (a_i)
0	66,25	1	52 668,75	1	52 669
1	—0,02273	1	0,30		
2	—0,00774	1	0,72		
3	1,96018	1	19 780,16	1	19 780
4	0,01349	1	1,46		
5	—0,02137	1	7,26		
6	—0,06551	1	19,26		
Остаток		5	13,09	10	42
Общий		12	72 491,00	12	72 491

Все a_i, кроме a_6, элиминируются при сжатии и не значимы при уровне $\alpha = 0{,}05$. Однако включение a_6 увеличивает R^2 всего только от 0,9979 до 0,9988, так что мы считаем a_6 статистически значимым, но численно пренебрежимым. Соответствующее значение $F = 19{,}36/\{13{,}09/5\} = 7{,}39$, что совсем не много превышает табличное значение $F\ (1;\ 5;\ 0{,}95) = 6{,}61$.

Вот выбранная модель:

$$\hat{Y} = 66{,}25 + 1{,}96\left\{\frac{2}{3}\left(X^2 - \frac{85}{4}\right)\right\}X.$$

Критерий Дарбина—Уотсона $d = 2{,}33$ не значим.

25. Решение не приводится.

26. $\widehat{\log Y} = 1{,}9929 + 0{,}5428 \log X_1 + 0{,}2740 \log X_2$.

Эта модель объясняет только 37,52 % вариации. Регрессия не значима, поскольку значение общего F-критерия равно $(6{,}49/2)/(12{,}03/13) = 3{,}51$, что меньше, чем $F(2; 13; 0{,}95) = 3{,}81$.

Седьмой остаток крайне велик и отрицателен (— 2, 626), так что соответствующее значение Y кажется слишком маленьким.

Графики остатков в зависимости от $\log X_1$ и от $\widehat{\log Y}$ имеют криволинейный характер.

Кажется весьма сомнительным, что ошибки аддитивны. Так ли это, можно было бы почувствовать при построении нелинейной модели $Y = \alpha X_1^{\beta} X_2^{\gamma} + \varepsilon$, для чего требуются методы из гл. 10. А в качестве начальных оценок (которые нужны при нелинейном оценивании) можно было бы взять те, что получились в подобранной выше модели, а именно $\alpha_0 = 10^{1{,}9929} \doteqdot 100$, $\beta_0 = 0{,}5428$ и $\gamma_0 = 0{,}2740$.

27. Для учета номеров побегов мы введем два фиктивных фактора Z_1 и Z_2. Переменная Z_1 равна 1 для первого побега и 0 для всех остальных, а Z_2 равна 1 для второго побега и 0 в иных случаях. Третий фиктивный фактор Z_3 принимает значение 1 для сорта «Уолдрон» и 0 для сорта «Киано». Переменную, соответствующую количеству азота, обозначим N. Теперь мы можем получить методом наименьших квадратов следующее уравнение:

$$\hat{Y} = -333{,}5 + 185{,}01X - 7{,}801X^2 - 0{,}1294N + 106{,}0Z_1 + 33{,}7Z_2 + 19{,}6Z_3.$$

(Мы сохраняем достаточное число значащих цифр, чтобы получать предсказания с точностью того же порядка, что и в исходных данных, т. е. до ближайшей единицы.) Таблица дисперсионного анализа выглядит так:

ANOVA

Источник	Число степеней свободы	SS	MS	F
$b_1 \mid b_0$	1	1 112 031	1 112 031	176,71
$b_{11} \mid b_0, b_1$	1	33 968	33 968	5,40
$b_2 \mid b_0, b_1, b_{11}$	1	20 392	20 392	3,24
$a_1, a_2, a_3 \mid$остальные	3	14 994	4 998	0,79
Остаток	11	69 225	6 293	
Общий (скорректированный)	17	1 250 610		

$F(1; 11; 0{,}95) = 4{,}48$; $F(1; 11; 0{,}99) = 9{,}65$.

Мы видим, что суммарный вклад фиктивных переменных не значим. Исключение этих переменных ведет к $R^2 = 0{,}933$ для подобранного уравнения $\hat{Y} = -338{,}3 + 200{,}3X - 7{,}594X^2 - 0{,}369N$, в то время как сохранение фиктивных факторов дает $R^2 = 0{,}945$, что несущественно больше. Правда, стоит отметить, что если взять одни *только* фиктивные переменные, то модель для трех таких переменных объяснит $R^2 = 0{,}571$ вариации относительно среднего. Это обусловлено корреляцией между фиктивными переменными и X. Отметим, например, низкие отклики для третьих побегов.

Нет ничего удивительного в том, что количество азота N сильно коррелировано с X ($r = 0{,}672$) и с X^2 ($r = 0{,}629$). Такой незначимый вклад можно было бы и исключить из уравнения, как было сделано с фиктивными факторами. Если это действительно сделать, то мы получим $\hat{Y} = -337 + 196{,}0X - 0{,}13X^2$ при $R^2 = 0{,}916$. Если нанести данные на график, то можно увидеть то, что видно и из таблицы дисперсионного анализа, а именно что квадратичный эффект проявляется совсем слабо и что член первого порядка вносит наибольший вклад в вариацию. Значения Y, $\hat{Y}$, стандартное отклонение [35] ($\hat{Y}$), e и стандартизированные остатки табулированы ниже для дальнейших исследований читателей.

Y	$\hat{Y}$	Стандартное отклонение ($\hat{Y}$)	e	Стандартное отклонение (e)	Y	$\hat{Y}$	Стандартное отклонение ($\hat{Y}$)	e	Стандартное отклонение (e)
370	343	25	27	0,34	188	148	30	40	0,51
659	652	32	7	0,09	632	546	33	86	1,12
935	803	50	132	1,97	538	690	31	—152	—1,96
390	263	24	127	1,59	27	22	46	5	0,07
753	683	32	70	0,91	141	148	30	—7	—0,09
733	819	62	—86	—1,53	262	263	24	—1	—0,01
182	192	26	—10	—0,13	34	133	31	—99	—1,28
417	462	31	—45	—0,58	222	192	26	30	0,37
686	697	31	—11	—0,14	242	355	26	—113	—1,43

28. См. с. 304

29. Решение не приводится.

[35] В оригинале «стандартная ошибка», а должно быть «стандартное отклонение». — *Примеч. пер.*

СОДЕРЖАНИЕ

Предисловие к русскому изданию 5
Предисловие к первому изданию 16
Предисловие ко второму изданию 18

Глава 1. **Подбор прямой методом наименьших квадратов** 20
1.0. Введение. Потребность в статистическом анализе 20
1.1. Прямолинейная зависимость между двумя переменными 26
1.2. Линейная регрессия: подбор прямой 29
1.3. Точность оценки регрессии 38
1.4. Исследование уравнения регрессии 43
1.5. Неадекватность и «чистая» ошибка 54
1.6. Корреляция между X и Y 63
1.7. Обратная регрессия (случай прямой линии) 69
1.8. Некоторые следствия из гл. 1, имеющие практическое значение 72
Упражнения 76
Ответы к упражнениям 89

Глава 2. **Матричный подход к линейной регрессии** 104
2.0. Введение 104
2.1. Подбор уравнения прямой в матричных обозначениях; оценки параметров β_0 и β_1 104
2.2. Дисперсионный анализ в матричных обозначениях 116
2.3. Дисперсия и ковариация коэффициентов на основе матричных вычислений 118
2.4. Дисперсия величины $\hat{Y}$ в матричных обозначениях 118
2.5. Резюме к матричному подходу при подборе прямой 119
2.6. Случай общей регрессии 121
2.7. Принцип «дополнительной суммы квадратов» 133
2.8. Ортогональные столбцы в матрице **X** 135
2.9. Частные и последовательные F-критерии 137
2.10. Проверка общей линейной гипотезы в регрессионных задачах 139
2.11. Взвешенный метод наименьших квадратов 145
2.12. Смещение регрессионных оценок 153
2.13. Метод наименьших квадратов при наличии ограничений 159
2.14. Некоторые замечания относительно ошибок в предикторах (одновременно с ошибками в откликах) 159
2.15. Обратная регрессия (в случае многомерного предиктора) 162
Приложение 2А. Некоторые полезные сведения из теории матриц 164
Приложение 2Б. Математическое ожидание дополнительной суммы квадратов 165
Приложение 2В. Насколько значимой должна быть регрессия? 167
Приложение 2Г. Неопределенные множители Лагранжа 171
Упражнения 173
Ответы к упражнениям 177

Глава 3. **Исследование остатков** 186
3.0. Введение 186
3.1. Общий график 188
3.2. График временной последовательности 191

3.3. График зависимости остатков от $\hat{Y}_i$ 193
3.4. График зависимости остатков от предикторных переменных X_{ji}, $i=1, 2, \ldots, n$ 194
3.5. Другие графики остатков 195
3.6. Статистики для исследования остатков 196
3.7. Корреляция между остатками 197
3.8. Выбросы 199
3.9. Сериальная корреляция остатков 199
3.10. Исследование серий на графиках временной последовательности остатков 204
3.11. Критерий Дарбина—Уотсона для некоторых видов сериальной корреляции 209
3.12. Определение влияющих наблюдений 216
Приложение 3А. Нормальные и полунормальные графики 224
Упражнения 231
Ответы к упражнениям 239

Глава 4. **Две предикторные переменные** 244

4.0. Введение 244
4.1. Сведение множественной регрессии с двумя предикторными переменными к последовательности простых линейных регрессий 248
4.2. Исследование уравнения регрессии 254
Упражнения 261
Ответы к упражнениям 266

Глава 5. **Более сложные модели** 273

5.0. Введение 273
5.1. Полиномиальные модели различных порядков по X_j 274
5.2. Модели, включающие преобразования, отличные от целых степеней . . 277
5.3. Семейства преобразований 281
5.4. Использование «фиктивных» переменных в множественной регрессии 298
5.5. Центрирование и масштабирование. Представление регрессии в корреляционной форме 314
5.6. Ортогональные полиномы 324
5.7. Преобразование матрицы X для получения ортогональных столбцов 334
5.8. Регрессионный анализ усредненных данных 337
Упражнения 340
Ответы к упражнениям 353

Дрейпер Н., Смит Г.

Д73 Прикладной регрессионный анализ: В 2-х кн. Кн. 1/ Пер. с англ.— 2-е изд., перераб. и доп.— М.: Финансы и статистика, 1986.— 366 с., ил.— (Математико-статические методы за рубежом).

Работа американских ученых посвящена регрессионному анализу, применяемому во всех отраслях народного хозяйства и научных исследованиях. Второе издание (1-е изд. перевода — 1973 г.— вышло в одной книге) значительно переработано и дополнено новыми алгоритмами и сравнением их достоинств. Кн. 1 содержит классическое описание модели линейной регрессии, включая описание алгоритмов для ЭВМ.

Для специалистов — статистиков, экономистов, социологов, научных работников.

Д $\frac{0702000000-118}{010(01)-86}$ 108—86

ББК 22.172

Монография

Норман Дрейпер, Гарри Смит

ПРИКЛАДНОЙ РЕГРЕССИОННЫЙ АНАЛИЗ. Кн. 1

Книга одобрена на заседании редколлегии серии
«Математико-статистические методы за рубежом» 26.05.83

Зав. редакцией *К. В. Коробов*
Редактор *Е. В. Крестьянинова*
Мл. редакторы *О. Г. Виноградова, А. С. Шиманская*
Техн. редактор *И. В. Завгородняя*
Худож. редактор *Ю. И. Артюхов*
Корректоры *Г. В. Хлопцева, Г. А. Башарина, Н. П. Сперанская и Е. В. Люминарская*

ИБ № 1594

Сдано в набор 20.03.86. Подписано в печать 5 09 86. Формат $60\times90^1/_{16}$. Бум. тип. № 1. Гарнитура «Литературная». Печать высокая. Усл. п. л. 23,0. Усл. кр.-отт. 23,0. Уч.-изд. л. 25,12. Тираж 12 000 экз. Заказ 949. Цена 2 р. 30 к.
Издательство «Финансы и статистика», 101000, Москва, ул. Чернышевского, 7

Ленинградская типография № 4 ордена Трудового Красного Знамени Ленинградского объединения «Техническая книга» им. Евгении Соколовой Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли. 191126, Ленинград, Социалистическая, 14