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 · Singapore

Н. Дрейпер
Г. Смит

ПРИКЛАДНОЙ РЕГРЕССИОННЫЙ АНАЛИЗ

Издание второе,
переработанное и дополненное

КНИГА 2

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

МОСКВА «ФИНАНСЫ И СТАТИСТИКА» 1987

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., Боши С., Бросье Ж. и др. Методы анализа данных.
16. Бартоломью Д. Стохастические модели социальных процессов.

ГОТОВЯТСЯ К ПЕЧАТИ

Дэйвисон М. Многомерное шкалирование. Методы наглядного представления данных.

Д $\frac{0702000000-006}{010(01)-87}$ 109—86

● ПРЕДИСЛОВИЕ К РУССКОМУ ИЗДАНИЮ

Вторая книга монографии Н. Дрейпера и Г. Смита «Прикладной регрессионный анализ» включает вычислительные аспекты регрессионного анализа, примеры его применения, принципы организации регрессионных исследований; в ней демонстрируется связь с дисперсионным анализом, рассмотрены вопросы нелинейного оценивания.

Благодаря регрессионному анализу в вычислительной математике возникло целое направление, связанное главным образом с решением плохо обусловленных задач. Появилось огромное число подходов, алгоритмов и программ, позволяющих в этих нелегких условиях более или менее рационально организовать вычислительные процедуры. Интерес к вычислительным аспектам возник еще в «домашинную» эру, что привело к появлению схемы Дуллиттла, регрессии на ортогональные полиномы Чебышева, преобразованию с использованием метода Грама-Шмидта. С появлением современной вычислительной техники арсенал вычислительных методов резко расширился. Появились регрессия на главные компоненты, ридж-регрессия, регрессия на основе G-обращения, регуляризация по Тихонову, регрессия с использованием сингулярных разложений, схемы Холецкого и Хаусхолдера и т. д.

Все эти методы нашли отражение на страницах данной книги. Если у читателя возникнет потребность ознакомиться с более подробным описанием конкретных процедур, он найдет соответствующие ссылки в примечаниях переводчиков к гл. 5. Подобные методы обычно основаны на преобразовании исходной задачи, на приведении ее к виду, удобному для вычислений. Как правило, их использование демонстрируется в применении к процедуре построения линейных регрессий.

При оценивании параметров нелинейных регрессий приходится прибегать к поисковым методам, имеющим итерационный характер. Число таких методов и их модификаций столь велико, что даже для их перечисления и краткой характеристики потребовался бы значительный объем. В книге по существу речь идет лишь о градиентном методе и о методе Маркуардта. В последнее время на практике все более широкое применение находят методы без вычисления производных. К ним относятся, например, последовательный симплексный метод и его модификации, овражные методы и методы случайного поиска, тоже имеющие большое число модификаций.

На первых этапах развития регрессионного анализа его широкое применение ограничивалось большим объемом вычислений, необходимых для получения результата. Развитие вычислительной техники

кардинально изменило ситуацию. Появилась возможность автоматизировать регрессионные вычисления. Были созданы известные методы включения и исключения. На их базе были написаны многочисленные программы, развитие которых вылилось в метод всех возможных регрессий, а затем и в шаговый регрессионный анализ, ставший к настоящему времени наиболее массовым методом решения регрессионных задач. На этой основе были написаны многочисленные программы и пакеты программ для большинства известных типов вычислительных машин и на разных алгоритмических языках. Число таких пакетов только у нас в стране превосходит 1000. (Отечественные пакеты можно условно разделить на три большие группы: исследовательские пакеты, промышленные пакеты и пакеты многоцелевого назначения.)

В данной книге описаны и охарактеризованы наиболее известные зарубежные пакеты, содержащие программы регрессионного анализа.

В последние годы все большее распространение получают диалоговые системы, позволяющие работать с ЭВМ в интерактивном режиме.

Несмотря на огромное число публикаций по регрессионному анализу, продолжается активное развитие этого направления. Проследим лишь некоторые из основных тенденций, предполагая, что читатель знаком с содержанием предисловия к книге 1, где об этом уже шла речь. Прежде всего надо отметить, что происходит пересмотр, размывание довольно жестких базовых предпосылок классического регрессионного анализа. Это касается таких предположений, как нормальность распределения ошибок, детерминированность факторов, аддитивность учитываемых в модели ошибок, однородность, независимость (точнее — отсутствие корреляции между ошибками). Отказ хотя бы от одного из перечисленных предположений фактически приводит к созданию новой модели. А последствия отказа сразу от нескольких предположений во многих случаях не исследованы. К тому же у каждого из базовых предположений есть не одна альтернатива, а целый спектр возможностей.

Вторая тенденция состоит в вовлечении в регрессионный анализ более тонких математических методов. Это методы функционального анализа, теории групп, топологии и т. д Так, например, представляет интерес обобщение регрессионной задачи на бесконечномерные пространства. При исследовании идентифицируемости моделей находят применение методы теории групп.

Третья тенденция состоит в том, что развитие теории регрессионного анализа стимулируется, помимо всего прочего, обращением ко все более сложным объектам исследования. Помимо упоминавшихся ранее модификаций многомерной регрессии, речь идет, например, о моделях в форме обыкновенных дифференциальных уравнений, а также уравнений математической физики Сюда же относятся интегральные и интегро-дифференциальные уравнения, системы таких уравнений и вообще операторные уравнения.

Кроме моделей, структуры которых выражаются формулами, в последнее время в рамках регрессионного анализа стали рассматриваться модели, которые задаются только алгоритмически. Это приводит к широкому внедрению имитационных моделей.

Время ставит все более сложные задачи, и регрессионный анализ становится одним из первых инструментов, применяемых в процессе поиска их решения. Вот почему один из крупнейших современных специалистов по математической статистике Р. Рао назвал регрессионный анализ «методом века».

Перейдем теперь к четвертой тенденции. Классический регрессионный анализ основан на том, что вид математической модели задан априори с точностью до параметров. Предполагается также, что уже реализован эксперимент, выполненный по некоторому плану. Таким образом, задача сводится к выбору наилучшей процедуры обработки этих данных. В последнее время получает развитие новый подход, в рамках которого предлагается одновременно выбирать наилучшую триаду: модель—план—метод оценивания, отвечающую, насколько возможно, рассматриваемой задаче.

Не меньший интерес, с нашей точки зрения, представляет и концепция анализа данных, вытекающая из работ Дж. Тьюки. В отличие от предыдущего случая здесь предполагается, что выбор триады должен осуществляться не однажды, а многократно, поскольку процесс обработки данных предполагается перманентным: с появлением новых экспериментальных данных (как в модели текущего регрессионного анализа) возникают новые идеи, подходы и методы, уточняется понимание происходящих процессов и т. д. Анализ данных свел воедино изначально как бы несвязанные друг с другом элементы, подчинив их единому механизму решения задачи, открыв тем самым дорогу новому взгляду на возможности сбора (в том числе целенаправленного), анализа и интерпретации данных различной природы.

Особого внимания заслуживают методы планирования эксперимента. Они образуют теперь целое направление в математической статистике. Наряду с регрессионным анализом их применяют в разнообразных областях современной науки от теории игр до распознавания образов. По мере развития теории планирования эксперимента усиливается ее воздействие на регрессионный анализ, благодаря чему создаются новые специальные процедуры обработки данных и проверки статистических гипотез, а иногда и новые подходы. Характерным примером может служить предложение пользоваться методами планирования эксперимента для выбора оптимального значения параметра регуляризации в ридж-регрессии (см.: Vuchkov I. A ridgetipe procedure for design of experiments.— Biometrika, 1977, **64**, № 1, p. 147—150).

Одно из естественных направлений развития планирования эксперимента приводит к идее управления выборкой в процессе обработки данных. Данные, собранные в связи с решением конкретной задачи, часто рассматриваются как выборка из некоторой генеральной совокупности, свойства которой и интересуют исследователя. Если эта выборка достаточно велика и представительна, то полученные на ее основе оценки могут характеризовать всю генеральную совокупность. Однако трудно найти критерий, который прояснил бы ситуацию для данной конкретной выборки и для избранного способа обобщения или прогноза. Остается ждать появления новой информации, чтобы

сравнить ее с предсказаниями, полученными на основе модели. Расхождение между эмпирическими наблюдениями и прогнозом может служить естественной мерой качества прогноза, а значит, и модели. В тех случаях, когда оценка качества модели должна быть получена до поступления дополнительной информации, прибегают к делению имеющихся данных на две группы: первую используют для построения модели, а вторую для проверки ее качества. Хотя такой подход давно известен в теории распознавания образов, его проникновение в статистику было нелегким, поскольку искусственное уменьшение объема выборки ведет к уменьшению числа степеней свободы и потому отрицательно сказывается на мощности критериев, на величине доверительных интервалов и т. д., т. е. увеличивается неопределенность результатов.

Более оправданная, но и более трудоемкая процедура, называемая методом «складного ножа», появилась в статистике в 50-е годы. Ее разработка связана с именами М. Кенуя и Дж. Тьюки. Эта процедура начинается с отбрасывания одного из наблюдений, построения модели на массиве оставшихся данных и ее проверки на отброшенном наблюдении. Так последовательно перебираются все наблюдения. Процесс можно продолжить, отбрасывая по два наблюдения, затем по три и так до тех пор, пока не останется «насыщенная» выборка. При этом нет необходимости в полном переборе всех вариантов, достаточно произвести рандомизированную случайную выборку. Слово «выборка» употребляется здесь не по отношению к эксперименту, который фиксирован, а по отношению к вариантам отбрасываемых наблюдений, т. е. происходит управление процессом обработки данных. Так возникла новая область планирования эксперимента.

Это направление получило дополнительный импульс в 1979 г., когда Б. Эфроном был предложен метод «бутстреп», предполагающий многократное тиражирование эмпирической выборки и рандомизированный отбор из такой совокупности большого числа выборок того же объема, что и эмпирическая. По каждой из отобранных таким образом выборок решается та конкретная задача, ради которой проводился эксперимент, а на множестве решений строятся «эмпирические» распределения статистик, интересующих экспериментатора, что дает гораздо больше информации, чем непосредственная оценка.

Таков краткий очерк проблем, связанных с развитием теории и практики регрессионного анализа.

Предлагаемая вниманию читателя книга прежде всего предназначена для специалистов, связанных с приложениями регрессионного анализа. Вместе с тем она может представить интерес и для тех, кто ищет новые пути в такой более широкой и содержательной области, которой является анализ данных.

Ю. АДЛЕР, В. ГОРСКИЙ

Глава 6. ВЫБОР «НАИЛУЧШЕГО» УРАВНЕНИЯ РЕГРЕССИИ

6.0. Введение

Мы отложим обсуждение общей процедуры построения модели до гл. 8, а в данной главе ограничимся рассмотрением только нескольких статистических методов отбора переменных в регрессионном анализе. Предположим, что мы хотим построить линейное регрессионное уравнение для некоторого отклика Y, связанного с главными «независимыми» или предикторными переменными $X_1, X_2, \ldots, X_k$. Предположим далее, что $Z_1, Z_2, \ldots, Z_r$ — все функции от одной или нескольких переменных X и эти функции образуют полный набор переменных, из которых должно формироваться уравнение. Допустим еще, что этот набор включает любые функции, скажем, такие, как квадраты, парные произведения, логарифмы, обратные величины и степени, которые, как можно предположить, желательны и необходимы. Существует два противоположных по смыслу критерия для выбора окончательного уравнения.

1. Если мы хотим сделать уравнение полезным для прогноза, мы должны стремиться включить в него как можно больше переменных Z, с тем чтобы определение прогнозируемых величин стало более надежным.

2. Поскольку затраты, связанные с получением информации и ее последующим контролем при большом числе переменных Z велики, мы должны стремиться к тому, чтобы модель включала как можно меньше величин Z.

Компромисс между этими крайностями как раз и есть то, что обычно называется *выбором «наилучшего» уравнения регрессии*. Для реализации такого выбора нет однозначной статистической процедуры. Если бы мы знали величину σ^2 (истинную дисперсию наблюдений, т. е. дисперсию воспроизводимости) для некоторой хорошо определенной задачи, то выбор наилучшего уравнения регрессии был бы намного легче. К сожалению, мы этого никогда не знаем, и потому субъективные суждения оказываются необходимой составной частью любого из рассматриваемых статистических методов. В этой главе мы опишем несколько предложенных методов. Все они, по-видимому, применяются в настоящее время. Для полноты картины добавим также, что в одной и той же задаче их применение не обязательно ведет к получению одинакового решения, хотя во многих случаях будет получаться тот же самый ответ. Мы обсудим: 1) метод всех возможных регрессий с использованием трех критериев: R^2, s^2 и критерия Маллоуза C_p; 2) метод наилучшего подмножества регрессий с примене-

нием критериев R^2, $\bar{R}^2$ (приведенного) и C_p; 3) метод исключения; 4) шаговый регрессионный метод; 5) некоторые вариации предыдущих методов; 6) гребневую регрессию; 7) ПРЕСС; 8) регрессию на главные компоненты; 9) регрессию на собственные числа и 10) ступенчатый регрессионный анализ. После обсуждения каждого метода мы сформулируем наше мнение о нем.

Некоторые предостережения относительно использования данных пассивного эксперимента

Если регрессионный анализ проводится по данным пассивного эксперимента (т. е. по данным, которые получаются при обычном функционировании объекта, а не в результате специально спланированных экспериментов), то могут возникать некоторые потенциально опасные ситуации, описанные в статье: Box G. E. P. Use and abuse of regression. Technometrics, 8, 1966, p. 625—629. Ошибка в модели может не быть случайной, а оказаться следствием совместного влияния нескольких переменных, не содержащихся в регрессионном уравнении, а возможно, и вовсе неизмеряемых (они называются *скрытыми* (латентными) переменными). Из-за возможного смещения оценок параметров (см. 2.12) наблюдаемый ложный эффект некоторой переменной может провоцироваться фактически неизмеряемой скрытой переменной. Если система продолжает действовать в том же режиме, в котором производилась запись данных, это не вводит в заблуждение. Однако поскольку эта скрытая переменная не измерялась, ее изменения не были видны и не регистрировались; в дальнейшем они могут привести к тому, что предсказания по модели станут ненадежными. Другой дефект данных пассивного эксперимента зачастую состоит в том, что наиболее существенные предикторные переменные изменяются в весьма узких пределах, вследствие чего отклики поддерживаются в определенных границах. Малость этих изменений может стать причиной того, что некоторые коэффициенты регрессии окажутся «статистически незначимыми». Подобный вывод к тому же не удовлетворит и практиков, поскольку они «знают», что эти переменные существенны. Обе точки зрения, конечно совместимы: если эффективная предикторная переменная не варьируется сильно, она будет выглядеть малоэффективной или неэффективной. Третья проблема, возникающая при использовании данных пассивного эксперимента, состоит в том, что распространенная на практике стратегия управления объектами (например, если X_1 повышается, то надо для компенсации снижать X_2) зачастую вызывает значительные корреляции предикторов[1]. Из-за этого невозможно понять, с X_1 или X_2 или с той и другой переменными связано изменение Y. Тщательно спланированный эксперимент может избавить от этих неприятностей. Эффекты скрытых переменных могут быть «рандомизированы», можно выбрать эффективные пределы изменения предикторных переменных и можно избежать корреляций между пре-

[1] Правильнее было бы сказать о большой степени сопряженности между предикторами, поскольку они предполагаются неслучайными. См. примечание к гл. 2 на с. 138, кн. 1.— *Примеч. пер.*

дикторами. В тех случаях, когда планирование экспериментов невозможно, данные случайного происхождения все же можно анализировать с помощью регрессионных методов. Однако надо иметь в виду, что при этом появляются дополнительные обстоятельства, благоприятствующие ошибочным заключениям.

6.1. МЕТОД ВСЕХ ВОЗМОЖНЫХ РЕГРЕССИЙ

Это самая громоздкая процедура. Она вообще не реализуема без быстродействующих вычислительных машин. Поэтому данный метод стал применяться лишь после того, как появились быстродействующие ЭВМ. Он требует прежде всего построения каждого из всех возможных регрессионных уравнений, которые содержат Z_0 и некоторое число переменных $Z_1, \ldots, Z_r$ (где мы, как обычно, добавили фиктивную переменную $Z_0 = 1$ к набору величин Z). Поскольку для каждой переменной Z_i есть всего две возможности: либо входить, либо не входить в уравнение, и это относится ко всем Z_i, $i = 1, 2, \ldots, r$, то всего будет 2^r уравнений. (Будем предполагать, что член Z_0 всегда содержится в уравнении). Если $r = 10$, это вовсе не так много, то надо исследовать $2^r = 1024$ уравнений. Каждое регрессионное уравнение оценивается с помощью некоторого критерия. Мы обсудим далее три критерия:

1) величина R^2, получаемая по методу наименьших квадратов,
2) величина s^2, остаточный средний квадрат и
3) C_p-статистика.

(Все эти критерии фактически связаны друг с другом.) Выбор наилучшего уравнения в таком случае делается на основе оценки наблюдаемой картины, что мы покажем на примере.

Воспользуемся данными для четырехфакторной задачи ($k = 4$), приведенной Хальдом на с. 647 его книги[2] (см.: Hald A. Statistical Theory with Engineering Applications.— New York: J. Wiley, 1952). Именно эта задача была выбрана потому, что она иллюстрирует некоторые типичные трудности регрессионного анализа. Исходные данные приведены на машинных распечатках в приложении Б. Предикторные переменные здесь X_1, X_2, X_3 и X_4. В данной задаче нет никаких преобразований, так что $Z_i = X_i$, $i = 1, 2, 3, 4$. Откликом служит переменная $Y = X_5$. Член β_0 *всегда* включается в модель. Таким образом, имеется $2^4 = 16$ возможных регрессионных уравнений, которые включают X_0 и X_i, $i = 1, 2, 3, 4$. Все они фигурируют в приложении Б. Теперь мы применим процедуры, указанные выше.

Статистика R^2

1. Разделим все варианты на 5 серий (наборов).
Серия А включает только один случай (модель $E(Y) = \beta_0$).
Серия Б состоит из 4 однофакторных уравнений (модель $E(Y) = \beta_0 + \beta_i X_i$).

---

[2] Имеется перевод этой книги на русский язык: Хальд А. Математическая статистика с техническими приложениями/Пер. с англ. Под ред. Ю. В. Линника.— М.: ИЛ, 1956.— 664 с.— *Примеч. пер.*

Серия В включает все двухфакторные уравнения (модель $E(Y) = \beta_0 + \beta_i X_i + \beta_j X_j$).
Серия Г состоит из всех трехфакторных уравнений (модель строится аналогично).
Серия Д — из всех уравнений с четырьмя факторами.

2. Упорядочим варианты внутри каждого набора по значению квадрата множественного коэффициента корреляции R^2.

3. Выявим лидеров и рассмотрим, имеется ли какая-нибудь закономерность среди переменных, входящих в лидирующие уравнения каждой серии. В данном примере мы имеем:

Серия	Переменные в уравнениях	100 R^2, %
Б	$\hat{Y} = f(X_4)$	67,50
В	$\hat{Y} = f(X_1, X_2)$; $\hat{Y} = f(X_1, X_4)$	97,9; 97,2
Г	$\hat{Y} = f(X_1, X_2, X_4)$	98,234
Д	$\hat{Y} = f(X_1, X_2, X_3, X_4)$	98,237

(Заметим, что в серии В имеется 2 лидера с практически одинаковыми значениями величины R^2.) Если мы рассмотрим эти результаты, то увидим, что после введения двух переменных дальнейший прирост величины R^2 мал. Исследуя корреляционную матрицу[3] для этих данных (приложение Б), можно обнаружить, что X_1 и X_3, а также X_2 и X_4 сильно коррелированы. В самом деле (если округлить до третьего знака после запятой)

$$r_{13} = -0{,}824 \quad \text{и} \quad r_{24} = -0{,}973.$$

Следовательно, если X_1 и X_2 или X_1 и X_4 уже содержатся в регрессионном уравнении, дальнейшее добавление переменных очень мало снижает необъясненную вариацию отклика. Отсюда становится ясным, почему величина R^2 так слабо увеличивается при переходе от серии В к серии Г. Прирост R^2 при переходе от серии Г к серии Д совсем уже мал. Это просто объясняется, если заметить, что X есть количества ингредиентов смеси и сумма их значений для любой заданной точки практически постоянна и заключена между 95 и 99.

Какое уравнение следует отобрать для более внимательного рассмотрения. Одно из уравнений серии В, но какое? Если выбрать $f(X_1, X_2)$, то это не совсем оправдано, поскольку наилучшее однофакторное уравнение включает X_4. По этой причине многие авторы отдали бы предпочтение зависимости $f(X_1, X_4)$. Исследование всех возможных уравнений не дает четкого ответа на этот вопрос. Чтобы можно было принять решение, всегда требуется дополнительная инфор-

[3] Эту матрицу правильнее именовать не корреляционной, а матрицей сопряженности (см. примечание к гл. 2 на с. 138, кн. 1).— *Примеч. пер.*

мация, такая, как, например, сведения о характерных особенностях изучаемого продукта и о физической природе переменных X.

Алгоритм The (Algol 60) Algorithm AS 38 (из работы: Garside M. J. Best subset search.— Applied statistics, 1971, **20**, p. 112—115) позволяет быстро найти из всех возможных подмножеств регрессионных моделей те, которые имеют наибольший коэффициент множественной корреляции. Этот метод описан полностью Гарсайдом (Garside) в том же номере журнала на с. 8—15.

Остаточный средний квадрат s^2

Если для некоторой большой задачи построены все регрессионные уравнения, то, рассматривая зависимость величины остаточного среднего квадрата от числа переменных, иногда можно наилучшим образом выбрать число переменных, которые следует сохранить в уравнении регрессии. Различные значения остаточного среднего квадрата по данным Хальда для всех наборов из p переменных, где p—число параметров в модели, *включая* β_0, указаны в распечатках, приведенных в приложении Б.

p	Остаточные средние квадраты	Средний
2	115,06; 82,39; 176,31; 80,35	113,53
3	5,79; 122,71; 7,48; 41,54; 86,89; 17,57*	47,00
4	5,35; 5,33; 5,65; 8,20	6,13
5	5,98	5,98

* Например, 17,57 — остаточный средний квадрат, который получается, если модель содержит X_3 и X_4.

Если число потенциальных переменных для модели велико, скажем, r больше 10, и если число экспериментальных точек значительно больше r, например от $5r$ до $10r$, то график s^2 (p) обычно довольно информативен. Подгонка регрессионных уравнений, которые включают больше предикторных переменных, чем нужно для удовлетворительного согласия экспериментальных и расчетных данных, называется переподгонкой (overfitting). По мере того как к «переподогнанному» уравнению добавляется все больше и больше предикторных переменных, остаточный средний квадрат имеет тенденцию стабилизироваться и приближается к истинной величине σ^2 с ростом числа переменных (при условии, что все важные переменные включены в модель, а число наблюдений значительно, т. е. в пять — десять раз, как указано выше, превосходит число переменных в уравнении). Эта ситуация показана на рис. 6.1. При меньших по объему наборах данных, таких, как в нашем примере, мы не можем, конечно, ожидать, что эта идея окажется плодотворной, но она может привести к полезным заключениям. График зависимости средней величины s_p^2 от p

показан на рис. 6.2. Из него следует, что превосходная оценка величины σ^2 равна примерно 6,00 и что в модель надо включить 4 параметра (т. е. три предикторные переменные). Однако при более детальном рассмотрении остаточных средних квадратов (см. таблицу выше)

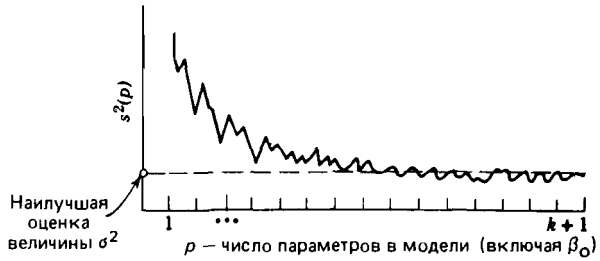


Рис. 6.1. Переподгонка, показывающая типичную стабилизацию s^2

мы видим, что в одном из вариантов при $p = 3$ остаточный средний квадрат составляет 5,79. Отсюда вытекает, что существует лучший вариант с тремя параметрами (т. е. двумя предикторными переменными), чем это вытекает из среднего по остаточным средним квадратам для $p = 3$, которое равно 47,00. Он соответствует фактически уравнению с переменными X_1 и X_2. Следующий лучший вариант при $p = 3$ есть уравнение с переменными X_1 и X_4, для которого $s^2 = 7{,}48$. Таким образом, эта процедура дала нам «асимптотическую» оценку σ^2, на основании которой мы можем выбрать модель или модели, у которых оценка остаточной дисперсии примерно равна 6 и которые включают наименьшее число предикторных переменных.

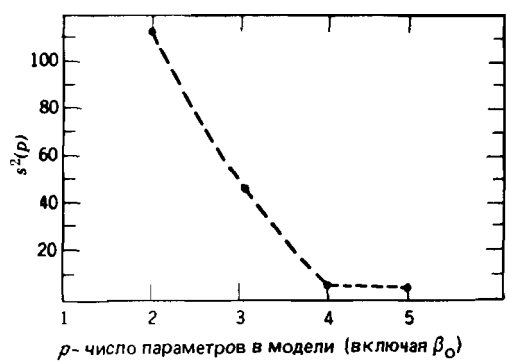


Рис. 6.2. График зависимости среднего из остаточных средних квадратов от p

Критерий Маллоуза C_p

Альтернативная статистика, которая в последние годы получила популярность,— это C_p-статистика, первоначально предложенная Маллоузом. Она имеет вид

$$C_p = \mathrm{RSS}_p/s^2 - (n - 2p), \qquad (6.1.1)$$

где RSS_p — остаточная сумма квадратов для модели, содержащей p параметров, *включая* β_0, а s^2 — остаточный средний квадрат для урав-

нения, содержащего все переменные Z. При этом предполагается, что s^2 является надежной несмещенной оценкой дисперсии σ^2. Как показал Кеннард, величина C_p тесно связана с приведенной R^2-статистикой, R_a^2 и с самой R^2-статистикой; см. уравнения (6.1.1), (2.6.11б) и (2.6.11а). Кроме того, если уравнение с p параметрами адекватно, т. е. наблюдается удовлетворительное согласие экспериментальных и расчетных данных, то $E\,(\mathrm{RSS}_p) = (n-p)\,\sigma^2$. Поскольку мы также предполагаем, что $E\,(s^2) = \sigma^2$ *приблизительно* верно, что отношение RSS_p/s^2 имеет математическое ожидание, равное $(n-p)\,\sigma^2/\sigma^2 = n-p$, откуда опять-таки вытекает, что для адекватной модели приблизительно верно соотношение

$$E\,(C_p) = p.$$

Отсюда следует, что график зависимости C_p от p для адекватной модели будет иметь вид кривой, точки которой довольно близко примыкают к прямой $C_p = p$. В случае уравнений с существенной неадекватностью, т. е. *смещенных уравнений*, возрастает число точек, которые расположены выше (а нередко и заметно выше) линии $C_p = p$. Благодаря случайным вариациям точки для хорошо подогнанных уравнений могут также оказаться ниже линии $C_p = p$. Фактическая величина C_p для каждой точки графика тоже имеет значение, поскольку (это можно показать) она представляет собой оценку полной суммы квадратов расхождений (обусловленных ошибками вариаций плюс ошибки смещения) расчетных значений откликов по подогнанной модели и откликов по истинной, но неизвестной модели. Когда к модели добавляют новые члены, чтобы уменьшить RSS_p, величина C_p обычно возрастает. *Наилучшая* модель выбирается после визуального анализа графика C_p. Мы будем искать регрессию с малым значением C_p, примерно равным p. Если выбор не очевиден, то руководствуются частными соображениями или отдают предпочтение:

1) смешанному уравнению, которое не представляет фактические данные так же хорошо из-за того, что ему соответствует большее значение RSS_p (так что $C_p > p$), но меньшая величина оценки C_p общего расхождения (обусловленного ошибками вариаций и ошибками смещения) с откликами истинной, но неизвестной модели или

2) уравнению с большим числом параметров, которое описывает фактические данные лучше (т. е. $C_p \doteqdot p$), но имеет большее общее расхождение (обусловленное ошибками вариаций и ошибками смещения) с откликами истинной, но неизвестной модели.

Иными словами, «более короткая» модель имеет меньшую величину C_p, но для «более длинной» модели (которая содержит больше членов) величина C_p ближе к p.

Дополнительные указания. Более детальное рассмотрение подобных ситуаций можно найти в книге Даниэля и Вуда (Daniel C., Wood F. S. Fitting Equations to Data. 2nd edition.— New York, J. Wiley, 1980) и в статье Гормана и Томана (Gorman J. W., Toman R. J. Selection of variables for fitting equations to data.— Technometrics, 1966, 8, p. 27—51); см. также работу

Маллоуза (M a l l o w s C. L. Some comments on C_p. — Technometrics, 1973, **15**, p. 661—675). Приведем цитату из последней работы, заслуживающую внимания: «Не следует ожидать, что критерий C_p позволит выбрать одно наилучшее уравнение, если данные существенно неадекватны для такого строгого вывода». Не существует *никакой* другой альтернативы. Все процедуры выбора по существу представляют собой методы упорядоченного представления и рассмотрения данных. Если их применять, руководствуясь здравым смыслом, можно получить полезные результаты. Необдуманное и/или механическое их применение может привести к бесполезным и даже бессмысленным результатам.

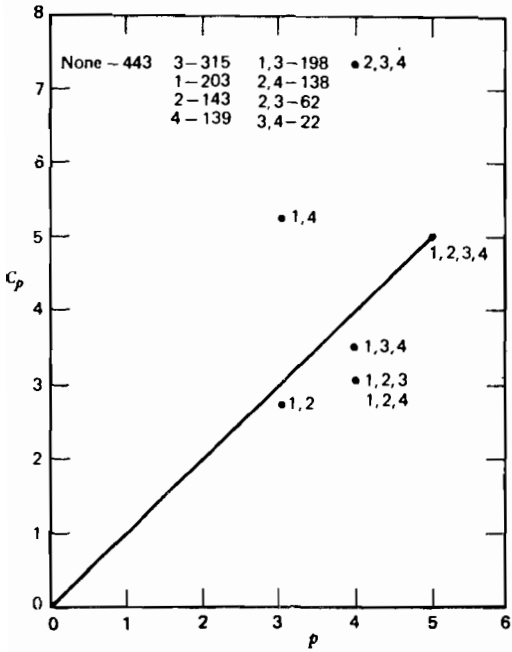


Рис. 6.3. График статистики C_p для данных Хальда

Пример использования C_p-статистики. Согласно данным Хальда (см. приложение Б) мы имеем $n = 13$ и $s^2 = 5{,}983$ для оцениваемой модели, содержащей все 4 предикторные переменные. Так, например, для модели $Y = \beta_0 + \beta_1 X_1 + \varepsilon$ (заметим, что в данном случае $p = 2$) мы получим

$$C_p = 1265{,}687/5{,}983 - (13-4) = 202{,}5.$$

Это значение и все остальные значения критерия C_p указаны в табл. 6.1. Заметим, что для уравнения, содержащего все возможные предикторы, $C_p = p$, что и должно быть справедливо по определению, так как в этом случае $\mathrm{RSS}_p = (n-p)\, s^2$. На рис. 6.3 приведены точки, которым отвечают меньшие значения C_p-статистики. Точки, имеющие большие значения критерия C_p, заметно отстоят от прямой по сравнению с остальными. Поэтому мы можем исключить их из рассмотрения. На основе C_p-статистики мы можем заключить, что уравнение с предикторами X_1 и X_2 является наиболее предпочтительным по сравнению с остальными. Ему не только соответствует наименьшее значение величины C_p, но оно имеет также преимущество по сравнению с моделью, содержащей предикторы X_1 и X_4, которая проявляет признаки смещения. Вывод о том, что уравнение с X_1 и X_2 является предпочтительным, согласуется с тем, что мы решили бы, производя отбор с использованием критериев R^2 и $s^2(p)$, как описано выше. Однако в данном примере такой вывод вытекает до некоторой степени более естественно из графика C_p.

Таблица 6.1. **Величины C_p и p для уравнений по данным Хальда**

Индексы переменных в уравнении	C_p при одинаковом числе параметров	p
	443,2	1
1, 2, 3, 4	202,5; 142,5; 315,2; 138,7	2
12, 13, 14	2,7; 198,1; 5,5	3
23, 24, 34	62,4; 138,2; 22,4	3
123, 124, 134, 234	3,0; 3,0; 3,5; 7,3	4
1234	5,0	5

Пример пользования таблицей. Для уравнения с предикторами X_2 и X_4 в левом столбце указаны индексы 24; величины C_p и p для этого уравнения соответственно равны 138,2 и 3.

Общие замечания. Ранее упоминалось, что данные, использованные в этом примере, подвержены теоретическому ограничению $X_1 + X_2 + X_3 + X_4 = \text{const}$. Из него вытекает, что X_4 теоретически зависит от X_1, X_2 и X_3. Следовательно, если бы в модель были включены все четыре фактора и ограничение выполнялось бы строго, то матрица $\mathbf{X'X}$ была бы вырожденной и имела бы детерминант, равный нулю, как до, так и после преобразования. Как мы видим из соответствующей машинной распечатки, см. с. 301, преобразованный детерминант действительно имеет очень малое значение, 0,0010677. Когда детерминант имеет такое малое значение, нередко оказывается, что вычисления содержат главным образом ошибки округления и потому бессмысленны. И хотя в данном случае этого не произошло, появление малых значений детерминанта всегда должно настораживать (см. 5.5).

Избранные ссылки на работы, где отражены различные аспекты метода всех регрессий, приведены в библиографии.

Мнение. В общем анализ всех уравнений регрессии — довольно ненадежная процедура. Хотя она означает, что статистик «рассмотрел все возможности», но одновременно это значит, что он исследовал большое число регрессионных уравнений, многие из которых при здравом размышлении могли бы быть отвергнуты сразу. Если исследуется более пяти переменных, то затраты машинного времени становятся чрезмерными. Резко возрастают также усилия, связанные с анализом результатов всех вычислений, выведенных на печать. Поэтому более предпочтительны другие методы выбора регрессионного уравнения, которые менее трудоемки.

6.2. МЕТОД ВЫБОРА «НАИЛУЧШЕГО ПОДМНОЖЕСТВА» ПРЕДИКТОРОВ

Существуют прекрасные вычислительные алгоритмы выбора наилучших наборов предикторных переменных в регрессии. Среди них популярен алгоритм, предложенный в статье: Furnival G. M.,

Wilson R. W. Regression by leaps and bounds. — Technometrics, 1974, **16**, p. 499—511. В этом алгоритме обрабатывается только часть всех возможных регрессий при определении наилучшего набора, включающего K уравнений, так называемого «K-подмножества». Для определения этого наилучшего K-подмножества могут использоваться три критерия, а именно:

1) максимум величины R^2;
2) максимум приведенной величины R^2 (см. уравнение (2.6.11в));
3) критерий C_p Маллоуза.

В пакете BMDP (см. с. 60) соответствующая программа обозначена как P9R, All Possible Subsets Regression. Пользователь назначает число K, т. е. число отбираемых наилучших регрессий, и сам критерий, по которому будет производиться отбор. Программа определяет наилучшее подмножество, включающее K регрессий, из всего множества возможных регрессий. (На машинных распечатках указываются все три критерия, но выбор наилучшего подмножества производится на основе какого-нибудь одного из них.) В распечатке приводятся также наилучшие выборы из K регрессий, включающих одну, две и более предикторных переменных. Вплоть до единственного уравнения, содержащего все предикторные переменные. В каждом из этих частных подмножеств выделяется наилучшее уравнение (уравнения), из которых формируется наилучшее общее подмножество, содержащее K регрессий. Если выбранное число K превосходит число уравнений, из которых может быть образовано некоторое частное подмножество, то в K-подмножество включаются все эти уравнения. Это будет ясно из примера, построенного на данных Хальда (см. приложение Б), где мы выбрали $K = 5$. Там приведены значения всех трех критериев, но выбор наилучших уравнений производится с помощью критерия C_p. В конце программа указывает характеристики наилучшего уравнения из всех наилучших K-подмножеств. По данным Хальда при $K = 5$ с помощью программы BMDP9R получена следующая машинная распечатка.

Критерии (R^2, приведенный R^2 и C_p), регрессионные коэффициенты и их t-статистики указываются для пяти наилучших наборов предикторов. Критерии вычислены также для многих других наборов, некоторые из которых могут быть также довольно хорошими. Однако они не обязательно лучше, чем некоторые из наборов, которые здесь не отражены.

. . . Регрессии с одной предикторной переменной . . .

R^2	Приведенный R^2	C_p	Переменные в наборе
0,67454	0,64495	138,73	4
0,66627	0,63593	142,49	2
0,53395	0,49158	202,55	1
0,28587	0,22095	315,15	3

. . . Регрессии с двумя предикторными переменными . . .

R^2	Приведенный R^2	C_p	Переменные в наборе
0,97868	0,97441	2,68	1,2

Это один из 5 наилучших наборов предикторов

Переменная	Коэффициент	t-статистика
1×1	0,146831D01	12,10
2×2	0,662250D00	14,44

Свободный член 0,525773D02

R^2	Приведенный R^2	C_p	Переменные в наборе
0,97247	0,96697	5,50	1,4
0,93529	0,92235	22,37	3,4
0,68006	0,61607	138,23	2,4
0,54823	0,45787	198,07	1,3

. . . Регрессии с тремя предикторнымн переменными

R^2	Приведенный R^2	C_p	Переменные в наборе
0,98234	0,97645	3,02	1, 2, 4

Это один из 5 наилучших наборов предикторов

Переменная	Коэффициент	t-статистика
1×1	0,145194D01	12,41
2×2	0,416110D00	2,24
4×4	—0,236540D00	—1,37

Свободный член 0,716483D02

0,98228	0,97638	3,04	1, 2, 3

Это один из 5 наилучших наборов предикторов

Переменная	Коэффициент	t-статистика
1×1	0,169589D01	8,29
2×2	0,656915D00	14,85
3×3	0,250017D00	1,35

Свободный член 0,481936D02

0,98128	0,97504	3,50	1, 3, 4

Это один из 5 наилучших наборов предикторов

Переменная	Коэффициент	t-статистика
1×1	0,105185D01	4,70
3×3	—0,410043D00	—2,06
4×4	—0,642796D00	—14,43

Свободный член 0,111684D03

0,97282	0,96376	7,34	2, 3, 4

. . . Регрессии с четырьмя предикторными переменнымн . . .

R^2	Приведенный R^2	C_p	Переменные в наборе
0,98238	0,97356	5,00	1, 2, 3, 4

Это один из 5 наилучших наборов предикторов

Переменная	Коэффициент	t-статистика
1×1	0,155110D01	2,08
2×2	0,510170D00	0,70
3×3	0,101911D00	0,14
4×4	—0,144059D00	—0,20

Свободный член 0,624052D02

В процессе нахождения наилучших наборов предикторов было вычислено 10 рег-

рессий. Было выполнено 38 умножений и делений (исключая вычисления, связанные с ковариационной матрицей).

Статистики для наилучшего набора	
Статистика Маллоуза	2,68
Квадрат множественного коэффициента корреляции	0,97868
Множественный коэффициент корреляции	0,98928
Квадрат приведенного множественного коэффициента корреляции	0,97441
Остаточный средний квадрат	0,579044D01
Стандартная ошибка оценки	0,240633D01
F-статистика	229,50
Число степеней свободы для числителя	2
Число степеней свободы для знаменателя	10
Значимость	0,0000

№	Переменная Обозначение	Коэффициент регрессии	Стандартная ошибка	Стандартизированный коэффициент	t-статистика	Двусторонняя значимость	Толерантность
	Свободный член	0,525773D02	0,228617D01	3,495	23,00	0,000	
1	X1	0,146831D01	0,121301D00	0,574	12,10	0,000	0,947751
2	X2	0,662250D00	0,458547D—01	0,685	14,44	0,000	0,947751

Мнение. Эта процедура имеет некоторые недостатки: (1) Она имеет склонность к выбору уравнений (входящих в наилучшее общее подмножество), которые содержат слишком много предикторов. (2) Если величина K выбирается малой, то наиболее подходящее уравнение может не войти в наилучшее общее подмножество моделей, хотя оно может фигурировать где-то в машинной распечатке. (3) В распечатке не содержится никакой подходящей информации относительно того, как получались различные наборы. Однако если принять во внимание эти особенности процедуры, программа такого типа может иметь большую ценность, и мы рекомендуем использовать этот метод в сочетании с методом шаговой регрессии, если желательно исследовать уравнения, «близкие» к наилучшему.

6.3. МЕТОД ИСКЛЮЧЕНИЯ

Метод исключения более экономичен, чем метод всех регрессий, поскольку в нем *делается попытка* исследовать только наилучшие регрессионные уравнения, содержащие определенное число переменных [4]. Основные шаги этого метода сводятся к следующему.

1. Рассчитывается регрессионное уравнение, включающее все переменные.

2. Вычисляется величина частного F-критерия для каждой предикторной переменной[*1] *в предположении как будто бы она была последней переменной, введенной в регрессионное уравнение.*

[4] Точнее, определенное число слагаемых в модели.— *Примеч. пер.*

[*1] Частный F-критерий связан с проверкой гипотезы $H_0 : \beta = 0$ против альтернативы $H_1 : \beta \neq 0$ для любого отдельного коэффициента регрессии. Однако это слишком вольная формулировка, в которой мы говорим об использовании F-статистики для отдельной предикторной переменной. Тем не менее она удобна, и мы иногда ею пользуемся.

3. Наименьшая величина частного F-критерия, обозначаемая, скажем, как F_L, сравнивается с заранее выбранным критическим значением, например F_0.

а) Если $F_L < F_0$, то переменная Z_L, которая обеспечила достижение только уровня F_L, исключается из рассмотрения и производится перерасчет уравнения регрессии с учетом остающихся переменных; затем переходят к следующему шагу.

б) Если $F_L > F_0$, то регрессионное уравнение оставляют таким, как оно было рассчитано.

На тех же данных Хальда (Hald, 1952), что и в предыдущем параграфе, мы проиллюстрируем теперь этот метод. Поскольку никакие преобразования предикторных переменных здесь не используются, $Z_i = X_i$, мы будем применять те же обозначения для переменных X, что и ранее.

Сначала получим полное регрессионное уравнение для всех предикторных переменных. В примере, который рассмотрен в 6.1, мы таким образом нашли МНК-уравнение $\hat{Y} = f(X_1, X_2, X_3, X_4)$. Анализ этой модели показан в приложении Б, с. 301. Поскольку матрица $\mathbf{X'X}$ невырожденная, полученная в итоге остаточная дисперсия служит хорошей оценкой величины σ^2 в асимптотическом смысле *2, как об этом говорилось в связи с рис. 6.1. Метод исключения по существу реализует попытку удалить все ненужные переменные X без существенного увеличения значения «асимптотической» оценки σ^2. Чтобы проверить переменные на этом шаге, необходимо определить вклад каждой переменной из набора X_1, X_2, X_3 и X_4 в регрессионную сумму квадратов так, как будто данная переменная была включена в уравнение последней. Значения частных F-критериев, служащих мерами вкладов этих переменных, указаны в последнем столбце машинной распечатки.

Теперь мы выберем наименьшую величину частного F-критерия и сравним ее с критическим значением F-статистики, основанным на определенном уровне значимости α. В данном случае критическая величина, например, для $\alpha = 0{,}10$ равна $F\ (1;\ 8;\ 0{,}90) = 3{,}46$. Наименьшее значение частного F-критерия отвечает переменной X_3 и равно $F = 0{,}018$. Так как вычисленное значение F меньше критической величины, равной 3,46, переменная X_3 исключается.

Затем найдем МНК-уравнение $\hat{Y} = f(X_1, X_2, X_4)$. Оно показано на с. 298. Полный F-критерий для уравнения равен $F = 166{,}83$. Эта величина статистически значима, поскольку она превосходит $F\ (3;\ 9;\ 0{,}999) = 13{,}90$. Исследуя это уравнение с целью последующего возможного исключения переменных, мы увидим, что величине X_4 соответствует наименьшее значение частного F-критерия, и эта переменная является кандидатом на исключение. Процедура такого элиминирования подобна описанной выше с одним лишь отличием: кри-

*2 Заметим, что метод, основанный на использовании критерия C_p, реализует ту же самую идею. Оценки величины σ^2 по различным моделям сравниваются с оценкой этой величины для полной модели. Удовлетворительной моделью признается та, для которой эти оценки примерно равны.

тическое значение величины F составляет F (1; 9; 0,90) = 3,36. Поскольку рассчитанная величина частного F-критерия, связанного с X_4, равна 1,86 (что меньше 3,36), мы исключаем X_4.

Теперь мы найдем МНК-уравнение $\hat{Y} = f(X_1, X_2)$, показанное на с. 289. Полное уравнение статистически значимо, поскольку соответствующая ему величина F равна 229,50 и превосходит критическое значение F (2; 10; 0,999) = 14,91. При этом значимы обе переменные X_1 и X_2 безотносительно к порядку, в котором они входят в модель, поскольку частные F-критерии в обоих случаях превосходят 14,91. Таким образом, процедура выбора уравнения методом исключения закончена и получено уравнение

$$\hat{Y} = 52{,}58 + 1{,}47X_1 + 0{,}66X_2.$$

Мнение. Это вполне удовлетворительная процедура, особенно для статистиков, которые любят видеть все переменные в уравнении, чтобы «чего-то не упустить». Этот метод значительно более экономичен по затратам машинного времени и труда, чем метод всех регрессий. Однако если из исходных данных получается плохо обусловленная матрица $\mathbf{X'X}$, т. е. почти вырожденная, то уравнение может быть бессмысленным из-за ошибок округления. Если использовать современные методы обращения матриц, то это обычно не становится серьезной проблемой. Важно иметь в виду, что, как только переменная исключается из уравнения с помощью этого метода, она элиминируется безвозвратно. Таким образом, все другие методы, основанные на использовании исключаемых переменных, здесь непригодны.

(Примечание. Резюмируем положения, о которых шла речь в тексте.

1) В некоторых программах, базирующихся на этом методе, вместо F-критерия используется t-критерий, представляющий собой корень квадратный из величины частного F-критерия. Это связано с тем фактом, что если F (1, ν) — случайная величина, подчиняющаяся F-распределению с 1 и ν степенями свободы, а t (ν) — случайная величина, подчиняющаяся t-распределению с ν степенями свободы, то F (1, ν) = $= t^2$ (ν) (см. с. 138, кн. 1).

2) В некоторых программах используется термин «F-критерий для исключения» («F to remove»). Он идентичен используемому нами термину «частный F-критерий» (см. с. 138, кн. 1.)

6.4. ШАГОВЫЙ РЕГРЕССИОННЫЙ МЕТОД

Метод исключения начинается с наиболее полного уравнения, включающего все переменные, и состоит в последовательном уменьшении числа переменных до тех пор, пока не принимается решение об использовании уравнения с оставшимися членами. Шаговый метод представляет собой попытку прийти к тем же результатам, действуя в обратном направлении, т. е. включая переменные по очереди в уравнение до тех пор, пока уравнение не станет удовлетворительным. Порядок включения определяется с помощью частного коэффициента корреляции как меры важности переменных, еще не включенных в

уравнение. Основная процедура состоит в следующем. Прежде всего выбирается величина Z, наиболее сильно коррелированная с Y (предположим, что это Z_1), и находится линейное, первого порядка регрессионное уравнение $\hat{Y} = f(Z_1)$. Затем мы проверяем, значима ли эта переменная. Если это не так, то мы должны согласиться с выводом, что наилучшая модель выражается уравнением $Y = \bar{Y}$. В противном случае мы должны найти вторую предикторную переменную [5], которую следует включить в модель. Мы определяем частные коэффициенты корреляции [*3] для всех предикторов, не включенных в уравнение на этом шаге, а именно для Z_j, $j \neq 1$, с Y с учетом поправки на Z_1. В математическом отношении это эквивалентно нахождению корреляции между (1) остатками от регрессии $\hat{Y} = f(Z_1)$ и (2) остатками от каждой из регрессий $\hat{Z}_j = f_j(Z_1)$ (которые мы фактически не определяли). Теперь выбирается величина Z_j (предположим, что это Z_2), которая имеет наибольшее значение частного коэффициента корреляции с Y, и находится второе регрессионное уравнение $\hat{Y} = f(Z_1, Z_2)$. Полное уравнение проверяется на значимость. Отмечается улучшение величины R^2 и исследуются частные F-критерии для *обеих переменных, содержащихся в уравнении*, а не только для той, которая только что была введена в уравнение [*4]. Наименьшая величина из этих двух частных F-критериев сравнивается затем с подходящей процентной точкой F-распределения. Соответствующая предикторная переменная сохраняется в уравнении или исключается из него в зависимости от результатов проверки. Такая проверка «наименее полезного предиктора в уравнении на данном этапе» проводится на каждом шаге этого метода. Может оказаться, что предиктор, который на предыдущем шаге был наилучшим кандидатом для включения в уравнение, на более позднем шаге оказывается ненужным. Это может быть вызвано теми связями, которые существуют между этой и другими переменными, содержащимися теперь в уравнении. Чтобы проверить это, на каждом шаге для каждой предикторной переменной, содержащейся в уравнении, вычисляется частный F-критерий и находится наименьший из них (он может быть связан с любой предикторной переменной, включениой в модель только что или ранее), ко-

---

[5] Авторы называют здесь предикторными, переменными величины Z_i. Это верно в данном случае, поскольку предполагается, что $Z_i = X_i$. — *Примеч. пер.*

[*3] Во многих пакетах программ по регрессионному анализу частные коэффициенты корреляции или их квадраты не вычисляются и на печать не выводятся. Вместо них вычисляется соответствующая F-статистика для включения каждого предиктора, не содержащегося в модели на данном шаге. Это дает по существу ту же самую информацию — наибольшее значение F-критерия для включения, которое связано со следующим кандидатом на введение в модель.

[*4] Проще, но менее эффективна процедура, в которой проверяется только тот предиктор, который включен в уравнение последним. Эту процедуру называют методом включения. Она была описана в первом издании данной книги, и ее можно рассматривать как один из вариантов шагового метода. Метод включения приводит к тому, что переменные, введенные в модель, не исключаются из нее в дальнейшем, хотя это может быть желательным в определенных приложениях.

торый затем сравнивается с заранее выбранной процентной точкой соответствующего F-распределения. Это позволяет судить о вкладе наименее ценной переменной в регрессию на данном шаге в предположении, что она только что была введена в модель безотносительно к тому, как это было на самом деле. Если проверяемая переменная показывает незначимый вклад в регрессию, она исключается из уравнения. После этого регрессионное уравнение пересчитывается с учетом всех оставшихся в нем предикторных переменных. Наилучшие переменные из тех, которые не вошли на данном шаге в модель (т. е. те, для которых коэффициент частной корреляции с Y при наличии предикторов в уравнении получился наибольшим), затем проверяются, чтобы убедиться, удовлетворяют ли они частному F-критерию для включения. Если удовлетворяют, их включают в уравнение и снова возвращаются к проверке всех частных F для переменных. Если же они не выдерживают этой проверки, переходят к следующей операции исключения. В конечном счете (если только уровень значимости α не выбран плохо, что приводит к зацикливанию *5) процесс прекращается, если никакие из переменных, содержащихся в текущем уравнении, не удается исключить из него, а ближайший наилучший предиктор-претендент не в состоянии занять место в уравнении. Когда переменная включается в регрессию, ее влияние на R^2, квадрат множественного коэффициента корреляции, обычно указывается в машинной распечатке.

Мы снова воспользуемся данными Хальда, чтобы проиллюстрировать, как работает шаговая процедура. (См. распечатку, где указано, что $Y = X_5$ и $Z_i = X_i$, $i = 1, 2, 3, 4$.) Для обоих критериев включения и исключения принят уровень значимости $\alpha = 0{,}10$.

1. Вычислим коэффициенты корреляции между каждой предикторной переменной и откликом. Выберем в качестве первой переменной для включения в регрессию ту, которая коррелирована с откликом наиболее сильно. Исследование корреляционной матрицы в приложении Б показывает, что X_4 наиболее сильно коррелирована с откликом Y или X_5; $r_{45} = -0{,}821$. Следовательно, X_4 это первая переменная, которая должна быть включена в регрессионное уравнение.

2. Построим регрессию Y в зависимости от X_4 и получим МНК-уравнение, приведенное на с. 288. Полный F-критерий показывает, что регрессия значима. Таким образом, переменная X_4 сохраняется в уравнении.

3. Вычислим частные коэффициенты корреляции между всеми переменными, не входящими в уравнение, и откликом. Их квадраты указаны внизу на с. 288. Выберем в качестве следующей переменной

---

*5 Обычно лучше всего выбирать одинаковые уровни значимости α для включения и исключения. Если для критерия исключения принимается меньший уровень значимости α, чем для критерия включения, то может возникать зацикливание процедуры. Использование больших значений α для критерия исключения также нецелесообразно, поскольку это может привести к сохранению в уравнении переменных, вклад которых в регрессию незначителен. Некоторые авторы считают такое положение желательным, однако это дело вкуса (см. также с. 26—28).

для включения в регрессионное уравнение переменную с наибольшим значением частного коэффициента корреляции. Это переменная X_1; $r^2_{15.4} = 0{,}915$.

4. Получим МНК-уравнение $Y = f(X_1, X_4)$, содержащее как X_1, так и X_4, см. с. 291. Этому уравнению соответствует $R^2 = 97{,}2\ \%$, и оно явно значимо, поскольку величина полного F-критерия равна $F = 176{,}63$. А это превосходит $F\ (2;\ 10;\ 0{,}90) = 4{,}10$. То, что новая переменная X_1 дает значимое снижение остаточной суммы квадратов, показывает ее частный F-критерий, который равен 108,22, что превосходит величину $F\ (1;\ 10;\ 0{,}90) = 4{,}96$. Таким образом, X_1 остается в уравнении. Мы проверим также вклад X_4 в предположении, что величина X_1 *будто бы* была включена в модель первой, а переменная X_4 — второй. Поскольку величина частного F-критерия равна 159,295 (см. с. 291), что значительно превосходит $F\ (1;\ 10;\ 0{,}90) = 4{,}96$, переменная X_4 сохраняется в уравнении. (На практике в большинстве программ не проверяются обе переменные, как это делается здесь, а выбирают переменную с наименьшим значением частного F-критерия и проверяют ее. Принимается решение об исключении или сохранении соответствующей предикторной переменной. При исключении уравнение пересчитывается, а при сохранении ищется следующий кандидат.)

5. Согласно шаговому методу теперь для включения в уравнение выбирается следующая переменная, которая имеет наиболее высокий частный коэффициент корреляции с откликом (при условии, что переменные X_4 и X_1 уже содержатся в регрессии). Как видно, это переменная X_2. (Квадрат частного коэффициента корреляции предиктора X_2 с откликом равен 0,358 — см. с. 291.)

6. Новое уравнение $\hat{Y} = f(X_4, X_1, X_2)$ приведено на с. 298. Квадрат множественного коэффициента корреляции, R^2, выраженный в %, увеличился с 97,2 до 98,2 %. Затем на этом шаге исследовались частные F-критерии для переменных X_1 X_2 и X_4. Наименьшее значение $F = 1{,}863$ (см. с. 298) соответствует X_4. И поскольку эта величина меньше, чем $F\ (1;\ 9;\ 0{,}90) = 3{,}36$, переменная X_4 отвергается. В уравнении, которое пересчитывается (с. 289), сохраняются переменные X_1 и X_2 как значимые.

7. Единственная остающаяся переменная, которая может рассматриваться на этом этапе, есть X_3. Поскольку эта переменная немедленно отвергается, шаговая регрессионная процедура заканчивается, и как наилучшее выбирается уравнение $\hat{Y} = f(X_1, X_2)$, показанное на с. 289, а именно

$$\hat{Y} = 52{,}58 + 1{,}47X_1 + 0{,}66X_2.$$

Мнение. Мы считаем этот метод одним из лучших среди обсуждавшихся выше и рекомендуем его применять. Он наиболее экономичен при обработке данных на ЭВМ. К тому же он позволяет избежать манипуляций с бо́льшим числом предикторов, чем это необходимо, хотя уравнение продолжает улучшаться с каждым шагом. Однако шаговый метод может легко стать обузой для профессиональ-

ного статистика. Как и во всех других обсуждавшихся методах, здесь требуются все же разумные суждения при первоначальном выборе переменных и при критическом анализе модели путем исследования остатков. Можно полагать, что использование этого метода для автоматического выбора наилучшего уравнения с помощью ЭВМ будет слишком затруднительным. Обсуждение этого метода дано в статье: Efroymson M. A. Multiple regression analysis, в книге: Mathematical Methods for Digital Computers/Ralston A. and Wilf H. S., ed.— New York: J. Wiley, 1962.

УРОВНИ ЗНАЧИМОСТИ В ШАГОВОМ РЕГРЕССИОННОМ МЕТОДЕ

При выполнении предыдущего примера мы назначали процентную точку для F-критерия исходя из $\alpha = 0{,}10$. Однако на четвертом шаге величина F была достаточно большой, чтобы превзойти даже 99,9 %-ную точку ($\alpha = 0{,}001$), в то время как на шестом шаге F-критерий для X_4 при наличии в уравнении X_1 и X_2 был недостаточно велик даже для того, чтобы превысить 90 %-ную точку ($\alpha = 0{,}10$). Обычно выбирается некоторый фиксированный уровень, такой, как, скажем, 95 % ($\alpha = 0{,}05$), и он используется без изменения на всех шагах процедуры. Можно также устанавливать различные уровни для критериев включения и исключения. Однако при этом неразумно принимать «α для исключения» меньше «α для включения», а также исключать предикторы, только что включенные в уравнение. Некоторые авторы предпочитают выбирать «α для исключения» на более высоком уровне, чем «α для включения», чтобы отдать предпочтение тем предикторам, которые уже включены в уравнение. Такие вариации зависят от личных вкусов исследователей, однако они наряду с фактически выбранными значениями величин α оказывают большое воздействие на то, каким образом будет вести себя данная процедура отбора и как много предикторов останется в итоговом уравнении. Некоторые исследователи вообще игнорируют таблицы F-критерия и просто сравнивают значение F-критерия с произвольным числом, скажем, 4. Мы предлагаем читателям, не имеющим каких-то обоснованных суждений на этот счет, принимать $\alpha = 0{,}05$ или $\alpha = 0{,}10$ как для критерия включения, так и для критерия исключения, если используемый пакет программ регрессионного анализа допускает такой выбор, как это имеет, например, место в пакете SAS. Эти уровни потом можно изменять, как подскажет опыт. В следующем параграфе мы увидим, что уровни значимости α — это по крайней мере не строгие характеристики, так что едва ли стоит прилагать большие усилия для их точного выбора. Обычно принято выбирать α на уровне $\alpha = 0{,}05$, т. е. фактическая величина α гораздо больше, чем 0,05 (это вытекает из некоторых предварительных исследований, посвященных данному вопросу), что влечет за собой тенденцию включать в уравнение больше предикторов, чем это может предвидеть пользователь.

Существует ряд пакетов программ шаговой регрессии. В частности, пакет BMDP дает замечательные результаты, он допускает раз-

личные возможности выбора. Шаговая процедура применительно к данным Хальда, обработанным в соответствии со специальными критериями, указанными ранее, описана в приложении В, с. 303—312.

6.5. НЕДОСТАТОК, КОТОРЫЙ СЛЕДУЕТ ПОНЯТЬ, НЕ ПРИДАВАЯ ЕМУ СЛИШКОМ БОЛЬШОГО ЗНАЧЕНИЯ

Как метод исключения, так и шаговый регрессионный метод страдают недостатком, который не очевиден на первый взгляд. Так, например, в шаговой процедуре проверка по частному F-критерию на стадии включения переменной производится для того предиктора, который имеет наибольшее значение частного F среди всех предикторов, не входящих в регрессию в данный момент. Корректное «нуль-распределение» (т. е. распределение в случае справедливости нулевой гипотезы) этой статистики, как мы полагаем, будет отнюдь не обычным F-распределением и для выборочной, и для теоретической статистики, а получить его очень трудно, за исключением нескольких простейших ситуаций. Исследования показывают, например, что в некоторых случаях, когда проверка с помощью F-критерия при включении переменных производилась при уровне значимости α, соответствующая вероятность была равна $q\alpha$, где q — число кандидатов на включение, которые имелись на этой стадии. Что можно сделать в связи с этим? Одна возможность состоит в определении правильных уровней значимости для любого данного случая. Другая заключается в использовании иных статистических критериев вместо частного F-критерия. Обе эти возможности обсуждались в недавних публикациях, но к моменту написания нашей книги проблема полностью и надлежащим образом еще не была решена, в том смысле, чтобы можно было гарантировать улучшение процедуры. Пока такое решение не найдено, мы предлагаем читателю применять процедуры в описанном виде, не придавая слишком большого значения действительным уровням вероятности, а просто рассматривая весь метод как проведение серии сравнений, которые позволяют выявлять, по-видимому, наиболее полезные наборы предикторов. Для тех, кто желает углубиться в проблему более основательно, мы приводим в конце книги некоторые избранные ссылки. Этим читателям следует также просмотреть последние выпуски основных статистических журналов, где могут содержаться другие работы.

Преодолеть указанную выше трудность проще всего, назначив в программах заранее значения F для включения и исключения (так, например, можно принять $F = 4$). Такой подход описывается Форсайтом в сборнике программ (BMDP-79, Biomedical Computer Programs, P-Series/D i x o n W. J., B r o w n M. B., Eds.— Berkeley: University of California Press, 1979, Appendix C, p. 855) следующим образом:

«Некоторые пользователи, применяющие программы шагового регрессионного и дискриминантного анализа, спрашивают: почему мы всюду используем термины «F для включения» и «F для исключения» вместо того, чтобы просто называть их величинами F. Другие предла-

гают, чтобы мы запрашивали у пользователя уровень значимости α и имели программу, позволяющую преобразовать эту величину в соответствующее значение F. Составить такую программу для ЭВМ в принципе достаточно просто, но это нелегко сделать статистически корректно, поскольку при выборе «наилучшей» переменной обычные таблицы F-критерия неприменимы. Подходящее критическое значение есть функция числа вариантов, числа переменных и, к несчастью, характера коррелированности предикторных переменных. Это означает, что уровень значимости, соответствующий F-критерию для включения, зависит от конкретного набора данных. Так, например, в случае нескольких сотен опытов и 50 потенциальных предикторов F-критерий для включения, равный 11, приближенно соответствовал бы $\alpha = 5$ %, если бы все 50 предикторов были некоррелированными. В обычно используемых F-таблицах ошибочно предлагается величина F, равная 4.»

(Добавим к этому, что использование значения, равного 4, не будет неправильным, если просто принимать α более высоким, чем мы обсуждали ранее. Однако зачастую принимают значения *по существу* более высокие. Так, например, в указанном выше случае при номинальном значении $\alpha = 0{,}05$ и $m = 50$ (50 некоррелированных предикторов) «действительная» величина α, определяемая по формуле $1 - (1-\alpha)^m$, равнялась бы 0,923. Эта формула может быть полезна в качестве грубого ориентира. Заметим, что пакет программ BMDP периодически пересматривается. В более поздних версиях учитываются возможные изменения метода.)

6.6. ВАРИАЦИИ ПРЕДЫДУЩИХ МЕТОДОВ

Хотя обсуждавшиеся ранее методы и не обеспечивают с абсолютной точностью выбор наилучшей модели, обычно они тем не менее позволяют выбрать подходящую модель. Поэтому для улучшения выбора модели были предложены некоторые другие методы, основанные на комбинации рассмотренных приемов. Обсудим теперь два таких метода.

1. Первое предложение сводится к следующему: проведите шаговую регрессионную процедуру с заданными уровнями значимости для включения и исключения. По окончании процедуры определите число переменных в итоговой модели. Используя это число, равное, например, q, найдите возможные наборы, содержащие q переменных из r исходных переменных, и выберите наилучший набор.

Мнение. Этот метод позволяет обнаружить ситуацию, отмеченную при обсуждении данных Хальда для двухфакторных случаев, а именно когда имеются два «кандидата для включения в модель» вместо одного. Если это имеет место, то можно сказать, что данные содержат недостаточно информации для однозначного выбора. Чтобы окончательно выбрать модель, требуются дополнительные априорные соображения и здравый смысл экспериментатора. Этот метод становится также несостоятельным, если модель можно улучшить за счет добавления переменных, которые не исследовались с помощью

данного алгоритма. Наш опыт показывает, что преимущества, которые дает эта процедура, незначительны, а дополнительных вычислений здесь много.

2. Второе предложение состоит в использовании шагового метода с менее ограниченными уровнями значимости для включения и исключения (т. е. с большими значениями α), что приводит к включению в модель нескольких дополнительных переменных сверх тех, которые были бы включены при меньших значениях уровней значимости. Это позволяет исследовать дополнительные переменные, которые не включаются в модель при использовании обычного шагового метода, и может привести к получению другой итоговой модели.

Мнение. В некоторых случаях такая процедура оказалась полезной, т. е. она приводила к получению другой модели с приблизительно такими же характеристиками в смысле предсказания. Наш опыт, однако, показывает, что это имеет место, если задача почти не допускает решения из-за очень высокой взаимной корреляции между предикторными переменными (Z) и, следовательно, требует большего, чем только использования статистических методов, отсеивания. См. также комментарии в § 6.5.

6.7. ГРЕБНЕВАЯ (РИДЖ) РЕГРЕССИЯ

Процедура с использованием *«следа гребня»* (ridge trace) была впервые предложена Херлом в 1962 г. и обсуждалась через некоторое время Херлом и Кеннардом в статье: Hoerl A. E., Kennard R. W. Ridge regression: biased estimation for nonorthogonal problems. — Technometrics, 1970, **12**, p. 55—67. Вторая статья этих авторов с примерами была опубликована в том же номере журнала на с. 69—82. Эта процедура предназначена для «плохо обусловленных» ситуаций, когда имеются значительные корреляции [6] между разными предикторами, входящими в модель, вследствие чего матрица $\mathbf{X'X}$ становится почти вырожденной и оценки параметров — неустойчивыми [7]. Оценки могут иметь, например, неправильный знак или иметь значения, которые намного превосходят те, что приемлемы из физических или практических соображений. С обсуждением такого рода ситуации можно познакомиться в статье Мулле: Mullet G. M. Way regression coefficients have the wrong sign.— Journal of Quality Technology, 1976, 8; p. 121—126.

Метод гребневой регрессии в его простейшей форме состоит в следующем. Пусть $\mathbf{Z}$ представляет собой подходящим образом центрированную и нормированную матрицу $\mathbf{X}$, соответствующую случаю, когда исследуемая регрессионная задача выражена в «корреляционной форме» (см. обсуждение в гл. 5). Тогда для модели, содержащей все возможные r предикторов $Z_1, Z_2, \ldots, Z_r$, можно получить оценки параметров $\mathbf{b}_z(\theta)$ по формуле

$$\mathbf{b}_z(\theta) = (\mathbf{Z'Z} + \theta \mathbf{I}_r)^{-1} \mathbf{Z'Y}, \qquad (6.7.1)$$

[6] См. примечание к гл. 2 на с. 138, кн. 1.— *Примеч. пер.*

[7] В этом случае крайне затруднительно получить единственное решение.— *Примеч. пер.*

где θ — положительное число. (В приложениях эта величина обычно лежит в интервале (0,1).) Исключения описаны в статье Брауна и Пэйна (B r o w n R. J., P a y n e C. Election night forecasting.— Journal of Royal Statistical Society, **A**—138, p. 463—483, с дискуссией на с. 483—498). Заметим, что в формуле (6.7.1) вектор $\mathbf{Y}$ представлен в обычной, а не в корреляционной форме, а $\mathbf{b}_z(\theta)=\{b_{1z}(\theta),\ b_{2z}(\theta),\ \ldots,\ b_{rz}(\theta)\}'$ есть $(r\times 1)$-вектор, не содержащий оценки свободного члена. Легко понять, что никакой поправки на свободный член не требуется, поскольку замена $\mathbf{Y}$ в уравнении (6.7.1) на $\mathbf{Y}-\bar{Y}\mathbf{1}$ (где $\mathbf{1}=(1, 1, \ldots, 1)'$) совсем не влияет на результаты, поскольку «центрирование» предикторов обеспечивает выполнение соотношения $\mathbf{Z1}=\mathbf{0}$. Как указано на с. 323, кн. 1, поскольку вектор $\mathbf{Y}$ выражен не в корреляционной форме *6, коэффициент $S^{1\,2}_{YY}$ использовать не надо, и мы можем выполнить преобразования

$$b_j(\theta)=b_{jz}(\theta)/S_{jj}^{1/2},\quad j=1,\ 2,\ \ldots,\ r, \tag{6.7.2}$$

где

$$S_{jj}=\Sigma_{i=1}^{n}(Z_{ji}-\bar{Z}_j)^2 \quad \text{и}$$

$$b_0(\theta)=\bar{Y}-\Sigma_{j=1}^{r}b_j(\theta)\bar{Z}_j, \tag{6.7.3}$$

чтобы получить вектор $\mathbf{b}(\theta)=\{b_0(\theta),\ b_1(\theta),\ \ldots,\ b_r(\theta)\}'$ размерностью $(r+1)\times 1$. При $\theta=0$ компоненты этого вектора $b_j(0)$, $j=0,\ 1,\ 2,\ \ldots,\ r$ — обычные МНК-оценки, как это вытекает из уравнения (6.7.1) при подстановке в него $\theta=0$. Выделяя сомножитель $(\mathbf{Z}'\mathbf{Z})^{-1}$ из правой части (6.7.1), мы можем выразить ридж-оцениватель через МНК-оцениватель $\mathbf{b}_z(0)=(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$, а именно

$$\mathbf{b}_z(\theta)=\{\mathbf{I}+\theta(\mathbf{Z}'\mathbf{Z})^{-1}\}^{-1}\mathbf{b}_z(0)=\mathbf{Q}\mathbf{b}, \tag{6.7.4}$$

так что ридж-оценки оказываются линейными комбинациями МНК-оценок с коэффициентами, определяемыми матрицей

$$\{\mathbf{I}+\theta(\mathbf{Z}'\mathbf{Z})^{-1}\}^{-1}.$$

Мы можем теперь построить график зависимости $b_{jz}(\theta)$ или $b_j(\theta)$ от θ для $j=1,\ 2,\ \ldots,\ r$ и исследовать его. Точка пересечения кривой с осью ординат обычно не изображается[8]. Такой график называют *следом гребня* (ridge trace). Обычно этот график строится в «корреляционных» единицах (т. е. используются величины $b_j(\theta)$), чтобы можно было проводить прямое сравнение относительных эффектов различных коэффициентов и исключить влияние масштаба измерения для различных X, от которого зависят значения первоначальных коэффициентов. В нашем примере ниже мы привели, однако, график к исходным единицам (т. е. мы использовали величины $b_j(\theta)$, так что если $\theta=0$, то мы получим МНК-оценки, соответствующие

*6 Однако вычисления можно провести тем способом, какой предпочтет исследователь. Коэффициент $S^{1/2}_{YY}$ можно затем учесть в полученных результатах (см. 5.5).

[8] Точнее, не точка пересечения, а прилегающий к ней участок кривой.— *Примеч. пер.*

ненормированным исходным величинам X.) По мере увеличения параметра θ оценки уменьшаются по абсолютной величине и стремятся к 0, когда θ стремится к бесконечности. Затем выбирается определенная величина θ, которую обозначим, скажем, буквой θ^*. Херл и Кеннард (Hoerl, Kennard. — Technometrics, 1970, **12**, p. 65) на этот счет говорят следующее:

«При выборе величины θ можно руководствоваться следующими обстоятельствами:

1. При определенном значении θ система стабилизируется и приобретает обычные свойства ортогональной системы.

2. Коэффициенты не могут иметь непомерно высокие абсолютные значения по сравнению с факторами, по отношению к которым они представляют собой скорости изменения.

3. Коэффициенты с явно неправильными знаками при $\theta = 0$ могут быть изменены, чтобы знак стал подходящим.

4. Остаточная сумма квадратов не должна увеличиваться до непомерно высоких значений. Она не должна быть слишком большой по отношению к минимальной остаточной сумме квадратов или по отношению к той величине, которой соответствуют приемлемые вариации процесса».

После того как значение θ выбрано (равным θ^*), величины b_j (θ^*) используются в предсказывающем уравнении. Результирующее уравнение содержит оценки, которые не являются оценками метода наименьших квадратов, имеют смещение, но оказываются более устойчивыми в указанном выше смысле. Они (как можно надеяться, см. уравнение (6.7.6) далее) приводят к более низкому значению полного среднего квадрата ошибки, поскольку вызванное ими уменьшение дисперсии ошибок будет больше того, которое нужно для компенсации введенного смещения.

(Заметим, что оценки, выбираемые согласно процедуре Маллоуза, смещены из-за коэффициентов, не учтенных в подгоняемой модели. Оценки, полученные согласно процедуре Херла и Кеннарда, оказываются смещенными, когда в выражения для них входит величина θ, и это смещение имеет место даже если постулируемое уравнение включает все «правильные» предикторные переменные. Иными словами, две указанные разновидности смещения имеют разную природу.)

Средний квадрат ошибки

Гребневую регрессию нередко оправдывают тем, что это практический прием, с помощью которого при желании можно получить меньшее значение среднего квадрата ошибки. Основной результат состоит в следующем (см., например, статью Херла и Кеннарда в журнале Technometrics, 1970, **12**, p. 62). Средний квадрат ошибки для гребневого оценивателя может быть записан в виде

$$\begin{aligned} \mathrm{MSE}(\theta) &= E\{\mathbf{b}_z(\theta)-\boldsymbol{\beta}_z\}'\{\mathbf{b}_z(\theta)-\boldsymbol{\beta}_z\} = E\{\mathbf{Qb}-\boldsymbol{\beta}_z\}'\{\mathbf{Qb}-\boldsymbol{\beta}_z\} = \\ &= E\{(\mathbf{b}-\boldsymbol{\beta}_z)'\mathbf{Q}'\mathbf{Q}(\mathbf{b}-\boldsymbol{\beta}_z)+\boldsymbol{\beta}_z'(\mathbf{Q}-\mathbf{I})'(\mathbf{Q}-\mathbf{I})\boldsymbol{\beta}_z. \end{aligned} \qquad (6.7.5)$$

Чтобы получить этот результат, надо воспользоваться выражением

(6.7.4), где $\mathbf{Q} = \{\mathbf{I} + \theta (\mathbf{Z}'\mathbf{Z})^{-1}\}^{-1}$. Затем надо выделить квадратичную форму относительно $(\mathbf{b}-\boldsymbol{\beta}_z)$, а оставшиеся члены перегруппировать и выполнить некоторые упрощения. Применяя далее матричные результаты приложения 2Б, получим

$$\mathrm{MSE} = \sigma^2 \operatorname{tr}\{\mathbf{Q}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Q}'\} + \boldsymbol{\beta}_z'(\mathbf{Q}-\mathbf{I})'(\mathbf{Q}-\mathbf{I})\boldsymbol{\beta}_z. \qquad (6.7.6)$$

Первый член есть сумма квадратов диагональных элементов матрицы $V(\mathbf{Qb}) = \sigma^2\mathbf{Q}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Q}'$, т. е. он представляет собой сумму дисперсий элементов гребневой оценки $\mathbf{Qb}$. Второй элемент — «квадрат гребневого смещения». (Заметим, что если $\theta = 0$, то $\mathbf{Q} = \mathbf{I}$ и первый член становится суммой дисперсий МНК-оценок коэффициентов, в то время как второй обращается в ноль. Величина, которая при этом достигается, равна MSE (0)). Имеет место теорема, в силу которой всегда существует такое $\theta^* > 0$, что MSE (θ^*) < MSE (0). Особенность этого результата состоит в том, что величина θ^* зависит от σ^2 и $\boldsymbol{\beta}$, которые неизвестны. Таким образом, хотя θ^* и существует, нет способа, позволяющего при решении конкретной практической задачи убедиться, что перед нами значение, которому отвечает величина MSE, меньшая, чем MSE (0).

Гребневая регрессия для данных Хальда

Теперь мы применим этот метод к данным Хальда, чтобы проиллюстрировать его особенности. Из-за связей между четырьмя предикторными переменными эти данные могут привести к повышению неустойчивости оценок, как это уже обсуждалось выше.

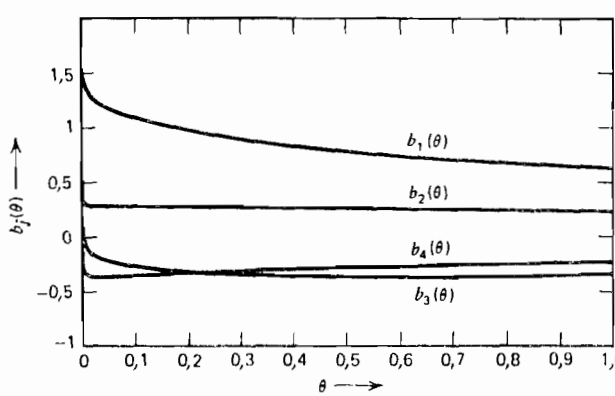


Рис. 6.4. Гребневый след для данных Хальда в интервале $0 \leqslant \theta \leqslant 1$ (по материалам доктора Кеннарда)

На рис. 6.4 показан гребневый след для данных Хальда в интервале $0 \leqslant \theta^* \leqslant 1$, а на рис. 6.5 дана детализация этого графика для интервала $0 \leqslant 0 \leqslant 0{,}03$. Какое значение θ^* следует выбрать?*[7]

*[7] Мы признательны доктору Кеннарду (R. W. Kennard) за любезно предоставленные нам результаты, на основании которых построены эти графики.

Один возможный автоматический способ выбора величины θ^* был предложен Херлом, Кеннардом и Болдуином в работе: Hoerl A. E., Kennard R. W., Baldwin K. F. Ridge regression: some simulation.— Communications in Statistics, 1975, **4**, p. 105—123. Они показали, что целесообразно выбирать эту величину согласно формуле

$$\theta^* = rs^2/\{\mathbf{b}_z(0)\}'\{\mathbf{b}_z(0)\}, \qquad (6.7.7)$$

где r — число параметров в модели, *не считая* $\hat{\beta}_0$; s^2 — остаточный средний квадрат, входящий в таблицу дисперсионного анализа и получаемый в стандартной МНК-процедуре;

$$\{\mathbf{b}_z(0)\}' = \\ = \{b_{1z}(0), \ \ldots, b_{rz}(0)\} = \\ = \{\sqrt{S_{11}}\, b_1(0), \ \ldots, \sqrt{S_{rr}}\, b_r(0)\}. \qquad (6.7.8)$$

(На практике эти выражения могут слегка отличаться из-за ошибок округления.)

(Заметим, что θ^* в уравнении (6.7.7) есть s^2 (оценка величины σ^2), деленная на среднее значение квадрата МНК-оценок $b_{jz}(0)$. Последняя величина может рассматриваться как оценка σ^2_{β} — дисперсии истинных, но неизвестных величин β_z. Поэтому с байесовской точки зрения (см. с. 35)[9] выбор величины θ^* согласно (6.7.7) выглядит разумным.)

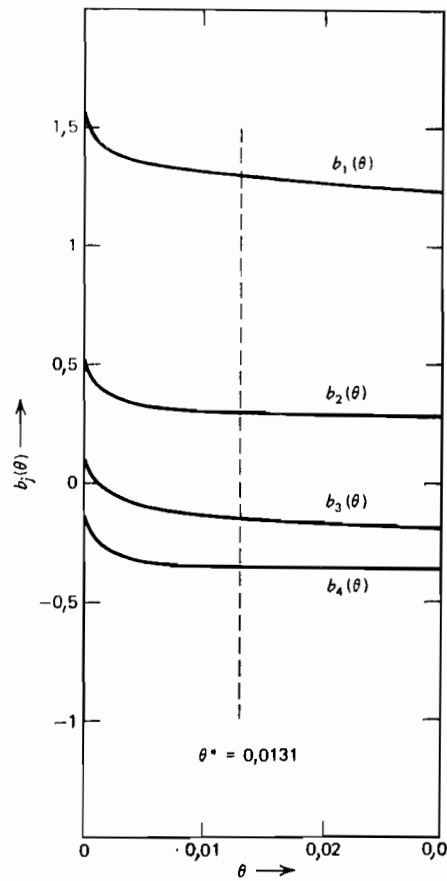


Рис. 6.5. Гребневый след для данных Хальда в интервале $0 \leqslant \theta \leqslant 0{,}03$ (по материалам доктора Кеннарда)

Для этих данных мы имеем $r = 4$, $s^2 = 5{,}983$ (см. с. 302) и $\mathbf{b}_z(0) = (31{,}633;\ 27{,}516;\ 2{,}241;\ -8{,}388)'$, так что $\theta^* = 4\,(5{,}983)/1832{,}4 = 0{,}0131$. Полученному значению величины θ^* соответствует вертикальная линия на рис. 6.5, а значения коэффициентов можно прочи-

[9] Полезные сведения по байесовским методам содержатся также в книгах: Box G. E. P., Chao G. C. Bayesian Inference in Statistical Analysis.— Addison Wesley, Reading, Massachusetts, 1973.— 588 p.; Лимер Э. Статистический анализ неэкспериментальных данных: Выбор формы связи/Пер с англ. Под ред. А. А. Рывкина. — М.: Финансы и статистика, 1983. — 381 с. — *Примеч. пер.*

тать прямо на рисунке или вычислить более точно. В итоге получаем уравнение

$$\hat{Y} = 83{,}414 + 1{,}300X_1 + 0{,}300X_2 - 0{,}142X_3 - 0{,}349X_4. \quad (6.7.9)$$

Это уравнение можно сразу применять в таком виде.

Возможно использование гребневой регрессии как процедуры выбора. Можно высказать соображения о том, как удалить одну или несколько предикторных переменных. Очевидно, в первую очередь следует выбрать X_3. Коэффициент b_3 (0,0131) $= -0{,}142$ наименьший по абсолютной величине. К тому же переменная X_3 такова, что *максимальное* (по абсолютной величине) *изменение* отклика от вариации X_3 составляет всего лишь 0,142 (23) = 3,266. По-видимому, наиболее разумно, чтобы первый шаг состоял в исключении X_3 с последующим исследованием гребневого следа для уравнения, содержащего X_1, X_2 и X_4. Мы не будем продолжать эту процедуру дальше, а рекомендуем читателю обратиться к приложениям в статье Херла и Кеннарда (Technometrics, 1970, **12**, p. 69—82), где содержатся замечания на этот счет. Следует заметить, что на практике гребневая регрессия обычно не применяется как процедура выбора наилучшего уравнения. Мы упоминаем об этом только как о некоторой возможности.

Имеются другие способы выбора величины θ^*. Один из них состоит в использовании итерационной процедуры. Основная идея этого метода состоит в следующем. При выборе θ^* по методу, указанному выше, в знаменателе берется величина $\{\mathbf{b}_z(0)\}'\{\mathbf{b}_z(0)\}$. По этой причине обозначим ее как θ_0^*. Рассмотрим итерационную формулу

$$\theta_{j+1}^* = rs^2/[\{\mathbf{b}_z(\theta_j^*)\}'\{\mathbf{b}_z(\theta_j^*)\}]. \quad (6.7.10)$$

Подставим величину θ_0^* в правую часть этой формулы и найдем θ_1^*, которая, в свою очередь, может быть подставлена в правую часть, чтобы получить θ_2^*, и т. д. Процедура продолжается до тех пор, пока не будет выполняться неравенство

$$(\theta_{j+1}^* - \theta_j^*)/\theta_j^* \leqslant \delta, \quad (6.7.11)$$

где δ — малое число, выбранное априори. Для более полного знакомства с этой процедурой см. статью Херла и Кеннарда: Hoerl A. E., Kennard R. W. — Communications in Statistics, 1976, **A5**, p. 77—87. Эти авторы показывают, что всегда $\theta_{j+1}^* \geqslant \theta_j^*$, так что $\delta \geqslant 0$, и предлагают выбирать подходящее значение δ по формуле

$$\delta = 20\{\mathrm{tr}\,(\mathbf{Z}'\mathbf{Z})^{-1}/r\}^{-1{,}30}, \quad (6.7.12)$$

которую они обосновывают в своей статье на с. 79—80.

Заметим, теперь что в общем не существует какого-то наилучшего способа выбора параметра θ^*. Как мы уже отмечали, уравнение (6.7.7) является до некоторой степени эмпирическим, поскольку оно может трактоваться как формула для грубой оценки отношения σ^2/σ_β^2, применяемого в байесовском подходе, о чем речь пойдет ниже. (Итеративная формула (6.7.10) также может рассматриваться как формула, дающая другую такую оценку.)

При каких ограничениях гребневая регрессия будет абсолютно корректной?

Вокруг гребневой регрессии возникает много споров. Прежде чем излагать нашу точку зрения, мы опишем две весьма ограниченные ситуации, в которых можно вполне уверенно рекомендовать гребневую регрессию как наилучший прием обработки данных.

1) Байесовская формулировка регрессионной задачи при наличии определенных априорных данных о параметрах. Она обсуждается (см. с. 291) в работе Голдштейна и Смита (Goldstein M., Smith A. F. M. Ridge-type estimators for regression analysis.— Journal of the Royal Statistical Society, 1974, **B-36**, p. 284—291), и упоминается также в исходной публикации Херла и Кеннарда. По существу метод гребневой регрессии может рассматриваться как процедура оценивания β_z из данных при условии, что имеются априорные сведения или предположения о том, что меньшие значения β_z (по модулю, т. е. по численному значению, игнорируя знак) более вероятны, чем большие значения, и что чем больше значения β_z по абсолютной величине, тем они более невероятны. Более точно эти априорные сведения могут быть выражены с помощью многомерного нормального распределения априорных величин β_z, т. е. с помощью распределения, размах которого зависит от параметра σ^2_β. Параметр θ гребневой регрессии фактически должен быть равен отношению σ^2/σ^2_β (где σ^2 есть обычная дисперсия отдельного наблюдения). Такой выбор величины θ в гребневой регрессии эквивалентен предположению о том, насколько большими могут быть величины β_z. Использование очень малых θ (например, $< 0{,}01$) означает, что мы не исключаем возможности для величины β_z стать довольно большими. (МНК-оценки, соответствующие $\theta = 0$ или $\sigma^2_\beta = \infty$, таковы, что *априори* мы никак не ограничиваем величины β_z.) Выбор больших значений θ (например, $\theta > 1$) означает, что мы исходим из предположения, что наиболее вероятны довольно малые β_z. Из такого понимания гребневой регрессии следует, что «стабилизация» гребневого следа с ростом θ фактически *не* вытекает из экспериментальных данных; она обусловлена *априорными* ограничениями на возможные значения параметров.

2) Формулировка регрессионной задачи как задачи МНК-оценивания при ограничениях определенного типа на параметры. Предположим, что такая задача решается при ограничениях «сферического» типа

$$\boldsymbol{\beta}_z'\boldsymbol{\beta}_z \leqslant c^2, \tag{6.7.13}$$

где c^2 — известная величина [10]. Если воспользоваться методом неопределенных множителей Лагранжа (см. приложение 2Г), то можно сформировать следующую целевую функцию:

$$F = (\mathbf{Y} - \bar{Y}\mathbf{1} - \mathbf{Z}\boldsymbol{\beta}_z)'(\mathbf{Y} - \bar{Y}\mathbf{1} - \mathbf{Z}\boldsymbol{\beta}_z) + \theta(\boldsymbol{\beta}_z'\boldsymbol{\beta}_z - c^2) \tag{6.7.14}$$

[10] Имеется в виду, что c^2 не превосходит величины, соответствующей $\theta = 0$.— *Примеч. пер.*

с учетом ограничения типа равенства из (6.7.13). Полагая далее $\partial F/\partial \boldsymbol{\beta}_z = 0$, получим уравнения

$$(\mathbf{Z}'\mathbf{Z} + \theta\,\mathbf{I})\,\boldsymbol{\beta}_z = \mathbf{Z}'\,(\mathbf{Y} - \bar{Y}\mathbf{1}) = \mathbf{Z}'\mathbf{Y}, \qquad (6.7.15)$$

решение которых выражается формулой (6.7.1). Однако в данном случае необходимо принимать во внимание ограничение $\boldsymbol{\beta}_z'\boldsymbol{\beta}_z = c^2$. Если мы решим (6.7.15) относительно $\boldsymbol{\beta}_z$ и подставим полученный результат в выражение для ограничения, то получим

$$\mathbf{Y}'\mathbf{Z}\,(\mathbf{Z}'\mathbf{Z} + \theta\mathbf{I}_r)^{-2}\,\mathbf{Z}'\mathbf{Y} = c^2, \qquad (6.7.16)$$

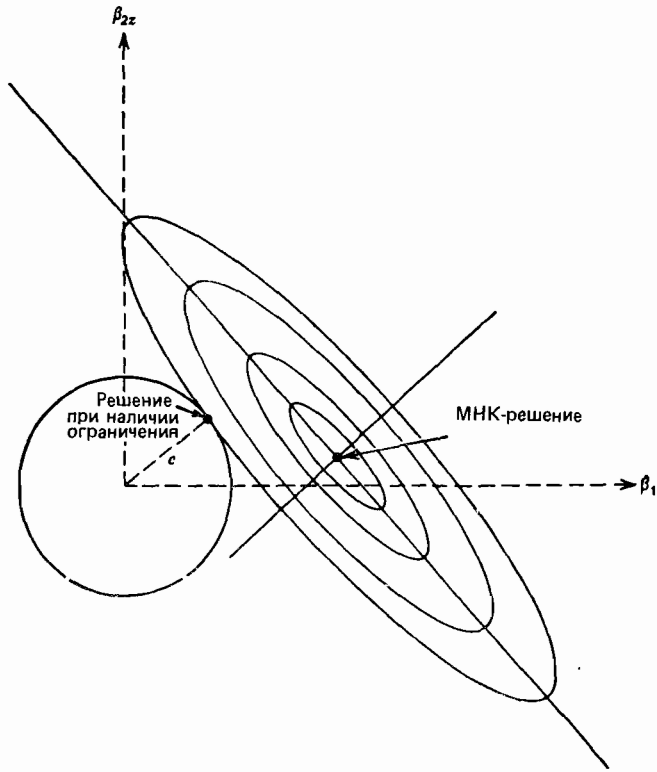


Рис. 6.6. Гребневая регрессия как МНК-решение задачи при ограничениях

т. е. уравнение, позволяющее определить θ в зависимости от заданной величины c^2. Следовательно, гребневая регрессия может рассматриваться как МНК-регрессия при сферических ограничениях на параметры и подходящим выборе величины θ, которая зависит от радиуса ограничений c. На рис. 6.6 дана геометрическая интерпретация двухпараметрического случая (в модели содержатся два параметра, не считая β_0). Обычное, без ограничений, МНК-решение соответствует «дну чаши», выраженной эллиптическими контурами суммы квадратов. (Более полное толкование подобной ситуации дается в гл. 10. Мы предлагаем читателю, не посвященному в эти

подробности, присоединиться к нам и сформулировать общие выводы об интересующих нас случаях, не заботясь здесь о деталях.) Сферические (в данном случае круговые) ограничения изображены на рис. 6.6 в виде окружности. Решение при наличии ограничения соответствует точке касания эллиптического контура наибольшего размера и окружности. Мы получим всю последовательность решений для гребневой регрессии, изменяя длину радиуса окружности от величины, при которой окружность в точности проходит через точку, отвечающую МНК-решению (чему соответствует $\theta = 0$) до нуля, $c = 0$ (чему соответствует $\theta = \infty$). Чем больше радиус окружности, тем «ближе» мы к обычному МНК-решению. Гребень ридж-регрессии выражается линией, точки которой представляют собой последовательность указанных выше решений при изменении радиуса окружности.

Альтернативная точка зрения на гребневую регрессию, когда последняя рассматривается как результат минимаксного решения задачи оценивания при ограничениях $\boldsymbol{\beta}'\boldsymbol{\beta} \leqslant \sigma^2/\theta$ излагается в статье Бунке (B u n k e O. Minimax linear, ridge and shrunken estimators for linear parametrs.— Math. Operationsforsch. u. Statist., 1975, **6**.)

К о м м е н т а р и и. Допуская ту или иную ситуацию, можно получить точное решение для гребневой регрессии. Однако при этом важны два обстоятельства.

1) Можем ли мы в действительности определить априорные ограничения, выраженные тем или иным способом, указанным выше? Так, например, если мы захотим ввести эллиптические ограничения

$$\boldsymbol{\beta}_z'\mathbf{T}\boldsymbol{\beta}_z \leqslant c^2 \tag{6.7.17}$$

вместо указанных выше сферических (6.7.13), то мы получим гребневое решение, отличающееся от (6.7.15), а именно

$$\mathbf{b}_z = (\mathbf{Z}'\mathbf{Z} + \theta\mathbf{T})^{-1}\mathbf{Z}'\mathbf{Y}. \tag{6.7.18}$$

Таким образом, точная форма ограничений на параметры может быть «критичной» по отношению к решению, которое мы получаем.

2) Применяя стандартную форму гребневой регрессии, хотим мы того или нет, по существу мы опираемся на один из указанных выше типов ограничений, не имея на то реальных оснований. Гребневая регрессия не есть панацея от всех бед. Это обычное решение задачи МНК-оценивания при наличии некоторой дополнительной информации о параметрах. Огульное применение гребневой регрессии без понимания ее предпосылок может быть опасным и приводить к заблуждениям. Если дополнительная информация разумная и не противоречит имеющимся данным, то гребневая регрессия оправдана. Простой тест для проверки их совместимости состоит в выяснении того, лежат ли ридж-оценки исходных параметров в пределах 95 %-ной доверительной области эллиптического вида для этих параметров. Области такого вида выражаются неравенствами типа (2.6.15). Если точка, соответствующая вектору ридж-оценок параметров, попадает внутрь доверительной области, то выполняется строгое неравенство. Когда она попадает на внешнюю поверхность (контур) доверительной области, выполняется строгое равенство. И если она выходит за пределы доверительной области, то неравенство не удовлетворяется.

Дополнительное беспокойство вызывает следующая особенность гребневой регрессии. Характерный эффект ридж-процедуры в отличие от обычного МНК-оценивания состоит в изменении значений незначимых оценок параметров. Они увеличиваются во всяком случае больше, чем значимые оценки параметров. Поэтому сомнительно, что при такой процедуре произойдет улучшение оценивания. В большинстве методов выбора предикторных переменных при тех же условиях незначимые переменные скорее бы исключались из уравнения, чем их незначимым коэффициентам придавались бы какие-либо численные значения. Такого рода стратегия более оправдана, так как она позволяет сконцентрировать внимание на наиболее важных предикторных переменных.

С другими соображениями относительно ценности гребневой регрессии как рутинного метода оценивания можно познакомиться в работе: Thisted R. A. Ridge regression, minimax estimation and empirical Bayes methods.— Stanford University Department of Biostatistics Technical Report, 1976, № 28. Некоторые предостережения относительно использования гребневой регрессии в экономических задачах изложены в статье Брауна и Битти (Brown W. G., Beattie B. R. Improving estimates of economic parameters by use of ridge regression with production function applications.— American Journal of Agricultural Economics, 1975, **57**, p. 21—32). Замечания по процедурам определения гребневого следа содержатся в двух статьях Конниффа и Стоуна (Coniffe D., Stone J. A critical view of ridge regression. — The Statistician, 1973, **22**, p. 181—187; A reply to Smith and Goldstein.—The Statistician, 1975, **24**, p. 67—68). Для ознакомления с общей дискуссией см.: Draper N. R, Van Nostrand R. C. Ridge regression and James—Stein estimation: review and comments.— Technometrics, 1979, **21**, p. 451—466 и Smith G., Campbell F. A critique of some ridge regression methods.— Journal of the American Statistical Association, 1980, **75**, p. 74—81, discussion p. 81—103 [11]. В большинстве опубликованных работ по гребневой регрессии изложение ведется в «канонической форме», которая возникает после приведения матрицы $\mathbf{Z}'\mathbf{Z}$ к диагональному виду. Это упрощает выражения и дает более удобный способ вычисления гребневого следа. Каноническая форма гребневой регрессии обсуждается в приложении 6А.

Точка зрения на фальшивые данные

Возможна еще одна, третья интерпретация гребневой регрессии. Мы можем ввести дополнительную информацию в регрессионную задачу путем добавления «априорных данных» в виде $\mathbf{Z}_0$, $\mathbf{Y}_0$. Если мы придадим вес, равный единице, каждой реальной точке и вес, равный θ, каждой точке, отвечающей априорным данным, то для $\boldsymbol{\beta}_z$ МНК-оцениватель [*8] с весами будет иметь вид

---

[11] Из отечественных работ к этому списку добавим книгу Демиденко Е. З. Линейная и нелинейная регрессии. — М.: Финансы и статистика, 1981.— 304 с.— *Примеч. пер.*

[*8] См. § 2.11.

$$\mathbf{b}_{zPD} = (\mathbf{Z}'\mathbf{Z} + \theta\mathbf{Z}_0'\mathbf{Z}_0)^{-1}(\mathbf{Z}'\mathbf{Y} + \theta\mathbf{Z}_0'\mathbf{Y}_0). \qquad (6.7.19)$$

Предположим теперь, что наши априорные сведения являются «фальшивыми» и сформированы так, что $\mathbf{Z}_0'\mathbf{Z}_0 = \mathbf{I}$ и $\mathbf{Y}_0 = \mathbf{0}$. Тогда мы получим из (6.7.19) гребневый оцениватель (6.6.1). Таким образом, другая точка зрения на гребневую регрессию состоит в разбавлении исходных данных некоторым количеством нулевых наблюдений в точках ортогонального плана, расположенных на сфере единичного радиуса, с соответствующими весами.

Можно также выбрать $\mathbf{Z}_0'\mathbf{Z} = \theta\mathbf{I}$ для данного θ и $\mathbf{Y}_0 = \mathbf{0}$. В таком случае пригоден обычный, невзвешенный регрессионный анализ [12]. Следовательно, читатели, которые не имеют пакета программ для гребневой регрессии, могут получить те же самые результаты с помощью стандартной программы регрессионного анализа типа BMDP2R после добавления к исходным данным фиктивных значений для r предикторных переменных, входящих в регрессионную модель. В каждом из r фиктивных опытов соответствующие предикторные переменные поочередно принимают значение, равное $\theta^{1/2}$, тогда как остальные переменные приравниваются нулю. Отклики приравниваются нулю. Иными словами, условия фиктивных опытов выражаются скалярной матрицей $\mathbf{Z}_0 = \theta^{1/2}\mathbf{I}_r$ а результаты — нулевым вектором $\mathbf{Y}_0 = \mathbf{0}$.

Этот метод следует применять осторожно. Правильные значения коэффициентов гребневой регрессии здесь получаются, но большинство других результатов традиционных регрессионных расчетов не имеют смысла для исходных данных, что представляет собой ловушку для неосмотрительных исследователей. (С другой стороны, можно доказать, что эти результаты корректно отражают влияние сделанных предположений.)

Моделирование гребневой регрессии — предостережение

Во многих работах, где в качестве средства моделирования используются регрессионные задачи, делается попытка показать, что оценки гребневой регрессии лучше, чем обычные МНК-оценки параметров, если выносить суждение на основании среднего квадрата ошибок. К подобным утверждениям следует относиться с осторожностью. Тщательное исследование показывает, что моделирование, выполненное при эффективных ограничениях, наложенных на значения параметров, точно воспроизводят ситуации, где гребневая регрессия есть метод, подходящий теоретически. Общий вывод о том, что гребневая регрессия «всегда» лучше, чем обычная регрессия, как правило, совершенно неправомерен.

Резюме

Из приведенных выше рассуждений видно, что применение гребневой регрессии совершенно оправдано в тех случаях, когда предпо-

[12] Наблюдения не обязательно должны быть нулевыми. Достаточно лишь чтобы было равно нулю скалярное произведение векторов $\mathbf{Z}_0$ и $\mathbf{Y}_0$, т. е. $\mathbf{Z}_0'\mathbf{Y}_0 = 0$. — *Примеч. пер.*

лагается, что большие значения β нереалистичны с практической точки зрения. Однако необходимо понимать, что выбор параметра θ по существу эквивалентен представлению о том, насколько большими могут быть эти β [13]. В тех случаях, когда нельзя согласиться с предположением об ограничениях на параметры, гребневая регрессия совершенно не подходит.

Заметим, что для многих наборов данных, когда величины МНК-оценок параметров приемлемы, процедура гребневого следа в своем обычном виде приводит к выбору значения $\theta = 0$. Значение $\theta \neq 0$ выбирают только тогда, когда результаты обычного оценивания не могут рассматриваться как удовлетворительные.

Имеется большое количество публикаций, число которых возрастает, относительно многих аспектов и обобщений метода гребневой регрессии. Некоторые избранные ссылки приведены в конце книги.

М н е н и е. Гребневая регрессия полезна и полностью обоснована в тех случаях, когда считается маловероятным, чтобы значения параметров регрессии были большими (как это интерпретировалось выше с помощью σ_β^2 или c^2). При анализе гребневых следов необходимо иметь субъективное представление о том, с какой ситуацией мы имеем дело. Либо 1) ее определяют априорные байесовские предположения о том, каковы вероятные значения параметров, либо 2) она обусловлена сферическими ограничениями в пространстве параметров. Процедура очень проста, и квалифицированный программист легко адаптирует для ее реализации стандартную программу обычного регрессионного анализа. В общем, мы тем не менее предостерегаем против неразборчивого применения гребневой регрессии, если не учитывается и не принимается во внимание все связанное с ее ограничениями. (Читатель должен знать, что многие авторы не согласны с нашей пессимистической оценкой гребневой регрессии.)

6.8. ПРЕСС

ПРЕСС-процедура [14] выбора предикторных переменных была предложена в работе Аллена: A l l e n D. M. The prediction sum of squares as a criterion for selecting predictor variables.— University of Kentucky, Department of Statistics, Technical Report 1971, 23. Это комбинация метода всех возможных регрессий, анализа остатков и метода перепроверки. (Последний метод обсуждается в гл. 8.)

Предположим, что модель содержит p параметров, *включая* β_0, и имеется всего n измерений. Основные вычисления сводятся к следующему:

1. Вычеркнем условия и результаты первого опыта, т. е. соответствующие значения предикторных переменных и значение отклика.

---

[13] Точнее было бы говорнть о том, насколько мала или велика величина $\beta_z'\beta_z$.— *Примеч. пер.*

[14] Название процедуры связано с первыми буквами английских слов «prediction sum square» (предсказанная сумма квадратов.).— *Примеч. пер.*

2. Построим все возможные регрессионные модели, используя условия и результаты оставшихся n—1 опытов.

3. По каждой модели подсчитаем предсказываемое значение отклика $\hat{Y}_{1p}$ в условиях первого опыта и вычислим предсказываемое расхождение $(Y_1 - \hat{Y}_{1p})$.

4. Повторим шаги 1, 2 и 3, исключив из обработки условия и результаты второго опыта, чтобы получить значения $(Y_2 - \hat{Y}_{2p})$. Затем исключается третий опыт и находятся значения $(Y_3 - \hat{Y}_{3p})$ для каждой модели. И так вплоть до исключения последнего n-го опыта.

5. Для каждой регрессионной модели вычислим сумму квадратов предсказываемых расхождений:

$$\sum_{i=1}^{n} (Y_i - \hat{Y}_{ip})^2.$$

6. Выберем «наилучшую» регрессионную модель. Она должна иметь сравнительно малую сумму квадратов предсказываемых расхождений, но не включать слишком много предикторов.

Чтобы проиллюстрировать, как работает этот алгоритм, исследуем результаты его применения к данным Хальда, собранным в табл. 6.2.

Таблица 6.2. **Значения сумм квадратов предсказываемых расхождений (округленные до целых чисел) для всех возможных моделей по данным Хальда**

Индексы переменных в модели	Соответствующие значения $\sum_{i=1}^{n}(Y_i - \hat{Y}_{ip})^2$
1, 2, 3, 4	1700; 1202; 2616; 1917
12, 13, 14, 23, 24, 34	95; 2220; 121; 793; 1464; 264
123, 124, 134, 234	91; 85; 87; 294
1234	110

В данном случае можно сделать совершенно однозначный выбор. «Наилучшей моделью» следует признать модель, включающую предикторы X_1 и X_2. Этой модели отвечает одно из самых малых значений суммы квадратов предсказываемых расхождений, равное 95. Имеются еще три суммы, которые даже меньше, но они получены для моделей, содержащих уже три предиктора. К тому же эти суммы совсем ненамного меньше.

Мы можем теперь получить дополнительную информацию, исследуя слагаемые, входящие в сумму для модели с предикторами X_1 и X_2. Эти результаты содержатся в табл. 6.3.

Таблица 6.3. **Вклады 13 слагаемых, образующих сумму квадратов предсказываемых расхождений для модели с предикторами X_1 и X_2, которые вычислены по данным Хальда**

Индекс i отбрасываемого наблюдения, результат которого предсказывается	$(Y_i - \hat{Y}_{ip})^2$	Индекс i отбрасываемого наблюдения, результат которого предсказывается	$(Y_i - \hat{Y}_{ip})^2$
1	4	8	8
2	2	9	5
3	3	10	9
4	5	11	16
5	2	12	1
6	22	13	14
7	4		
		Всего	95

Из таблицы видно, что наблюдение 6 хуже всего предсказывается по модели, содержащей предикторы X_1 и X_2 и построенной по остальным точкам. Наблюдения 11 и 13 также плохо предсказываются. При одном наборе исходных данных это может свидетельствовать о наличии выбросов. В других случаях это может служить указанием на то, что подобные точки чрезвычайно информативны. Они не должны бездумно выбрасываться из набора данных, поскольку содержат много информации о согласии модели с экспериментальными данными *[9]. В данном случае, сопоставляя шестое наблюдение с откликом 109,2 в точке (11; 55; 9; 22) с третьим наблюдением, где отклик равен 104,3 при значениях предикторов (11; 56; 8; 20), можно предположить, что результат шестого опыта несколько завышен.

Мнение. ПРЕСС-процедура имеет то преимущество, что дает массу детальной информации об устойчивости различных построенных в пространстве данных моделей и позволяет сконцентрировать внимание на наиболее важных точках в пространстве предикторов. Основной недостаток процедуры — необходимость выполнения громадного объема вычислений. К тому же нет точных правил для выбора наилучшей модели. (Другие процедуры тоже имеют свои недостатки.) Мы считаем, что для решения типовых задач выбора модели на практике все эти работы вовсе не обязательны. Однако устойчивые вычисления могут дать полезную дополнительную информацию, как только модель выбрана. Важно стремиться к более полному пониманию проблемы. Некоторые другие относящиеся к данному вопросу обсуждения содержатся в гл. 8.

*[9] См. 3.12. ПРЕСС-остатки («вычеркнутые») вычисляются в программе BMDP9R.

6.9. РЕГРЕССИЯ НА ГЛАВНЫХ КОМПОНЕНТАХ

До сих пор в этой главе, посвященной методам выбора предикторов, мы имели дело с моделями предсказания, используя только переменные Z в качестве основы для представления исходных данных. Возникавшие при этом трудности обычно ограничивались тем, что мы не знали значение величины σ^2, а также наблюдалась высокая степень коррелированности переменных Z (проблема мультиколлинеарности). Гребневая регрессия представлялась как метод, позволяющий преодолеть последнее затруднение. Альтернативная процедура, которая основана на детальном анализе корреляционной структуры, была впервые предложена Хотеллингом [15] в его ставшей уже классической работе: H o t t e l i n g Harold. Analysis of a complex of statistical variables into principal components.— Journal of Educational Psychology, 1933, **24**, p. 417—441, 489—520.

Используя обозначения, ранее применявшиеся в 6.7, посвященном гребневой регрессии, начнем с введения «центрированной и нормированной матрицы **X**», которую мы будем теперь называть «матрицей **Z**». В таком случае $\mathbf{Z'Z}$ будет корреляционной матрицей. Характеристические корни (нередко их называют скрытыми (латентными) корнями или собственными числами) [16] корреляционной матрицы представляют собой r решений $\lambda_1, \lambda_2, \ldots, \lambda_r$ детерминантного уравнения

$$|\mathbf{Z'Z}-\lambda\mathbf{I}|=0. \tag{6.9.1}$$

Можно показать, что сумма характеристических корней корреляционной матрицы равна следу $\mathbf{Z'Z}$, т. е. сумме ее диагональных элементов. Она равна: $r = p-1$, если использовать обозначения, применяемые в данной главе. (Напомним, что p есть число параметров в модели, *включая* β_0. Стандартизируем данные, т. е. перейдем к новым переменным, из которых формируются вектор-столбцы матрицы:

$$z_{ji}=(Z_{ji}-\bar{Z}_j)/S_{jj}^{1/2}, \tag{6.9.2}$$

где

$$n\bar{Z}_j=\sum_{i=1}^{n} Z_{ji}, \quad S_{jj}=\sum_{i=1}^{n}(Z_{ji}-\bar{Z}_j)^2. \tag{6.9.3}$$

Сумма элементов каждого такого столбца равна нулю, а сумма квадратов элементов — единице. В результате мы ортогонализировали новый β_0' коэффициент и перевели предикторы в «корреляционную форму». Ранг невырожденной корреляционной матрицы равен: $p-1 = r$.) Очевидно, сумма всех сумм квадратов элементов по столб-

---

[15] Метод главных компонент был предложен впервые в 1901 г. К. Пирсоном, а затем развит, доработан, описан и обоснован в работах Г. Хотеллинга, Г. Хармана, С. Рао, П. Ф. Андруковича, С. А. Айвазяна и др. (см: Д у б р о в А. М. Обработка статистических данных методом главных компонент.— М.: Статистика, 1978.— 136 с.; см. с. 26). Сжатое, но четкое последовательное изложение этого метода приведено в книге: А й в а з я н С. А., Е н ю к о в И. С., М е ш а л к и н Л. Д. Прикладная статистика. Исследование зависимостей.— М.: Финансы и статистика, 1985.— 488 с.; см. с. 350—354.— *Примеч. пер.*

[16] В отечественной литературе их чаще всего называют собственными числами.— *Примеч. пер.*

цам z_j равна r. Назовем ее полной дисперсией переменных. Разложим эту величину, используя преобразованные переменные W.

С каждым характеристическим корнем λ_j связан характеристический вектор $\boldsymbol{\gamma}_j$, который удовлетворяет системе однородных уравнений:

$$(\mathbf{Z}'\mathbf{Z} - \lambda_j \mathbf{I})\,\boldsymbol{\gamma}_j = \mathbf{0}. \tag{6.9.4}$$

Решения $\boldsymbol{\gamma}_j = (\gamma_{1j}, \gamma_{2j}, \ldots, \gamma_{rj})'$ выбираются из бесконечного множества «пропорциональных» решений, соответствующих каждому j, таким образом, чтобы соблюдалось условие $\boldsymbol{\gamma}_j' \boldsymbol{\gamma}_j = 1$ [17]. Можно показать далее, что если все λ_j различны, то характеристические векторы попарно ортогональны. (Если не все λ_j различны, то надо выполнить некоторые преобразования, которые мы здесь не обсуждаем, поскольку такой случай обычно не возникает при обработке реальных данных.) Векторы $\boldsymbol{\gamma}_j$ используются для того, чтобы перейти от переменных Z к главным компонентам W в форме

$$W_j = \gamma_{1j} z_1 + \gamma_{2j} z_2 + \ldots + \gamma_{rj} z_r. \tag{6.9.5}$$

Сумма квадратов элементов W_{jt}, из которых формируются вектор-столбцы $\mathbf{W}_j$, равна λ_j. Иными словами, каждой главной компоненте $\mathbf{W}_j$ соответствует величина λ_j, представляющая собой часть полной дисперсии переменных: $\Sigma_j \lambda_j = r$ и $\Sigma_j \Sigma_t W_{jt}^2 = r$.

Таким образом, при реализации этой процедуры формируются новые искусственные переменные W_j [18] с помощью линейного преобразования исходных переменных, выражаемого уравнением (6.9.5), причем так, что вектор-столбцы $\mathbf{W}_j$ взаимно ортогональны. Переменная W_j, соответствующая наибольшему характеристическому корню λ_j, называется первой главной компонентой. Она «объясняет» наибольшую часть вариации в наборе стандартизированных данных [19]. Последующие главные компоненты $\mathbf{W}_j$ объясняют все меньшие и меньшие доли вариации. В итоге

$$\sum_{i=1}^{p} \lambda_j = r.$$

Обычно используют не все главные компоненты, а выбирают некоторые из них по определенному правилу. Однако не существует универсальной процедуры, которой придерживались бы все. Некоторые психологи используют, например, правило, согласно которому принимаются в расчет лишь компоненты, для которых собственные числа больше единицы. Моррисон (Morrison D. F.) во 2-м издании Multivariate Statistical Methods (New York: Mc Graw-Hill, 1976,

---

[17] Иными словами, выбираются характеристические векторы единичной длины.— *Примеч. пер.*

[18] Конечно, неудачно, что одной и той же буквой с одинаковым индексом обозначены переменная W_j и вектор-столбец $\mathbf{W}_j$. Однако различить их можно, так как в последнем случае используется полужирная буква.— *Примеч. пер.*

[19] Собственное число λ_j есть не что иное, как дисперсия j-й главной компоненты. Их сумма равна полной (суммарной) дисперсии переменных.— *Примеч. пер.*

р. 273) отмечает, что «. . . компоненты целесообразно вычислять до тех пор, пока они не «объяснят» некоторую заранее назначенную большую долю (скажем, 75 % или более того) суммарной дисперсии». Иными словами, мы выбираем первые k главных компонент, для которых выполняется условие $\sum_{j=1}^{k} \lambda_j / r > 0{,}75$. В некоторых таких правилах автоматически получается набор из k главных компонент, и исходные переменные Z_i представляются теперь с помощью этого набора из k новых предикторных переменных. Метод наименьших квадратов используется затем для получения уравнения, связывающего Y с выбранными главными компонентами. Порядок их включения в уравнение роли не играет, так как все они «ортогональны» друг к другу. Коль скоро такое уравнение получено в виде функции от выбранных главных компонентов W_j, его можно преобразовать и выразить через исходные предикторы Z_i, если это желательно, или проинтерпретировать в терминах выбранных переменных W_j.

Мы проиллюстрируем этот метод, используя данные Хальда (с. 283) и программу BMDP4R programs, Regression on Principal Components. Машинная распечатка сокращена. Сначала приведем контрольные данные.

ЗАВИСИМАЯ ПЕРЕМЕННАЯ (ЫЕ) 5
НЕЗАВИСИМЫЕ ПЕРЕМЕННЫЕ 1 2 3 4
ВЫЧИСЛЕНИЯ ОСНОВАНЫ НА КОРРЕЛЯЦИОННОЙ МАТРИЦЕ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ
ГЛАВНЫЕ КОМПОНЕНТЫ ВВОДЯТСЯ В ПОРЯДКЕ СЛЕДОВАНИЯ ЗНАЧЕНИЙ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ С ЗАВИС. ПЕРЕМ.
МАКСИМАЛЬНОЕ ЧИСЛО КОМПОНЕНТОВ ДЛЯ ВКЛЮЧЕНИЯ . . 4
ЧИСЛО ВКЛЮЧАЕМЫХ КОМПОНЕНТОВ, ОТГРАНИЧЕННЫХ ВЕЛИЧИНОЙ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ, БОЛЬШЕЙ 0,0100, ОТ МАКСИМАЛЬНОГО СОБСТВЕННОГО ЧИСЛА 0,0100
(эта величина назначается экспериментатором произвольно)
ЧИСЛО ОПЫТОВ 13

Для удобства приведем суммарные статистики исходных данных и корреляционную матрицу.

ПЕРЕМЕННАЯ №, ИМЯ	СРЕДНЕЕ	СТАНДАРТНОЕ ОТКЛОНЕНИЕ	КОЭФФИЦИЕНТ ВАРИАЦИИ
Z_1	7,46154	5,88239	0,78836
Z_2	48,15379	15,56089	0,32315
Z_3	11,76923	6,40512	0,54423
Z_4	29,99995	16,73816	0,55794
Y	95,42302	15,04373	0,15765

КОРРЕЛЯЦИОННАЯ МАТРИЦА

	Z_1	Z_2	Z_3	Z_4
Z_1	1,0000			
Z_2	0,2286	1,0000		
Z_3	—0,8241	—0,1392	1,0000	
Z_4	—0,2454	—0,9730	0,0295	1,0000

Далее приводятся собственные числа λ_i и соответствующие им собственные векторы. Здесь лишь только три собственных числа, которые превышают установленный порог 0,01. Показано также, какая часть суммарной дисперсии независимых переменных Z_i объясняется собственными числами.

СОБСТВЕННЫЕ ЧИСЛА

λ_1	λ_2	λ_3	λ_4	$\Sigma_{i=1}^{4}\lambda_i=r$
2,23569	1,57606	0,18661	0,00162	4,00

НАКОПЛЕННАЯ ДОЛЯ СУММАРНОЙ ДИСПЕРСИИ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ

$^1/_4\,\lambda_1$	$^1/_4\,(\lambda_1+\lambda_2)$	$^1/_4\,(\lambda_1+\lambda_2+\lambda_3)$	$^1/_4\,(\lambda_1+\lambda_2+\lambda_3+\lambda_4)$
0,55893	0,95294	0,99959	1,00000

СОБСТВЕННЫЕ ВЕКТОРЫ

	W_1	W_2	W_3
Z_1	0,4760	—0,5090	0,6755
Z_2	0,5639	0,4139	—0,3144
Z_3	—0,3941	0,6050	0,6377
Z_4	—0,5479	—0,4512	—0,1954

Интерпретация приведенных выше результатов такова:

$$W_1 = 0{,}4760\,z_1 \quad +0{,}5639\,z_2 \quad -0{,}3941\,z_3 \quad -0{,}5479\,z_4 \qquad (6.9.6)$$

и т. д. Заметим, кстати, что это есть преобразование, позволяющее перейти от Z к W. Для того чтобы получить зависимость W от Z, надо воспользоваться соотношением типа (6.9.2). Значения главных компонент вычисляются для каждой точки исходных данных и печатаются построчно вместе со значениями откликов.

ЗНАЧЕНИЯ ГЛАВНЫХ КОМПОНЕНТ И ИСХОДНОЙ ЗАВИСИМОЙ ПЕРЕМЕННОЙ

Опыт	W_1	W_2	W_3	Y
1	—0,9813	—1,5159	—1,2269	78,5
2	—1,4284	—0,1899	—0,6718	74,3
3	0,7557	—0,1465	—0,0248	104,3
4	—0,4413	—1,2560	0,4148	87,6
5	0,2399	—0,3852	—1,7133	95,9
6	0,6465	—0,1354	0,1984	109,2
7	0,6225	1,7005	—0,4005	102,7
8	—1,4928	0,5510	1,0642	72,5
9	—0,2351	1,1409	—0,0731	93,1
10	1,1119	—1,4562	1,9704	115,9
11	—1,0969	1,0316	1,1440	83,8
12	1,1320	0,3124	—0,0459	113,3
13	1,1675	0,3485	—0,6357	109,4

Таблица 6.4. **Таблица дисперсионного анализа данных Хальда, обработанных методом главных компонент**

Источник вариации	Степени свободы	SS	MS	Обычный F-критерий	Последовательный F-критерий
b_1'	1	2620,4866	2620,4866	486,00	302,53
b_2'	1	46,4569	46,4569	8,62	9,52
b_3'	1	0,2954	0,2954	0,05	0,05
Остаток	9	48,5277	5,3920		
Общий (скорректированный)	12	2715,7666			

Коэффициенты корреляции W_j и Y таковы:

$$r_{W_1Y} = 0{,}982, \quad r_{W_2Y} = 0{,}010, \quad r_{W_3Y} = 0{,}131. \tag{6.9.7}$$

Модель

$$Y_i = \beta_0' + \beta_1' W_{1i} + \beta_2' W_{2i} + \beta_3' W_{3i} + \varepsilon_i \tag{6.9.8}$$

была подогнана к исходным данным по методу наименьших квадратов. В итоге получили

$$\hat{Y} = 95{,}42302 + 9{,}88312 W_1 + 0{,}12498 W_2 + 4{,}5583 W_3. \tag{6.9.9}$$

Табл. 6.4 соответствует этой модели. В последнем столбце приведены значения F-критерия, полученные при введении компонент W_1, W_2 и W_3 в модель поодиночке, последовательно в порядке расположения их вкладов. Так, например, если в модель вводится только одна первая компонента, остаточная сумма квадратов равна:

$$48{,}5277 + 0{,}2954 + 46{,}4569 = 95{,}2800$$

с 11 степенями свободы. Величина последовательного F-критерия в таком случае равна:

$$F = (2620{,}4866/1)/(95{,}28/11) = 302{,}53.$$

Величина R^2 растет от 0,9649 до 0,9820 и до 0,9821 по мере того, как добавляются ортогональные переменные W_1, W_2 и W_3. Переходя обратно от переменных W к переменным Z для этих трех случаев получим следующие подогнанные уравнения.

Только с компонентой W_1:

$$\hat{Y} = 89{,}0731 + 0{,}7997 Z_1 + 0{,}3581 Z_2 - 0{,}6080 Z_3 - 0{,}3235 Z_4. \tag{6.9.10}$$

С компонентами W_1 и W_2:

$$\hat{Y} = 85{,}8603 + 1{,}3227 Z_1 + 0{,}2661 Z_2 - 0{,}1546 Z_3 - 0{,}3767 Z_4. \tag{6.9.11}$$

С компонентами W_1, W_2 и W_3

$$Y = 85{,}7430 + 1{,}3119Z_1 + 0{,}2694Z_2 - 0{,}1428Z_3 - 0{,}3801Z_4. \quad (6.9.12)$$

Из всего приведенного выше анализа ясно, что только две компоненты W_1 и W_2 играют существенную роль в подбираемом уравнении. В силу этого целесообразно принять уравнение (6.9.11). Заметим, что во всех полученных уравнениях участвуют все переменные Z, ни одна из них не исключается при реализации данной процедуры. Укажем также, что уравнения, выраженные через Z, имеют смещенные оценки параметров, тогда как применяя метод наименьших квадратов непосредственно к уравнению, выраженному через исходные переменные Z, получим несмещенные оценки.

Мнение. Основной вопрос, который возникает здесь, таков: имеют или нет реальный смысл переменные W или по крайней мере имеют ли они больший смысл, чем переменные Z для тех систем, для которых получены данные? Если да или в том случае, когда переменные W могут использоваться при эксплуатации системы, регрессия на главных компонентах чрезвычайно полезна. Если же переменные W не несут в себе большего смысла, чем Z, то метод главных компонент не настолько ценен, чтобы его рекомендовать. В целом этот метод, по-видимому, малоприменим в физических, технических и биологических науках. Он может быть полезным иногда в общественных науках как метод отыскания эффективных комбинаций переменных.

Попытки улучшить регрессию на главных компонентах обсуждаются в § 6.10.

6.10. РЕГРЕССИЯ НА СОБСТВЕННЫХ ЗНАЧЕНИЯХ

Дальнейшее развитие регрессии на главных компонентах для исследования альтернативных регрессионных моделей и для исключения предикторных переменных было дано Уэбстером, Гунстом и Масоном в работе: Webster J. T., Gunst R. F., Mason R. L. Latent root regression analysis.— Technometrics, 1974, **16**, p. 513—522. Эти авторы расширили матрицу данных, содержащую центрированные и нормированные предикторные переменные, дополнив ее центрированными и нормированными значениями отклика, разместив их первыми по порядку, т. е.

$$\mathbf{Z}^* = (\mathbf{y},\ \mathbf{Z}), \quad (6.10.1)$$

где $\mathbf{Z}$ — центрированная и нормированная «$\mathbf{X}$-матрица», $\mathbf{y} = (\mathbf{Y} - \mathbf{1}\bar{Y})/S_{YY}^{1/2}$, а $S_{YY} = \Sigma (Y_i - \bar{Y})^2$. Отсюда следует, что $\mathbf{Z}^{*\prime}\mathbf{Z}^*$ есть расширенная корреляционная матрица. Как и в методе главных компонент, здесь вычисляются скрытые корни [*10] и соответствующие

[*10] Как мы уже указывали ранее в 6.9, термины «скрытые корни», «характеристические корни», «собственные значения» равнозначны. То же самое справедливо и для соответствующих им векторов, которые называют «скрытыми», «характеристическими» или «собственными векторами».

им собственные векторы. Однако в данном методе первый элемент («Y-коэффициент») каждого собственного вектора используется как мера предсказуемости отклика с помощью данного собственного вектора. Чем больше абсолютная величина первого элемента собственного вектора, тем больший вклад этого вектора в предсказание отклика, и наоборот. Наличие малых собственных значений указывает на возможное существование линейных связей между предикторными переменными. Чем меньше собственные значения, тем сильнее выражены эти связи. МНК-оценивание приводит к получению наилучшей линейной комбинации всех этих собственных векторов. Опуская скрытые векторы, для которых собственные значения по абсолютной величине малы, и первые элементы этих собственных векторов, мы получим модифицированное МНК-уравнение. Такой модифицированный метод приводит к смещенным оценивателям. Как только модифицированное уравнение получено, можно воспользоваться методом исключения, чтобы удалить из него малозначимые предикторные переменные.

Проиллюстрируем рассматриваемый метод на данных Хальда, см. приложение Б.

Шаг 1. Прежде всего получаем расширенную корреляционную матрицу.

	Y	Z_1	Z_2	Z_3	Z_4
Y	1,0000				
Z_1	0,7307	1,0000			
Z_2	0,8163	0,2286	1,0000		
Z_3	—0,5347	—0,8241	—0,1392	1,0000	
Z_4	—0,8213	—0,2454	—0,9730	0,0295	1,0000

(6.10.2)

Шаг 2. Затем определяем собственные числа λ_j расширенной корреляционной матрицы и соответствующие им собственные векторы γ_j. Ниже собственные числа указаны первыми в каждой строке. Остальные элементы строк образуют соответствующие собственные векторы.

Собственные числа λ_j	Y γ_{0j}	Z_1 γ_{1j}	Z_2 γ_{2j}	Z_3 γ_{3j}	Z_4 γ_{4j}
$\lambda_4 = 3{,}2116$	0,5534	0,4012	0,4682	—0,3189	—0,4603
$\lambda_3 = 1{,}5761$	—0,0034	0,5125	—0,4096	—0,6080	0,4471
$\lambda_2 = 0{,}1990$	0,2112	0,5809	—0,3899	0,6747	—0,1039
$\lambda_1 = 0{,}0117$	—0,8047	0,4129	0,1884	—0,0531	—0,3791
$\lambda_0 = 0{,}0016$	0,0408	—0,2617	—0,6523	—0,2657	—0,6586

(6.10.3)

Шаг 3. Теперь исследуем собственные числа λ_j и соответствующие им величины γ_{0j}. Если для некоторого j обе эти величины малы, то это указывает на то, что данные близки к вырожденным и отклик плохо предсказуем. Уэбстер и его соавторы (1974, p. 518) рекомендуют такие условия малости [20]: $\lambda_j \leqslant 0{,}05$ и $\gamma_{0j} \leqslant 0{,}10$. Для данных

[20] Имеется в виду малость абсолютного значения этой величины, т. е. малость величины $|\gamma_{0j}|$ — *Примеч. пер.*

Хальда имеем $\lambda_0 = 0{,}0016 < 0{,}05$ и $\gamma_{00} = 0{,}0408 < 0{,}10$. Это указывает на то, что последняя вектор-строка должна быть исключена из рассмотрения. Следующее наименьшее собственное число равно: $\lambda_1 = 0{,}0117 < 0{,}05$, что свидетельствует о вырожденности задачи. Однако при этом $\gamma_{01} = -0{,}8047$, что указывает на достаточно высокую предсказуемость, так что этот вектор должен быть сохранен. Все другие λ_j превосходят пороговое значение, так что мы сохраняем соответствующие векторы, несмотря на невысокие значения γ_{0j}.

Шаг 4. Затем выполняем процедуру оценивания. Сначала мы должны решить на основании шага 3, какие векторы мы желаем сохранить. Затем следует вычислить модифицированные МНК-оценки параметров по формуле (см.: Webster et al. 1974, p. 514—515, в частности формулу (4.6)):

$$\mathbf{b}^* = \begin{bmatrix} b_1^* \\ b_2^* \\ \cdot \\ \cdot \\ \cdot \\ b_r^* \end{bmatrix} = c\Sigma_j^* \gamma_{0j}\lambda_j^{-1} \begin{bmatrix} \gamma_{1j} \\ \gamma_{2j} \\ \cdot \\ \cdot \\ \cdot \\ \gamma_{rj} \end{bmatrix} \tag{6.10.4}$$

где c — константа, определяемая по формуле

$$c = -\{\Sigma_j^* \gamma_{0j}^2 \lambda_j^{-1}\}^{-1} \left\{\sum_{i=1}^{n} (Y_i - \bar{Y})^2\right\}^{1/2}, \tag{6.10.5}$$

а Σ^* означает суммирование, но лишь по тем индексам j, которым соответствуют векторы, сохраняемые на данном шаге процедуры. Параметр b_0^* для этой модели определяется как $b_0^* = \bar{Y}$.

Предположим, например, что мы сохраняем все векторы. В таком случае первый элемент вектора $\mathbf{b}^*$ будет равен:

$$b_1^* = c\sum_{j=0}^{4} \gamma_{0j}\lambda_j^{-1}\gamma_{1j} = c\{(0{,}5534)(0{,}4012)/3{,}2116 + \ldots + (0{,}0408) \times$$
$$\times(-0{,}2617)/0{,}0016\} = c\{0{,}069 - 0{,}001 + 0{,}617 - 28{,}398 - 6{,}674\} =$$
$$= (-34{,}387)\,c, \tag{6.10.6}$$

где

$$c = \{(0{,}5534)^2/3{,}2116 + \ldots + (0{,}0408)^2/0{,}0016\}^{-1}(2715{,}7636)^{1/2} =$$
$$= -0{,}919. \tag{6.10.7}$$

Таким образом, $b_1^* = -0{,}919\ (-34{,}387) = 31{,}6017$. Это коэффициент при стандартизированной переменной

$$z_1 = (X_1 - \bar{X}_1)/S_{11}^{1/2} = (X_1 - 7{,}462)/20{,}3772, \tag{6.10.8}$$

так что коэффициент при X_1 равен: $31{,}6017/20{,}3772 = 1{,}551$. Обращаясь к приложению Б, с. 301, можно обнаружить, что фактически это МНК-коэффициент. Это есть проявление общего правила: если

уравнения (6.10.4) и (6.10.5) применяются по отношению ко всем векторам, то модифицированные МНК-коэффициенты в точности совпадают с обычными МНК-коэффициентами. (На практике из-за ошибок округления они, конечно, могут слегка отличаться, в зависимости от числа знаков, которые сохраняются при вычислениях.)

Остаточная сумма квадратов для некоторого модифицированного МНК-уравнения может быть записана (см.: Webster et al., 1974, p. 515, формула (4.7)) в виде

$$\text{Остаточная сумма SS} = \left\{\sum_{i=1}^{n} (Y_i - \overline{Y})^2\right\} \{\Sigma_j^* \gamma_{0j}^2 \lambda_j^{-1}\}^{-1} =$$

$$= -c \{\Sigma (Y_i - \overline{Y})^2\}^{1/2}. \qquad (6.10.9)$$

Используя ее для проверки правильности решения задачи МНК-оценивания, найдем остаточную сумму квадратов: $0{,}919\ (2715{,}7636)^{1\,2} = 47{,}892$. Правильная величина равна 47,863.

Если мы теперь произведем параллельные вычисления, при которых вектор, соответствующий наименьшему собственному числу, опущен, то мы получим следующие результаты (в сравнении с обычным МНК-оцениванием). Символ OLS соответствует обычному, а символ MLS — модифицированному МНК-оцениванию:

	b_1	b_2	b_3	b_4	Остаточная сумма SS
OLS	1,551	0,5102	0,1019	—0,1441	47,86
MLS	1,273	0,2308	—0,1812	—0,4179	48,76

(6.10.10)

Мы видим, что имеют место заметные отклонения коэффициентов от их несмещенных МНК-оценок. Однако остаточная сумма квадратов совсем ненамного больше минимального значения, достигаемого при обычном МНК-оценивании. Таким образом, модифицированное МНК-уравнение может расцениваться как уравнение, которое будет по крайней мере почти таким же хорошим, как МНК-уравнение.

Шаг 5. Теперь можно применить процедуру исключения, предложенную Уэбстером и соавторами (1974, p. 517, в частности см. формулу (6.1)). Остаточная сумма квадратов, которая получается после вычеркивания X_l, $l = 1, 2, \ldots, r$, из модифицированного МНК-уравнения, может быть вычислена по формуле

$$\left\{\sum_{i=1}^{n} (Y_i - \overline{Y})^2\right\} \left\{t_{00} - \frac{t_{l0}^2}{t_{ll}}\right\}^{-1}, \qquad (6.10.11)$$

где

$$t_{pq} = \Sigma_j^* \frac{\gamma_{pj}\gamma_{qj}}{\lambda_j} \qquad (6.10.12)$$

и где Σ^* снова обозначает оператор суммирования по элементам, остающимся при вычислении модифицированного МНК-уравнения. Согласно Уэбстеру и его коллегам (1974, p. 521): «Основное достоинство этого метода состоит в выявлении непредсказуемого эффекта почти вырожденности, в более четком представлении влияния неза-

висимых переменных на зависимую». Как это проявляется на данных Хальда? Покажем, чему равна остаточная сумма квадратов, получаемая обычным и модифицированным методом наименьших квадратов, после удаления указанной X-переменной.

Удаляемый предиктор

	X_1	X_2	X_3	X_4
OLS	73,81	50,84	47,97	48,11
MLS без λ_0	299,39	242,42	55,39	1234,11

(6.10.13)

В обоих случаях в первую очередь представляется целесообразным удалять X_3, но это значительно более очевидно для модифицированной процедуры оценивания. Приближенный F-критерий для удаления был предложен Уэбстером и его соавторами (1974, p. 517, формула (6.2)). Модифицированная процедура исключения может затем выполняться параллельно с обычной МНК-процедурой исключения, с теми различиями, которые уже указывались. Окончательное модифицированное МНК-уравнение, полученное по данным Хальда, имеет вид

$$\hat{Y} = 95{,}947 + 1{,}435X_1 + 0{,}1993X_2 - 0{,}4396X_4. \qquad (6.10.14)$$

Упрощенное модифицированное МНК-уравнение

Из вычислений, приведенных выше, ясно, что наибольший вклад в регрессию вносит вектор, соответствующий λ_1. Так что мы можем исследовать подгоняемое уравнение, содержащее только этот вектор. Символ Σ^* теперь уже нам не требуется, а в общих уравнениях (6.10.4) и (6.10.5.) нужно положить $j = 1$. Поэтому

$$\hat{Y} = \overline{Y} - \{\Sigma\,(Y_i - \overline{Y})^2\}^{1/2} \sum_{l=1}^{4} (\gamma_{l1}/\gamma_{00})\, z_l \qquad (6.10.15)$$

или

$$\hat{Y} = 95{,}423075 + 26{,}7397z_1 + 12{,}2009z_2 - 3{,}4388z_3 - 24{,}5508z_4. \qquad (6.10.16)$$

Теперь воспользуемся для подстановки выражениями:

$$z_1 = (X_1 - 7{,}461538)/20{,}3772, \quad z_2 = (X_2 - 48{,}153845)/53{,}9045,$$
$$z_3 = (X_3 - 11{,}769230)/22{,}1880, \quad z_4 = (X_4 - 30)/57{,}9827 \qquad (6.10.17)$$

и получим модифицированное МНК-уравнение

$$\hat{Y} = 89{,}2611 + 1{,}3122X_1 + 0{,}2263X_2 - 0{,}1550X_3 - 0{,}4234X_4. \qquad (6.10.18)$$

Эта смещенная модель объясняет 0,9819 вариации данных относительно

среднего значения $\bar{Y}$, что очень близко к величине 0,9824 для полной МНК-модели. В этом смысле можно говорить, что подгонка отличная. Теперь было бы желательно выполнить модифицированную процедуру исключения, принимая полученное уравнение за исходное.

Мнение. Этот метод оставляет хорошее впечатление, когда его иллюстрируют на примере, но преимущества становятся сомнительными при детальном рассмотрении. Он снабжает нас той же основной информацией, которую мы получаем при использовании других методов. Вместе с тем можно легко не заметить произвола в процедуре смещенного оценивания (о природе которого подробно говорится в исходной статье: Webster et al., 1974, p. 514). Такая процедура может быть полезной для исследователя, который умудрен опытом и постоянно использует данный метод, но для большинства исследователей мы не можем ее рекомендовать.

6.11. СТУПЕНЧАТЫЙ РЕГРЕССИОННЫЙ МЕТОД

Этот метод не дает правильного МНК-решения для переменных, включенных в итоговое уравнение. Основная его идея состоит в следующем. После того как получено регрессионное уравнение для переменной X, наиболее сильно коррелированной с Y, находят остатки $Y_i - \hat{Y}_i$. Эти остатки теперь рассматриваются как значения отклика, и строится регрессия этого отклика на предикторную переменную X, которая наиболее сильно коррелирована с этим новым откликом. Процесс продолжается до любой желаемой стадии. Новые предикторы X не корректируются на исходные X. Так как на каждой стадии отклик = предсказываемый отклик + (отклик — предсказываемый отклик), конечное регрессионное уравнение можно получить путем последовательных подстановок регрессионных уравнений. Оно *не* будет МНК-уравнением для включенных в него переменных. Для иллюстрации снова воспользуемся примером Хальда. Вычисления выполняются довольно легко на настольных калькуляторах и не требуют привлечения ЭВМ. (Распечатки, приведенные в приложении Б, подходят только для первой стадии этого метода.)

Шаг 1. Построим графики зависимости отклика Y от каждой из предикторных переменных по очереди, или вычислим коэффициенты корреляции отклик-предиктор. Выберем предикторную переменную, наиболее сильно коррелированную с откликом. Для данных Хальда это X_4.

Шаг 2. Построим регрессию Y в зависимости от X_4 и получим $\hat{Y}(X_4) = 117{,}57 - 0{,}74\, X_4$. Вычислим остатки $V_i = Y_i - \hat{Y}_i(X_4)$ для каждого значения X_4. Они показаны на с. 289.

Шаг 3. Остатки V_i рассматриваем как новые отклики и выбираем из оставшихся переменных ту, которая наиболее сильно коррелирована с этими остатками. Часто оказывается полезным исходный график (см. рис. 6.7).

Шаг 4. Хотя и X_1 и X_3 — возможные кандидаты, вычисления соответствующих коэффициентов корреляции показывают, что X_1 наи-

более сильно коррелирована с остатками среди всех X. Поэтому мы теперь построим регрессию, связывающую V_i из шага 2 с переменной X_1. Данные для этой регрессии имеют вид

X_1	V	X_1	V
7	5,22	1	—12,59
1	—4,88	2	—8,22
11	1,50	21	17,52
11	4,73	1	—8,67
7	2,69	11	4,59
11	7,87	10	0,69
3	—10,44		

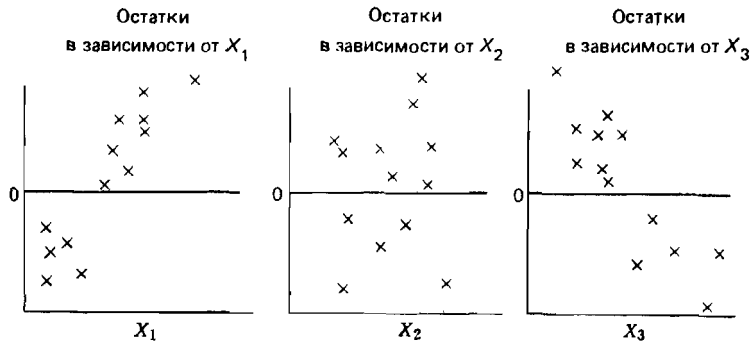


Рис. 6.7. Графики остатков $Y-\hat{Y}$ (X_4) в зависимости от X_1, X_2 и X_4

Наилучшее уравнение прямой есть

$$\hat{V} = -10{,}10 + 1{,}35X_1.$$

На первом этапе мы можем записать

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

Затем, поскольку

$$V_i = \hat{V}_i + (V_i + \hat{V}_i),$$

мы можем записать

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

Следовательно, на втором этапе наше уравнение представляется с помощью $\hat{Y}_i + \hat{V}_i$, т. е.

$$117{,}57 - 0{,}74X_4 - 10{,}10 + 1{,}35X_1,$$

или

$$107{,}47 - 0{,}74X_4 + 1{,}35X_1.$$

Теперь остатками служат, как и ожидалось, $V_i - \hat{V}_i = Y_i - \hat{Y}_i - \hat{V}_i$. Другие переменные могут быть добавлены аналогичным образом.

На каждом этапе строится регрессия текущих остатков в зависимости от новой переменной до тех пор, пока регрессия станет незначимой. Процесс заканчивается без включения последней переменной. Завершая наш пример на втором этапе, мы видим, что окончательное уравнение не совпадает с МНК-уравнением, содержащим X_1 и X_4.

Для сравнения укажем МНК-уравнение для этой модели (см. приложение Б, с. 291):

$$\hat{Y} = 103{,}10 - 0{,}61X_4 + 1{,}44X_1.$$

Ниже приводятся разные формулы для вычисления b-коэффициентов.

МНК	Ступенчатый метод
$b_4 = \dfrac{\sum_{i=1}^{n} x_{4i}y_i - \dfrac{\sum_{i=1}^{n} x_{4i}x_{1i} \sum_{i=1}^{n} x_{1i}y_i}{\sum_{i=1}^{n} x_{1i}^2}}{\sum_{i=1}^{n} x_{4i}^2 - \dfrac{\left(\sum_{i=1}^{n} x_{4i}x_{1i}\right)^2}{\sum_{i=1}^{n} x_{1i}^2}}$	$b_4 = \dfrac{\sum_{i=1}^{n} x_{4i}y_i}{\sum_{i=1}^{n} x_{4i}^2}$
$b_1 = \dfrac{\sum_{i=1}^{n} x_{1i}y_i - \dfrac{\sum_{i=1}^{n} x_{4i}x_{1i} \sum_{i=1}^{n} x_{4i}y_i}{\sum_{i=1}^{n} x_{4i}^2}}{\sum_{i=1}^{n} x_{1i}^2 - \dfrac{\left(\sum_{i=1}^{n} x_{4i}x_{1i}\right)^2}{\sum_{i=1}^{n} x_{4i}^2}}$	$b_1 = \dfrac{\sum_{i=1}^{n} x_{1i}y_i - \dfrac{\sum_{i=1}^{n} x_{4i}x_{1i} \sum_{i=1}^{n} x_{4i}}{\sum_{i=1}^{n} x_{4i}^2}}{\sum_{i=1}^{n} x_{1i}^2}$

где

$$x_{ji} = X_{ji} - \overline{X}_j,$$
$$y_i = Y_i - \overline{Y}.$$

Сравним коэффициент b_1, полученный по методу наименьших квадратов (обозначен $b_{1,\,\mathrm{LS}}$), с b_1, полученным ступенчатым методом (обозначен $b_{1,\,\mathrm{SW}}$):

$$b_{1,\,\mathrm{SW}} = b_{1,\,\mathrm{LS}} \left[\frac{\Sigma x_1^2 - \dfrac{(\Sigma x_4 x_1)^2}{\Sigma x_4^2}}{\Sigma x_1^2} \right] = b_{1,\,\mathrm{LS}} \left[1 - \frac{(\Sigma x_4 x_1)^2}{\Sigma x_1^2 \Sigma x_4^2} \right].$$

Мы можем переписать теперь эту формулу в виде

$$b_{1,\,SW} = b_{1,\,LS}[1 - r_{14}^2],$$

где r_{14}^2 — квадрат обычного коэффициента корреляции между X_1 и X_4. Таким образом, оценки ступенчатого регрессионного метода по абсолютной величине меньше, чем МНК-оценки. Коэффициент пропорциональности зависит от квадрата коэффициента корреляции. Это, в свою очередь, показывает, что введенная переменная фактически более важна, чем это видно из регрессионной зависимости, построенной по остаткам.

Несмотря на то что этот метод будет всегда менее точным, т. е. будет давать больший остаточный средний квадрат, чем метод наименьших квадратов, он имеет следующие преимущества. Он позволяет выбрать первую переменную не на основе ее корреляции с Y, а исходя из других соображений. Допустим, например, что имеется набор переменных X, сильно коррелированных друг с другом; согласно обычным процедурам отбора в качестве первой переменной выбирается та, скажем, X_1, которая наиболее сильно коррелирована с Y. На следующей стадии все другие переменные, не входящие в уравнение, корректируются на X_1. Поэтому, если, скажем, X_2 сильно коррелирована с X_1, она может быть отвергнута как возможная переменная. Однако величина X_2 может быть как раз той переменной, которую желает использовать экспериментатор. Например, X_2 может быть непосредственно управляемой переменной, в то время как X_1 может быть неуправляемой. К примеру, при моделировании торговой ситуации X_2 может представлять собой собственные расходы на рекламу, тогда как X_1 — расходы на рекламу со стороны конкурента. Таким образом, экспериментатор может принять регрессионную зависимость $\hat{Y} = f(X_2)$ и затем продолжить рассмотрение задачи, используя остатки от этой регрессии как значения зависимой переменной при построении следующих регрессий.

Дополнительно укажем, что существуют ситуации, когда нужно устранить тренд в данных, прежде чем пытаться составить уравнение для предсказания. Экономисты часто корректируют данные относительно тренда или сезонности и затем переходят к анализу результирующих вариаций с помощью метода наименьших квадратов.

Мнение. Одна из особенностей ступенчатого регрессионного метода состоит в том, что переменные могут вводиться таким образом, чтобы можно было сохранить ожидаемое направление действия любых эффектов (в противном случае их не следует вводить). Линейное МНК-уравнение не всегда позволяет этого добиться, ввиду корреляции между переменными из-за специфических особенностей области пространства переменных X, в которой находятся наши данные. Это не приносит вреда при условии, что мы не пытаемся выделять определенные члены в модели. Настоящее МНК-уравнение обычно обладает лучшими свойствами в отношении предсказания, чем уравнение, полученное ступенчатым способом. По этой причине ступенчатый рег-

рессионный метод не рекомендуется для типичных производственных задач. О его пригодности в экономических ситуациях можно прочесть в статьях, указанных в библиографии.

6.12. РЕЗЮМЕ

Как мы видели, все МНК-процедуры отбора переменных привели к уравнению $\hat{Y} = 52{,}58 + 1{,}47X_1 + 0{,}66X_2$ как «наилучшему» для данных Хальда. Метод всех возможных регрессий также дал уравнение $\hat{Y} = 103{,}10 + 1{,}44X_1 - 0{,}61X_4$ как второй подходящий вариант, но последующие процедуры показали, что это уравнение менее желательно, чем модель, содержащая X_1 и X_2. (Процедура гребневого следа привела нас к другой модели, отличной от модели метода наименьших квадратов, которая включает все предикторы X, но имеет смещенные оценки параметров.)

Хотя во многих случаях все МНК-процедуры приводят к одному и тому же уравнению, это не всегда имеет место, как показано Хамакером (H a m a k e r H. C. On multiple regression analysis.— Statistica Nederlandica, 1962, **16**, p. 31—56).

В теоретическом отношении наилучшим будет метод всех возможных регрессий, поскольку он позволяет «исследовать все». Благодаря существованию метода выбора «наилучшего подмножества» предикторов (см. 6.2) теперь можно не прибегать к методу всех возможных регрессий. Однако в случае данных Хальда как метод исключения, так и шаговый регрессионный метод приводят к одному и тому же уравнению. Если рассматриваются все регрессии, то выбор предикторов во многом зависит от уровней, при которых отвергаются различные гипотезы с помощью F-критерия, а также от отношения статистиков к желаемой степени увеличения величины R^2. Необоснованный выбор может привести к совершенно различным уравнениям при использовании разных методов, и в этом нет ничего удивительного.

М н е н и е. Мы предпочитаем использовать в практических ситуациях шаговый регрессионный метод. Если есть желание исследовать уравнение, полученное шаговым методом, мы отдаем предпочтение *[11] процедуре выбора «наилучшего подмножества» предикторов с использованием C_p-статистики. Применять метод всех возможных регрессий неразумно, исключая тот случай, когда число предикторов мало́. В целом мы пришли к выводу, что процедуры, не основанные на методе наименьших квадратов, мало полезны на практике, хотя мы поддерживаем метод гребневой регрессии, если он применяется при соответствующих условиях, описанных в § 6.7. Если возникают трудности с трактовкой прогноза по построенной модели или функций от оценок регрессионных коэффициентов, то желательно проверить устойчивость модели на основе фактических данных и попытаться выяснить, в чем суть дела, вместо того чтобы вслепую применять

*[11] Некоторые исследователи сначала применяют метод выбора «наилучшего подмножества» предикторов, а затем — шаговый метод или какой-либо его аналог.

метод, ограничения которого не совсем понятны. Конечно, располагая современными ЭВМ, теперь можно без особого труда выполнить все расчеты, которые мы обсуждали, но может потребоваться несметное количество бумаги. Лучше всего работать с каким-то одним методом и овладеть его специфическими особенностями. Такой подход вероятно выгоднее, чем в течение длительного времени выбирать подходящий метод. Никакой метод не будет хорошо работать при всех условиях, как бы хорошо он не проявил себя на частном примере. Нет метода, который был бы всегда лучше всех остальных. (Если бы это было так, то данная глава была бы куда короче!) Необходимо постоянно иметь в виду, что, если данные беспорядочны, а не получены с помощью специально спланированных экспериментов, любая модель отражает ограничения, которые вытекают из структуры данных и практических ограничений задачи. Методы, описанные в этой главе, могут быть полезными. Однако никакой из них не может компенсировать здравый смысл и жизненный опыт.

6.13. ВЫЧИСЛИТЕЛЬНЫЕ АСПЕКТЫ ШАГОВОЙ РЕГРЕССИИ

Этот параграф был в первом издании, но во втором издании мы его опустили ввиду широкого распространения программ шаговой регрессии. Если вас интересует содержание данного параграфа, обратитесь к первому изданию.

6.14. РОБАСТНАЯ (УСТОЙЧИВАЯ) РЕГРЕССИЯ

Если мы строим МНК-регрессию, используя n наблюдений, в виде p-параметрической модели $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, то мы принимаем определенные идеализированные предположения о векторе ошибок $\boldsymbol{\varepsilon}$, а именно считаем, что они распределены согласно $N(\mathbf{0}, \mathbf{I}\sigma^2)$. На практике отклонения от этих предположений имеют место. Если такие отклонения существенны, то можно надеяться опознать их по поведению остатков и затем подходящим образом скорректировать модель и/или метрику переменных. Часто отклонения, если они имеют место вообще, не слишком серьезны, чтобы осуществлять коррекцию, тогда анализ проводится обычным путем.

Определенные типы отклонений от идеализированных предположений выдвигаются как наиболее вероятные чаще, чем другие. Распределение случайных ошибок может быть симметричным, но не быть нормальным. Оно может быть «более заостренное», чем нормальное, и с «более легкими хвостами» или «менее заостренное», чем нормальное, и с «более тяжелыми хвостами». Или, даже если распределение нормальное, данные могут содержать выбросы, т. е. наблюдения, кокоторые нетипичны для обычного нормального распределения, возможно, из-за того, что они имеют другое среднее, или потому, что они принадлежат к нормальному распределению, но со значительно большей дисперсией, чем σ^2.

Разработаны рекомендации, как бороться с теми или другими возможными недостатками. Вместо МНК-процедуры предлагается ис-

пользовать методы построения *робастной регрессии*. Основное достоинство, приписываемое этим методам, состоит в том, что они менее чувствительны, нежели обычный МНК, к типичным отклонениям от принятых предположений, которые встречаются на практике. Эта идея очень привлекательна на первый взгляд, но она имеет определенные недостатки. Мы кратко обсудим робастные М-оцениватели в регрессии, выскажем свое мнение и укажем ссылки на литературу для более глубокого ознакомления с данной темой.

M-Оцениватели

M-оцениватели относятся к оценивателям «максимально правдоподобного» типа. Предположим, что ошибки распределены независимо и все принадлежат одному и тому же распределению $f(\varepsilon)$. Тогда оцениватель, основанный на методе максимального правдоподобия (MLE), для $\boldsymbol{\beta}$ выражается вектором $\hat{\boldsymbol{\beta}}$, который максимизирует величину

$$\prod_{i=1}^{n} f(Y_i - \mathbf{x}_i' \boldsymbol{\beta}), \tag{6.14.1}$$

где $\mathbf{x}_i'$ есть i-я вектор-строка матрицы $\mathbf{X}$, $i = 1, 2, \ldots, n$, в модели $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Если $f(\varepsilon)$ есть плотность нормального распределения, то выражение (6.14.1) совпадает с (2.6.5). Эта же оценка максимизирует

$$\sum_{i=1}^{n} \ln f(Y_i - \mathbf{x}_i' \boldsymbol{\beta}). \tag{6.14.2}$$

В более общем случае мы можем следовать логике, содержащейся в работе Эндрьюса (A n d r e w s D. F. A robust method for multiple linear regression.— Technometrics, 1974, **16**, p. 523—531), и определить как *M*-оцениватель вектора параметров $\boldsymbol{\beta}$ вектор, который максимизирует величину

$$\sum_{i=1}^{n} \psi(e_i/s), \tag{6.14.3}$$

где ψ есть функция, которая придает более высокие веса некоторым величинам e_i/s (обычно меньшим единицы) и более низкие веса другим e_i/s (обычно большим единицы). В общем случае эта функция максимизируется численно итерационными методами. В качестве примера такой функции может служить

$$\psi(x) = \begin{cases} [1 + \cos(x/c)]\, c & \text{для} \quad |x| \leqslant c\pi, \\ 0 & \text{для прочих } x \end{cases} \tag{6.14.4}$$

при некотором определенном выборе c. Очевидно, существует много других выражений этой функции.

Когда робастные оцениватели оправданы?

Положение здесь во многом сходно с тем, которое имеет место в случае гребневой регрессии. Любой определенный робастный оцениватель оправдан, если он приводит (точно или приближенно) к оценкам максимального правдоподобия для параметров при тех предположениях об ошибках, которые считаются верными, если предположение $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2)$ *неверно*. Безрассудное применение робастных оценивателей подобно безрассудному использованию гребневых оценивателей. Они могут быть пригодными, но могут быть и неуместными. Основная проблема в том, что мы не знаем, какие робастные оцениватели и при каких типах предположений об ошибках целесообразно применять; хотя некоторая работа в этом направлении уже проведена (см., например: Chen G. G. Studies in Robust Estimation.— University of Wisconsin at Madison, Ph. D. Thesis, 1979).

Что же надо делать на практике?

Если мы предполагаем, что ошибки не имеют «идеального» распределения $N(0, \mathbf{I}\sigma^2)$, то надо попытаться сформулировать альтернативный закон распределения. Затем надо постулировать структуру параметрической модели, отвечающую этим альтернативным предположениям об ошибках. После этого для оценивания параметров целесообразно применить метод максимального правдоподобия или эквивалентный ему байесовский метод.

Мнение. Мы считаем, что использование методов робастной регрессии нецелесообразно в настоящее время, до тех пор пока не будут сформулированы правила, позволяющие решать при *каких* обстоятельствах, и *какой* робастный метод следует использовать, и пока не будет выполнена их проверка. Если модель (которая включает предположения о распределении ошибок) неправильна, то можно подходящим образом изменить модель и использовать снова метод максимального правдоподобия, т. е. *не* следует менять метод оценивания. (Как и в случае гребневой регрессии, следует применять метод лишь тогда, когда он соответствует обстоятельствам.) [21]

6.15. НЕКОТОРЫЕ ЗАМЕЧАНИЯ О ПАКЕТАХ ПРИКЛАДНЫХ ПРОГРАММ ПО СТАТИСТИКЕ

BMDP-79. Пакет биомедицинских программ

(BMDP-79: Biomedical Compnter programs — P Series, 1979 Edition)

В версии этого издания имеются шесть линейных и три нелинейные регрессионные программы.

[21] В добавление к указанным выше работам по робастному оцениванию приведем книги на русском языке: Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия/Пер. с англ. Под ред. Ю. П. Адлера.— М.: Финансы и статистика, 1982, вып. 1; 1982, гл. 10; Хьюбер П. Робастность в статистике/Пер. с англ. Под ред. И. Г. Журбенко.— М.: Мир, 1984.— 304 с.; Устойчивые статистические методы оценки данных/Под ред. Р. Л. Лонера, Г. Н. Уилкинсона)/Пер. с англ. Под ред. Н. Г. Волкова.— М.: Машиностроение, 1984.— 232 с.— *Примеч. пер.*

Линейные регрессионные программы:
P1R Множественная линейная регрессия.
P2R Шаговая регрессия.
P4R Регрессия на главные компоненты.
P5R Регрессия полиномиальная.
P6R Частные корреляции и многомерная регрессия.
P9R Все возможные подмножества регрессий.
Нелинейные регрессионные программы:
P3R Нелинейная регрессия.
PAR Нелинейная регрессия без вычисления производных.
PLR Шаговая логистическая регрессия.

Руководство к пакету биомедицинских программ BMDP-79 является более трудным для чтения, чем руководство к статистическому пакету для общественных наук SPSS. (Чтобы модифицировать инструкции к программам, требуется некоторое знакомство с Фортраном.) Однако руководство к пакету биомедицинских программ гораздо лучше руководства к статистическому пакету для общественных наук при работе с регрессиями. Регрессионные программы приведены в гл. 13 и 14, в каждой из которых имеется обычно введение к содержащимся в ней программам. Вместо того чтобы обсуждать каждую программу в отдельности, мы укажем общие достоинства и недостатки программ, выявившиеся в процессе пользования.

В случае большого массива разнородных данных, когда число предикторных переменных превышает 30, пригодна только программа P2R, да и то лишь тогда, когда используется шаговая процедура включения. Программа P2R, в принципе применима и для процедуры исключения, но мы пришли к выводу, что от нее мало пользы. К тому же, если матрица $\mathbf{X}'\mathbf{X}$, составленная для *всех* предикторов, почти вырожденная, P2R не является хорошей программой для применения процедуры исключения, поскольку в этом случае вычисления выполняются с одинарной точностью. Однако это не ограничение, когда применяется шаговая процедура включения.

Программа P9R превосходна. В своей регулярной форме она ограничена использованием максимум 27 предикторов. Можно увеличить число предикторов, но это потребует модификации программы и увеличения объема памяти. Как и в предыдущем случае, если матрица $\mathbf{X}'\mathbf{X}$ для входных предикторных переменных почти вырожденная, программа не работает. Большим достоинством этого пакета программ является широкий выбор процедур анализа остатков.

Полезное качество всех этих программ состоит в том, что они позволяют использовать взвешенную регрессию со специально определенными весами. Так, например, назначая вес, равный единице, для одних опытов, и равный нулю — для других, программа позволяет построить регрессию по подмножеству опытов с единичными весами, и проверить ее по опытам с нулевыми весами. Это процедура проверки достоверности модели, которая обсуждается в гл. 8.

Если в наборе данных имеются недостающие значения, то каждый опыт, содержащий по меньшей мере одно недостающее значение, следует исключать из обработки. Если желают использовать только

те предикторные переменные X, для которых нет отсутствующих значений, то необходимо тщательно сформулировать описание формата, чтобы исключить такие переменные. Если данные полные, можно исключать переменные и можно установить произвольный порядок расположения переменных в регрессии, используя определенные правила выбора.

Нелинейные программы удобны для использования и показали себя очень полезными.

SPSS: Статистический пакет для общественных наук (Statistical Package for the Social Sciences)

Эта система содержит одну основную регрессионную программу с заголовком «Множественный регрессионный анализ: подпрограмма регрессии». До сентября 1980 г. этот пакет содержал только процедуру включения *[12]. Однако система SPSS обладает некоторыми приятными особенностями, благодаря которым пользователь может получать дополнительные результаты.

Каждая потенциальная предикторная переменная может быть сразу включена в модель, исключена из уравнения, может быть введена в модель согласно шаговой процедуре, причем в заранее определенном порядке. Все эти альтернативы могут быть реализованы с помощью соответствующей формулировки задания на решение регрессионной задачи. Приведем в качестве примера некоторые возможные формулировки задания на построение регрессии по данным Хальда, где отклик Y обозначен как X_5, предикторами являются X_1, X_2, X_3, X_4:

а) Регрессия = X_5 с X_1 (8), X_2 (4), X_3 (2), X_4 (6).
б) Регрессия = X_5 (3; 4,00; 0,01) с X_1 до X_4 (1)
в) Регрессия = X_5 с X_1 до X_4 (1).
г) Регрессия = X_5 с X_1 до X_2 (2), X_3 до X_4 (1).
д) Регрессия = X_5 с X_1 (6), X_2 (0), X_3 (2), X_4 (4).
е) Регрессия = X_5 с X_1 до X_2 (2). Остат. = 0.

Задание а) означает, что сначала подгоняется модель вида $\hat{X}_5 = f(X_1)$, так как цифра (8) в круглых скобках есть наибольшее четное число. Затем подгоняется модель $\hat{X}_5 = f(X_1, X_4)$, поскольку после X_4 в скобках приведено наибольшее четное число из оставшихся, т. е. не считая 8. Потом подгоняется модель $\hat{X}_5 = f(X_1, X_4, X_2)$ и, наконец, модель $\hat{X}_5 = f(X_1, X_4, X_2, X_3)$. В итоге мы получили запись выражения модели, где предикторы в скобках расположены в порядке уменьшения четных чисел, приведенных в задании после каждого предиктора в круглых скобках.

Задание б) означает шаговую процедуру включения в модель всех предикторных переменных от X_1 до X_4. Это обозначено с помощью

*[12] Как следует из публикаций, в пакет SPSS ввели программу «Новая регрессия» (New Regression), которая имеет многие характерные особенности других систем.

единицы в скобках после X_4. Цифры в скобках после зависимой переменной X_5 определяют три параметра (n, F, T). Здесь $n = 3$ обозначает максимальное число предикторов, которые могут быть введены в модель с использованием F- и T-критериев; $F = 4{,}00$ есть пороговое значение для включения новой переменной, если величина F превосходит указанный порог. T есть та доля вариации, которая относится к новой переменной и не может быть отнесена на счет переменных, которые ранее содержались в модели. Переменная сохраняется в уравнении, если T превосходит величину 0,01, или больше, чем 1 %.

Задание в) соответствует шаговой процедуре включения всех переменных в модель, от X_1 до X_4, при условии, что реализация этой процедуры осуществляется с использованием определенных значений параметров (n, F, T). В данном пакете программ эти параметры по умолчанию имеют значения $n = 80$, $F = 0{,}01$ и $T = 0{,}001$. По существу это означает, что ограничений нет совсем.

Задание г) означает, что необходимо построить комбинированную регрессию, в которой часть переменных сразу вводится в модель, а оставшиеся — пошагово. А именно переменные X_1 и X_2 включаются в модель сразу, а переменные X_3 и X_4 — пошагово. Причем делается это с использованием численных значений F и T, указанных при расшифровке предыдущего задания.

Задание д) означает, что в модель вводятся три переменные — X_1, X_4 и X_3, причем порядок их введения, как и в задании а), определяется цифрами в круглых скобках. Итоговая модель имеет вид $\hat{X}_5 = f(X_1, X_4, X_3)$. Переменная X_2 в процедуре не участвует, поскольку в следующих за ней скобках стоит 0.

Задание е) предполагает, что в модель сразу вводятся переменные X_1 и X_2 и вычисляются остатки. Могут быть построены графики стандартизированных остатков, стандартизированных остатков в зависимости от значений стандартизированного отклика. Чтобы выполнить это, надо добавить слово «statistics». Могут быть вычислены также средние, стандартные отклонения и корреляции *всех* входных данных.

Если строится множество регрессий, то вычисления остатков и построение графиков проводятся в конце реализации программы на ЭВМ, а не в том порядке, как сказано в задании.

Укажем лишь одно предостережение: *наш* опыт использования говорит о том, что регрессионная программа требует большого объема машинной памяти.

Руководство для пользователей легко читается и воспринимается.

SAS: **Статистический анализ систем** (Statistical Analysis System)

Программы SAS превосходны, и, в частности, эта система включает очень простые и мощные процедуры обработки данных. Преимущества при обработке данных плюс возможность доступа к использованию пакета BMDP делают этот пакет достойным внимания для тех,

кто располагает компьютерами из серии IBM или компьютером, который может быть подсоединен к IBM.

Программы SAS существуют в двух формах. В виде стандартной библиотеки программ, включающей программы SAS полностью, и в виде дополнительной библиотеки программ, которая не полностью опирается на SAS.

Стандартная библиотека программ содержит 4 программы линейной регрессии и одну нелинейную программу.

К линейным программам относятся:

GLM. Общие линейные модели.

R Square. Все возможные регрессии.

Stepwise. Шаговая регрессия.

SYS Reg. Системы регрессий.

Нелинейная программа имеет имя NLIN.

GLM-процедура есть общая программа целевого назначения для обработки и анализа определенной модели. Процедура *R* Square представляет собой метод всех возможных регрессий с выводом на печать R^2- и C_p-статистики Маллоуза для каждой модели. Программа Stepwise содержит метод включения, метод исключения и шаговый метод, усовершенствованные процедуры выбора максимального (MAXR) и минимального (MINR) критерия R^2. Последняя программа из линейных позволяет строить одно-, двух- и трехшаговые регрессии и делать некоторые другие стандартные выводы.

В дополнительной библиотеке программ содержится много программ специального назначения.

Logist. Логистическая регрессия.

PLglm. Регрессионная модель Кокса для таблиц дожития [22].

LAV. Линейная модель, основанная на минимизации максимального по абсолютной величине отклонения.

Minitab: **Интерактивная (и пакетная) статистическая система вычислений на ЭВМ**

Эта система содержит две регрессионные команды: REGRESS и STEPWISE, каждая из которых допускает работу со 100 переменными на большинстве ЭВМ.

По команде REGRESS строится множественная регрессия. Должны быть определены переменные, которые включаются в модель, а также порядок, в котором они должны включаться (он соответствует порядку, в котором они записываются). Существует несколько вариантов представления конечных и промежуточных результатов. BRIEF output дает (19 + число предикторов) компактных строк с результатами, которые размещаются на экране дисплея. Вариант NOBRIEF позволяет представить больше конечных результатов, включая предсказываемые значения, стандартные отклонения предсказываемых значений, стандартизированные остатки, метки для то-

---

[22] См. книгу: Cox P. R. Life Tables: The Measure of Mortality, 1975.— *Примеч. пер.*

чек с большими остатками или большим влиянием на регрессию, отдельные числа степеней свободы, разложения сумм квадратов отклонений и элементы матрицы $(\mathbf{X}'\mathbf{X})^{-1}$. Полезные промежуточные подробности выдаются в соответствии с вариантом BRIEF относительно маркированных точек. Стандартизированные остатки, предсказываемые значения, коэффициенты и элементы матрицы $(\mathbf{X}'\mathbf{X})^{-1}$ можно еще и сохранить для дальнейшего анализа. С помощью простых манипуляций со стандартизированными остатками можно строить различные диагностические графики и доверительные интервалы для предсказываемых откликов и других наблюдений. Могут задаваться веса. Возможно использование в модели фиктивных переменных. Модели могут не содержать свободного члена. Данные вычеркиваются только в том случае, когда пропущенные наблюдения встречаются у переменных, фактически используемых в регрессии.

STEPWISE содержит процедуры включения и исключения, а также обычную шаговую процедуру. FENTER и FREMOVE можно задавать (по умолчанию=4). Пользователь может сразу включить некоторые отобранные переменные, другие могут вводиться, но все же быть кандидатами на удаление, тогда как некоторые могут быть исключены (REMOVED) на определенном шаге, но все же оставаться кандидатами для включения на следующих шагах. Ближайшие K наилучших (BEST) предикторов можно вывести на печать, и они будут перечисляться на каждом шаге вместе со значениями t-критерия, для включения предикторов в модель. Число шагов (STEPS) можно задавать с пульта управления (терминала). Если работа с программой проводится в режиме диалога (в интерактивном режиме), пользователь может вмешиваться в конце каждого кадра дисплея. Он может, например, вводить или исключать переменные или изменять значения F-критерия. Выходные данные представляются в виде компактной итоговой таблицы. Наиболее ответственные вычисления выполняются с двойной точностью. Данные не приводятся, если пропущенные наблюдения относятся к переменным, которые состоят в списке кандидатов на включение. Свободный член может быть опущен.

Общей особенностью системы Minitab является легкость, с которой проводятся манипуляции с данными и их анализ. Например, можно произвести преобразование, опустить некоторые опыты, построить регрессию, график остатков, составить гистограмму, построить другую регрессию, рассчитать расхождения между двумя наборами остатков и работать с представлением типа «опора и консоль» [23] для их разностей, причем все с помощью команд на языке, похожем на английский.

Система программ Minitab пригодна для эксплуатации на самых различных ЭВМ, как в режиме диалога, так и в пакетном режиме. Документация включает элементарное руководство на 348 с., а также инструкции для наиболее квалифицированных пользователей и инструкции по оказанию оперативной помощи (HELP).

[23] Этот прием описан, в частности, в книге: Мостеллер Ф., Тьюкн Дж. Анализ данных и регрессия/Пер. с англ. Под ред. Ю. П. Адлера.— М.: Финансы и статистика, 1982, вып. 1, гл. 3.— *Примеч. пер.*

Мнение. Мы считаем, что пакет регрессионных программ BMDP чрезвычайно полезен. В особенности нам нравится формат выходных данных. Однако простота SAS — в манипулировании с набором данных, преобразованиях переменных и др.— дает пользователям этой системы возможность ее присоединения к программам BMDP. А это объединяет все преимущества обеих систем. Пакет SPSS мы применяли мало. Однако он очень полезен при анализе таблиц сопряженности и категоризованных данных. А эти направления очень важны в общественных науках. Minitab отличается компактным представлением выходных данных, что делает его удобным для работы в диалоговом режиме. В пакетном режиме Minitab несколько легче использовать, чем SAS, но его возможности более ограничены. Он требует меньше затрат при работе, чем SAS, и потому годится для значительно большего числа типов ЭВМ [24].

ПРИЛОЖЕНИЕ 6А. КАНОНИЧЕСКАЯ ФОРМА ГРЕБНЕВОЙ РЕГРЕССИИ

Для получения канонической формы гребневой регрессии надо произвести поворот β_z-осей и совместить их с новыми осями, которые параллельны главным осям контуров поверхности суммы квадратов отклонений (см. рис. к упражнению 6.1). Пусть $\lambda_1, \lambda_2, \ldots, \lambda_r$ есть собственные числа матрицы $\mathbf{Z'Z}$ и пусть

$$\mathbf{G} = (\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_r) \tag{6A.1}$$

представляет собой $(r \times r)$-матрицу, $(r \times 1)$-столбцы которой $\mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_r$ есть собственные векторы матрицы $\mathbf{Z'Z}$. Это означает, что числа λ_i и векторы $\mathbf{g}_j$ удовлетворяют соотношениям

$$\mathbf{Z'Z}\mathbf{g}_i = \lambda_i \mathbf{g}_i, \quad i = 1, 2, \ldots, r. \tag{6A.2}$$

Чтобы получить эти векторы, решим полиномиальное уравнение

$$\mathrm{Det}\,(\mathbf{Z'Z} - \lambda \mathbf{I}) = 0. \tag{6A.3}$$

Значения $\lambda_1, \lambda_2, \ldots, \lambda_r$ являются его корнями. (В реальных практических задачах все корни, как правило, различны. Если они не таковы, что иногда бывает при планировании эксперимента, а также и в других случаях, то это означает, что определенные сечения поверхности суммы квадратов получились круговыми или сферическими, и не надо никак поворачивать соответствующие оси, поскольку они уже имеют нужное направление. Теперь подставим λ_i снова в уравнения (6А.2) и выберем соответствующие решения $\mathbf{g}_i$, чтобы они были

[24] Большая подборка программ регрессионного анализа собрана в книге: Песаран М., Слейтер Л. Динамическая регрессия: теория и алгоритмы/Пер. с англ. Под ред. Э. Б. Ершова.— М.: Финансы и статистика, 1984.— 310 с.; обстоятельный обзор как зарубежных, так и отечественных программ такого типа содержится в книге: Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Исследование зависимостей.— М.: Финансы и статистика, 1985.— 488 с.; интересные программы можно найти также в книге: Алгоритмы и программы восстановления зависимостей/Под ред. В. Н. Вапника.— М.: Наука, 1984.— 816 с.— *Примеч. пер.*

нормализованными так: $\mathbf{g}_i'\mathbf{g}_i = 1$. При этом выполняются следующие соотношения:

$$\mathbf{G}'\mathbf{G} = \mathbf{G}\mathbf{G}' = \mathbf{I}, \tag{6A.4}$$

$$\mathbf{G}'\mathbf{Z}'\mathbf{Z}\mathbf{G} = \mathbf{\Lambda}, \tag{6A.5}$$

где $\mathbf{\Lambda} = \text{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_r)$ есть диагональная матрица, содержащая числа λ_i на главной диагонали, тогда как остальные ее элементы равны нулю. Введем $\mathbf{Q}$ и $\boldsymbol{\gamma}$ следующим образом:

$$\mathbf{Q} = \mathbf{Z}\mathbf{G}, \quad \boldsymbol{\gamma} = \mathbf{G}'\boldsymbol{\beta}_z, \tag{6A.6}$$

так что

$$\mathbf{Z} = \mathbf{Q}\mathbf{G}', \quad \boldsymbol{\beta}_z = \mathbf{G}\boldsymbol{\gamma}. \tag{6A.7}$$

Тогда

$$\mathbf{Z}'\mathbf{Z} + \theta\mathbf{I} = \mathbf{G}\,(\mathbf{\Lambda} + \theta\mathbf{I})\,\mathbf{G}', \tag{6A.8}$$

и, поскольку $\mathbf{G}' = \mathbf{G}$ в силу (6A.4), имеем

$$(\mathbf{Z}'\mathbf{Z} + \theta\mathbf{I})^{-1} = \mathbf{G}\,(\mathbf{\Lambda} + \theta\mathbf{I})^{-1}\,\mathbf{G}'. \tag{6A.9}$$

Гребневый оцениватель выражается, таким образом, формулой

$$\mathbf{b}_z(\theta) = \mathbf{G}\,(\mathbf{\Lambda} + \theta\mathbf{I})^{-1}\,\mathbf{G}'\mathbf{Z}'\mathbf{Y}. \tag{6A.10}$$

Умножая левую и правую части на $\mathbf{G}'$ и учитывая, что $\mathbf{G}' = \mathbf{G}^{-1}$, найдем

$$\hat{\boldsymbol{\gamma}}(\theta) = (\mathbf{\Lambda} + \theta\mathbf{I})^{-1}\,\mathbf{Q}'\mathbf{Y}. \tag{6A.11}$$

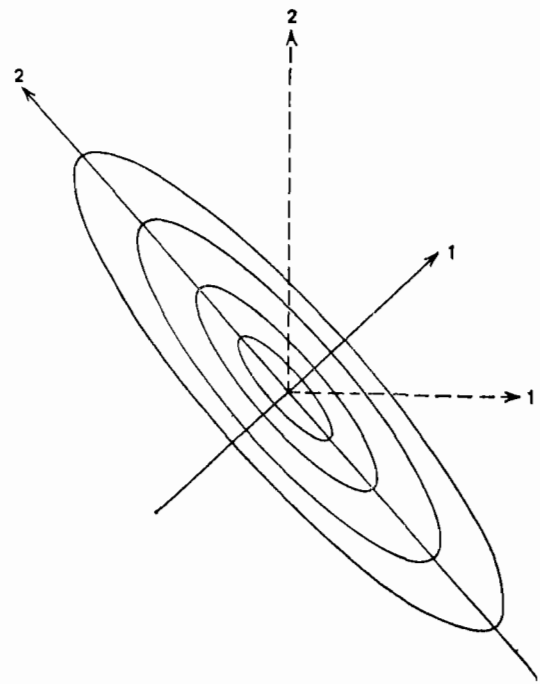


Рис. 6A.1. Контуры поверхности суммы квадратов отклонений для $r = 2$. (Заметим, что центр системы координатных осей смещен в точку β_z, соответствующую МНК-оценкам параметров.) Исходные оси изображены пунктиром, а новые — сплошными линиями

Если мы представим теперь i-й компонент вектора $\mathbf{Q}'\mathbf{Y}$ в виде

$$c_i = (\mathbf{Q}'\mathbf{Y})_i \tag{6A.12}$$

и обратим диагональную матрицу $\mathbf{\Lambda} + \theta\mathbf{I}$, то получим выражение для элементов вектора $\hat{\boldsymbol{\gamma}}(\theta)$, т. е.

$$\hat{\gamma}_i(\theta) = c_i/(\lambda_i + \theta), \quad i = 1, 2, \ldots, r. \tag{6A.13}$$

Это соотношение представляет собой *каноническую форму* гребневой регрессии. Однако на графиках гребневого следа обычно изобра-

жается зависимость $b_{jz}(\theta)$ или $b_j(\theta)$. На основании (6А.7) можно записать

$$\mathbf{b}_z(\theta) = \mathbf{G}\hat{\boldsymbol{\gamma}}(\theta) \tag{6А.14}$$

и далее

$$\mathbf{b}(\theta) = \{\mathrm{diag}\,(S_{jj}^{-1/2})\}\,\mathbf{G}\hat{\boldsymbol{\gamma}}(\theta). \tag{6А.15}$$

Каноническая форма гребневой регрессии часто встречается в работах, посвященных этому типу регрессий. Она удобна в вычислительном отношении, когда отыскивается гребневый след.

Остаточная сумма квадратов

Она может быть записана в виде

$$\mathrm{RSS}_\theta = \mathrm{RSS}_0 + \theta^2 \sum_{i=1}^{r} \hat{\gamma}_i^2(\theta)/\lambda_i, \tag{6А.16}$$

где

$$\mathrm{RSS}_0 = \{\Sigma Y_i^2 - n\overline{Y}^2\} - \sum_{i=1}^{r} c_i^2/\lambda_i \tag{6А.17}$$

есть минимальное значение, достигаемое МНК-оценивателем $\hat{\boldsymbol{\gamma}}(\theta)$ при $\theta = 0$. Возможный способ выбора значения θ^* для θ основан на том, что разность $\mathrm{RSS}_\theta - \mathrm{RSS}_0$ приравнивается к некоторому априори выбранному числу или выражению. Можно, например, приравнять эту разность к величине $s^2(r+1)$. Это оправдано, поскольку $\sigma^2(r+1)$ есть математическое ожидание разности $E(\mathrm{RSS}_\theta - \mathrm{RSS}_0)$, где используется истинная величина $\boldsymbol{\gamma}$ вместо $\hat{\boldsymbol{\gamma}}(\theta)$. Это означает, что θ выбирается таким образом, чтобы квадрат расстояния, выражаемый величиной $\mathrm{RSS}_\theta - \mathrm{RSS}_0$, имел значение, которое мы «ожидаем». (Таким образом мы пытаемся «удержать» оценку примерно на том самом доверительном контуре, на который в среднем попадает истинная величина. Направление $\hat{\boldsymbol{\gamma}}_0$, на котором будет лежать точка, соответствующая оценке, однако, может отличаться от направления, на котором лежит точка, соответствующая истинной величине.) Следуя изложенному способу, получим уравнение

$$\theta^* = \left[s^2(r+1) \Big/ \left\{ \sum_{i=1}^{r} \gamma_i^2(\theta^*)/\lambda_i \right\} \right]^{1/2}, \tag{6А.18}$$

которое, вероятно, можно решить относительно θ^* с помощью итераций, по аналогии с тем, как была реализована процедура (6.7.10), при условии, что достигается сходимость в соответствии с неравенством типа (6.7.11). Подстановка $\theta = 0$ в правую часть уравнения, (6А.18) дает значение θ^* в результате первой итерации, которое может оказаться при некоторых обстоятельствах достаточно хорошим.

Средний квадрат ошибки

Уравнение (6.7.5) в новых обозначениях приобретает вид

$$\mathrm{MSE}(\theta) = \sum_{i=1}^{r} (\lambda_i \sigma^2 + \theta^2 \gamma_i^2)/(\lambda_i + \theta)^2. \qquad (6A.19)$$

Некоторые альтернативные формулы

Если мы введем обозначения для вектора МНК-оценок

$$\hat{\boldsymbol{\gamma}}(0) = \boldsymbol{\Lambda}^{-1}\mathbf{Q}'\mathbf{Y}, \qquad (6A.20)$$

а диагональную матрицу с элементами $\delta_i = \lambda_i/(\lambda_i + \theta)$ представим в виде

$$\boldsymbol{\Delta} = (\boldsymbol{\Lambda} + \theta\mathbf{I})^{-1}\boldsymbol{\Lambda}, \qquad (6A.21)$$

то увидим, что уравнения (6A.11), (6A.13), (6A.14) и (6A.19) приобретают соответственно вид:

$$\hat{\boldsymbol{\gamma}}(\theta) = \boldsymbol{\Delta}\hat{\boldsymbol{\gamma}}(0), \qquad (6A.22)$$

$$\hat{\gamma}_i(\theta) = \delta_i c_i/\lambda_i, \qquad (6A.23)$$

$$\hat{\mathbf{b}}_z(\theta) = \mathbf{G}\boldsymbol{\Delta}\hat{\boldsymbol{\gamma}}(0), \qquad (6A.24)$$

$$\mathrm{MSE}(\theta) = \mathrm{tr}\,\{\sigma^2\boldsymbol{\Delta}^2\boldsymbol{\Lambda}^{-1} + (\mathbf{I} - \boldsymbol{\Delta})\boldsymbol{\gamma}\boldsymbol{\gamma}'(\mathbf{I} - \boldsymbol{\Delta})\}. \qquad (6A.25)$$

В этих формулах проявляется связь между МНК-оценивателем и гребневым оценивателем.

Упражнения

1. Инженер-экономист, работающий на производстве, отвечает за снижение затрат. Наиболее важная статья расходов на его заводе — ежемесячные затраты на воду, используемую в производстве. Он решил исследовать потребление воды с помощью 17 наблюдений за ее расходом и другими переменными. Экономист уже слышал о множественном регрессионном анализе, но, поскольку он был настроен довольно скептически, добавил к исходным наблюденням столбец из случайных чисел. Полный набор данных приведен после упражнения 3 со средними, стандартными отклонениями и коэффициентами корреляции. По этой информации было построено 7 из $2^5 = 32$ возможных вариантов регрессионных уравнений (для доверительного оценивания коэффициентов β был принят уровень значимости $\alpha = 0{,}05$).

1) Выполнив все вычисления, покажите, что результаты машинных расчетов для регрессионного уравнения [25] первого порядка $\hat{Y} = \hat{X}_6 = f(X_2)$ правильны.

---

[25] В заданиях к этому и ряду других упражнений авторы записывают подлежащие оцениванию регрессии в виде $Y = f(X_1, \ldots)$, однако правильнее представлять их в форме $\hat{Y} = f(X_1, \ldots)$. Мы внесли в текст соответствующие исправления.— *Примеч. пер.*

2) Используя информацию, содержащуюся в приведенных результатах вычислений для регрессин первого порядка $\hat{Y} = \hat{X}_6 = f(X_2, X_4)$, постройте 95 %-ный доверительный интервал для β_4. Сравните его с результатом, указанным в машинных распечатках.

3) Получите обратную матрицу от корреляционной матрицы для регрессионного уравнения первого порядка $\hat{Y} = \hat{X}_6 = f(X_1, X_2, X_4)$. Проведнте все вычисления.

2. Используя информацию н результаты машинных расчетов для упражнения 1, выполните следующее:

1) Выберите возможную модель для предсказания расхода воды.

2) Обоснуйте свой выбор и рассмотрите вопрос о неадекватностн модели.

3) Что вы можете сказать относительно случайного вектора, каким является столбец из элементов, обозначенных X_5?

3. Если для решения задачи в упражнении 1 используется шаговый метод с критическим уровнем для включения и исключения переменных, равным $F = 3{,}74$, то процедура заканчивается получением модели первого порядка $\hat{Y} = f(X_2, X_4)$. Если бы критическое значение для включеиия переменных было равно $F = 2{,}00$, то шаговый метод привел бы к выбору модели первого порядка $\hat{Y} = f(X_1, X_2, X_3, X_4)$. Применение метода нсключения привело бы к выбору уравнения $\hat{Y} = f(X_1, X_2, X_3, X_4)$.

Обобщите результаты Вашего анализа и выводы, используя информацию, содержащуюся в упражнениях 1, 2 и 3.

Сводка данных

X_1 — среднемесячная температура, F °,
X_2 — количество продукцни, фунты,
X_3 — число рабочнх дней в месяце,
X_4 — численность занятых на пронзводстве по месячной ведомости зарплаты,
X_5 — двузначное случайное число,
$X_6 = Y$ — месячное потребление воды, галлоны.

Исходные данные

	X_1	X_2	X_3	X_4	X_5	$X_6 = Y$
1	58,8	7 107	21	129	52	3067
2	65,2	6 373	22	141	68	2828
3	70,9	6 796	22	153	29	2891
4	77,4	9 208	20	166	23	2994
5	79,3	14 792	25	193	40	3082
6	81,0	14 564	23	189	14	3898
7	71,9	11 964	20	175	96	3502
8	63,9	13 526	23	186	94	3060
9	54,5	12 656	20	190	54	3211
10	39,5	14 119	20	187	37	3286
11	44,5	16 691	22	195	42	3542
12	43,6	14 571	19	206	22	3125
13	56,0	13 619	22	198	28	3022
14	64,7	14 575	22	192	7	2922
15	73,0	14 556	21	191	42	3950
16	78,9	18 573	21	200	33	4488
17	79,4	15 618	22	200	92	3295

Средние значения переменных

64,852940 12900,470 21,470588 181,823520 45,470587 3303,7058

Стандартные отклонения переменных

13,5100930 3526,78600 1,46277340 21,9949850 27,4775310 446,698370

Корреляционная матрица

1,00000000	—0,02410741	0,43762975	—0,08205777	0,10762982	0,28575758
—0,02410741	1,00000000	0,10573055	0,91847987	—0,11145872	0,63074956
0,43762975	0,10573055	1,00000000	0,03188120	0,03768543	—0,08882581
—0,08205777	0,91847987	0,03188120	1,00000000	—0,15900788	0,41324613
0,10762982	—0,11145872	0,03768543	—0,15900788	1,00000000	—0,06562381
0,28575758	0,63074956	—0,08882581	0,41324613	—0,06562381	1,00000000

Регрессии

1. Анализ модели $\hat{Y} = \hat{X}_6 = f(X_2)$:

Номер включаемой переменной	2
Доля объясненной вариации R^2 в %	39,7845000
Стандартное отклонение остатков	357,9998900
Среднее значение отклика	3303,7058000
Стандартное отклонение в % от среднего отклика	10,836
Число степеней свободы	15
Величина детерминанта	1,0000000

ANOVA (Дисперсионный анализ)

Источник	Число степеней свободы	SS	MS	Полный F-критерий
Общий	16	3 192 631,00		
Регрессия	1	1 270 172,00	1 270 172,00	9,91
Остаток	15	1 922 459,00	128 163,92	

***B*-коэффициенты и доверительные пределы**

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
2	12900,47	0,0798899	0,1339688 0,0258111	0,0253772	9,91

Свободный член в регрессионном уравнении равен 2273,0881000.

Анализ остатков

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	3067,000	2840,8659	226,1341	0,6316597
2	2828,000	2782,2266	45,7734	0,1278587
3	2891,000	2816,0201	74,9799	0,2094411
4	2994,000	3008,7146	—14,7146	—0,0411021
5	3082,000	3454,8200	—372,8200	—1,0413969
6	3898,000	3436,6051	461,3949	1,2888129
7	3502,000	3228,8913	273,1087	0,7628737
8	3060,000	3353,6794	—293,6794	—0,8203338
9	3211,000	3284,1751	—73,1751	—0,2043998
10	3286,00	3401,0541	—115,0541	—0,3213803
11	3542,000	3606,5310	—64,5310	—0,1802542
12	3125,000	3437,1644	—312,1644	—0,8719679
13	3022,000	3361,1091	—339,1091	—0,9472324
14	2922,000	3437,4839	—515,4839	—1,4398996
15	3950,000	3435,9660	514,0340	1,4358496
16	4488,000	3756,8839	731,1161	2,0422243
17	3295,000	3520,8091	—225,8091	—0,6307519

2. Анализ модели $\hat{Y} = \hat{X}_6 = f(X_4)$:

Номер включаемой переменной	4
Доля объясненной вариации R^2, в %	17,0772400
Стандартное отклонение остатков	420,1124900
Среднее значение отклика	3303,7058000
Стандартное отклонение в % от среднего отклика	12,716
Число степеней свободы	15
Величина детерминанта	1,0000000

ANOVA

Источник	Число степеней свободы	SS	MS	Полный F-критерий
Общий	16	3 192 631,00		
Регрессия	1	545 213,00	545 213,00	3,09
Остаток	15	2 647 148,00	176 494,50	

***B*-коэффициенты и доверительные пределы**

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
4	181,8235200	8,3926569	18,5683810 —1,7830690	4,7750943	3,09

Свободный член в регрессионном уравнении равен 1777,7234000.

Анализ остатков

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	3067,000	2860,3761	206,6239	0,4918299
2	2828,000	2961,0880	—133,0880	—0,3167913
3	2891,000	3061,7999	—170,7999	—0,4065575
4	2994,000	3170,9044	—176,9044	—0,4210882
5	3082,000	3397,5061	—315,5061	—0,7510039
6	3898,000	3363,9355	534,0645	1,2712416
7	3502,000	3246,4383	255,5617	0,6083173
8	3060,000	3338,7575	—278,7575	—0,6635306
9	3211,000	3372,3282	—161,3282	—0,3840119
10	3286,000	3347,1502	—61,1502	—0,1455567
11	3542,000	3414,2914	127,7086	0,3039867
12	3125,000	3506,6107	—381,6107	—0,9083536
13	3022,000	3439,4694	—417,4694	—0,9937086
14	2922,000	3389,1135	—467,1135	—1,1118772
15	3950,000	3380,7208	569,2792	1,3550637
16	4488,000	3456,2547	1031,7453	2,4558786
17	3295,000	3456,2547	—161,2547	—0,3838370

3. Анализ модели $\hat{Y} = \hat{X}_6 = f(X_1, X_2)$:

Номера переменных, включаемых в модель	1,2
Доля объясненной вариации R^2 в %	48,8476600
Стандартное отклонение остатков	341,5412200
Среднее значение отклика	3303,7058000
Стандартное отклонение в % от среднего отклика	10,338
Число степеней свободы	14
Величина детерминанта	0,9994190

ANOVA

Источник	Число степеней свободы	SS	MS	Полный F-крнтерий
Общий	16	3 192 631,00		
Регрессия	2	1 559 525,00	779 762,70	6,68
Остаток	14	1 633 106,00	116 650,40	

***B*-коэффициенты и доверительные пределы**

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошнбка	Частный F-критерий
1	64,8529400	9,9568521	23,5174330 —3,6037290	6,3219494	2,48
2	12 900,4700000	0,0808094	0,1327561 0,0288628	0,0242176	11,13

Свободный член в регрессионном уравнении равен 1615,4950000.

Анализ остатков

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	3067,000	2775,2705	291,7295	0,8541561
2	2828,000	2779,6803	48,3197	0,1414755
3	2891,000	2870,6167	20,3833	0,0596804
4	2994,000	3130,2486	—136,2486	—0,3989229
5	3082,000	3600,4065	—518,4065	1,5178446
6	3898,000	3598,9086	299,0914	—0,8757110
7	3502,000	3298,1968	203,8032	0,5967163
8	3060,000	3344,7662	—284,7662	—0,8337682
9	3211,000	3180,8676	30,1324	0,0882248
10	3286,000	3149,7390	136,2610	0,3989592
11	3542,000	3407,3652	134,6348	0,3941978
12	3125,000	3227,0880	—102,0880	—0,2989039
13	3022,000	3273,6224	—251,6224	—0,7367263
14	2922,000	3437,5008	—515,5008	—1,5093369
15	3950,000	3518,6075	431,3926	1,2630762
16	4488,000	3901,9643	586,0357	1,7158564
17	3295,000	3668,1508	—373,1508	—1,0925498

4. Анализ модели $\hat{Y} = \hat{X_6} = f(X_2, X_4)$:

Номера переменных, включаемых в модель	2,4
Доля объясненной вариации R^2, в %	57,4219500
Стандартное отклонение остатков	311,6041600
Среднее значение отклика	3303,7058000
Стандартное отклонение в % от среднего отклика	9,432
Число степеней свободы	14
Величина детерминанта	0,1563949

ANOVA

Источник	Число степеней свободы	SS	MS	Полный F-критерий
Общий	16	3 192 631,00		
Регрессия	2	1 833 271,00	916 635,40	9,44
Остаток	14	1 359 360,00	97 097,15	

***B*-коэффициенты и доверительные пределы**

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
2	12 900,4700000	0,2034310	0,3232374 0,0836246	0,0558538	13,27
4	181,8235200	—21,5673720	—2,3570090 —40,7777350	8,9558802	5,80

Свободный член в регрессионном уравнении равен 4600,8056000.

Анализ остатков

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклоненне
1	3067,000	3264,3988	—197,3988	—0,6334922
2	2828,000	2856,2720	—28,2720	—0,0907305
3	2891,000	2683,5148	207,4852	0,6658615
4	2994,000	2893,8146	100,1854	0,3215150
5	3082,000	3447,4543	—365,4543	—1,1728158
6	3898,000	3487,3415	410,6585	1,3178851
7	3502,000	3260,3641	241,6359	0,7754579
8	3060,000	3340,8823	—280,8823	—0,9014074
9	3211,000	3077,6278	133,3722	0,4280180
10	3286,000	3439,9495	—153,9495	—0,4940547
11	3542,000	3790,6350	—248,6350	—0,7979194
12	3125,000	3122,1202	2,8798	0,0092419
13	3022,000	3100,9929	—78,9929	—0,2535040
14	2922,000	3424,8771	—502,8771	—1,6138330
15	3950,000	3442,5793	507,4207	1,6284143
16	4488,000	4065,6553	422,3447	1,3553885
17	3295,000	3464,5167	—169,5167	—0,5440130

5. Анализ модели: $\hat{Y} = \hat{X}_6 = f(X_1, X_2, X_3)$:

Номера переменных, включаемых в модель	1, 2, 3
Доля объясненной вариации R^2, в %	59,2865800
Стандартное отклонение остатков	316,2070100
Среднее значение отклика	3303,7058000
Стандартное отклонение в % от среднего отклика	9,571
Число степеней свободы	13
Величина детерминанта	0,7944891

ANOVA

Источник	Число степеней свободы	SS	MS	Полный F-критерий
Общий	16	3 192 631,00		
Регрессия	3	1 892 802,00	630 933,833	6,31
Остаток	13	1 299 829,00	99 986,869	

***B*-коэффициенты и доверительные пределы**

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний нижний	Стандартная ошибка	Частный F-критерий
1	64,8529400	15,2339490	29,3340350 1,1338630	6,5278179	5,45
2	12900,4700000	0,0861496	0,1349897 0,0373095	0,0226112	14,52
3	21,4705880	—110,6610800	20,2625400 —241,5847000	60,6127880	3,33

Свободный член в регрессионном уравнении равен 3580,3279000.

Анализ остатков

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	3067,000	2764,4665	302,5335	0,9567577
2	2828,000	2688,0689	139,9311	0,4425300
3	2891,000	2811,3436	79,6564	0,2519122
4	2994,000	3339,4792	—345,4792	—1,0925728
5	3082,000	3296,1776	—214,1776	—0,6773355
6	3898,000	3523,7554	374,2446	1,1835430
7	3502,000	3493,1207	8,8793	0,0280807
8	3060,000	3173,8316	—113,8316	—0,3599908
9	3211,000	3287,6656	—76,6656	—0,2424538
10	3286,000	3185,1931	100,8069	0,3188003
11	3542,000	3261,6175	280,3825	0,8867055
12	3125,000	3397,2530	—272,2530	—0,8609961
13	3022,000	3172,1564	—150,1564	—0,4748674
14	2922,000	3387,0508	—465,0508	—1,4707162
15	3950,000	3622,5168	327,4832	1,0356607
16	4488,000	4058,4599	429,5401	1,3584142
17	3295,000	3700,8438	—405,8438	—1,2834750

6. Анализ модели $\hat{Y} = \hat{X}_6 = f(X_1, X_2, X_4)$:

Номера переменных, включаемых в модель	1, 2, 4
Доля объясненной вариации R^2, в %	63,1905300
Стандартное отклонение остатков	300,6647200
Среднее значение отклика	3303,7058000
Стандартное отклонение в % от среднего отклика	9,101
Число степеней свободы	13
Величина детерминанта	0,1527141

ANOVA

Источник	Число степеней свободы	SS	MS	Полный F-критерий
Общий	16	3 192 631,00		
Регрессия	3	2 017 440,00	672 480,100	7,44
Остаток	13	1 175 191,00	90 399,276	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
1	64,8529400	8,0364126	20,1979790 —4,1251550	5,6303554	2,04
2	12 900,4700000	0,1933407	0,3107467 0,0759347	0,0543546	12,65
4	181,8235200	—19,6762750	—0,7925920 —38,5599580	8,7424461	5,07

Свободный член в регрессионном уравнении равен 3865,9449000.

Анализ остатков

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	3067,000	3174,3186	—107,3186	—0,3569378
2	2828,000	2847,7243	—19,7243	—0,0656023
3	2891,000	2739,1996	151,8004	0,5048826
4	2994,000	3001,9824	—7,9824	—0,0265492
5	3082,000	3565,6065	—483,6065	—1,6084577
6	3898,000	3613,8919	284,1081	0,9449333
7	3502,000	3313,5425	188,4575	0,6268028
8	3060,000	3334,8104	—274,8104	—0,9140095
9	3211,000	3012,3566	198,6434	0,6606808
10	3286,000	3233,6966	52,3034	0,1739592
11	3542,000	3613,7407	—71,7407	—0,2386070
12	3125,000	2980,1867	144,8133	0,4816438
13	3022,000	3053,1881	—31,1881	—0,1037305
14	2922,000	3425,9961	—503,9961	—1,6762728
15	3950,000	3508,7012	441,2988	1,4677438
16	4488,000	4155,6790	332,3210	1,1052876
17	3295,000	3588,3755	—293,3755	—0,9757563

7. Анализ модели $\hat{Y} = \hat{X}_6 = f(X_1, X_2, X_3, X_4)$:

Номера переменных, включаемых в модель	1, 2, 3, 4
Доля объясненной вариации R^2 в %	76,7027100
Стандартное отклонение остатков	248,9639800
Среднее значение отклика	3303,7058000
Стандартное отклонение в % от среднего отклика	7,536
Число степеней свободы	12
Величина детерминанта	0,1198925

ANOVA

Источник	Число степеней свободы	SS	MS	Полный F-критерий
Общий	16	3 192 631,00		
Регрессия	4	2 448 834,00	612 208,570	9,88
Остаток	12	743 797,00	61 983,083	

***B*-коэффициенты и доверительные пределы**

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
1	64,8529400	13,8688640	25,1120960 2,6256320	5,1598130	7,22
2	12 900,4700000	0,2117030	0,3109414 0,1124646	0,0455431	21,61
3	21,4705880	—126,6903600	—22,0497400 —231,3309800	48,0223160	6,96
4	181,8235200	—21,8179740	—5,9450100 —37,6909380	7,2845180	8,97

Свободный член в регрессионном уравнении равен 6360 3385000.

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	3067,000	3205,3847	—138,3847	—0,5558423
2	2828,000	2750,2493	77,7507	0,3122970
3	2891,000	2657,0365	233,9635	0,9397484
4	2994,000	3227,5588	—233,5588	—0,9381229
5	3082,000	3213,5221	—131,5221	—0,5282776
6	3898,000	3529,4835	368,5165	1,4802000
7	3502,000	3538,3717	—36,3717	—0,1460922
8	3060,000	3138,0322	—78,0332	—0,3134277
9	3211,000	3116,2823	94,7177	0,3804474
10	3286,000	3283,4248	2,5752	0,0103437
11	3542,000	3469,3447	72,6553	0,2918306
12	3125,000	3148,1257	—23,1257	—0,0928877
13	3022,000	2913,0311	108,9689	0,4376894
14	2922,000	3366,9861	—444,9861	—1,7873513
15	3950,000	3626,5837	323,4163	1,2990485
16	4488,000	4362,4591	125,5409	0,5042533
17	3295,000	3617,1208	—322,1208	—1,2938449

4. Ниже приводится совокупность данных из книги [26]: Brownlee K. A. Statistical Theory and Methodology in Science and Engineering (second edition).— New York, Wiley, 1965, p. 545.

Воспроизводятся результаты наблюдений за 21 день работы производства по получению азотной кислоты окислением аммиака [27].

Переменные:

X_1 — скорость воздуха;

X_2 — температура охлаждающей воды в змеевике абсорбционной башни для поглощения окислов азота;

X_3 — концентрация HNO_3 в абсорбционной жидкости (кодированная величина, полученная умножением на 10 числа, равного настоящей концентрации в процентах минус 50);

Y — потери аммиака с неабсорбированными окислами азота (умноженная на 10 доля в процентах от всего поступающего аммиака). Эта величина по смыслу противоположна выходу азотной кислоты.

Из анализа производственных условий следует, что можно объединить и рассматривать как повторные нижеследующие наборы наблюдений: (1, 2), (4, 5, 6); (7, 8); (11, 12) и (18, 19). Хотя наблюдения в каждом таком наборе не строго параллельные, однако соответствующие им точки в X-пространстве лежат близко друг к другу и их можно считать совпадающими. Машинные распечатки для всех возможных регрессий приведены на последующих страницах.

---

[26] Имеется русский перевод: К. А. Браунли. Статистическая теория и методология в науке и технике/Пер. с англ. Под ред. Л. Н. Большева.— М.: Наука, 1977.— 407 с. Однако в этом переводе опущена глава, в которой содержится данный пример.— *Примеч. пер.*

[27] Читатели, желающие получить более полные сведения по технологии производства азотной кислоты, могут обратиться к следующему источнику: Общая химическая технология/Под ред. И. П. Мухленова. 2-е изд. перераб. и доп.: Учебник для химико-технологических специальностей вузов.— М.: Высшая школа, 1970.— 600 с.; см. гл. 11, с. 342—352.— *Примеч. пер.*

Исходные данные. Производство азотной кислоты окислением аммиака

Номер наблюдения	Скорость воздуха X_1	Температура охлаждающей воды X_2	Концентрация кислоты X_3	Потери аммиака Y	Номер наблюдения	Скорость воздуха X_1	Температура охлаждающей воды X_2	Концентрация кислоты X_3	Потери аммиака Y
1	80	27	89	42	12	58	17	88	13
2	80	27	88	37	13	58	18	82	11
3	75	25	90	37	14	58	19	93	12
4	62	24	87	28	15	50	18	89	8
5	62	22	87	18	16	50	18	86	7
6	62	23	87	18	17	50	19	72	8
7	62	24	93	19	18	50	19	79	8
8	62	24	93	20	19	50	20	80	9
9	58	23	87	15	20	56	20	82	15
10	58	18	80	14	21	70	20	91	15
11	58	18	89	14					

(Примечание. Ошибки округления и правила останова печати вызывают непостоянство в пятом знаке после запятой.)

1) Полагая, что линейная множественная регрессионная модель пригодна для описания данных, какую модель вы исследовали бы в первую очередь?

2) Вычислите «чистую» ошибку.

3) Постройте предварительную схематичную таблицу дисперсионного анализа для этого эксперимента (т. е. укажите только источники рассеяния и числа степеней свободы).

4) Укажите, каким статистическим критерием можно воспользоваться для проверки гипотезы об адекватности модели.

5) Достаточную ли чувствительность имеет критерий для проверки адекватности модели при отклонении нуль-гипотезы?

6) Исследуйте все приведенные регрессионные уравнения. Выберите модель, которая, по-видимому, будет наилучшей для предсказания.

7) Исследуйте остатки для проверки, не пропущена ли какая-нибудь существенная переменная.

8) При каких условиях наиболее желательно использовать выбранное уравнение? Сформулируйте рекомендации для дальнейшего экспериментирования.

Средние значения для преобразованных переменных

1	60,42857000	21,09523700	86,28571300	17,52380900

Стандартные отклонения для преобразованных переменных

1	9,16826650	3,16077100	5,35857090	10,17162000

Корреляционная матрица

1	1,00000000	0,78185250	0,50014295	0,91966375
2	0,78185250	1,00000000	0,39093959	0,87550465
3	0,50014295	0,39093959	1,00000000	0,39982969
4	0,91966375	0,87550465	0,39982969	1,00000000

Исходные и/или преобразованные данные. Окисление аммиака до азотной кислоты

	X_1	X_2	X_3	$X_4 = Y$
1	80,00000000	27,00000000	89,00000000	42,00000000
2	80,00000000	27,00000000	88,00000000	37,00000000
3	75,00000000	25,00000000	90,00000000	37,00000000
4	62,00000000	24,00000000	87,00000000	28,00000000
5	62,00000000	22,00000000	87,00000000	18,00000000
6	62,00000000	23,00000000	87,00000000	18,00000000
7	62,00000000	24,00000000	93,00000000	19,00000000
8	62,00000000	24,00000000	93,00000000	20,00000000
9	58,00000000	23,00000000	87,00000000	15,00000000
10	58,00000000	18,00000000	80,00000000	14,00000000
11	58,00000000	18,00000000	89,00000000	14,00000000
12	58,00000000	17,00000000	88,00000000	13,00000000
13	58,00000000	18,00000000	82,00000000	11,00000000
14	58,00000000	19,00000000	93,00000000	12,00000000
15	50,00000000	18,00000000	89,00000000	8,00000000
16	50,00000000	18,00000000	86,00000000	7,00000000
17	50,00000000	19,00000000	72,00000000	8,00000000
18	50,00000000	19,00000000	79,00000000	8,00000000
19	50,00000000	20,00000000	80,00000000	9,00000000
20	56,00000000	20,00000000	82,00000000	15,00000000
21	70,00000000	20,00000000	91,00000000	15,000000000

1. Контрольные данные

Число наблюдений	21
Номер отклика	4
Уровень значимости для доверительных интервалов *B*-коэффициентов	0,5 %
Перечень невключенных переменных	2,3
Включаемая переменная	1
Последовательный *F*-критерий	104,2013300
Доля объясненной вариации R^2 в %	84,5780900
Стандартное отклонение остатков	4,0982407
Среднее значение отклика	17,5238090
Стандартное отклонение в % от среднего отклика	23,387
Число степеней свободы	19
Величина детерминанта	1,00000000

ANOVA

Источник	Число степеней свободы	SS	MS	Полный *F*-критерий
Общий	20	2069,2370000		
Регрессия	1	1750,1211000	1750,1211000	104,20
Остаток	19	319,1159700	16,7955770	
Неадекватность	13	238,4493033	18,3422541	1,36
«Чистая» ошибка	6	80,6666667	13,4444445	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный *B*-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный *F*-критерий
1	60,4285700	1,0203093	1,2295106 0,8111080	0,0999529	104,20

Свободный член в регрессионном уравнении равен — 44,1320220.

Квадраты частных коэффициентов корреляции для переменных, не включенных в уравнение

Переменные	Квадраты коэффициентов
2	0,40838
3	0,03127

Анализ остатков для $\hat{X}_4 = f(X_1)$

Номер наблюдения	Наблюдаемый *Y*	Предсказываемый *Y*	Остаток	Нормальное отклонение
1	42,00000000	37,4927220	4,5072780	1,0998080
2	37,00000000	37,4927220	—0,4927220	—0,1202277
3	37,00000000	32,3911750	4,6088250	1,1245862
4	28,00000000	19,1271540	8,8728460	2,1650377
5	18,00000000	19,1271540	—1,1271540	—0,2750336
6	18,00000000	19,1271540	—1,1271540	—0,2750336
7	19,00000000	19,1271540	—0,1271540	—0,0310265
8	20,00000000	19,1271540	0,8728460	0,2129807
9	15,00000000	15,0459170	—0,0459170	—0,0112041
10	14,00000000	15,0459170	—1,0459170	—0,2552112
11	14,00000000	15,0459170	—1,0459170	—0,2552112
12	13,00000000	15,0459170	—2,0459170	—0,4992184
13	11,00000000	15,0459170	—4,0459170	—0,9872326
14	12,00000000	15,0459170	—3,0459170	—0,7432255
15	8,00000000	6,8834430	1,1165570	0,2724479
16	7,00000000	6,8834430	0,1165570	0,0284407
17	8,00000000	6,8834430	1,1165570	0,2724479
18	8,00000000	6,8834430	1,1165570	0,2724479
19	9,00000000	6,8834430	2,1165570	0,5164550
20	15,00000000	13,0052980	1,9947020	0,4867215
21	15,00000000	27,2896290	—12,2896290	—2,9987572

2. Контрольные данные

Число наблюдений	21
Номер отклика	4
Уровень значимости для доверительных интервалов *B*-коэффициентов	5 %
Перечень невключенных переменных	1,3
Включаемая переменная	2
Последовательный *F*-критерий	62,3732090
Доля объясненной вариации R^2 в %	76,6507900
Стандартное отклонение остатков	5,0427155
Среднее значение отклика	17,5238090
Стандартное отклонение в % от среднего отклика	28,776
Число степеней свободы	19
Величина детерминанта	1,00000000

ANOVA

Источник	Число степеней свободы	SS	MS	Полный *F*-критерий
Общий	20	2069,2370000		
Регрессия	1	1586,0865000	1586,0865000	62,37
Остаток	19	483,1506100	25,4289790	
Неадекватность	13	402,4839433	30,9603033	2,30
«Чистая» ошибка	6	80,6666667	13,4444445	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный *B*-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный *F*-критерий
2	21,0952370	2,8174450	3,5641096 2,0707804	0,3567438	62,37

Свободный член в регрессионном уравнении равен — 41,9103610.

Квадраты частных коэффициентов корреляции для переменных, не включенных в уравнение

Переменные	Квадраты коэффициентов
1	0,60924
3	0,01675

Анализ остатков для $\hat{X}_4 = f(X_2)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	42,00000000	34,1061540	7,8398460	1,5546873
2	37,00000000	34,1061540	2,8398460	0,5631581
3	37,00000000	28,5252640	8,4747360	1,6805897
4	28,00000000	25,7078190	2,2921810	0,4545529
5	18,00000000	20,0729290	—2,0729290	—0,4110740
6	18,00000000	22,8903740	—4,8903740	—0,9697898
7	19,00000000	25,7078190	—6,7078190	—1,3301997
8	20,00000000	25,7078190	—5,7078190	—1,1318939
9	15,00000000	22,8903740	—7,8903740	—1,5647073
10	14,00000000	8,8031490	5,1968510	1,0305659
11	14,00000000	8,8031490	5,1968510	1,0305659
12	13,00000000	5,9857040	7,0142960	1,3909759
13	11,00000000	8,8031490	2,1968510	0,4356484
14	12,00000000	11,6205940	0,3794060	0,0752384
15	8,00000000	8,8031490	—0,8031490	—0,1592691
16	7,00000000	8,8031490	—1,8031490	—0,3575750
17	8,00000000	11,6205940	—3,6205940	—0,7179850
18	8,00000000	11,6205940	—3,6205940	—0,7179850
19	9,00000000	14,4380390	—5,4380390	—1,0783949
20	15,00000000	14,4380390	0,5619610	0,1114402
21	15,00000000	14,4380390	0,5619610	0,1114402

3. Контрольные данные

Число наблюдений	21
Номер отклика	4
Уровень значимости для доверительных интервалов B-коэффициентов	5 %
Перечень невключенных переменных	1, 2
Включаемая переменная	3
Последовательный F-критерий	3,6153784
Доля объясненной вариации R^2 в %	15,9863400
Стандартное отклонение остатков	9,5654028
Среднее значение отклика	17,5238090
Стандартное отклонение в % от среднего отклика	54,585
Число степеней свободы	19
Величина детерминанта	1,0000000

ANOVA

Источник	Число степеней свободы	SS	MS	Полный F-критерий
Общий	20	2069,2370000		
Регрессия	1	330,7952600	330,7952600	3,62
Остаток	19	1738,4417000	91,4969310	
Неадекватность	13	1657,7750333	127,5211563	9,49
«Чистая» ошибка	6	80,6666667	13,4444445	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
3	86,2857130	0,7589552	1,5943822 —0,0764717	0,3991529	3,62

Свободный член в регрессионном уравнении равен — 47,9631840.

Квадраты частных коэффициентов корреляции для переменных, не включенных в уравнение

Переменные	Квадраты коэффициентов
1	0,82218
2	0,72673

Анализ остатков для $\hat{X}_4 = f(X_3)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	42,00000000	19,5838310	22,4161690	2,3434631
2	37,00000000	18,8248760	18,1751240	1,9000897
3	37,00000000	20,3427860	16,6572140	1,7414022
4	28,00000000	18,0659210	9,9340790	1,0385426
5	18,00000000	18,0659210	—0,0659210	—0,0068916
6	18,00000000	18,0659210	—0,0659210	—0,0068916
7	19,00000000	22,6196520	—3,6196520	—0,3784108
8	20,00000000	22,6196520	—2,6196520	—0,2738674
9	15,00000000	18,0659210	—3,0659210	—0,3205219
10	14,00000000	12,7532340	1,2467660	0,1303412
11	14,00000000	19,5838310	—5,5838310	—0,5837528
12	13,00000000	18,8248760	—5,8248760	—0,6089525
13	11,00000000	14,2711440	—3,2711440	—0,3419766
14	12,00000000	22,6196520	—10,6196520	—1,1102148
15	8,00000000	19,5838310	—11,5838310	—1,2110134
16	7,00000000	17,3069650	—10,3069650	—1,0775254
17	8,00000000	6,6815920	1,3184080	0,1378309
18	8,00000000	11,9942790	—3,9942790	—0,4175756
19	9,00000000	12,7532340	—3,7532340	—0,3923759
20	15,00000000	14,2711440	0,7288560	0,0761971
21	15,00000000	21,1017410	—6,1017410	—0,6378969

4. Контрольные данные

Число наблюдений	21
Номер отклика	4

Уровень значимости для доверительных интервалов *B*-коэффициентов	5 %
Перечень невключенных переменных	3
Включаемые переменные	1,2
Последовательный *F*-критерий	12,4249510
Доля объясненной вариации R^2 в %	90,8761000
Стандартное отклонение остатков	3,2386150
Среднее значение отклика	17,5238090
Стандартное отклонение в % от среднего отклика	18,481
Число степеней свободы	18
Величина детерминанта	0,3887071

ANOVA

Источник	Число степеней свободы	SS	MS	Полный *F*-критерий
Общий	20	2069,2370000		
Регрессия	2	1880,4418000	940,2209000	89,64
Остаток	18	188,7952900	10,4886270	
Неадекватность	12	108,1286233	9,0107186	0,67
«Чистая» ошибка	6	80,6666667	13,4444445	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный *B*-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный *F*-критерий
1	60,4285700	0,6711545	0,9373324 0,4049767	0,1266910	28,06
2	21,0952370	1,2953510	2,0674378 0,5232642	0,3674854	12,42

Свободный член в регрессионном уравнении равен — 50,3588360.

Квадраты частных коэффициентов корреляции для переменных, не включенных в уравнение

Переменные	Квадраты коэффициентов
3	0,05278

Анализ остатков для $\hat{X}_4 = f(X_1, X_2)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	42,00000000	38,3080040	3,6919960	1,1399922
2	37,00000000	38,3080040	—1,3080040	—0,4038776
3	37,00000000	32,3615290	4,6384710	1,4322390
4	28,00000000	22,3411690	5,6588310	1,7472996
5	18,00000000	19,7504670	—1,7504670	—0,5404986
6	18,00000000	21,0458180	—3,0458180	—0,9404693
7	19,00000000	22,3411690	—3,3411690	—1,0316660
8	20,00000000	22,3411690	—2,3411690	—0,7228920
9	15,00000000	18,3612000	—3,3612000	—1,0378510
10	14,00000000	11,8844450	2,1155550	0,6532283
11	14,00000000	11,8844450	2,1155550	0,6532283
12	13,00000000	10,5890940	2,4109060	0,7444250
13	11,00000000	11,8844450	—0,8844450	—0,2730936
14	12,00000000	13,1797960	—1,1797960	—0,3642903
15	8,00000000	6,5152090	1,4847910	0,4584648
16	7,00000000	6,5152090	0,4847910	0,1496908
17	8,00000000	7,8105600	0,1894400	0,0584941
18	8,00000000	7,8105600	0,1894400	0,0584941
19	9,00000000	9,1059110	—0,1059110	—0,0327026
20	15,00000000	13,1328380	1,8671620	0,5765310
21	15,00000000	22,5290010	—7,5290010	—2,3247595

5. Контрольные данные

Число наблюдений	21
Номер отклика	4
Уровень значимости для доверительных интервалов B-коэффициентов	5 %
Перечень невключенных переменных	2
Включаемые переменные	1,3
Последовательный F-критерий	0,5810138
Доля объясненной варнации R^2 в %	85,0603200
Стандартное отклонение остатков	4,1441891
Среднее значение отклика	17,5238090
Стандартное отклонение в % от среднего отклика	23,649
Число степеней свободы	18
Величина детерминаита	0,7498572

ANOVA

Источник	Число степеней свободы	SS	MS	Полный F-критерий
Общий	20	2069,2370000		
Регрессия	2	1760,0996000	880,0498000	51,24
Остаток	18	309,1374600	17,1743030	
Неадекватность	12	228,4707933	19,0392328	1,42
«Чистая» ошибка	6	80,6666667	13,4444445	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
1	60,4285700	1,0648068	1,3100370 0,8195766	0,1167207	83,22
3	86,2857130	—0,1522227	0,2673549 —0,5718004	0,1997038	0,58

Свободный член в уравнении регрессии равен — 33,6862970.

Квадраты частных коэффициентов корреляции для переменных, не включенных в уравнение

Переменные	Квадраты коэффициентов
2	0,42152

Анализ остатков для $\hat{X}_4 = f(X_1, X_3)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	42,00000000	37,9504250	4,0495750	0,9771695
2	37,00000000	38,1026470	—1,1026470	—0,2660706
3	37,00000000	32,4741680	4,5258320	1,0920910
4	28,00000000	19,0883470	8,9116530	2,1503972
5	18,00000000	19,0883470	—1,0883470	—0,2626200
6	18,00000000	19,0883470	—1,0883470	—0,2626200
7	19,00000000	18,1750110	0,8249890	0,1990713
8	20,00000000	18,1750110	1,8249890	0,4403730
9	15,00000000	14,8291200	0,1708800	0,0412336
10	14,00000000	15,8946790	—1,8946790	—0,4571893
11	14,00000000	14,5246750	—0,5246750	—0,1266050
12	13,00000000	14,6768970	—1,6768970	—0,4046381
13	11,00000000	15,5902340	—4,5902340	—1,1076314
14	12,00000000	13,9157840	—1,9157840	—0,4622820
15	8,00000000	6,0062210	1,9937790	0,4811023
16	7,00000000	6,4628890	0,5371110	0,1296058
17	8,00000000	8,5940070	—0,5940070	—0,1433349
18	8,00000000	7,5284480	0,4715520	0,1137863
19	9,00000000	7,3762250	1,6237750	0,3918197
20	15,00000000	13,4606200	1,5393800	0,3714551
21	15,00000000	26,9979110	—11,9979110	—2,8951166

6. Контрольные данные

Число наблюдений	21
Номер отклика	4
Уровень значимости для доверительных интервалов *B*-коэффициентов	5 %
Перечень невключенных переменных	1
Включаемые переменные	2,3
Последовательный *F*-критерий	0,3066295
Доля объясненной вариации R^2 в %	77,0418800
Стандартное отклонение остатков	5,1373253
Среднее значение отклика	17,5238090
Стандартное отклонение в % от среднего отклика	29,316
Число степеней свободы	18
Величина детерминанта	0,8471664

ANOVA

Источник	Число степеней свободы	SS	MS	Полный *F*-критерий
Общий	20	2069,2370000		
Регрессия	2	1594,1790000	797,0895000	30,20
Остаток	18	475,0580100	26,3921110	
Неадекватность	12	394,3913433	32,8659453	2,44
«Чистая» ошибка	6	80,6666667	13,4444445	

***B*-коэффициенты и доверительные пределы**

Номер переменной	Среднее	Натуральный *B*-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный *F*-критерий
2	21,0952370	2,7319656	3,5615693 1,9023619	0,3948614	47,87
3	86,2857130	0,1289721	0,6183167 —0,3603725	0,2329103	0,31

Свободный член в регрессионном уравнении равен — 51,2360990.

Квадраты частных коэффициентов корреляции для переменных, не включеиных в уравнение

Переменные	Квадраты коэффициентов
1	0,62356

Аиализ остатков для $\widehat{X}_4 = f(X_2, X_3)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклоиение
1	42,00000000	34,0054870	7,9945130	1,5561625
2	37,00000000	33,8765150	3,1234850	0,6079983
3	37,00000000	28,6705280	8,3294720	1,6213635
4	28,00000000	25,5516450	2,4483550	0,4765817
5	18,00000000	20,0877140	—2,0877140	—0,4063815
6	18,00000000	22,8196790	—4,8196790	—0,9381689
7	19,00000000	26,3254780	—7,3254788	—1,4259322
8	20,00000000	26,3254780	—6,3254780	—1,2312784
9	15,00000000	22,8196790	—7,8196790	—1,5221303
10	14,00000000	8,2570470	5,7429530	1,1178877
11	14,00000000	9,4177960	4,5822040	0,8919435
12	13,00000000	6,5568590	6,4431410	1,2541820
13	11,00000000	8,5149910	2,4850090	0,4837165
14	12,00000000	12,6656500	—0,6656500	—0,1295713
15	8,00000000	9,4177960	—1,4177960	—0,2759794
16	7,00000000	9,0308790	—2,0308790	—0,3953184
17	8,00000000	9,9572360	—1,9572360	—0,3809835
18	8,00000000	10,8600410	—2,8600410	—0,5567179
19	9,00000000	13,7209790	—4,7209790	—0,9189566
20	15,00000000	13,9789230	1,0210770	0,1987565
21	15,00000000	15,1396720	—0,1396720	—0,0271877

7. Контрольные данные

Число наблюдений	21
Номер отклика	4
Уровень значимости для доверительных интервалов B-коэффициентов	5 %
Включаемые переменные	1, 2, 3
Последовательный F-критерий	0,9473322
Доля объясненной вариации R^2, в %	91,3576900
Стандартное отклонение остатков	3,2433636
Среднее значение отклика	17,5238090
Стандартное отклонение в % от среднего отклика	18,508
Число степеней свободы	17
Велнчина детерминанта	0,2914747

ANOVA

Источник	Число степеней свободы	SS	MS	Полный F-критерий
Общнй	20	2069,2370000		
Регрессия	3	1890,4071000	630,1357000	59,90
Остаток	17	178,8299200	10,5194070	
Неадекватность	11	98,1632533	8,9239321	0,66
«Чистая» ошнбка	6	80,6666667	13,4444445	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный *B*-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный *F*-критерий
1	60,4285700	0,7156403	1,0001910 0,4310896	0,1348582	28,16
2	21,0952370	1,2952857	2,0718168 0,5187546	0,3680242	12,39
3	86,2857130	—0,1521225	0,1776579 —0,4819029	0,1562940	0,95

Свободный член в регрессионном уравнении равен — 39,9196680.

Анализ остатков для $\hat{X}_4 = f(X_1, X_2, X_3)$

Номер наблюдения	Наблюдаемый *Y*	Предсказываемый *Y*	Остаток	Нормальное отклонение
1	42,00000000	38,7653640	3,2346360	0,9973091
2	37,00000000	38,9174870	—1,9174870	—0,5912032
3	37,00000000	32,4444690	4,5555310	1,4045699
4	28,00000000	22,3022260	5,6977740	1,7567484
5	18,00000000	19,7116550	—1,7116550	—0,5277407
6	18,00000000	21,0069410	—3,0069410	—0,9271057
7	19,00000000	21,3894910	—2,3894910	—0,7367324
8	20,00000000	21,3894910	—1,3894910	—0,4284105
9	15,00000000	18,1443800	—3,1443800	—0,9694812
10	14,00000000	12,7328090	1,2671910	0,3907027
11	14,00000000	11,3637060	2,6362940	0,8128272
12	13,00000000	10,2205430	2,7794570	0,8569674
13	11,00000000	12,4285640	—1,4285640	—0,4404576
14	12,00000000	12,0505020	—0,0505020	—0,0155709
15	8,00000000	5,6385840	2,3614160	0,7280762
16	7,00000000	6,0946520	0,9050480	0,2790461
17	8,00000000	9,5199530	—1,5199530	—0,4686348
18	8,00000000	8,4550960	—0,4550960	—0,1403161
19	9,00000000	9,5982590	—0,5982590	—0,1844563
20	15,00000000	13,5878550	1,4121450	0,4353952
21	15,00000000	22,2377170	—7,2377170	—2,2315465

5. Потребность в некотором продукте широкого потребления изменяется под действием многих факторов. В одном исследовании были проведены измерения относительной урбанизации, образовательного уровня и относительного заработка для девяти географических районов с целью определить влияние этих факторов на потребление рассматриваемого продукта. Были получены следующие данные.

Исходные данные

Номер района	Относительная урбанизация X_1	Образовательный уровень X_2	Относительный заработок X_3	Потребление продукта $X_4 = Y$
1	42,2	11,2	31,9	167,1
2	48,6	10,6	13,2	174,4
3	42,6	10,6	28,7	160,8
4	39,0	10,4	26,1	162,0
5	34,7	9,3	30,1	140,8
6	44,5	10,8	8,5	174,6
7	39,1	10,7	24,3	163,7
8	40,1	10,0	18,6	174,5
9	45,9	12,0	20,4	185,7
Средние	41,8555	10,6222	22,4222	167,0666
Стандартные отклонения	s_1 4,176455	s_2 0,7462871	s_3 7,927921	s_4 12,645157

Корреляционная матрица

	X_1	X_2	X_3	X_4
X_1	1	0,683742	—0,615790	0,801752
X_2	0,683742	1	—0,172493	0,769950
X_3	—0,615790	—0,172493	1	—0,628746
X_4	0,801752	0,769950	—0,628746	1

1) Используя шаговый метод, определите подходящую модель первого порядка для прогнозирования при следующих критических значениях:
$F = 2{,}00$ для включения переменных;
$F = 2{,}00$ для исключения переменных.
2) Выпишите таблицы дисперсионного анализа для каждого шага.
3) Приведите ваши соображения по поводу адекватности окончательного регрессионного уравнения на основе исследования остатков.

6. Бригада по упаковке посылок состоит из пяти рабочих — им присвоены номера от 1 до 5 — и мастера, который работает все время. В приведенных ниже данных
$X_j = 1$, если рабочий j на дежурстве и 0 — в противном случае,
Y — число посылок, отправленных в этот день.
Используя выбранный вами метод подбора переменных, постройте регрессионное уравнение в виде $\hat{Y} = b_0 + \Sigma\, b_j X_j$ (знак суммирования может и не распространяться на все j), которое по вашему мнению лучше всего объясняет имеющиеся данные. Получите также обычную таблицу дисперсионного анализа и выполните другие обычные процедуры, а также прокомментируйте настолько подробно, насколько вы можете, выбранное вами уравнение и все прочие соотношения, которые вы видите. В частности, располагая полученными результатами, какой совет вы могли бы дать?

Исходные данные

Опыт	X_1	X_2	X_3	X_4	X_5	Y
1	1	1	1	0	1	246
2	1	0	1	0	1	252
3	1	1	1	0	1	253
4	0	1	1	1	0	164
5	1	1	0	0	1	203
6	0	1	1	1	0	173
7	1	1	0	0	1	210
8	1	0	1	0	1	247
9	0	1	0	1	0	120
10	0	1	1	1	0	171
11	0	1	1	1	0	167
12	0	0	1	1	0	172
13	1	1	1	0	1	247
14	1	1	1	0	1	252
15	1	0	1	0	1	248
16	0	1	1	1	0	169
17	0	1	0	0	0	104
18	0	1	1	1	0	166
19	0	1	1	1	0	168
20	0	1	1	0	0	148
Сумма	9	16	16	9	9	3 880
SS	9	16	16	9	9	795 364

Повторные опыты: (1, 3, 13, 14), (2, 8, 15), (4, 6, 10, 11, 16, 18, 19), (5, 7).

7. (Источники: Статья Гормана и Томана, Gorman J. W., Toman R. J. Selection of variables for fitting equations to data.— Technometrics, 1966, 8, 27—51, и книга Daniel C., Wood F. S. Fitting Equations to Data, 2 nd edition.— New York: 1980, p. 95—103, 109—117.) Горман и Томан обсуждали эксперимент, в котором измерялась «скорость образоваиия колеи» иа тридцати одной экспериментальной площадке из асфальта. Чтобы выявить условия, при которых происходят изменения асфальта, были использованы пять иезависимых (предикторных) переменных. Между тем как шестая «фиктивиая» переменная примеиялась для того, чтобы отразить различие между двумя отдельными «блоками» опытов, иа которые был разделен весь эксперимент. Для подгоики данных использовалось уравнение

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \varepsilon, \qquad (1)$$

где Y — логарифм изменения глубины колеи в дюймах за миллиои проходов колеса;
X_1 — логарифм вязкости асфальта,
X_2 — содержание асфальта в дорожном покрытии в %,
X_3 — содержание асфальта в основании «подушки» в %,
X_4 — фиктивная переменная для различения двух блоков опытов,
X_5 — доля мелких неровиостей в дорожном покрытии в %,
X_6 — доля трещин в дорожном покрытии в %.

Таблица к упражнению 6.7. Остаточные суммы квадратов (RSS) для всех возможных вариантов подгоняемых моделей, с точностью до третьего знака после запятой

Переменные *	RSS	Переменные *	RSS	Переменные *	RSS	Переменные *	RSS
—	11,058	5	9,922	6	9,196	56	7,680
1	0,607	15	0,597	16	0,576	156	0,574
2	11,795	25	9,479	26	9,192	256	7,679
12	0,499	125	0,477	126	0,367	1 256	0,364
3	10,663	35	9,891	36	8,848	356	7,678
13	0,600	135	0,582	136	0,567	1 356	0,561
23	10,168	235	9,362	236	8,838	2 356	7,675
123	0,498	1 235	0,475	1 236	0,365	12 356	0,364
4	1,522	45	1,397	46	1,507	456	1,352
14	0,582	145	0,569	146	0,558	1 456	0,553
24	1,218	245	1,030	246	1,192	2 456	1,024
124	0,450	1 245	0,413	1 246	0,323	12 456	0,313
34	1,453	345	1,383	346	1,437	3 456	1,342
134	0,581	1 345	0,561	1 346	0,555	13 456	0,545
234	1,041	2 345	0,958	2 346	0,995	23 456	0,939
1234	0,441	12 345	0,412	12 346	0,311	123 456	0,307

* Переменные, включенные в уравнение; свободный член β_0 содержится во всех уравнениях.

Вы можете исходить из того, что уравнение (1) является полным в том смысле, что оно включает все переменные, относящиеся к делу. Ваша задача — выбрать подходящий список из трех переменных, отвечающих «наилучшему» регрессионному уравнению при данных обстоятельствах.

В приведенной выше таблице вы найдете остаточные суммы квадратов для всех возможных регрессий. Эта информация позволит вам выполнить любую из следующих процедур:

1. Выбор регрессии методом исключения.
2. Подбор регрессия методом включения.
3. Шаговую регрессию.
4. Исследование C_p-критерия.
5. Вариации процедур 1—4.

Похвально, если вы пустите в ход свою изобретательность вместо или вместе с использованием некоторых из указанных «стандартных» процедур.

Подсказка. Остатки от регрессии, не содержащей никаких предикторов, а только β_0, позволят вам найти скорректированную сумму квадратов.

8. (Источник. Торнбулл и Уильямс, Turnbull P., Williams G. Sex differentials in teachers' pay.— Journal of the Royal Statistical Society, 1974, A—137, p. 245—258.)

Предостережение. Данные, которые мы здесь даем, были получены первоначально путем систематического выбора 100 наблюдений (каждое 30-е из 3000 наблюдений, входящих в 3414 первоначальных наблюдений. Затем из них было отобрано 90 наблюдений, а 10 были отброшены. На этой стадии исключились также две предикторные переменные. Таким образом, хотя данные, представленные ниже, и представительны в некотором смысле, они не обязательно отражают надлежащим образом поведение (характеристики) полного набора исходных данных. Для их анализа надо обратиться к указанному источнику.

В прилагаемой таблице [28] содержатся данные о 90 британских учителях, оказавшихся в выборке, проведенной в 1971 г.:

Y — оклад в фунтах стерлингов,

X_1 — продолжительность службы в месяцах минус 12,

X_2 — фиктивная переменная, отражающая пол (1 — для мужчин и 0 — для женщин),

X_3 — фиктивная переменная, связанная с полом (1 — для одинокой женщины и 0 — для замужней)

(Заметим, что комбинация $(X_2, X_3) = (1, 0)$ никогда не встречается; оставшиеся три комбинации соответствуют мужчине (1,1), одинокой женщине (0,1) и замужней женщине (0,0).),

X_4 — квалификационный уровень в соответствии с дипломом, кодируется от 0 до 5,

X_5 — тип школы, в которой работает учитель, кодируется 0 или 1,

$X_6 =$ 1 для дипломированного учителя, имеющего опыт работы, 0 — для дипломированного учителя, не имеющего опыта работы, или недипломированного учителя, имеющего опыт работы,

$X_7 =$ 1 в случае перерыва в работе в течение более двух лет, 0 — в прочих случаях.

1) Используя ту или иную процедуру выбора предикторных переменных, постройте «наилучшую» модель в форме

$$\log_{10} Y = \beta_0 + \Sigma \beta_j X_j + \varepsilon$$

по данным, где суммирование может распространяться не обязательно на все возможные X.

2) Проверьте остатки. Какой другой возможный член предложили бы вы включить в модель?

3) Добавьте этот новый член в уравнение и определите снова коэффициенты модели, используя ту же процедуру выбора переменных.

4) Какую модель вы стали бы использовать на практике, если бы хотели объяснить основную часть вариаций окладов? Почему? (Мы благодарим доктора Торнбулла, который любезно предоставил в наше распоряжение исходный набор данных.)

Ответы к упражнениям

1. 2) $b_4 \pm t\,(14;\ 0{,}95)$; станд. ош. $(b_4) = -21{,}567372 \pm (2{,}145)$ $(8{,}9558802)$ $-40{,}777735 \leqslant \beta_{4Y \cdot 2} \leqslant 2{,}357009$.

3) $$\mathbf{C}^{-1} = \begin{bmatrix} 1{,}024102384 & -0{,}335666762 & 0{,}392338723 \\ -0{,}335666762 & 6{,}504097791 & -6{,}001426973 \\ 0{,}392338723 & -6{,}001426973 & 6{,}544384320 \end{bmatrix}$$

2. 1) $\hat{Y} = \hat{X}_6 = f(X_1,\ X_2,\ X_3,\ X_4)$ или $\hat{Y} = 6360{,}3385 + 13{,}868864X_1 + 0{,}211703X_2 - 126{,}690360X_3 - 21{,}817974X_4$.

2) Приведенное выше МНК-уравнение дает значение R^2, равное 76,702710, которое будет наибольшим среди имеющихся. Полный $F = 9{,}8770256$ статистически значим. Частные значения F-критерия также все статистически значимы. Никакой из 95 %-ных доверительных пределов для β-коэффициентов не включает нуль. Стандартное отклонение в процентах от среднего отклика равно 7,536 %, что ниже, чем для любого другого уравнения.

---

[28] В Великобритании существует довольно сложная система ученых степеней для педагогов. Различается диплом с отличием по усложненной программе первого, второго и третьего класса. Возможен и диплом без отличия после экзаменов по облегченной программе. Да еще особые отличия присуждаются в университетах Оксфорда и Кембриджа. Вот и получается шесть вариантов, используемых в данной задаче, где предпринимается попытка исследовать успехи в педагогической карьере в зависимости от ряда переменных, среди которых и успехи в обучении.— *Примеч. пер.*

Исходные данные к задаче 6.8

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
998	7	0	0	0	0	0	0	2201	158	1	1	4	0	1	1
1015	14	1	1	0	0	0	0	2992	159	1	1	5	1	1	1
1028	18	1	1	0	1	0	0	1695	162	0	1	0	0	0	0
1250	19	1	1	0	0	0	0	1792	167	1	1	0	1	0	0
1028	19	0	1	0	1	0	0	1690	173	0	0	0	0	0	1
1028	19	0	0	0	0	0	0								
1018	27	0	0	0	0	0	1								
1072	30	0	0	0	0	0	0	1827	174	0	0	0	0	0	1
1290	30	1	1	0	0	0	0	2604	175	1	1	2	1	1	0
1204	30	0	1	0	0	0	0	1720	199	0	1	0	0	0	0
								1720	209	0	0	0	0	0	0
								2159	209	0	1	4	1	0	0
1352	31	0	1	2	0	1	0	1852	210	0	1	0	0	0	0
1204	31	0	0	0	1	0	0	2104	213	1	1	0	1	0	0
1104	38	0	0	0	0	0	0	1852	220	0	0	0	0	0	1
1118	41	1	1	0	0	0	0	1852	222	0	0	0	0	0	0
1127	42	0	0	0	0	0	0	2210	222	1	1	0	0	0	0
1259	42	1	1	0	1	0	0								
1127	42	1	1	0	0	0	0								
1127	42	0	0	0	1	0	0	2266	223	0	1	0	0	0	0
1095	47	0	0	0	0	0	1	2027	223	1	1	0	0	0	0
1113	52	0	0	0	0	0	1	1852	227	0	0	0	1	0	0
								1852	232	0	0	0	0	0	1
								1995	235	0	0	0	0	0	1
1462	52	0	1	2	0	1	0	2616	245	1	1	3	1	1	0
1182	54	1	1	0	0	0	0	2324	253	1	1	0	1	0	0
1404	54	0	0	0	1	0	0	1852	257	0	1	0	0	0	1
1182	54	0	0	0	0	0	0	2054	260	0	0	0	0	0	0
1594	55	1	1	2	1	1	0	2617	284	1	1	3	1	1	0
1459	66	0	0	0	1	0	0								
1237	67	1	1	0	1	0	0								
1237	67	0	1	0	1	0	0	1948	287	1	1	0	0	0	0
1496	75	0	1	0	0	0	0	1720	290	0	1	0	0	0	1
1424	78	1	1	0	1	0	0	2604	308	1	1	2	1	1	0
								1852	309	1	1	0	1	0	1
								1942	319	0	0	0	1	0	0
1424	79	0	1	0	0	0	0	2027	325	1	1	0	0	0	0
1347	91	1	1	0	1	0	0	1942	326	1	1	0	1	0	0
1343	92	0	0	0	0	0	1	1720	329	1	1	0	1	0	0
1310	94	0	0	0	1	0	0	2048	337	0	0	0	0	0	0
1814	103	0	0	2	1	1	0	2334	346	1	1	2	1	1	1
1534	103	0	0	0	0	0	0								
1430	103	1	1	0	0	0	0								
1439	111	1	1	0	1	0	0	1720	355	0	0	0	0	0	1
1946	114	1	1	3	1	1	0	1942	357	1	1	0	0	0	0
2216	114	1	1	4	1	1	0	2117	380	1	1	0	0	0	1
								2742	387	1	1	2	1	1	1
								2740	403	1	1	2	1	1	1
1834	114	1	1	4	1	1	1	1942	406	1	1	0	1	0	0
1416	117	0	0	0	0	0	1	2266	437	0	1	0	0	0	0
2052	139	1	1	0	1	0	0	2436	453	0	1	0	0	0	0
2087	140	0	0	2	1	1	1	2067	458	0	1	0	0	0	0
2264	154	0	0	2	1	1	1	2000	464	1	1	2	1	1	0

3) Случайный вектор X_5 не вносит значимого вклада в вариации отклика. Действительно, его вклад составляет менее 1 %, и он увеличивает стандартное отклонение, выраженное в процентах от среднего отклика, с 7,536 до 7,759 %. Частный F-критерий также показывает, что эта переменная статистически незначима.

3. Регрессионная модель $\hat{Y} = f(X_2, X_4)$ объясняет только 57,4 % общего разброса. Стандартное отклонение составляет 9,4 % от среднего отклика. Все величины F статистически значимы.

Регрессионная модель $\hat{Y} = f(X_1, X_2, X_3, X_4)$ дает $R^2 = 76{,}7$ %; стандартное отклонение снижается до 7,5 % от среднего отклика; все переменные статистически значимы. Однако при большом потреблении воды возникают дополнительные проблемы, даже если модель работает достаточно хорошо.

4. (Для более детального и всестороннего анализа этих данных см. гл. 5 и 7 (с. 138) книги:: D a n i e l C., W o o d F. S. Fitting Equations to Data, 2 nd ed.— New York: Wiley, 1980.

1) $Y = f(X_1)$, поскольку переменная X_1 наиболее коррелирована с $\hat{Y}$.

2) $s^2 = 13{,}444444$.

3)

ANOVA

Источник	Число степеней свободы
Общий	21
Среднее (b_0)	1
Общий (скорректированный)	20
Обусловленный регрессией	3
$b_1 \mid b_0$	1
$b_2 \mid b_0, b_1$	1
$b_3 \mid b_0, b_1, b_2$	1
Остаток	17
Неадекватность	11
«Чистая» ошибка	6

4) Проверка адекватности с помощью F-критерия применима, так как для оценки «чистой» ошибки могут служить повторные опыты.

5) Да, хотя 6 степеней свободы для «чистой» ошибки это и немного, но все же достаточно.

6) $\hat{Y} = -50{,}358836 + 0{,}6711545X_1 + 1{,}295351X_2$.

7) Графики не дают оснований считать, что какая-то очевидная переменная пропущена.

8) Остатки от уравнения из 6 показывают, что оно наименее удовлетворительно при $(X_1, X_2, X_3) = (70, 20, 91)$. Следовательно в окрестности данной точки уравнением следует пользоваться с осторожностью. Будущие опыты должны быть выбраны так, чтобы обеспечить более сбалансированное покрытие X-пространства.

5. 2) Шаг 1 (ввести X_1).

ANOVA

Источник	Число степеней свободы	SS	MS	F-критерий	Частный F-критерий
Общий (скорректированный)	8	1279,20010			
$b_1 \mid b_0$	1	822,27852	822,27852	12,597	12,597
Остаток	7	456,92166	65,27452		

Шаг 2 (ввести X_2).

ANOVA

Источник	Число степеней свободы	SS	MS	F-критерий	Частный F-критерий
Общий (скорректированный)	8	1279,200010			
Регрессия	2	938,29314	469,14657	8,257	
$b_1 \mid b_0$	1	822,27852	822,27852	14,472	3,236
$b_2 \mid b_0, b_1$	1	116,01462	116,01462	2,042	2,042
Остаток	6	340,90696	56,81783		

Шаг 3 (ввести X_3).

ANOVA

Источник	Число степеней свободы	SS	MS	F-критерий	Частный F-критерий
Общий (скорректированный)	8	1279,20010			
Регрессия	3	1081,34820	360,44940	9,109	
$b_1 \mid b_0$	1	822,27852	822,27852	20,780	0,056
$b_2 \mid b_0, b_1$	1	116,01462	116,01462	2,932	5,599
$b_3 \mid b_0, b_1, b_2$	1	143,05506	143,05506	3,615	3,615
Остаток	5	197,85190	39,57038		

Шаг 4 (отвергнуть X_1).

ANOVA

Источник	Число степеней свободы	SS	MS	F-критерий	Частный F-критерий
Общий (скорректированный)	8	1279,20010			
Регрессия	2	1079,12600	529,56300	16,181	
$b_2 \mid b_0$	1	754,40445	754,40445	22,624	17,197
$b_3 \mid b_0, b_2$	1	324,72155	324,72155	9,738	9,738
Остаток	6	200,07410	33,34568		

3) Конечное уравнение: $\hat{Y} = 63{,}021 + 11{,}517\ X_2 - 0{,}816\ X_3$.

4) График остатков не вызывает никаких вопросов.

6. Исследование корреляционной матрицы сразу показывает, что X_1 и X_5 полностью коррелированы. Это обусловлено тем, что рабочие 1 и 5 в каждом опыте либо одновременно *оба* на работе, либо *оба* отсутствуют. Таким образом, их эффекты нельзя оценить раздельно, и для построения регрессии можно опустить X_5 (или X_1, что безразлично). Среди этих данных содержится только 8 различных опытов, следовательно, имеется двенадцать степеней свободы для вычисления «чистой» ошибки. Сумма квадратов, связанная с «чистой» ошибкой, равна 131,929 (число степеней свободы равно 12), так что $s_e^2 = 10{,}994$. В нижеследующей таблице указаны остаточные суммы квадратов для различных моделей с переменными X_1, X_2, X_3 и X_4.

Переменные в модели*	Число степеней свободы для остатков	Остаточная SS**	100 R^2	Переменные в модели*	Число степеней свободы для остатков	Остаточная SS**	100 R^2
—	19	42 644,00	—	23	17	32 700,17	23,32
1	18	8 352,28	80,14	24	17	24 102,10	43,48
2	18	36 253,69	14,99	34	17	16 276,60	61,83
3	18	36 606,19	14,16	123	16	761,41	98,21
4	18	27 254,91	36,09	124	16	5 614,59	86,83
12	17	7 713,10	81,91	134	16	163,93	99,62
13	17	762,55	98,21	234	16	15 619,01	63,37
14	17	6 071,56	85,76	1234	15	163,10	99,62

* Дополнительно к свободному члену β_0, который присутствует во всех моделях.
** До исключения «чистой» ошибки.

Если принять шаговую процедуру, то мы получим такую последовательность результатов:

(а) Вводим X_1. $F_1 = (42\,644 - 8352{,}28)/(8352{,}28/18) = 73{,}90$.
Оставляем X_1.

(б) Добавляем X_3. $F_3 = (8352{,}28 - 762{,}55)/(762{,}55/17) = 169{,}20$.
Оставляем X_3.
$F_1 = (36606{,}19 - 762{,}55)/(762{,}55/17) = 799{,}08$.
Оставляем X_1.

(в) Добавляем X_4. $F_4 = (762{,}55 - 163{,}93)/(163{,}93/16) = 58{,}43$.
Оставляем X_4.
$F_3 = (6071{,}56 - 163{,}93)/(163{,}93/16) = 576{,}60$.
Оставляем X_3.
$F_1 = (16276{,}60 - 163{,}93)/(163{,}93/16) = 1572{,}64$.
Оставляем X_1.

(г) Добавляем X_2. Вклад X_2 очень мал, и потому эта перемеиная отклоняется. Так что мы возвращаемся к набору (X_1, X_3, X_4) и заканчиваем на этом. Итоговая таблица дисперсионного анализа имеет следующий вид:

ANOVA

Источник	Число степеней свободы	SS	MS	F-критерий
b_1, b_3, $b_4 \mid b_0$	3	42 480,07		
Неадекватность	4	32,00	8,00	0,73 незначимо
«Чистая» ошибка	12	131,93	10,99	
Общий (скорректированный)	19	42 644,00		

Заключение. Неадекватность модели не очевидна. Мы не смогли различить эффекты рабочих 1 и 5. Рабочий 2, по-видимому, не вносит заметного вклада по сравнению с другими. Оценка величины σ^2 равна:

$$s^2 = (32{,}00 + 131{,}93)/(4 + 12) = 10{,}25 = 3{,}2^2.$$

7. Мы имеем здесь результаты применения процедур исключения, включения и шаговой. Об анализе на основе критерия C_p см. приведенные литературные источники.

1. *Метод исключения*

Для полного уравнения (123456), SS (остаточная) = 0,307 с 31—7 = 24 степенями свободы. Остаточные суммы для каждой из пятифакторных моделей равны:

Переменные	SS (остат.)	
12 345	0,412	
12 346	0,311	← Это наилучший результат. Тест для переменной 5. Частный F-критерий для переменной 5 при наличии 12 346:
12 356	0,364	
12 456	0,313	
13 456	0,545	
23 456	0,939	

$$F_{1,24} = \frac{0{,}311 - 0{,}307}{0{,}307/24} = 0{,}31.$$

Частный F-критерий показывает, что переменная 5 не необходима для хорошего согласия. Исключим переменную 5.

Наша модель теперь включает только переменные (12346), так что SS (остат.) = 0,311 с 31—6 = 25 степенями свободы. Из этой модели могут быть образованы следующие четырехфакторные модели:

Переменные	SS (остат.)	
1234	0,441	
1236	0,365	
1246	0,323	← Тест для переменной 3. Частный F-критерий для переменной 3 при наличии 1246:
1346	0,555	
2346	0,995	

$$F_{1,25} = \frac{0{,}323 - 0{,}311}{0{,}311/25} = 0{,}96.$$

Переменная 3, по-видимому, не необходима. Исключаем ее.

Наша модель теперь содержит переменные (1246). SS (остат.) = 0,323 с 31—5 = 26 степенями свободы. Из этой модели образуем следующие трехфакторные модели:

Переменные	SS (остат.)	
124	0,450	
126	0,367	← Тест для переменной 4. Частный F-критерий для 4 при наличии 126:
146	0,558	
246	1,192	

$$F_{1,26} = \frac{0{,}367 - 0{,}323}{0{,}323/26} = 3{,}54.$$

Это F-отношение значимо при уровне значимости 0,10 и незначимо при уровне 0,05. Таким образом, это промежуточный случай. Если уравнение будет использоваться только для описания данных, разумно сохранить переменную 4 в уравнении. Процедура исключения в таком случае останавливается, и за окончательную модель принимается (1246). Если же уравнение предполагается использовать для предсказания, тот факт, что переменная 4 (это фиктивная переменная, которая отражает разницу между средними значениями откликов в двух блоках) не управляется, может служить, по-видимому, основанием для ее исключения из модели.

Если переменная 4 исключается, остающаяся модель имеет вид (126), SS (остат.) = 0,367 с 31—4 = 27 степенями свободы.

На основе этой модели могут быть сформированы следующие двухфакторные модели:

Переменные	SS (остат.)	
12	0,499	← Тест для переменной 6. Частный F-критерий для 6 при наличии 12:
16	0,576	
26	9,192	

$$F_{1,27} = \frac{0{,}499 - 0{,}367}{0{,}367/27} = 9{,}71.$$

Это F-отношение значимо при уровне 0,01. Следовательно, переменная 6 необходима и ее нельзя исключать. Процедура закончена, и итоговая модель — (126).

2. *Метод включения*

Чтобы найти переменные, наиболее сильно коррелированные с откликом, рассмотрим все 6 однофакторных уравнений:

Переменные	SS (остат.)	Переменные	SS (остат.)
1	0,607	4	1,522
2	10,795	5	9,922
3	10,663	6	9,196

Модель с переменной 1, бесспорно, наилучшая. F-статистика для переменной 1

$$F_{1,29} = \frac{11{,}058 - 0{,}607}{0{,}607/29} = 499{,}31.$$

Она, несомненно, значима. Теперь попытаемся добавить каждую из пяти оставшихся переменных:

Переменные	SS (остат.)	
12	0,499	← Тест для переменной 2. Частный F-критерий для 2 при наличии 1:
13	0,600	
14	0,582	
15	0,597	$F_{1,28} = \frac{0{,}606 - 0{,}499}{0{,}499/28} = 6{,}00.$
16	0,576	

F-отношение значимо при уровне значимости 0,05. Добавляем переменную 2. Наша модель теперь включает переменные (12). Попытаемся добавить теперь каждую из четырех оставшихся переменных:

Переменные	SS (остат.)	
123	0,498	
124	0,450	
125	0,477	
126	0,367	← Тест для переменной 6. Частный F-критерий для 6 при наличии 12:

$$F_{1,27} = \frac{0{,}499 - 0{,}367}{0{,}367/27} = 9{,}71.$$

F-отношение значимо при уровне 0,01. Вводим переменную 6 в модель. Наша модель теперь — (126). Попытаемся добавить каждую из трех оставшихся переменных:

Переменные	SS (остат.)	
1236	0,365	
1246	0,323	← Тест для переменной 4. Частный F-критерий для 4 при наличии 126:
1256	0,364	

$$F_{1,26} = \frac{0{,}367 - 0{,}323}{0{,}323/26} = 3{,}54.$$

Это промежуточный случай. (См. приведенные выше рассуждения по аналогичному поводу, которые были даны при выполнении процедуры исключения.) Если мы решим не включать переменную 4, то на этом процедура останавливается и окончательная модель имеет вид (126). Если мы включаем переменную 4 в модель, то процедура продолжается. Попытаемся ввести в модель (1246) по одной из оставшихся переменных:

Переменные	SS (остат.)	
12 346	0,311	← Тест для переменной 3. Частный F-критерий для 3 при наличии 1246:
12 456	0,313	

$$F_{1,25} = \frac{0{,}323 - 0{,}311}{0{,}311/25} = 0{,}96.$$

Это F-отношение незначимо. Переменную 3 не следует вводить в модель. Процедура останавливается. Окончательная модель — (1246).

(Примечание. После выполнения нескольких этапов наилучшее уравнение оказывается одинаковым как при использовании процедуры включения, так и при использовании процедуры исключения. Однако с другими экспериментальными данными такого может и не быть.)

3. *Шаговая регрессия*

Эта процедура начинается подобно процедуре включения, описанной выше, с включения переменной 1 и затем — переменной 2. Теперь на этом этапе мы используем частный F-критерий для 1 при наличии переменной 2. Так как SS (остат.) для 2 есть 10,795, а SS (остат.) для (12) равна 0,499, то

$$F_{1,28} = \frac{10{,}795 - 0{,}499}{0{,}499/28} = 577{,}73.$$

Это F-отношению, безусловно, значимо, так что мы не можем удалить переменную 1 из модели. Будем исходить из того, что критическое значение F для исключения переменной из модели не превосходит критического значения для включения (как рекомендовано на с. 24). Так что нет необходимости проверять, стоит ли исключать переменную 2, только что включенную в модель. Продолжая процедуру включения, мы введем переменную 6, а затем проверим, какую из двух ранее введениых в модель переменных 1 или 2 можно исключить из модели.

Частный F для 1 при наличии (26): $F_{1,27} = \dfrac{9{,}192 - 0{,}367}{0{,}367/27} = 649{,}25.$

Частный F для 2 при наличии (16): $F_{1,27} = \dfrac{0{,}576 - 0{,}367}{0{,}367/27} = 15{,}38.$

Оба эти F-отношения значимы при уровне значимости 0,01, так что мы не можем исключить из модели ни ту, ни другую переменную. Продолжая процедуру включения, в качестве следующего кандидата для введения в модель выберем переменную 4. Если мы решим, что частное F-отношение (3.51) не дает оснований для включения переменной 4 в модель, шаговая регрессионная процедура останавливается, и окончательная модель имеет вид (126). Если же мы решим включить переменную 4 в модель, шаговая процедура продолжается. Надо проверить, какая из ранее включенных в модель переменных (1, 2 или 6) может быть исключена из модели.

Частный F для 1 при наличии (246): $F_{1,26} = \dfrac{1{,}192 - 0{,}323}{0{,}323/26} = 69{,}95.$

Частный F для 2 при наличии (146): $F_{1,26} = \dfrac{0{,}558-0{,}323}{0{,}323/26} = 18{,}92.$

Частный F для 6 при наличии (124): $F_{1,26} = \dfrac{0{,}450-0{,}323}{0{,}323/26} = 10{,}22.$

Все эти F-отношения значимы при уровне значимости 0,01. Ни одиу из проверяемых переменных не надо исключать из модели. Продолжая, попытаемся ввести в модель переменную 3, но найдем, что она незначимо улучшает согласие. Таким образом, окончательная модель есть (1246).

8. Как шаговый метод, так и процедура исключения приводят к уравиению

$$\hat{Y} = 3{,}068360 + 0{,}0007259X_1 + 0{,}0446022X_4 \qquad (1)$$

с $R^2 = 0{,}7874$. Однако наблюдается неадекватность этой модели; соответствующее F-отношение равно: $F = \{0{,}27351/73\}/\{0{,}02154/14\} = 2{,}435$, тогда как критическое значение составляет $F\ (73;\ 14;\ 0{,}95) = 2{,}21$. График остатков указывает на квадратичную зависимость от X_1, а статистика Дарбина—Уотсоиа равна: $d = 0{,}8480$, что указывает на наличие сериальной корреляции. Введем в уравиение дополнительиое слагаемое $\beta_{11}X_1^2$, а также другие переменные. Реализуя, скажем, шаговую процедуру, получим такую последовательность результатов:

Предиктор	R^2	Изменение R^2	Уровень значимости (α)
X_1	0,5999	0,5999	0,000
X_4	0,7874	0,1875	0,000
X_1^2	0,8917	0,1043	0,000
X_6	0,9000	0,0083	0,009
X_7	0,9056	0,0056	0,028

В итоге получим модель

$$\hat{Y} = 2{,}997526 + 0{,}0019418X_1 - 0{,}00000289X_1^2 + 0{,}0222738X_4 + \\ + 0{,}0607877X_6 - 0{,}0224359X_7. \qquad (2)$$

Эта модель адекватно отражает экспериментальиые данные, соответствующее F-отиошение равно $F = \{0{,}10944/70\}/\{0{,}02154/14\} = 1{,}016 < F\ (70;\ 14;\ 0{,}95) = 2{,}21$. Статистика Дарбина—Уотсона равна $d = 1{,}60$. Это неубедительный результат при обычной проверке. Из таблицы, приведенной выше, видио, что переменные X_6 и X_7 вносят незначительный вклад в величину R^2, хотя они были включены в модель как «статистически значимые», поскольку мы использовали стандартные (но не совсем правильные, см. с. 26) тесты. Если мы исключим X_6 и X_7 из модели и найдем оценки параметров вновь, мы получим уравнение

$$\hat{Y} = 3{,}0016115 + 0{,}0018344X_1 - 0{,}000002641X_1^2 + 0{,}0390777X_4. \qquad (3)$$

F-отношение для проверки неадекватности теперь будет иметь вид $F = \{0{,}12877/72\}/\{0{,}02154/14\} = 1{,}162$. Эта величина незначима по сравнению с $F\ (72;\ 14;\ 0{,}95) = 2{,}21$. Статистика Дарбина—Уотсона равиа $d = 1{,}64$, она ие позволяет сделать убедительные выводы с помощью обычного теста.

Следовательно, если мы хотим использовать модель, полученную с помощью формальной шаговой процедуры, то это будет уравнение (2), однако с практической точки зрения, по-видимому, более целесообразно работать с моделью (3). Исследование соответствующих машинных распечаток показывает, что предсказании по этим моделим мало различаются между собой.

Дополнительные замечания.

(а) Член X_1^3 также можно рассмотреть. Если его ввести, то мы получим такую последовательность результатов: X_1 ($R^2 = 0{,}5999$), X_4 (приращение R^2 будет равно 0,1875); X_1^2 (0,1043), X_6 (0,0083), X_1^3 (0,0079), X_7 (0,0052).

(б) Если в качестве дополнительных рассматриваются X_4^2 и X_4^3 (так же как и X_1^2 и X_1^3), последовательность шаговых приращеиий величины R^2 будет в точности такой же, как и в случае (а), но с дополнительными результатами, а именно X_4^3 (0,0052), зато будет X_4 (— 0,0010).

В обоих случаях мы стали бы, вероятно, использовать только переменные X_1, X_4 и X_1^2, поскольку все другие значимые переменные увеличивают R^2 менее чем на 0,01.

Глава 7. ● ДВА ТИПИЧНЫХ ПРИМЕРА

7.0. ВВЕДЕНИЕ

Когда у статистика есть основания считать, что опыты были сделаны в определенной серии исследований, анализ данных может быть совсем простым [1]. Часто, однако, приходится анализировать результаты, которые накоплены как *часть* некоторой программы исследования или серии родственных (взаимосвязанных) подобных программ, и по полученным данным строить эмпирическую предсказывающую модель. При этом в такую модель могут включаться не только сами исходные факторы, но и такие переменные, как их смешанные произведения, квадраты или другие комбинации, а также преобразования исходных факторов. Наличие программ, подобных описанным в гл. 6, позволяет автоматизировать вычисления. Но независимо от того, имеются такие программы или нет, толковое использование графиков остатков в зависимости от новых переменных-кандидатов может оказаться очень полезным. В первом примере мы покажем теперь, как сконструировать эмпирическую модель с помощью исследования таких графиков. Второй пример этой главы относится к тем задачам, где данные получаются из соответствующим образом спланированного эксперимента, а для поверхности отклика подбирается и исследуется полиномиальная модель.

7.1. ПЕРВАЯ ЗАДАЧА

На характеристики ракетных двигателей влияет совокупность внешних условий. Обычно в камере сгорания устанавливается стандартное давление, соответствующее нормальной характеристике. В процессе изготовления и испытания двигателей производится регистрация внешних условий и давлений в камере. Типичный набор данных, соответствующий подобным наблюдениям, приведен в

---

[1] Точка зрения, согласно которой планирование эксперимента обеспечивает исключительную простоту анализа, постепенно начинает пересматриваться под воздействием практики и подходов вроде подхода Дж. Тьюки, называемого «анализом данных». Видимо, для «простоты» анализа нужны еще «простота» объекта исследования и «простота» задачи, стоящей перед исследователем. Верно, конечно, что планирование эксперимента облегчает изучение и объектов очень большой сложности, но лишь облегчает, а не делает анализ «совсем простым».— *Примеч. пер.*

Таблица 7.1. Данные по ракетным двигателям

Номер опыта	Температура цикла X_1	Вибрация X_2	Перепад (мгновенный) X_3	Статическое пламя X_4	Давление в камере Y	Номер опыта	Температура X_1	Вибрация X_2	Перепад (мгновенный) X_3	Статическое пламя X_4	Давление в камере Y
1	—75	0	0	—65	1,4	13	0	175	165	—65	11,8
2	175	0	0	150	26,3	14	—75	—75	—65	150	28,4
3	0	—75	0	150	26,5	15	175	175	165	—65	11,5
4	0	175	0	—65	5,8	16	0	—75	0	150	26,5
5	0	—75	0	150	23,4	17	0	175	0	—65	5,8
6	0	175	0	—65	7,4	18	0	0	—65	—65	1,3
7	0	0	—65	150	29,4	19	0	0	165	150	21,4
8	0	0	165	—65	9,7	20	0	—75	—65	—65	0,4
9	0	0	0	150	32,9	21	0	175	165	150	22,9
10	—75	—75	0	150	26,4	22	0	—75	—65	150	26,4
11	175	175	0	—65	8,4	23	0	175	165	—65	11,4
12	0	—75	—65	150	28,8	24	0	0	0	—65	3,7

табл. 7.1. Мы хотим воспользоваться этими данными, чтобы на их основе сконструировать эмпирическую модель, позволяющую предсказывать давление в камере [2].

Приведенные данные (за исключением давления в камере) заимствованы из книги Ллойда и Липова (Lloyd D. K., Lipow M. Reliability: Management, Methods and Mathematics, 2nd ed., Authors and Publishers, Redondo Beach. Ca., 1977, p. 360; есть русский перевод с первого издания, выпущенного в издательстве Prentice-Hall в 1962 г.: Ллойд Д., Липов М. Надежность. Управление, методы и математика/Пер. с англ. Под ред. Б. В. Гнеденко.— М.: Советское радио, 1964, с. 686).

---

[2] Приводимый пример относится к важной области испытания сложных систем, толчком к развитию которой послужили космические исследования. При таких испытаниях планирование и анализ результатов эксперимента, наряду с имитационным моделированием, играют решающую роль. См., например: 1) Шаракшанэ А. С., Железнов И. Г. Испытания сложных систем.— М.: Высшая школа, 1974.— 184 с.; 2) Железнов И. Г., Семенов Г. П. Комбинированная оценка характеристик сложных систем.— М.: Машиностроение, 1976.— 55 с.; 3) Шаракшанэ А. С., Железнов И. Г., Иваницкий В. А. Сложные системы.— М.: Высшая школа, 1977.— 247 с.; 4) Элементы теории испытаний и контроля технических систем/Под ред. Р. М. Юсупова.— Л.: Энергия, 1978.— 192 с.; 5) Миленко Н. П., Сердюк А. В. Моделирование испытаний ЖРД.— М.: Машиностроение, 1975.— 184 с.; 6) Тиме В. А. Оптимизация технико-экономических параметров гидротурбин.— Л.: Машиностроение, 1976.— 272 с.; 7) Нефедов А. Ф., Высочин Л. Н. Планирование эксперимента и моделирование при исследовании эксплуатационных свойств автомобиля.— Львов: Вища школа, 1976.— 160 с.; 8) Бажинов И. К., Почукаев В. Н. Оптимальное планирование навигационных измерений в космическом полете.— М.: Машиностроение, 1976.— 288 с.— *Примеч. пер.*

7.2. ИССЛЕДОВАНИЕ ДАННЫХ

Как можно видеть при заполнении табл. 7.2, не все данные хорошо сбалансированы по предикторам.

Таблица 7.2. **Альтернативная форма представления данных**

		X_1 (−75)			X_1 (0)			X_1 (175)		
		X_2 (−75)	X_2 (0)	X_2 (175)	X_2 (−75)	X_2 (0)	X_2 (175)	X_2 (−75)	X_2 (0)	X_2 (175)
X_3 (−65)	X_4 (−65)				0,4	1,3				
	X_4 (150)	28,4			28,8 26,4	29,4				
X_3 (0)	X_4 (−65)		1,4			3,7	5,8 7,4 5,8			8,4
	X_4 (150)	26,4			26,5 23,4 26,5	32,9			26,3	
X_3 (165)	X_4 (−65)					9,7	11,8 11,4			11,5
	X_4 (150)					21,4	22,9			

Тем не менее мы можем попытаться построить эмпирическую предсказывающую модель по этим данным. Если удастся найти подходящее уравнение, то его можно будет использовать при выборе дополнительных опытов для проверки пригодности построенной модели или для подбора совершенно другого экспериментального плана.

Так как наблюдается заметное возрастание давления в камере с ростом статического пламени, переменная X_4 представляется важным фактором. Повторение опытов позволяет оценить «чистую» ошибку, что создает большое преимущество для любого множества данных. Как обычно, все повторения производятся только при $X_1 = 0$ в предположении, что случайная ошибка постоянна в области эксперимента. Радикальные нарушения этого предположения, если они вообще существуют, будут видны на графиках остатков.

7.3. ВЫБОР ПЕРВОГО ФАКТОРА ДЛЯ ВКЛЮЧЕНИЯ В РЕГРЕССИЮ

Хотя, как мы указали ранее, X_4 — первый кандидат в наиболее важные факторы, исследуем тем не менее корреляционную матрицу. Поскольку эта матрица симметрична, приведена только ее правая верхняя часть (результаты округлены до третьего десятичного знака).

	X_1	X_2	X_3	X_4	Y
X_1	1,000	0,376	0,194	—0,157	—0,065
X_2		1,000	0,538	—0,597	—0,464
X_3			1,000	—0,225	—0,128
X_4				1,000	0,944
Y					1,000

Наибольшая корреляция, равная 0,944, существует между X_4 и Y. Теперь мы можем построить модель

$$Y = \beta_0 + \beta_4 X_4 + \varepsilon.$$

Опустим детали вычислений. Готовое уравнение имеет вид

$$Y = 12{,}614147 + 0{,}0932946 X_4,$$

ему соответствует следующая таблица дисперсионного анализа:

ANOVA

Источник	Степени свободы	SS	MS	F
Общий (скорректированный)	23	2711,60		
Регрессия	1	2414,02	2414,02	
Остаток	22	297,58	13,53	
Неадекватность	16	286,51	17,91	9,70
«Чистая» ошибка	6	11,07	1,85	

Вычисленное значение $R^2 = 2414{,}02/2711{,}60 = 0{,}890$. Это значит, что модель (по имеющимся данным) объясняет 89 % разброса относительно среднего. Однако значение F-критерия для оценки степени неадекватности, равное 9,70, превосходит F (16; 6; 0,975), т. е. модель должна считаться неадекватной. Чтобы разобраться в этом, мы можем исследовать остатки, вычерчивая, в частности, их графики в зависимости от «кандидатов» в новые факторы. Исходные наблюдения, предсказанные значения и остатки приведены в табл. 7.3. На рис. 7.1, 7.2 и 7.3 показаны графики остатков в зависимости от X_1, X_2 и X_3 соответственно.

Рис. 7.1. График остатков в зависимости от X_1

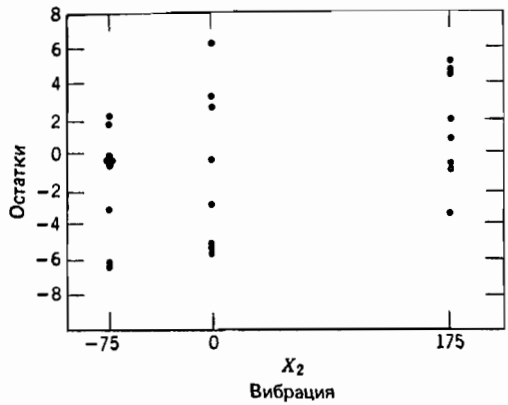


Рис. 7.2. График остатков в зависимости от X_2

Таблица 7.3. **Наблюдения, предсказанные значения и остатки**

Наблюдаемый Y	Предсказываемый $\hat{Y}$ из $\hat{Y} = f(X_4)$	Остаток
1,4	6,55	—5,15
26,3	26,61	—0,31
26,5	26,61	—0,11
5,8	6,55	—0,75
23,4	26,61	—3,21
7,4	6,55	0,94
29,4	26,61	2,79
9,7	6,55	3,15
32,9	26,61	6,29
26,4	26,61	—0,21
8,4	6,55	1,85
28,8	26,61	2,19
11,8	6,55	5,25
28,4	26,61	1,79
11,5	6,55	4,95
26,5	26,61	—0,11
5,8	6,55	—0,75
1,3	6,55	—5,25
21,4	26,61	—5,21
0,4	6,55	—6,15
22,9	26,61	—3,71
26,4	26,61	—0,21
11,4	6,55	4,85
3,7	6,55	—2,85

Из-за вертикального разброса нанесенных значений, указания на зависимость остатков от X_1, X_2 или X_3 не удивительны. (Заметим, однако, что факторы X_1, X_2 и X_3 не согласуются с фактором X_4, уже введенным в регрессию. Это иногда вуалирует связь, которая может проявиться *после* такого согласования.) Вместо попыток использовать наши переменные мы обратим внимание на возможность добавления в модель переменных другого типа.

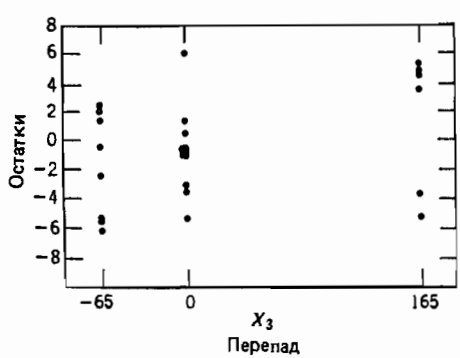


Рис. 7.3. График остатков в зависимости от X_3

7.4. ПОСТРОЕНИЕ НОВЫХ ПЕРЕМЕННЫХ

Теперь мы построим шесть возможных столбцов для смешанных произведений факторов вида X_iX_j. Члены такого типа позволяют учесть взаимодействие переменных X_i и X_j и его влияние на отклик. Элемент столбца X_iX_j — просто произведение соответствующих элементов столбцов X_i и X_j. Например, первый элемент столбца X_1X_4 есть (— 75) (— 65) = 4875 (см. табл. 7.4).

Таблица 7.4 **Исходные данные и столбцы эффектов взаимодействия**

№	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	—75	0	0	—65	0	0	4875	0	0	0	1,4
2	175	0	0	150	0	0	26250	0	0	0	26,3
3	0	—75	0	150	0	0	0	0	—11250	0	26,5
4	0	175	0	—65	0	0	0	0	—11375	0	5,8
5	0	—75	0	150	0	0	0	0	—11250	0	23,4
6	0	175	0	—65	0	0	0	0	—11375	0	7,4
7	0	0	—65	150	0	0	0	0	0	—9750	29,4
8	0	0	165	—65	0	0	0	0	0	—10725	9,7
9	0	0	0	150	0	0	0	0	0	0	32,9
10	—75	—75	0	150	5625	0	—11250	0	—11250	0	26,4
11	175	175	0	—65	30625	0	—11375	0	—11375	0	8,4
12	0	—75	—65	150	0	0	0	4 875	—11250	—9750	28,8
13	0	175	165	—65	0	0	0	28 875	—11375	—10725	11,8
14	—75	—75	—65	150	5625	4875	—11250	4 875	—11250	—9750	28,4
15	175	175	165	—65	30625	28875	—11375	28 875	—11375	—10725	11,5
16	0	—75	0	150	0	0	0	0	—11250	0	26,5
17	0	175	0	—65	0	0	0	0	—11375	0	5,8
18	0	0	—65	—65	0	0	0	0	0	4225	1,3
19	0	0	165	150	0	0	0	0	0	24750	21,4
20	0	—75	—65	—65	0	0	0	4 875	4875	4225	0,4
21	0	175	165	150	0	0	0	28 875	26250	24750	22,9
22	0	—75	—65	150	0	0	0	4 875	—11250	—9750	26,4
23	0	175	165	—65	0	0	0	28 875	—11375	—10725	11,4
24	0	0	0	—65	0	0	0	0	0	0	3,7

7.5. ВКЛЮЧЕНИЕ В МОДЕЛЬ ВЗАИМОДЕЙСТВИЯ

График остатков из табл. 7.3 в зависимости от соответствующих значений смешанных произведений столбца X_3X_4 из табл. 7.4 при-

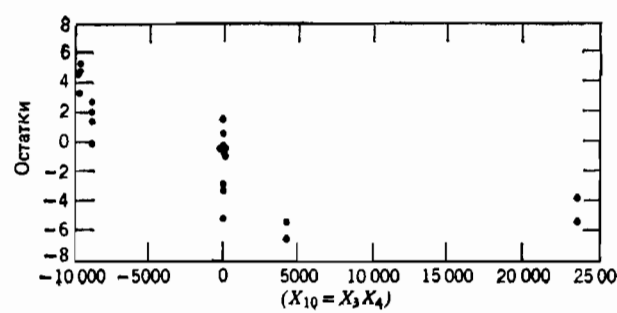


Рис. 7.4. График остатков в зависимости от X_3X_4

веден на рис. 7.4. Он обнаруживает бо́льшую зависимость, чем в случае других смешанных произведений, графики которых мы здесь не приводим. Теперь добавим в исходную модель член $\beta_{34}X_3X_4$ и все пересчитаем. Пересчитанное уравнение имеет вид

$$\hat{Y} = 12{,}153563 + 0{,}979107X_4 - 0{,}0002650X_3X_4,$$

а таблица дисперсионного анализа такова:

ANOVA

Источник	Степени свободы	SS	MS	F
Общий (скорректированный)	23	2711,60		
Регрессия	2	2553,93	1276,97	
Остаток	21	157,67	7,51	
Неадекватность	15	146,59	9,77	5,30
«Чистая» ошибка	6	11,07	1,85	

Теперь $R^2 = 2553{,}93/2711{,}60 = 0{,}942$, т. е. более чем 94 % вариации относительно среднего по всем данным объясняется с помощью модели. Значение F для оценки неадекватности, равное 5,30, превышает F (15; 6; 0,975); поэтому дальнейшее улучшение, видимо, возможно.

7.6. РАСШИРЕНИЕ МОДЕЛИ

На этом этапе были снова найдены остатки и снова нанесены на график, теперь уже в зависимости от соответствующих элементов столбцов, входящих в табл. 7.4 (исключая X_4 и X_3X_4, которые уже входят в модель). Мы опускаем детали, оставляя их читателю в качестве упражнений. Графики показывают, что теперь, после того как большая часть вариации с добавлением члена $\beta_{34}X_3X_4$ устранена, наиболее подходящим кандидатом, по-видимому, будет переменная X_2. Добавляя эту переменную в модель, имеем

$$Y = \beta_0 + \beta_4X_4 + \beta_{34}X_3X_4 + \beta_2X_2 + \varepsilon.$$

После вычислений мы получим

$$Y = 10{,}713137 + 0{,}112340X_4 - 0{,}0003140X_3X_4 + 0{,}0233483X_2$$

и сможем построить следующую таблицу дисперсионного анализа:

ANOVA

Источник	Степени свободы	SS	MS	F
Общий (скорректированный)	23	2711,60		
Регрессия	3	2641,59	880,53	251,55
Остаток	20	70,01	3,50	
Неадекватность	14	58,94	4,21	2,28
«Чистая» ошибка	6	11,07	1,85	

Неадекватность незначима, так что модель адекватна. Построенное уравнение обеспечивает $R^2 = 2641{,}59/2711{,}60 = 0{,}974$, или объясняет 97,4 % вариации относительно среднего по всем данным. Регрессия в целом значима, так же как значимы и все индивидуальные коэффициенты.

Коэффициенты b и 95 %-е доверительные пределы для β

Номер переменной	Среднее	Натуральный b-коэффициент	Пределы верхний/нижний
X_4	42,5	0,1123395	0,1220046/0,1026743
X_3X_4	—997,9	—0,0003140	—0,0002242/—0,0004038
X_2	33,3	0,02333483	0,0330812/0,0136155

Теперь мы получили эмпирическую модель, которую можно использовать для предсказания результатов. Модель воссоздает несколько произвольный образ того, почему отклик реагирует на вариацию в независимых переменных. Такая модель дает лишь эмпирическое истолкование данных, которое тем не менее может оказаться полезным в дальнейшей работе. С этой точки зрения может быть разумна перепроверка текущих остатков, чтобы посмотреть, работает ли модель удовлетворительно за пределами той области, в которой были выполнены наблюдения. Табл. 7.5 содержит остатки для последней модели в двух формах — в исходной форме и в форме нормальных отклонений (см. § 3.1). Приблизительно 5 %, или одно на двадцать нормальных отклонений остатков могут выходить за область (— 2;2). Так как за этими пределами лежит только одно значение, а именно 2,85 (для опыта № 9), то нет оснований для беспокойства. Более того, опыт № 9 был проведен при $X_1 = X_2 = X_3 = 0$ и $X_4 = 150$, когда наблюдаемое давление в камере было самым высоким из зарегистрированных. Фактически же наибольший интерес представляет низкое давление в камере, так что построенная эмпирическая модель хуже всего работает в «малоинтересной» области.

Таблица 7.5. **Наблюдения, предсказанные значения и остатки для окончательной модели**

Номер наблюдения	Y	$\hat{Y}$	Остаток $Y-\hat{Y}$	Нормальные отклонения остатков	Номер наблюдения	Y	$\hat{Y}$	Остаток $Y-\hat{Y}$	Нормальные отклонения остатков
1	1,4	3,41	—2,01	—1,07	13	11,8	10,86	0,94	0,50
2	26,3	27,56	—1,26	—0,68	14	28,4	28,87	—0,47	—0,25
3	26,5	25,81	0,69	0,37	15	11,5	10,86	0,64	0,34
4	5,8	7,50	—1,70	—0,91	16	26,5	25,81	0,69	0,37
5	23,4	25,81	—2,41	—1,29	17	5,8	7,50	—1,70	—0,91
6	7,4	7,50	—0,10	—0,05	18	1,3	2,08	—0,78	—0,42
7	29,4	30,63	—1,23	—0,66	19	21,4	19,79	1,61	0,86
8	9,7	6,78	2,92	1,56	20	0,4	0,33	0,07	0,04
9	32,9	27,56	5,34	2,85	21	22,9	23,88	—0,98	—0,52
10	26,4	25,81	0,59	0,31	22	26,4	28,87	—2,47	—1,32
11	8,4	7,50	0,90	0,48	23	11,4	10,86	0,54	0,29
12	28,8	28,87	—0,07	—0,04	24	3,7	3,41	0,29	0,15

7.7. ВТОРАЯ ЗАДАЧА. ЧИСЛЕННЫЕ ПРИМЕРЫ ПОВЕРХНОСТИ ВТОРОГО ПОРЯДКА, ПОСТРОЕННОЙ ДЛЯ ТРЕХ И ДЛЯ ДВУХ ФАКТОРОВ

В журнале Industrial and Engineering Chemistry, January 1961, **53**, p. 55—57, было опубликовано исследование Айя, Голдсмита и Муни (Aia M. A., Goldsmith R. L. and Mooney R. W.), выполненное на полупромышленной установке под названием «Предсказание стехиометрического $CaHPO_4 \cdot 2H_2O$» (Predicting Stoichiometric $CaHPO_4 2H_2O$). В настоящем параграфе эта работа воспроизводится с разрешения Американского химического общества (American Chemical Society). Мы опустили здесь подробности, касающиеся химии процесса, а также произвели некоторые незначительные изменения в авторском анализе [3].

В исследуемой задаче имелось всего семь кандидатов в предикторы, но четыре из них было решено зафиксировать на время эксперимента на постоянных уровнях. Три оставшиеся для исследования поверхности отклика, вместе с их областями экспериментирования, приведены ниже.

Всего представляло интерес семь откликов. Для каждого из них последовательно использовалась одна и та же идея, а именно, предпринималась попытка построить функцию второго порядка от r, t и pH. Мы воспользуемся здесь только первым откликом (который назовем Y); данные для четвертого отклика (обозначим его Y_4) оставим для упражнения 7.11.

[3] Соединение, которое получается в данном примере,— вторичный гидросульфат кальция, соль ортофосфорной кислоты — широко распространенное фосфатное удобрение. Мольное отношение — один из способов задания концентраций. Подсчитывается как отношение числа молей двух компонентов в растворе.— *Примеч. пер.*

Фактор	Обозначение	Область значений
Мольное отношение $NH_3/CaCl_2$ в растворе хлористого кальция	r	0,70—1,00
Время (в минутах) введения аммонийного соединения $NH_4H_2PO_4$ в раствор $CaCl_2$	t	10—90
Начальная кислотность (pH) раствора $NH_4H_2PO_4$	pH	2—5

Выбранный план [4] представлял собой «куб плюс звезда плюс семь нулевых точек», т. е. был композиционным планом со звездным плечом $\alpha = 5/3 = 1{,}667$. (Чтобы обладать свойством «ротатабельности», выбранный план должен был бы иметь значение звездного плеча $\alpha = 2^{3/4} = 1{,}6818$, а значит, этот план близок к ротатабельному, но не точно ротатабельный. Авторы пренебрегли этим обстоятельством и фактически использовали $\alpha = 1{,}6818$ в своих вычислениях. По этой причине наши вычисления несколько отличны от авторских.)

Выбранный план требует для каждого фактора пяти уровней. Натуральные значения факторов кодировались с помощью следующих преобразований:

$$X_1 = (r - 0{,}85)/0{,}09, \quad X_2 = (t - 50)/24, \quad X_3 = (\text{pH} - 3{,}5)/0{,}9.$$

Тогда в соответствии с планом фактические уровни переменных можно выразить следующей таблицей:

Кодированные уровни X_1, X_2 или X_3	Фактические уровни (натуральные)		
	r	t	pH
$\frac{5}{3}$	1,00	90	5,0
1	0,94	74	4,4
0	0,85	50	3,5
—1	0,76	26	2,6
$-\frac{5}{3}$	0,70	10	2,0

[4] Ротатабельные планы обладают тем приятным свойством, что дисперсия предсказанного значения отклика, в случае их использования, не зависит от направления от центра плана (центральной точки), в котором ведется предсказание, а зависит только от расстояния до центра. Отсюда и название «ротатабельный» — инвариантный относительно вращения координатной системы. Экспериментальные точки в таких планах располагаются на не менее чем двух концентрических сферах соответствующей размерности. Планы такого рода были предложены Дж. Боксом в США. Широко исследовались и применялись в СССР. См., например, коллективную монографию: Новые идеи в планировании эксперимента/Под ред. В. В. Налимова.— М.: Наука, 1969.— 336 с. Для начального знакомства см.: Налимов В. В., Голикова Т. И. Логические основания планирования эксперимента.— 2-е изд.— М.: Металлургия, 1981.— 152 с. и Адлер Ю. П., Грановский Ю. В., Маркова Е. В. Теория эксперимента: прошлое, настоящее, будущее.— М.: Знание, 1982.— 64 с.— *Примеч. пер.*

Таблица 7.6. **Численный пример: матрица X и два отклика**

	Матрица плана											
u	1	X_1	X_2	X_3	X_1^2	X_2^2	X_3^2	X_1X_2	X_1X_3	X_2X_3	Y	Y_4
1	1	−1	−1	−1	1	1	1	1	1	1	52,8	6,95
2	1	1	−1	−1	1	1	1	−1	−1	1	67,9	5,90
3	1	−1	1	−1	1	1	1	−1	1	−1	55,4	7,10
4	1	1	1	−1	1	1	1	1	−1	−1	64,2	7,08
5	1	−1	−1	1	1	1	1	1	−1	−1	75,1	5,64
6	1	1	−1	1	1	1	1	−1	1	−1	81,6	5,18
7	1	−1	1	1	1	1	1	−1	−1	1	73,8	6,84
8	1	1	1	1	1	1	1	1	1	1	79,5	5,67
9	1	$-\frac{5}{3}$	0	0	$\frac{25}{9}$	0	0	0	0	0	68,1	6,00
10	1	$\frac{5}{3}$	0	0	$\frac{25}{9}$	0	0	0	0	0	91,2	5,67
11	1	0	$-\frac{5}{3}$	0	0	$\frac{25}{9}$	0	0	0	0	80,6	5,52
12	1	0	$\frac{5}{3}$	0	0	$\frac{25}{9}$	0	0	0	0	77,5	6,47
13	1	0	0	$-\frac{5}{3}$	0	0	$\frac{25}{9}$	0	0	0	36,8	7,17
14	1	0	0	$\frac{5}{3}$	0	0	$\frac{25}{9}$	0	0	0	78,0	5,36
15	1	0	0	0	0	0	0	0	0	0	74,6	6,48
16	1	0	0	0	0	0	0	0	0	0	75,9	5,91
17	1	0	0	0	0	0	0	0	0	0	76,9	6,39
18	1	0	0	0	0	0	0	0	0	0	72,3	5,99
19	1	0	0	0	0	0	0	0	0	0	75,9	5,86
20	1	0	0	0	0	0	0	0	0	0	79,8	5,96

Фактический план в кодированных переменных показан как отмеченная часть матрицы **X** в табл. 7.6. Относительно кодированных переменных постулируется модель второго порядка, которую можно записать в виде

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{11} X_1^2 + \beta_{22} X_2^2 + \beta_{33} X_3^2 + \\ + \beta_{12} X_1 X_2 + \beta_{13} X_1 X_3 + \beta_{23} X_2 X_3$$

для отклика Y и аналогично для Y_4.

Опыты проводились с рандомизацией во времени, и в каждом из них регистрировались все семь откликов. Вот два из них: Y — выход в процентах от теоретического и Y_4 — насыпная плотность в граммах на кубический дюйм [5], которые как раз и приведены в табл. 7.6. Соответствующая матрица $\mathbf{X'X}$ будет одна и та же для любого отклика; она имеет вид

$$\mathbf{X'X} = \begin{bmatrix} N & 0 & 0 & 0 & B & B & B & 0 & 0 & 0 \\ 0 & B & 0 & 0 & & & & & & \\ 0 & 0 & B & 0 & & 0 & & & 0 & \\ 0 & 0 & 0 & B & & & & & & \\ B & & & & C & D & D & & & \\ B & & 0 & & D & C & D & & 0 & \\ B & & & & D & D & C & & & \\ 0 & & & & & & & D & 0 & 0 \\ 0 & & 0 & & & 0 & & 0 & D & 0 \\ 0 & & & & & & & 0 & 0 & D \end{bmatrix}$$

где в нашем частном случае:

$$N = 20,$$

$$B = 8 + 2\alpha^2 = \frac{122}{9},$$

$$C = 9 + 2\alpha^4 = \frac{1898}{81},$$

$$D = 8.$$

Такой тип разбиения матрицы часто встречается в исследованиях по планированию эксперимента для построения поверхности второго порядка. Для него легко получить обратную матрицу. В общем случае, когда число факторов равно k (а не трем, как в рассмотренном примере), используя те же обозначения как обычно, получим матрицу $\mathbf{X'X}$ бóльших размеров и сможем выписать для нее обратную в сле-

---

[5] Кубический дюйм равен 16,39 см³.— *Примеч. пер.*

дующем виде:

$$(\mathbf{X}'\mathbf{X})^{-1} =$$

	0	1	2	...	k	11	22	...	kk	12	13	...	$(k-1)\,k$	
0	P	0	0	...	0	Q	Q	...	Q	0	0	...	0	0
1	0	$1\,B$	0	...	0									1
2	0	0	$1/B$	...	0									2
...	...						0				0			...
k	0	0	0	...	$1\,B$									k
11	Q					R	S	...	S					11
22	Q					S	R	...	S					22
...	...		0			...					0			...
kk	Q					S	S	...	R					kk
12	0									$1/D$	0	...	0	12
13	0									0	$1/D$	...	0	13
...	...		0				0							...
$(k-1)\,k$	0									0	0	...	$1/D$	$(k-1)\,k$

Значения P, Q, R и S приведены в табл. 7.7 во втором столбце, обозначенном $C \neq 3D$. (Значения в третьем столбце относятся к наипростейшей форме, когда $C = 3D$, а это бывает, если план «ротатабелен», т. е. контуры дисперсии $V\{\hat{Y}(X)\}$ имеют сферическую форму. При таких обстоятельствах план был бы ротатабельным в пространстве предикторов (т. е. в пространстве X) и не давал бы никакого предпочтения в точности получаемой информации в зависимости от направления.)

В нашем случае $k = 3$, а $3D = 1944/81$, так что $C - 3D$ мало, но не равно нулю, как это было бы в случае ротатабельности плана. Следовательно, мы получим:

$$P = 1597/9614, \quad Q = -549/9614,$$
$$R = 685{,}3248/9614, \quad S = 62{,}3376/9614.$$

Таблица 7.7. **Формулы для получения элементов матрицы $(\mathrm{X}'\mathrm{X})^{-1}$**

Символ	Значение при $C \neq 3D$	Значение при $C = 3D$ (ротатабельный план)
P	$(C-D)(C+(k-1)D)/A$	$2(k+2)D^2/A$
Q	$-(C-D)B/A$	$-2DB/A$
R	$\{N(C+(k-2)D)-(k-1)B^2\}/A$	$\{N(k+1)D-(k-1)B^2\}/A$
S	$(B^2-ND)/A$	$(B^2-ND)/A$
A	$(C-D)\{N(C+(k-1)D)-kB^2\}$	$2D\{N(k+2)D-kB^2\}$

Заметьте, что в формулах для P, Q, R и S встречается величина A. Таким образом, $1/B = \frac{9}{122}$, а $1/D = \frac{1}{8}$. (Это точные значения. Они позволяют в дальнейшем избежать ошибок округления. Заключительное деление на 9614 имеет смысл отложить до тех пор, пока не будут выполнены последовательные перемножения матриц). Теперь нам нужен вектор $\mathbf{X'Y}$. Ниже приведены такие векторы для откликов Y и Y_4:

$$\mathbf{X'Y} = \frac{1}{9}\begin{bmatrix} 12941{,}1 \\ 671{,}4 \\ -87{,}0 \\ 1245{,}3 \\ 8935{,}2 \\ 8905{,}2 \\ 7822{,}7 \\ -63{,}9 \\ -105{,}3 \\ -20{,}7 \end{bmatrix}, \qquad \mathbf{X'Y}_4 = \frac{1}{9}\begin{bmatrix} 1108{,}26 \\ -29{,}25 \\ 41{,}43 \\ -60{,}45 \\ 744{,}99 \\ 752{,}99 \\ 766{,}49 \\ 2{,}88 \\ -5{,}04 \\ 3{,}24 \end{bmatrix}$$

Воспользовавшись обычной формулой $\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$ для оценок параметров регрессии, мы получим следующее уравнение для первого отклика Y:

$$Y = 76{,}022 + 5{,}503X_1 - 0{,}713X_2 + 10{,}207X_3 + 0{,}712X_1^2 + 0{,}496X_2^2 - \\ - 7{,}298X_3^2 - 0{,}888X_1X_2 - 1{,}463X_1X_3 - 0{,}288X_2X_3.$$

Когда реализован план второго порядка и получена матрица $\mathbf{X'X}$ такого типа, как на с. 116, таблица дисперсионного анализа выглядит так, как показано в табл. 7.8. В этой таблице приняты следующие обозначения:

$$(0y) = \sum_{u=1}^{N} Y_u,$$

$$(iy) = \sum_{u=1}^{N} X_{iu}Y_u,$$

$$(iiy) = \sum_{u=1}^{N} X_{iu}^2 Y_u,$$

$$(ijy) = \sum_{u=1}^{N} X_{iu}X_{ju}Y_u,$$

для всех скалярных произведений столбцов матрицы $\mathbf{X}$ по столбцу наблюдений $\mathbf{Y}$, так что все они оказываются элементами вектора $\mathbf{X'Y}$. Обычно нам приходится объединять суммы квадратов SS $(b_{ii}|b_0)$ и SS (b_{ij}), чтобы получить SS (члены второго порядка $|\ b_0$) с $^1/_2\ k\ (k+1)$ степенями свободы, однако в таблице они приводятся в отдельности. Тем самым подчеркивается, что только *дополнительная* сумма квад-

Таблица 7.8. Стандартная таблица дисперсионного анализа для некоторых типов планов второго порядка

Источник	Степени свободы	SS
b_0 (среднее)	1	$\left(\sum_{u=1}^{N} Y_u\right)^2 / N$
b_i (члены первого порядка)	k	$\sum_{i=1}^{k} b_i (iY)$
$b_{ii} \mid b_0$ (квадраты при данном b_0)	k	$b_0 (0Y) + \sum_{j=1}^{k} b_{ii} (iiY) - \left(\sum_{u=1}^{N} Y_u\right)^2 / N$
b_{ij} (взаимодействия — смешанные члены второго порядка)	$^1/_2 k\,(k-1)$	$\sum_{i=1}^{k} \sum_{j=1}^{k} b_{ij} (ijY)$ $i < j$
Неадекватность	$N - n_e - {}^1/_2(k+1) \times (k+2)$	По разности
«Чистая» ошибка	n_e	Обычный подсчет
Общий	N	$\sum_{u=1}^{N} Y_u^2$

ратов, которая представляет собой SS $(b_{ii}|b_0)$, обусловлена ортогональностью многих пар столбцов матрицы $\mathbf{X}$, причем это свойство присуще только некоторым частным видам планов. Теперь самым обычным образом мы можем проверить адекватность и оценить вклад членов первого и второго порядка.

Следует еще раз отметить, что многие особые свойства такого оценивания методом наименьших квадратов и дисперсионного анализа приложимы *только* к планам, матрицы $\mathbf{X}'\mathbf{X}$ которых имеют специальный вид, такой, что для $(\mathbf{X}'\mathbf{X})^{-1}$ можно воспользоваться теми формулами, какие были приведены выше. Более того, столбец «Источник» в приведенной выше таблице дисперсионного анализа создает основу для использования таких формул. Сумма квадратов, обусловленная чистой ошибкой, получается как обычно, а все значения для сумм квадратов оценок параметров будут получаться в виде дополнительных сумм квадратов, как описано в § 2.7.

Приведем теперь соответствующую нашему примеру таблицу дисперсионного анализа (см. табл. 7.9). Поскольку F (5; 5; 0,95) =

Таблица 7.9. **Таблица дисперсионного анализа для полученной модели**

Источник	Степени свободы	SS	MS	F
Среднее (b_0)	1	103 377,82		
Первый порядок	3	1 829,80	609,93	
Второй порядок $\mid b_0$	6	813,54	135,59	
Неадекватность	5	93,91	18,78	3,04
«Чистая» ошибка	5	30,86	6,17	
Общий	20	106 145,93		

$= 5{,}05 > 3{,}04$, нет указаний на неадекватность. Поэтому можно объединить суммы квадратов для неадекватности и для «чистой» ошибки и оценить дисперсию $V(Y_t) = \sigma^2$ так:

$$S^2 = (93{,}91 + 30{,}86)/(5+5) = 12{,}477.$$

Деление на полученную дисперсию среднего квадрата для членов первого порядка дает отношение $609{,}93/12{,}477 = 48{,}88$, что превосходит F (3; 10; 0,999) = 12,55; в то же время для членов второго порядка получается отношение $135{,}59/12{,}477 = 10{,}88$, что тоже превосходит F (6; 10; 0,999) = 9,93. Таким образом, в полученной модели нельзя обойтись ни без членов первого, ни без членов второго порядка.

А нужна ли нам переменная X_2?

В исходной работе авторы заметили, что все оцеики коэффициентов, которые относятся к фактору 2 и его функциям, имеют в сравнении со своими стандартными ошибками малые величины. Отсюда они заключили, что их модель вовсе не должна содержать X_2. Когда ситуация вполне прозрачна — все имеющиеся коэффициенты малы по сравнению со своими стандартными ошибками, такое заключение вряд ли может оказаться ошибочным. Правда, в таком случае надо было бы воспользоваться, вообще говоря, главной дополнительной суммой квадратов. Продемонстрируем здесь, как это делается.

Пусть мы хотим проверить нулевую гипотезу $H_0: \beta_2 = \beta_{22} = \beta_{12} = \beta_{23} = 0$, против альтернативной гипотезы о том, что по крайней мере один из этих коэффициентов не равен нулю. Из таблицы дисперсионного анализа найдем регрессионную сумму квадратов для полной модели второго порядка с X_1, X_2, X_3 при наличии b_0:

$S_1 = \text{SS (члены первого порядка)} + \text{SS (члены второго порядка}\mid b_0) =$

$= 1829{,}80 + 813{,}54 =$

$= 2643{,}34$ (с $3+6=9$ степенями свободы).

Теперь надо сформулировать гипотезу H_0 применительно к сокращенной модели

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_{11} X_1^2 + \beta_{33} X_3^2 + \beta_{13} X_1 X_3.$$

Соответствующую матрицу $\mathbf{X}$ можно получить из табл. 7.6 после вычеркивания столбцов X_2, X_2^2, X_1X_2, X_2X_3. Точно так же можно получить и матрицу $\mathbf{X'X}$, вычеркивая строки и столбцы, соответствующие фактору X_2. Аналогично получается и вектор $\mathbf{X'Y}$ после вычеркивания тех же строк. В результате получается уравнение:

$$\hat{Y} = 76{,}420 + 5{,}503X_1 + 10{,}207X_3 + 0{,}667X_1^2 - 7{,}343X_3^2 - 1{,}463X_1X_3.$$

Далее нам нужна регрессионная сумма квадратов при данном b_0 для сокращенной модели. Мы найдем, что она равна $S_2 = 2626{,}025$ (с 5 степенями свободы). Тогда дополнительная сумма квадратов, обусловленная b_2, b_{22}, b_{12} и b_{23}, будет равна:

$$S_1 - S_2 = 2643{,}34 - 2626{,}03 = 17{,}31 \text{ (с } 9-5=4 \text{ степенями свободы).}$$

Это приводит к среднему квадрату 17,31/4 = 4,33, который можно сравнить с оценкой остаточного среднего квадрата σ^2, полученной для исходной трехфакторной регрессии. Получается, что нуль-гипотезу о том, что $\beta_2 = \beta_{22} = \beta_{12} = \beta_{23} = 0$, нельзя отвергнуть. Значит, вполне резонно пользоваться сокращенной моделью, которая не включает члены, содержащие X_2. Составим теперь таблицу дисперсионного анализа, соответствующую сокращенной модели (см. табл. 7.10). Видно, что неадекватность отсутствует, а в регрессии высокозначимы члены и первого и второго порядка.

Таблица 7.10. **Дисперсионный анализ для сокращенной модели второго порядка с факторами X_1 и X_3**

Источник	Степени свободы	SS	MS	F
Первый порядок	2	1822,91	911,46	89,80
Второй порядок | b_0	3	803,12	267,71	26,38
Неадекватность	9 } 14	111,22 } 142,08	12,36 } 10,15	2,00
«Чистая» ошибка	5 }	30,86 }	6,17 }	
Общий (скорректированный)	19	2768,11		

Для изучения построенной поверхности второго порядка мы могли бы провести обычный «канонический анализ», в котором эта поверхность описывается в терминах координат, совпадающих с главными осями поверхности. Такой анализ крайне полезен и позволяет охватить всю ситуацию, даже в случае многих факторов [6]. Но когда фак-

[6] Канонический анализ — стандартный прием анализа поверхностей второго порядка, получаемых методом планирования экспериментов. За подробностями можно обратиться, например, к работе: Адлер Ю. П. Введение в планирование эксперимента.— М.: Металлургия, 1969.— 158 с. (особо с. 107—113) и к серии публикаций Р. И. Слободчиковой, например: Рузинов Л. П., Слободчикова Р. И. Планирование эксперимента в химии и химической технологии.— М.: Химия, 1980.— 280 с.— *Примеч. пер.*

торов всего два, как здесь, мы можем сразу построить контурные линии для $\hat{Y}$, переписав полученное уравнение в таком виде:

$$-7{,}343X_3^2+(10{,}207-1{,}463X_1)\,X_3+$$
$$+(0{,}667X_1^2+5{,}503X_1+76{,}420-\hat{Y})=0.$$

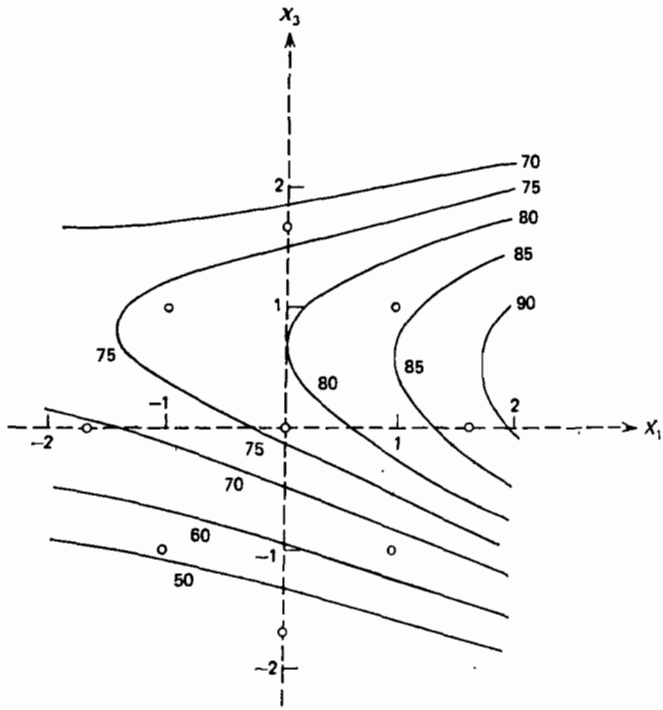


Рис. 7.5. Контуры равных значений Y для модели второго порядка относительно факторов X_1 и X_3

Если задаться значением $\hat{Y}$, то соответствующий ему контур можно вычертить, задаваясь значениями X_1 и вычисляя X_3. Полученные таким образом контуры показаны на рис. 7.5. Экспериментальные точки отмечены на рисунке кружочками. Правда, повторяющиеся точки никак не выделены, и, чтобы их найти, надо обратиться к табл. 7.11. Контуры показывают, что поверхность имеет вид гребня. Исследование системы этих контуров привело авторов к гипотезе о тех химических реакциях, которые могли бы привести к подобным контурам. (Довольно часто исследователи поверхностей отклика на начальных этапах проводят более основательное теоретическое изучение рассматриваемого объекта.)

Полученные контуры можно было бы еще рассмотреть в связи с остатками, приведенными в табл. 7.11. «Структура» графика остатков, на котором каждый остаток расположен рядом с той точкой

Таблица 7.11. Предсказанные значения и остатки, полученные для модели поверхности второго порядка $\hat{Y} = f(X_1, X_3)$

u	X_1	X_3	Y	$\hat{Y}$	$e=Y-\hat{Y}$	u	X_1	X_3	Y	$\hat{Y}$	$e=Y-\hat{Y}$
1	—1	—1	52,8	52,57	0,23	11	0	0	80,6	76,42	4,18
2	1	—1	67,9	66,50	1,40	12	0	0	77,5	76,42	1,08
3	—1	—1	55,4	52,57	2,83	13	0	$-\frac{5}{3}$	36,8	39,01	—2,21
4	1	—1	64,2	66,50	—2,30	14	0	$\frac{5}{3}$	78,0	73,04	4,97
5	—1	1	75,1	75,91	—0,81	15	0	0	74,6	76,42	—1,82
6	1	1	81,6	83,99	—2,39	16	0	0	75,9	76,42	—0,52
7	—1	1	73,8	75,91	—2,11	17	0	0	76,9	76,42	0,48
8	1	1	79,5	83,99	—4,49	18	0	0	72,3	76,42	—4,12
9	$-\frac{5}{3}$	0	68,1	69,10	—1,00	19	0	0	75,9	76,42	—0,52
10	$\frac{5}{3}$	0	91,2	87,44	3,76	20	0	0	79,8	76,42	3,38

плана, к которой он относится, показана на рис. 7.6. Из двадцати остатков шесть, наибольших по абсолютной величине, приходятся на точки $(X_1, X_3)=(0, 0)$ (три), $\left(\frac{5}{3}, 0\right)$, $\left(0, \frac{5}{3}\right)$ и (1,1) (по одному). Таким образом, модель выглядит подогнанной наименее хорошо в первом квадранте плоскости (X_1, X_3), и любые утверждения, которые считаются достоверными для полученной поверхности, в этой области надо брать под сомнение. (Может ли иметь место эффект, который следует из авторских выводов?— это скорее вопрос к инженеру-химику, чем к статистику, и мы не будем его здесь обсуждать.) Сомнения такого рода иногда могут стимулировать дальнейшие действия, связанные с экспериментированием в том районе, где форма полученной поверхности наиболее подозрительна, и с получением на этой основе новой модели в заданной более узкой области факторного пространства.

Теперь мы можем исследовать остатки другими способами, чтобы посмотреть, нет ли каких-нибудь иных аномалий. На рис. 7.7 приведены следующие стандартные графики остатков: (*а*) общий, (*б*) в зависимости от предсказанных значений $\hat{Y}_u$ (*в*), в зависимости от X_{1u} и (*г*) в зависимости от X_{3u}.

Общий график не выглядит так, что он отвергает предположение о нормальности, на которое опирается критерий отношения дисперсий в дисперсионном анализе. График остатков в зависимости от $\hat{Y}_u$ на первый взгляд демонстрирует тенденцию к «расширению», но это

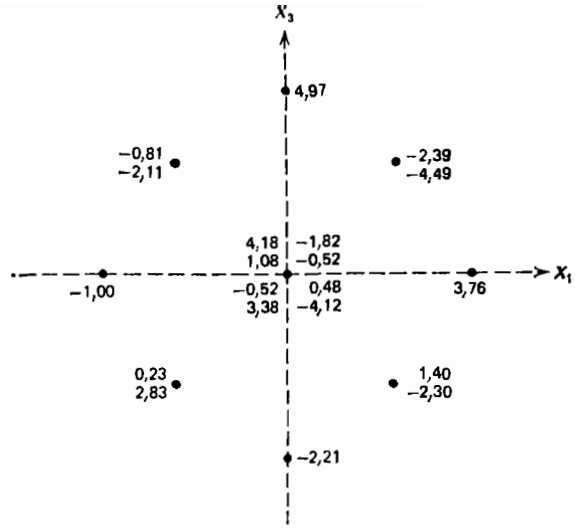


Рис. 7.6. «Структура» графика остатков для модели второго порядка с откликом Y и факторами X_1 и X_3

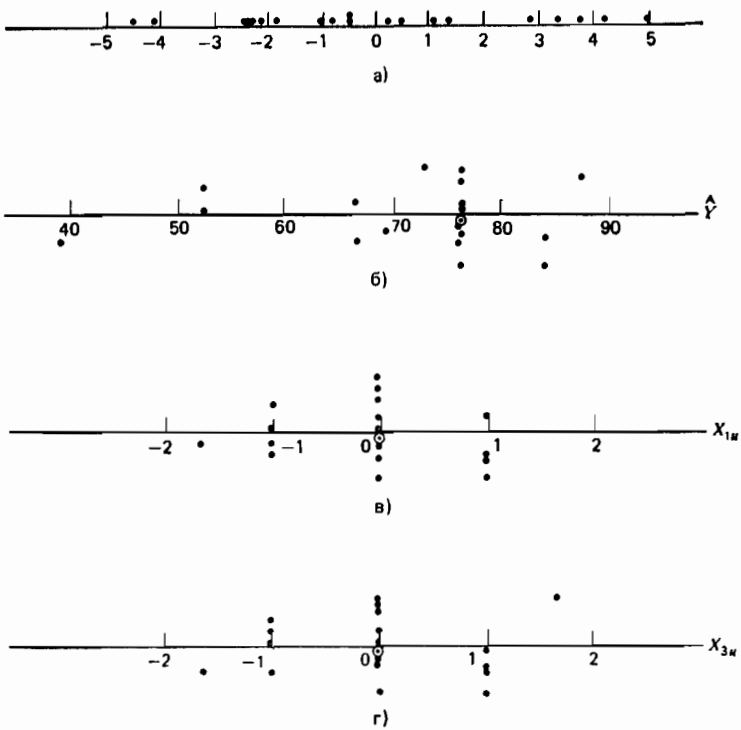


Рис. 7.7. Стандартные графики остатков для получения уравнения второго порядка с откликом Y и факторами X_1 и X_3:
(а) Общий. (б) В зависимости от $\hat{Y}$. (в) В зависимости от X_1 (г) В зависимости от X_3.

обманчиво, ибо большинство остатков велики и размеры полосы остатков плохо определены на нижнем конце шкалы $\hat{Y}$. Аналогичное поведение наблюдается и на графиках зависимостей от X_1 и X_3, где размеры полосы остатков плохо определены на краях диапазона. Следовательно, ни на одном из этих графиков нет ничего, что давало бы основания говорить об отклонении от нормальности. А значит, и нет оснований считать, что не выполняются основные регрессионные предпосылки. (Заметим, что, поскольку мы не знаем, в каком порядке были реализованы опыты, мы не можем проверить, нет ли временно́го тренда, влияющего на отклик.)

Если продолжить исследование, то дополнительные усилия могли бы включать попытки переоценить исходные данные в первом квадранте, где получились большие остатки, а также рассмотреть новые факторы, вариация которых, возможно, имела место, но которые до сих пор не рассматривались. Таким образом, быть может, удастся усовершенствовать модель. К тому же вопреки ожиданию область, в которой модель оказалась под вопросом, следовало бы изучить более подробно, чем предполагалось ранее.

Вычисление «чистой» ошибки в случае, когда факторы выпадают

Предыдущий анализ ставит один вопрос, которого мы до сих пор избегали: когда какой-нибудь фактор, вроде X_2, выпадает из модели, надо ли нам пересчитывать «чистую» ошибку? В табл. 7.6 параллельными опытами были только опыты 15—20, когда же фактор X_2 выпал, т. е. данные стали такими, как в табл. 7.11, то пары точек, а именно (1,3), (2,4), (5,7) и (6,8), превратились в пары параллельных опытов относительно факторов X_1 и X_3. Более того, опыты 11 и 12 превратились в новые центральные точки. Поэтому следовало бы учесть, что новый план надо рассматривать как повторенный дважды полный факторный план 2^2 (восемь точек: 1—8) плюс двухфакторная звезда (четыре точки: 9, 10, 13, 14) плюс восемь центральных точек (11, 12, 15, . . . , 20). Если так сделать, то пришлось бы пересмотреть табл. 7.10, чтобы получить новые значения:

$$\text{SS неадекватности} = 78{,}26 \text{ (3 степени свободы); MS} = 26{,}08,$$
$$\text{SS «чистой» ошибки} = 63{,}82 \text{ (11 степеней свободы); MS} = 5{,}80.$$

Соответствующее F-отношение равно: $4{,}50 > F\ (3;\ 11;\ 0{,}95) = 3{,}59$. Отсюда следует несколько неожиданный вывод о том, что *имеет* место неадекватность. В таком пересмотренном анализе получается, что, хотя фактор X_2 совершенно не обязателен в модели, его исключение ведет к неадекватности! Между тем ясно, что X_2 мало помогает объяснить вариацию в наблюдениях. Фактически величина среднего квадрата «чистой» ошибки обесценивается, когда X_2 *участвует* в модели, и в то же самое время уменьшение числа степеней свободы для «чистой» ошибки снижает чувствительность F-критерия к неадекватности.

Какой же анализ верен? Можно было бы защищать обе позиции. В целом, однако, мы предпочитаем пользоваться той «чистой» ошибкой, которую мы вычислили сначала, до всякого отбрасывания факторов. Видимо, повторные опыты в исходных данных *и в самом деле* параллельны, если можно так выразиться: но нельзя утверждать то же самое относительно тех опытов, которые выглядят как параллельные, когда выпал некоторый фактор. Таким образом, во многих наборах данных может возникнуть противоположная проблема, т. е. будет иногда пропадать «генетическая» неадекватность, поскольку в «новых» параллельных опытах будет проявляться больше вариации, чем ее было в исходных параллельных.

В качестве одного из надежных путей преодоления возникшей трудности можно предложить проводить анализ двумя способами с дальнейшим выяснением, нет ли между результатами согласия. Для многих наборов данных такое согласие будет наблюдаться. Ну а если его нет, то данные должны стать предметом дальнейшего тщательного рассмотрения.

А как же нам все-таки поступить в данном примере? Наша модель, безусловно, должна быть поставлена под сомнение. Хотя члены первого и второго порядка объясняют в общей вариации долю, равную $R^2 = (1822{,}91 + 803{,}12)/2758{,}11 = 0{,}949$ (с учетом коррекции на среднее), для проверки адекватности остается всего лишь пять степеней свободы. (При включении в модель фактора X_2 результат увеличивается всего лишь до 0,955.) Иными словами, эта модель объясняет 95 % вариации относительно среднего, даже если, в принципе, неадекватность и возможна. Чтобы выяснить, где же неадекватность могла бы проявиться, можно обратиться к совместному исследованию контурных линий и остатков. Если линии истинной поверхности отклика хорошо проявились в широкой области факторного пространства, то заключения на основе полученной модели могут оказаться в этой области вполне состоятельными. Дальнейшее исследование остатков позволяет также выявить, не нарушаются ли основные предпосылки регрессионного анализа, такие, как нормальность, постоянство дисперсии, независимость наблюдений, и нет ли еще каких-либо путей ревизии нашей модели.

Иногда в практической работе оказывается, что «чистая» ошибка «слишком мала» просто потому, что параллельные опыты не были рандомизированы (или хотя бы распределены) с остальными опытами. Если некоторые параллельные опыты делаются подряд или в течение короткого времени, то отклики проявляют тенденцию больше походить друг на друга, чем при рандомизации. Иначе говоря, при отсутствии рандомизации «чистая» ошибка оказалась бы непредставительной по отношению к разбросу данных, характерному для проводимого эксперимента. Иногда это приводит к ложным сигналам о наличии неадекватности, что нуждается в тщательном исследовании.

Комментарий. Пример, который мы только что обсуждали, в некотором отношении необычен. Когда имеются основания для исключения из модели некоторых членов, в упрощенной модели неадекватность обычно не проявляется, если только в данных нет ка-

ких-то особенностей. Между тем мы могли видеть, что такие особенности не проявились в нарушениях предпосылок метода наименьших квадратов. Их источник остается предметом для размышлений.

В широком смысле этот пример совсем не необычен. Хотя эксперимент и ответил на некоторые вопросы, другие он оставил неразрешенными. Они станут предметом дальнейших обдумываний и дальнейшей работы. В этом смысле такого рода пример типичен для большинства практических исследований.

Вычисление «чистой» ошибки, когда план разбит на блоки

План из табл. 7.6 не был разбит на блоки. Однако часто планы для изучения поверхностей отклика разбиваются на блоки, причем так, чтобы блоки были ортогональны к модели. Опыты, которые были бы параллельными в плане без разбиения на блоки, в таком случае часто разделяются между блоками. Тогда эти опыты перестают быть параллельными, *если только они не попадают в один блок*, и «чистую» ошибку надо вычислять, исходя из этого. Следовательно, дисперсионный анализ должен содержать некоторую сумму квадратов для блоков. Когда блоки ортогональны к модели, блоковая сумма квадратов получается обычно как $\text{SS (блоков)} = \sum_{w=1}^{m} \frac{B_w^2}{n_w} - \frac{G^2}{N}$ с $(m-1)$ степенями свободы из таблицы дисперсионного анализа, где B_w — сумма тех n_w наблюдений, которые попали в w-й блок (а всего имеется m блоков), а G — общая сумма по всем наблюдениям во всех m блоках. Когда же блоки не ортогональны по отношению к модели, используется принцип дополнительной суммы квадратов. (Его, конечно, можно применять во всех случаях, безотносительно к тому, ортогональны блоки или нет. При этом будет получаться тот ответ, который приведен выше для случая ортогональности.)

Упражнения

1. Каждое из двадцати бревен, полученных из взрослых деревьев сосны, было распилено на чурбаки толщиной 30 мм. Эти чурбаки морили в черной морилке «хлорозол Е», после чего для каждого дерева были определены следующие характеристики [7].

[7] В производстве бумаги качество готовой продукции по понятным причинам существенно зависит от качества исходного сырья — балансовой древесины («балансов»). Балансами называются круглые или колотые лесоматериалы длиной 1—3 м и диаметром 8—24 см. Такими габаритами обладают уже сложившиеся взрослые растения. Наиболее влияющими на качество бумаги считаются факторы, перечисленные в таблице к этому упражнению, но измерять легче всего удельный вес. Поэтому соблазнительно найти регрессионную модель, связывающую удельный вес с трудно определяемыми переменными. Для этого приходится ставить специальный эксперимент. Разрезая бревно на круглые чурбаки, мы выявляем структуру годовых колец, а после обработки морилкой хорошо видим весеннюю и летнюю части кольца, поскольку они окрашиваются различно. Ясно, что выборка столь малого объема не позволяет надеяться на получение достаточно надежных результатов, однако методическое значение работы несомненно.

Данные по анатомическим факторам и удельному весу (древесины) сосновых бревен

Число древесных волокон на 1 мм² в весенней древесине	Число древесных волокон на 1 мм² в летней древесине	Весенняя древесина, %	Адсорбция воздуха, %		Удельный вес древесины
			весенней древесиной	летней древесиной	
573	1059	46,5	53,8	84,1	0,534
651	1356	52,7	54,5	88,7	0,535
606	1273	49,4	52,1	92,0	0,570
630	1151	48,9	50,3	87,9	0,528
547	1135	53,1	51,9	91,5	0,548
557	1236	54,9	55,2	91,4	0,555
489	1231	56,2	45,5	82,4	0,481
685	1564	56,6	44,3	91,3	0,516
536	1182	59,2	46,4	85,4	0,475
685	1564	63,1	56,4	91,4	0,486
664	1588	50,6	48,1	86,7	0,554
703	1335	51,9	48,4	81,2	0,519
653	1395	62,5	51,9	89,2	0,492
586	1114	50,5	56,5	88,9	0,517
534	1143	52,1	57,0	88,9	0,502
523	1320	50,5	61,2	91,9	0,508
580	1249	54,6	60,8	95,4	0,520
448	1028	52,2	53,4	91,8	0,506
476	1057	42,9	53,2	92,9	0,595
528	1057	42,4	56,6	90,0	0,568

Источник. Van Buijtenen J. P. Anatomical Factors Influencing Wood Specific Gravity of Slash Pines and the Implications for the Development of a High-Quality Pulpwood.—Tappi, 1964, **47** (7), p. 401—404.

Получите предсказывающее уравнение для удельного веса древесины. Используйте шаговый метод и примите критическое значение F равным 2,00 как для включения, так и для исключения факторов. Исследуйте остатки для полученной модели и сделайте выводы.

2. При диспергировании сыворотки в центрифуге были собраны следующие данные [8].

Производство бумаги и вообще переработка древесины — традиционная область приложения статистических методов, и в частности планирования эксперимента. Вот основные отечественные публикации:— Пен Р. З., Менчер Э. М. Статистические методы в целлюлозно-бумажном производстве.— М.: Лесная промышленность, 1973.— 120 с.; Пен Р. З. Статистические методы моделирования и оптимизации процессов целлюлозно-бумажного производства: Учебное пособие.—Красноярск: Изд-во Красноярского университета, 1982.— 192 с.; Пижурин А. А. Современные методы исследований технологических процессов в деревообработке.— М.: Лесная промышленность, 1972.— 248 с.; Пижурин А. А. Оптимизация технологических процессов деревообработки.— М.: Лесная промышленность, 1975.— 312 с.; Фергин В. Р. Методы оптимизации в лесопильно-деревообрабатывающем производстве.— М.: Лесная промышленность, 1975.— 216 с.— *Примеч. пер.*

[8] Речь идет о фармакологической задаче, где устанавливается зависимость конечного размера частиц от условий центрифугирования и состояния исходной смеси. Желающие углубиться в эту область исследования могут начать, например, с работы: Беликов В. Г., Пономарев В. Д., Коковкин-

Значения экспериментально управляемых переменных и среднего размера частиц

Номер опыта	Скорость на входе на единицу длины (г/с/см) X_1	Окружная скорость ротора (см/с) X_2	Вязкость на входе (пуазы) X_3	Средний размер частиц, μ Y	Номер опыта	Скорость на входе на единицу длины (г/с/см) X_1	Окружная скорость ротора (см/с) X_2	Вязкость на входе (пуазы) X_3	Средний размер частиц, μ Y
1	0,0174	5 300	0,108	25,4	19	0,1010	5 700	0,098	39,7
2	0,0630	5 400	0,107	31,6	20	0,0622	6 200	0,102	31,5
3	0,0622	8 300	0,107	25,7	21	0,0622	7 700	0,102	26,9
4	0,0118	10 800	0,106	17,4	22	0,0170	10 200	0,100	18,1
5	0,1040	4 600	0,102	38,2	23	0,0118	4 800	0,102	28,4
6	0,0118	11 300	0,105	18,2	24	0,0408	6 600	0,102	27,3
7	0,0122	5 800	0,105	26,5	25	0,0622	8 300	0,102	25,8
8	0,0122	8 000	0,100	19,3	26	0,0170	7 700	0,102	23,1
9	0,0408	10 000	0,106	22,3	27	0,0408	9 000	0,613	23,4
10	0,0408	6 600	0,105	26,4	28	0,0170	10 100	0,619	18,1
11	0,0630	8 700	0,104	25,8	29	0,0408	5 300	0,671	30,9
12	0,0408	4 400	0,104	32,2	30	0,0622	8 000	0,624	25,7
13	0,0415	7 600	0,106	25,1	31	0,1010	7 300	0,613	29,0
14	0,1010	4 800	0,106	39,7	32	0,0118	6 400	0,328	22,0
15	0,0170	3 100	0,106	35,6	33	0,0170	8 000	0,341	18,8
16	0,0412	9 300	0,105	23,5	34	0,0118	9 700	1,845	17,9
17	0,0170	7 700	0,098	22,1	35	0,0408	6 300	1,940	28,4
18	0,0170	5 300	0,099	26,5					

Источник: Scott M. W., Robinson M. J., Pauls J. F., Lantz R. J. Spray congealing: particle size relationships using a centrifugal wheel atomizer. — Journal of Pharmacentical Sciences. June, 1964, **53** (6), p. 670—675. Воспроизводится с разрешения владельца авторских прав.

Предлагаемая модель, основанная на теоретических представлениях, такова:

$$Y = \alpha X_1^{\beta} X_2^{\gamma} X_3^{\delta} \varepsilon.$$

После линеаризующего преобразования постройте предлагаемую модель с помощью метода наименьших квадратов. Установите, какие из независимых переменных наиболее важны, и проверьте все коэффициенты на статистическую значимость (примите $\alpha = 0{,}05$). Удовлетворительна ли модель?

3. Приведенные ниже данные, относящиеся к исследованиям количества витамина B_2 в ботве репы [9], взяты из работы Уейкли из Университета Северной Каролины (Роли): Wakeley J. T. Annual Progress Report on the Soils-weather Project, 1948. University of North Carolina (Raleigh) Institute of Statistics Mimeo Series, **19**, 1949.

Щербак Н. И. Применение математического планирования и обработка результатов эксперимента в фармации.— М.: Медицина, 1973.— 232 с.— *Примеч. пер.*

[9] Витамин B_2 — рибофлавин или лактофлавин — относится к водорастворимым витаминам, входящим в состав окислительно-восстановительных ферментов (флавопротеидов). Содержится в мясо-молочных продуктах, салатных овощах, курином желтке, пивных дрожжах. Для получения препарата витамина B_2 полезно найти такие источники сырья, которые бы не использовались непосредственно как продукты питания. Этим отчасти и обусловлена постановка данной работы.— *Примеч. пер.*

Переменные:
X_1 — солнечная радиация в полдень, измеряемая в относительных граммкалориях в минуту (кодирована делением на 100);
X_2 — среднее напряжение влажной почвы (кодировано делением на 100);
X
X_3 — температура воздуха в градусах Фаренгейта (кодирована делением на 10);
Y — миллиграммы витамина B_2 на грамм ботвы репы.

X_1	X_2	X_3	Y	X_1	X_2	X_3	Y
1,76	0,070	7,8	110,4	1,80	0,020	7,3	75,3
1,55	0,070	8,9	102,8	1,80	0,020	6,5	92,0
2,73	0,070	8,9	101,0	1,77	0,020	7,6	82,4
2,73	0,070	7,2	108,4	2,30	0,020	8,2	77,1
2,56	0,070	8,4	100,7	2,03	0,474	7,6	74,0
2,80	0,070	8,7	100,3	1,91	0,474	8,3	65,7
2,80	0,070	7,4	102,0	1,91	0,474	8,2	56,8
1,84	0,070	8,7	93,7	1,91	0,474	6,9	62,1
2,16	0,070	8,8	98,9	0,76	0,474	7,4	61,0
1,98	0,020	7,6	96,6	2,13	0,474	7,6	53,2
0,59	0,020	6,5	99,4	2,13	0,474	6,9	59,4
0,80	0,020	6,7	96,2	1,51	0,474	7,5	58,7
0,80	0,020	6,2	99,0	2,05	0,474	7,6	58,0
1,05	0,020	7,0	88,4				

Эти данные были использованы в работе Андерсона и Банкрофта (Anderson R. L., Bancroft T. A. Statistical Theary in Research.— New York: McGraw-Hill, 1959) на с. 192 для построения модели.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_{12} X_1 X_2 + \varepsilon.$$

Постройте подходящую модель, используя эти данные, и сравните ее с той, которую получили Андерсон и Банкрофт.

4. Пользуясь данными из приложения А, получите модель для предсказания месячного расхода пара, исключив из рассмотрения переменную 6 и обеспечив:

1) R^2 выше чем 0,8.
2) Статистическую значимость всех b-коэффициентов (при $\alpha = 0{,}05$).
3) Отсутствие явных структур в остатках, полученных для построенной модели.
4) Величину остаточного стандартного отклонения, выраженную в процентах от среднего значения отклика, не более чем 7 %.

(Если воспользуетесь процедурой отбора, то возьмите критическое значение F-критерия, равное 3,00, как для включения, так и для исключения факторов.)

5. В отчете № 17 (CAED Report № 17) Университета штата Айова (Iowa State University) за 1963 г. приведены следующие данные[10] по штату Айова за 1930—1962 гг. (см. с. 130).

[10] Изучение урожайности сельскохозяйственных культур в зависимости от погоды — одна из постоянных тем статистических исследований. См., например: Применение статистических методов в агрометеорологии/Под ред. А. П. Федосеева, О. Д. Сиротенко//Труды института экспериментальной метеорологии. Вып. 18.— М.: Гидрометеоиздат. Московское отделение, 1970.— 80 с.; Юзбашев М. М., Манелля А. И. Статистический анализ тенденций и колеблемости.— М.: Финансы и статистика, 1983.— 207 с.; Полевой опыт.— 2-е изд./Под ред. П. Г. Найдина.— М.: Колос, 1968.— 328 с. (особо с. 120—130).— *Примеч. пер.*

Постройте предсказывающую модель для урожая зерна (бушелей на акр). Прокомментируйте относительную важность рассматриваемых независимых переменных и предложите программу дальнейших исследований.

Год		X_1 Предсезонные осадки, фунтов	X_2 Майская температура, °F	X_3 Июньские дожди, фунтов	X_4 Июньская температура, °F	X_5 Июльские дожди, фунтов	X_6 Июльская температура, °F	X_7 Августовские дожди, фунтов	X_8 Августовская температура, °F	X_9 Урожай зерна, бушелей на акр	Y
1930	1	17,75	60,2	5,83	69,0	1,49	77,9	2,42	74,4	34,0	
	2	14,76	57,5	3,83	75,0	2,72	77,2	3,30	72,6	32,9	
	3	27,99	62,3	5,17	72,0	3,12	75,8	7,10	72,2	43,0	
	4	16,76	60,5	1,64	77,8	3,45	76,1	3,01	70,5	40,0	
	5	11,36	69,5	3,49	77,2	3,85	79,7	2,84	73,4	23,0	
	6	22,71	55,0	7,00	65,9	3,35	79,4	2,42	73,6	38,4	
	7	17,91	66,2	2,85	70,1	0,51	83,4	3,48	79,2	20,0	
	8	23,31	61,8	3,80	69,0	2,63	75,9	3,99	77,8	44,6	
	9	18,53	59,5	4,67	69,2	4,24	76,5	3,82	75,7	46,3	
	10	18,56	66,4	5,32	71,4	3,15	76,2	4,72	70,7	52,2	
1940	11	12,45	58,4	3,56	71,3	4,57	76,7	6,44	70,7	52,3	
	12	16,05	66,0	6,20	70,0	2,24	75,1	1,94	75,1	51,0	
	13	27,10	59,3	5,93	69,7	4,89	74,3	3,17	72,2	59,9	
	14	19,05	57,5	6,16	71,6	4,56	75,4	5,07	74,0	54,7	
	15	20,79	64,6	5,88	71,7	3,73	72,6	5,88	71,8	52,0	
	16	21,88	55,1	4,70	64,1	2,96	72,1	3,43	72,5	43,5	
	17	20,02	56,5	6,41	69,8	2,45	73,8	3,56	68,9	56,7	
	18	23,17	55,6	10,39	66,3	1,72	72,8	1,49	80,6	30,5	
	19	19,15	59,2	3,42	68,6	4,14	75,0	2,54	73,9	60,5	
	20	18,28	63,5	5,51	72,4	3,47	76,2	2,34	73,0	46,1	
1950	21	18,45	59,8	5,70	68,4	4,65	69,7	2,39	67,7	48,2	
	22	22,0	62,2	6,11	65,2	4,45	72,1	6,21	70,5	43,1	
	23	19,05	59,6	5,40	74,2	3,84	74,7	4,78	70,0	62,2	
	24	15,67	60,0	5,31	73,2	3,28	74,6	2,33	73,2	52,9	
	25	15,92	55,6	6,36	72,9	1,79	77,4	7,10	72,1	53,9	
	26	16,75	63,6	3,07	67,2	3,29	79,8	1,79	77,2	48,4	
	27	12,34	62,4	2,56	74,7	4,51	72,7	4,42	73,0	52,8	
	28	15,82	59,0	4,84	68,9	3,54	77,9	3,76	72,9	62,1	
	29	15,24	62,5	3,80	66,4	7,55	70,5	2,55	73,0	66,0	
	30	21,72	62,8	4,11	71,5	2,29	72,3	4,92	76,3	64,2	
1960	31	25,08	59,7	4,43	67,4	2,76	72,6	5,36	73,2	63,2	
	32	17,79	57,4	3,36	69,4	5,51	72,6	3,04	72,4	75,4	
1962	33	26,61	66,6	3,12	69,1	6,27	71,6	4,31	72,5	76,0	
Средние	17	19,09	60,8	4,85	70,3	3,55	75,2	3,82	73,2	50,0	

6. Плотность готового продукта — это его важная рабочая характеристика. Ею можно в значительной мере управлять в условиях производства с помощью четырех основных факторов:

X_1 — количество воды в смеси продуктов,

X_2 — количество перерабатываемого материала в смеси продуктов,

X_3 — температура смеси,

X_4 — температура воздуха в сушильной камере.

Кроме того, для процесса важен состав поступающего сырья. Мерой его качества служит величина X_5 — подъем температуры. Были собраны следующие данные:

X_1	X_2	X_3	X_4	X_5	X_6	X_1	X_2	X_3	X_4	X_5	X_6
0	800	135	578	13,195	104	75	800	135	550	12,745	103
0	800	135	578	13,195	102	75	800	135	550	12,745	111
0	800	135	578	13,195	100	75	800	135	550	12,745	111
0	800	135	578	13,195	96	75	800	135	550	12,745	107
0	800	135	578	13,195	93	75	800	135	550	12,745	112
0	800	135	578	13,195	103	75	800	135	550	12,745	106
0	800	150	585	13,180	118	75	800	150	595	13,885	111
0	800	150	585	13,180	113	75	800	150	595	13,885	107
0	800	150	585	13,180	107	75	800	150	595	13,885	104
0	800	150	585	13,180	114	75	800	150	595	13,885	103
0	800	150	585	13,180	110	75	800	150	595	13,885	104
0	800	150	585	13,180	114	75	800	150	595	13,885	103
0	1000	135	590	13,440	97	75	1000	135	530	11,705	116
0	1000	135	590	13,440	87	75	1000	135	530	11,705	108
0	1000	135	590	13,440	92	75	1000	135	530	11,705	104
0	1000	135	590	13,440	85	75	1000	135	530	11,705	116
0	1000	135	590	13,440	94	75	1000	135	530	11,705	116
0	1000	135	590	13,440	102	75	1000	135	530	11,705	112
0	1000	150	590	13,600	104	75	1000	150	590	13,835	111
0	1000	150	590	13,600	102	75	1000	150	590	13,835	110
0	1000	150	590	13,600	101	75	1000	150	590	13,835	115
0	1000	150	590	13,600	104	75	1000	150	590	13,835	114
0	1000	150	590	13,600	98	75	1000	150	590	13,835	114
0	1000	150	590	13,600	101	75	1000	150	590	13,835	114

1) Тщательно рассмотрите данные и сделайте предварительные выводы.
2) Оцените коэффициенты модели

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon.$$

3) Адекватна ли приведенная выше модель?
4) Можете ли вы предложить какую-нибудь альтернативную модель?
5) Сделайте некоторые общие выводы относительно этого эксперимента.

7. Для определения зависимости скорости от шести факторов был поставлен некоторый эксперимент. По приведенным ниже данным постройте предсказывающее уравнение для скорости и воспользуйтесь им для предсказания некоторого множества рабочих значений, которые могут дать возрастание скорости.

X_1	X_2	X_3	X_4	X_5	X_6	Y (скорость)
149	66	—15	150	105	383	267
143	66	—5	115	105	383	269
149	73	—5	150	105	383	230
143	73	—15	115	105	383	233
149	73	—15	115	78	383	222
143	66	—15	150	78	383	267
143	73	—5	150	78	383	231
149	66	—5	115	78	383	260
149	73	—15	150	78	196	238
149	66	—5	150	78	196	262
143	73	—5	115	78	196	252
143	66	—15	115	78	196	263
143	66	—5	150	105	196	263
149	73	—5	115	105	196	236
149	66	—15	115	105	196	268
143	73	—15	150	105	196	242

8. Приведенные ниже данные содержат 16 наблюдений относительно индекса потребления резины [11] с 1948 по 1963 г. Используя эти данные, постройте подходящие уравнения при предсказаниях Y_1 и Y_2 в отдельности и с независимыми переменными X_1, X_2, X_3 и X_4.

Номер наблюдения	Общее потребление резины Y_1	Потребление в шинной промышленности Y_2	Производство автомобилей X_1	Валовой национальный доход X_2	Чистый годовой доход на душу населения X_3	Потребление моторного топлива X_4
1	0,909	0,871	1,287	0,984	0,987	1,046
2	1,252	1,220	1,281	1,078	1,064	1,081
3	0,947	0,975	0,787	1,061	1,007	1,051
4	1,022	1,021	0,796	1,013	1,012	1,046
5	1,044	1,002	1,392	1,028	1,029	1,036
6	0,905	0,890	0,893	0,969	0,993	1,020
7	1,219	1,213	1,400	1,057	1,047	1,057
8	0,923	0,918	0,721	1,001	1,024	1,034
9	1,001	1,014	1,032	0,996	1,003	1,014
10	0,916	0,914	0,685	0,972	0,993	1,013
11	1,173	1,170	1,291	1,046	1,027	1,037
12	0,938	0,952	1,170	1,004	1,001	1,007
13	0,965	0,946	0,817	1,002	1,014	1,008
14	1,106	1,096	1,231	1,049	1,032	1,024
15	1,011	0,999	1,086	1,023	1,020	1,030
16	1,080	1,093	1,001	1,035	1,053	1,029

9. Эксперимент, результаты которого сведены в таблицу, был выполнен на полупромышленной установке для выяснения воздействия изменений в концентрации одного из компонентов смеси (X_1), температуры смеси (X_2) и скорости потока через реактор (X_3) на три отклика — Y_1, Y_2 и Y_3. Исходные факторы были закодированы, но отклики сохранили первоначальные единицы измерения. Приведенный ниже экспериментальный план представляет собой композиционный план, образованный кубом из восьми точек (X_1, X_2, X_3) = ($\pm$ 1, $\pm$ 1, $\pm$ 1), шестью звездными точками ($\pm \alpha$, 0, 0), (0, $\pm \alpha$, 0) и (0, 0, $\pm \alpha$), где $\alpha = 1{,}2154$, и одной центральной точкой (0, 0, 0). Порядок, в котором записаны опыты,— это как раз тот порядок, какой получился в результате рандомизации плана.

Ротатабелен ли этот план? (см. 7.7). Используя методы множественного регрессионного анализа, постройте подходящие модели первого или второго порядка отдельно для Y_1, Y_2 и Y_3. Выполните полный анализ и сделайте практические выводы. Если бы было желательно получить наибольшие значения откликов, то в какой области факторного пространства было бы лучше работать?

---

[11] Индексы — относительные показатели, в частности, учитывающие колебания стонмости доллара. Благодаря этому они облегчают сопоставления, подобные тем, что предлагаются в данном упражнении. Полезную информацию о построении и использовании индексов производства и потребления можно найти в книге: Левшин Ф. М. Мировые товарные рынки.— М.: Международные отношения, 1978.— 359 с. Кстати, автор этой работы предлагает пользоваться в анализе конъюнктуры методами регрессионного анализа и цитирует перевод первого издания данной книги.— *Примеч. пер.*

X_1	X_2	X_3	Y_1	Y_2	Y_3
—1	—1	1	85,3	72,7	97,1
1	1	—1	72,3	57,6	96,9
0	1,2154	0	71,4	56,5	96,4
0	—1,2154	0	72,0	64,6	96,8
—1	—1	—1	87,0	79,2	97,0
1	1	1	55,6	32,6	96,2
0	0	—1,2154	85,0	75,9	97,2
1,2154	0	0	70,9	53,4	97,9
0	0	0	75,9	59,3	97,4
1	—1	1	76,1	63,2	97,4
—1	1	—1	85,0	75,3	97,2
0	0	1,2154	68,0	57,2	95,5
—1,2154	0	0	89,6	83,6	97,2
—1	1	1	75,0	61,5	96,5
1	—1	—1	74,2	61,0	98,2

10. Исследовательский отдел одной крупной корпорации разработал новое хлебобулочное изделие. Первейшей заботой было получение максимального припека от теста, подготовленного к выпечке из стандартных ингредиентов. Было признано, что на величину припека влияют четыре главные переменные: процент жира, процент воды, высота подъема теста и скорость вращения миксера в об/мин. Были проделаны эксперименты, указанные в таблице. В результате найдены величины припеков в каждом опыте (они показаны в самой таблице). Заметим, что в опыте с координатами (12, 50, 20, 130) были поставлены четыре параллельных эксперимента, они дали результаты: 492, 523, 530 и 590.

Эксперимеитальные данные по расширению области максимума припека

<table>
<tr><th rowspan="4">Процент жира</th><th rowspan="4">Подъем теста</th><th colspan="9">Процент воды</th></tr>
<tr><th colspan="3">46</th><th colspan="3">50</th><th colspan="3">54</th></tr>
<tr><th colspan="3">об/мин</th><th colspan="3">об/мин</th><th colspan="3">об/мин</th></tr>
<tr><th>90</th><th>130</th><th>170</th><th>90</th><th>130</th><th>170</th><th>90</th><th>130</th><th>170</th></tr>
<tr><td rowspan="3">8</td><td>10</td><td>833</td><td></td><td>540</td><td rowspan="3" colspan="3">537</td><td>673</td><td></td><td>493</td></tr>
<tr><td>20</td><td colspan="3"></td><td colspan="3"></td></tr>
<tr><td>30</td><td>577</td><td></td><td>547</td><td>660</td><td></td><td>512</td></tr>
<tr><td rowspan="3">12</td><td>10</td><td rowspan="3" colspan="3">653</td><td colspan="3">547</td><td rowspan="3" colspan="3">487</td></tr>
<tr><td>20</td><td>650</td><td>492
523
530
590</td><td>553</td></tr>
<tr><td>30</td><td colspan="3">595</td></tr>
<tr><td rowspan="3">16</td><td>10</td><td>802</td><td></td><td>477</td><td rowspan="3" colspan="3">575</td><td>710</td><td></td><td>520</td></tr>
<tr><td>20</td><td colspan="3"></td><td colspan="3"></td></tr>
<tr><td>30</td><td>568</td><td></td><td>401</td><td>572</td><td></td><td>483</td></tr>
</table>

1) Выбрав подходящие значения центрального (нулевого) уровня и интервалы варьирования, закодируйте все четыре предиктора так, чтобы их уровнями стали (— 1, 0,1). Запишите план в кодированных переменных и убедитесь, что это действительно план типа «куб плюс звезда плюс четыре центральные точки». Ротатабелен ли он?

2) Используя методы множественной регрессии, постройте подходящую модель первого или второго порядка для предсказания максимума припека. В ваших выводах укажите относительную важность предикторов и сделайте другие замечания, какие найдете уместными[12].

11. Постройте модель второго порядка и проделайте полный анализ, используя отклик Y_4 и данные из табл. 7.6.

Ответы к упражнениям

1. $\hat{Y} = 0{,}4368012 + 0{,}0001139X_1 - 0{,}0051897X_3 - 0{,}0018887X_4 + 0{,}0044263X_5$.
График остатков в зависимости от $\hat{Y}$ показывает, что дисперсия неоднородна. Стоит попробовать взвешенный метод наименьших квадратов или, быть может, преобразовать Y_i.

Эта модель объясняет только 76,9 % общего разброса, а доверительные пределы для $\beta_{1Y\cdot345}$ и $\beta_{4Y\cdot135}$ включают нуль. Стандартное отклонение остатков составляет 3,3 % от среднего отклика. Следовательно, моделью все-таки можно пользоваться для предсказания, хотя она и не так хороша, как нам бы хотелось. Если бы удалось избавиться от большой дисперсии для больших значений отклика, то модель стала бы гораздо лучше.

2. Модель $Y = \alpha X_1^{\beta} X_2^{\gamma} X_3^{\delta} \cdot \varepsilon$.

Логарифмированием по основанию e можно преобразовать модель в линейную форму:

$$\ln Y = \ln \alpha + \beta \ln X_1 + \gamma \ln X_2 + \delta \ln X_3 + \ln \varepsilon,$$

или

$$\widehat{\ln Y} = 8{,}5495297 + 0{,}1684244 \ln X_1 - 0{,}537137 \ln X_2 - 0{,}0144135 \ln X_3.$$

Переменная X_2 — окружная скорость.

Как показывает F-отношение, переменную X_3 — вязкость на входе — надо исключить:

$$2{,}15 < F(1;\ 31;\ 0{,}95) = 4{,}16.$$

То, что вариация объясняется на 96,52 % и что стандартное отклонение от среднего значения отклика, равное 1,563 %, мало, говорит о том, что получено хорошее предсказывающее уравнение. Графики остатков не выявляют никаких особенностей.

3. Выбранная модель имеет следующий вид:

$$\hat{Y} = 120{,}627 + 490{,}412X_2 - 5{,}716X_3 - 1107{,}847X_2^2.$$

График остатков выявляет опыты со знаками + и —, указывая на наличие нерассмотренных X-переменных.

---

[12] Речь идет о дрожжевом тесте. Сама постановка вопроса представляется сомнительной. С одной стороны, ясно, что за долгие столетия человечество методом «проб и ошибок», видимо, отыскало оптимальный вариант выпечки. С другой стороны, припек — «слишком» экономический отклик. Вряд ли это достаточная характеристика качества для изделий такого рода. Да и выбранный план загадочен. Впрочем, для исследователя эта задача, несомненно, очень интересна. Решения, к сожалению, авторы не приводят.— *Примеч. пер.*

Добавление члена второго порядка от X_2 и X_3 окажет лишь самую незначительную помощь.

Уравнение имеет $R^2 = 90{,}27\ \%$ со стандартным отклонением 6,2233.

Андерсон и Банкрофт получили такую модель:

$$\hat{Y} - 84{,}204 = 2{,}463\,(X_1 - 1{,}86) - 75{,}369\,(X_2 - 0{,}188) + 1{,}584\,(X_3 - 7{,}64) -$$

$$- 1{,}380\,(X_1X_2 - 0{,}3507).$$

Эта модель подобрана не так хорошо. Остатки имеют определенную структуру, а $R^2 = 75{,}49\ \%$.

Предупреждение: Пример из Андерсона и Банкрофта был использован для иллюстрации регрессионных расчетов, ведущихся без намерения построить наилучшую модель.

4. Модель: $\hat{Y} = b_0 + b_2X_2 + b_8X_8$,

или

$$\hat{Y} = 9{,}4742224 + 0{,}7616482X_2 - 0{,}0797608X_8, \qquad R^2 = 86{,}0\%.$$

Стандартное отклонение в процентах от среднего отклика равно 6,761 %.

5. 1) $\hat{Y} = 87{,}158859 + 0{,}8519104X_1 + 0{,}5988662X_2 + 2{,}3613018X_6 -$

$- 0{,}9755309X_9$,

где X_1 — год,

X_2 — колнчество осадков, выпавших в предшествующем сезоне, в дюймах,

X_6 — осадки в июле в дюймах,

X_9 — температура в августе.

2) Наиболее важной переменной, учитывающей тенденцию к росту урожая зерна, служит X_1. Из всех остальных переменных только количество осадков в предыдущем сезоне, осадки в июле и температура в августе вносят существенный вклад в регрессию.

3) Это уравнение с $R^2 = 72{,}06\ \%$ и стандартным отклонением от среднего отклика 14,903 % нуждается в улучшении. Было бы хорошо найти новые переменные, которые увеличили бы R^2 и уменьшили стандартное отклонение остатков. Исследование остатков может помочь в решении этой задачи.

6. 1) Имеется множество параллельных опытов. Следовательно, можно получить независимую оценку «чистой» ошибки. Дисперсионный анализ можно записать в следующем виде:

ANOVA

Источник рассеяния	Степени свободы
Общий	47
Регрессия	5
Остатки	42
Неадекватность	2
«Чистая» ошибка	40

2) $\hat{Y} = 134{,}258 + 0{,}050X_1 - 0{,}012X_2 + 0{,}834X_3 - 0{,}154X_4 - 3{,}804X_5$.

3) Модель неадекватна, так как критерий неадекватности значим при $\alpha = 0{,}05$.

ANOVA

Источник рассеяния	Степени свободы	SS	MS	F
Общий	47	2850,3107		
Регрессия	5	1817,1055		
Остаток	42	1033,2052		
Неадекватность	2	383,7052	191,8526	11,82 *
«Чистая» ошибка	40	649,5000	16,2375	

4) Модель объясняет только 63,75 % разброса, что нельзя считать достаточно хорошим вариантом. Остатки образуют определенную неслучайную картину.

5) Этот эксперимент плохо спланирован, здесь слишком много повторений и недостаточно различных точек плана.

7. Предсказывающее уравнение, полученное с помощью шаговой процедуры с использованием критического значения $F = 2{,}00$ при включении и исключении, имеет следующий вид:

$$\hat{Y} = 250{,}1875 - 2{,}3124998\left(\frac{X_1 - 146}{3}\right) -$$

$$- 14{,}687499\left(\frac{X_2 - 69{,}5}{3{,}5}\right) - 2{,}8124997\left(\frac{X_6 - 289{,}5}{93{,}5}\right).$$

Оптимальная скорость на основании предсказывающего уравнения будет в точке $\hat{Y} = 270$, $X_1 = 143$, $X_2 = 66$, $X_6 = 196$, причем другие переменные поддерживаются на средних уровнях, а именно: $X_3 = -10$, $X_4 = 132{,}5$, $X_5 = 91{,}5$.

8. $\hat{Y}_1 = -2{,}80512 + 0{,}15176X_1 + 3{,}60191X_3$, $\hat{Y}_2 = -2{,}84492 + 0{,}11344X_1 + 3{,}67343X_3$.

9. 10. Решение не приводится.

11. 1) Вот подобранная модель:

$$\hat{Y} = 6{,}087 - 0{,}240X_1 + 0{,}340X_2 - 0{,}495X_3 - 0{,}036X_1^2 + 0{,}021X_2^2 + 0{,}118X_3^2 + 0{,}040X_1X_2 - 0{,}070X_1X_3 + 0{,}045X_2X_3.$$

2) Таблица дисперсионного анализа имеет такой вид:

ANOVA

Источник	Степени свободы	SS	MS	*F*
Среднее (b_0)	1	758,173		
Первый порядок	3	5,671	1,890	
Второй порядок	6	0,299	0,050	
Неадекватность	5	0,839	0,168	2,37
«Чистая» ошибка	5	0,354	0,071	
Общий	20	765,336		

3) Неадекватность не проявилась и члены второго порядка не вносит заметного вклада в объясняемую долю общего разброса данных, поскольку средний квадрат, обусловленный членами второго порядка, сравним с остаточным средним квадратом $s^2 = (0{,}839 + 0{,}354)/(5 + 5) = 0{,}119$ при 10 степенях свободы.

4) Если подобрать по данным упрощенную модель первого порядка, то она будет иметь вид

$$\hat{Y} = 6{,}157 - 0{,}240X_1 + 0{,}340X_2 - 0{,}495X_3,$$

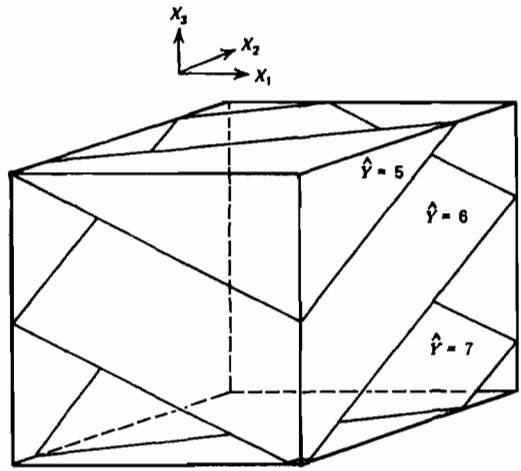


Рисунок к решению упражнения 11 (1)

причем значимая неадекватность не обнаруживается, а новая оценка для σ^2 $s^2 = 0{,}093$ с семнадцатью степенями свободы.

Примечание. Предсказанные (плоские) контуры имеют вид, показанный на рисунке к решению упражнения 11 (1). Изображенный на этом рисунке куб ограничен плоскостями (± 2, ± 2, ± 2).

Иной способ представления тех же данных приведен на рисунке к решению упражнения 11 (2). На трех частях этого рисунка проведены прямолинейные контуры для X_1 и X_2 в трех плоскостях, соответствующих значениям $X_3 = -1, 0$ и 1. Такие контуры надо представлять себе как непрерывные и расположенные между тремя указанными плоскостями. Точки плана, попавшие на одну из плоскостей, изображены кружочками. Две звездные точки, которые не лежат ни в одной из этих плоскостей, на рисунке не показаны. Заметим, что наши контуры покрывают меньшую область, чем отмеченная на предыдущем рисунке.

Наборы таких двумерных сечений весьма полезны при изучении контуров поверхности второго порядка. Когда же число предикторов больше трех, более полезным становится метод канонических преобразований.

Исследование остатков для этой модели мы оставляем читателю.

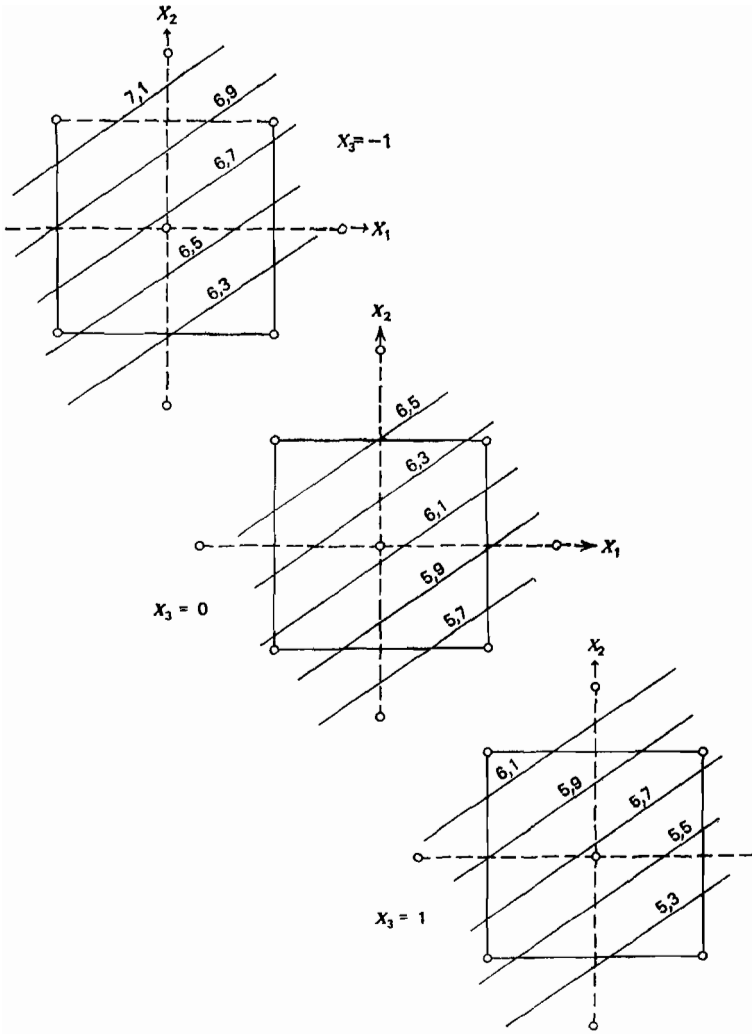


Рисунок к решению упражнения 11 (2)

Глава 8 ● МНОЖЕСТВЕННАЯ РЕГРЕССИЯ И ПОСТРОЕНИЕ МАТЕМАТИЧЕСКОЙ МОДЕЛИ

8.0. ВВЕДЕНИЕ

Методы множественной линейной регрессии, которые мы обсуждаем, могут быть очень полезными, но также и очень опасными, если они неверно используются или интерпретируются. Прежде чем приступать к большой задаче с применением методов множественной регрессии, имеет смысл, насколько это возможно, предварительно спланировать всю работу применительно к конкретной цели и наметить контрольные мероприятия, проводимые по ходу дела. Такое планирование будет предметом данной главы. Прежде, однако, мы обсудим три основных типа математических моделей, часто используемые в науке:

1. Функциональная модель.
2. Модель для управления.
3. Модель для предсказания.

ФУНКЦИОНАЛЬНАЯ МОДЕЛЬ

Если в некоторой задаче известна «истинная» функциональная связь между откликом и предикторами, то экспериментатор в силах понять и предсказать отклик, да и управлять им [1]. Однако в жизни редко встречаются ситуации, когда можно предложить подобную модель. Но даже и в этих случаях функциональные уравнения обычно очень сложны, трудны для понимания и применения и имеют чаще всего нелинейный вид. В наиболее сложных случаях может потребоваться численное интегрирование таких уравнений. Примеры нелинейных моделей упоминались в гл. 5, а их построение будет обсуждаться в гл. 10. Для таких моделей линейные регрессионные методы неприменимы или применимы только для аппроксимации истинных моделей в итеративных процедурах оценивания.

Модель для управления

Функциональная модель, даже если она известна полностью, не всегда пригодна для управления выходной переменной (откликом). Например, в задаче про пар, используемый на заводе, одна из наиболее важных переменных — наружная температура, а она неуправ-

[1] В советской литературе функциональные модели часто называют «детерминированными». Ведется спорадическая дискуссия о месте и роли такой модели, о ее взаимоотношениях с вероятностными (стохастическими) моделями. Дискуссия, впрочем, не слишком актуальна, ибо функциональные модели — большая редкость, хоть и весьма желанная.— *Примеч. пер.*

ляема в том смысле, в каком управляемы температура, давление и другие факторы процесса. Или еще: специалист по рекламе, желающий оценить влияние на продажу товаров его фирмы рекламных передач по телевидению, вполне сознает важность для любой функциональной модели сбыта такого фактора, как активность конкурентов. Однако эту активность определяют неуправляемые переменные, независимо от того, как они входят в функциональную модель. Для управления откликом нужна такая модель, которая содержит факторы, подконтрольные экспериментатору.

Модель, полезную для управления, можно построить методами множественной регрессии, если они применяются корректно. Когда возможно использование планирования экспериментов для управляемых переменных, влияние этих переменных на отклик при применении множественной регрессии можно найти подобно тому, как рассматривается в гл. 9. Однако есть ситуации, в которых планирование эксперимента невозможно. Например, условия эксперимента, проводимого на действующей промышленной установке, обычно нарушаются изо дня в день, и если не компенсировать достаточно хорошо возможные в результате этого изменения в измеряемом отклике, то такой эксперимент не будет показателен. Еще пример: эксперимент, проводимый на рынке, может быть спланирован и обработан, но неуправляемые факторы (хотя и известные) будут делать любые вычисленные математические эффекты управляемых факторов настолько запутанными, что результаты станут бесполезными. Эти ситуации вынуждают практиков применять модели для предсказания.

Модели для предсказания

Когда функциональная модель очень сложна и когда возможности для получения независимых оценок эффектов ограничены, часто удается построить линейную предсказывающую модель, которая хотя в некотором смысле и нереалистична, но по крайней мере воспроизводит основные черты поведения изучаемого отклика. Такая предсказывающая модель весьма полезна и при определенных условиях может вести к реальному проникновению в процесс или проблему. При построении предсказывающих моделей такого типа методы множественной регрессии оказываются наиболее ценными. Эти задачи обычно упоминаются как «задачи с неупорядоченными данными», т. е. данными, среди которых много коррелированных между собой. Модель для предсказания не обязательно функциональна и не обязательно полезна для управления, что вопреки мнению некоторых ученых вовсе не делает ее бесполезной [2]. Опираясь на нее, если нет

[2] Существует множество исследований, направленных на выявление дефектов регрессионного анализа, особенно в связи с его противопоставлением применению той же регрессии, но в ситуации активного, спланированного эксперимента. См., например: Налимов В. В., Чернова Н. А. Статистические методы планирования экстремальных экспериментов.— М.: Наука, 1965.— 340 с. (особо гл. 1, там есть и другие ссылки). Ясно, однако, что никакая критика не отменит данных пассивного эксперимента, которые тоже надо как-то обрабатывать.— *Примеч. пер.*

ничего лучшего, можно выбрать и линию поведения для дальнейшего экспериментирования, уточнив важные переменные, и, что очень полезно, отсеять несущественные переменные.

Вместе с тем применение множественной регрессии требует особой осторожности, чтобы избежать непонимания и неверных выводов. Организация схемы для решения задач с помощью методов множественного регрессионного анализа не только полезна, но и необходима.

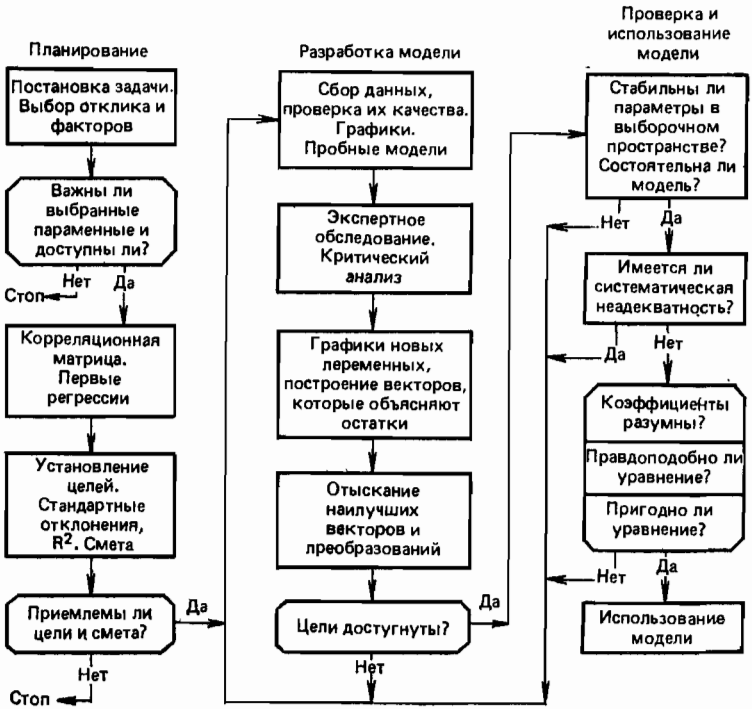


Рис. 8.1. Блок-схема процедуры построения модели

Эта глава — только план, а любое использование предложенной или подобной схемы будет требовать специальной «настройки» на конкретную ситуацию.

Хотя приведенный ниже план предназначен для разработки предсказывающей математической модели, он является достаточно общим; им можно воспользоваться при построении как функциональных, так и управляющих моделей. Особое внимание обратим на задачи с «неуправляемыми данными». Схема делится на три стадии — планирование, разработку и использование. Блок-схема приведена на рис. 8.1, и в дальнейшем она будет детально обсуждена.

8.1. ПЛАНИРОВАНИЕ ПРОЦЕССА ПОСТРОЕНИЯ МОДЕЛИ

Постановка задачи, выбор отклика и предполагаемых факторов

Конкретная постановка — наиболее важная фаза в процедуре решения любой задачи [3]. Важно, чтобы инженер, ученый и бизнесмен умели точно сформулировать условия своей задачи. Например, формулировки наподобие «Почему покупатель покупает мой продукт?» или «Почему сегодня линия № 5 работает не очень хорошо?» интересны, но недостаточно конкретны для принятия какого-либо решения. Постановка задачи должна быть очень четкой, и нужно точно установить как отклик, так и предикторы. В начале этой фазы планирования исследователь не должен связывать себя жесткими ограничениями; он может записать любые мыслимые факторы и отклики, которые, как он предполагает, оказывают какое-либо влияние на задачу. Список может получиться длинным, но в результате обсуждения он будет последовательно сведен к разумному числу переменных. Важно помнить, что *в любой статистической процедуре отсев факторов никогда не осуществляется однозначно*, в том числе и в процедурах множественной регрессии, описанных в гл. 6. В конце концов достигается постановка конкретной задачи с конкретным откликом или откликами, которые предстоит исследовать в связи с конкретным множеством потенциальных предикторов.

Действительно ли выбраны основные для данной задачи переменные и доступны ли они?

Полученный при формулировке задачи список факторов следует подвергнуть тщательному исследованию. Многие из этих предикторов можно исключить как неизмеримые, например температура капли в процессе может рассматриваться как важная переменная, но в настоящее время ее нельзя измерить. Такой фактор либо заменяют другим, который измерим и может использоваться вместо температуры капли, либо же находят новый измерительный инструмент. Вторая альтернатива требует затрат, и исследователю предстоит определить, какая из двух альтернатив предпочтительнее. Такая научная и практическая оценка всех переменных должна быть сделана именно на этой стадии планирования, т. е. до того как будет собрана основная масса данных.

Следующий вопрос таков: в состоянии ли мы получить полное множество фактических данных одновременно для всех выбранных предикторов и откликов? Будет ли наше множество данных полным? Есть много случаев, когда этого не удается достичь и приходится искать какие-то компромиссы. Вот одна из типичных ситуаций: все данные можно собрать одновременно, но измерения откликов нуж-

[3] Здесь авторы приступают к краткому описанию проблем, которые мы относим к области так называемого «предпланирования» эксперимента — одного из ключевых моментов в решении любой задачи. См., например: Адлер Ю. П. Предпланирование эксперимента.— М.: Знание, 1978.— 63 с.— *Примеч. пер.*

даются в дополнительной математической обработке или получаются после дополнительных лабораторных анализов. Из-за загруженности лаборатории текущей работой может оказаться, что пройдет несколько недель, прежде чем мы сможем, наконец, получить ожидаемые результаты. Надо ли дожидаться этих анализов? Не стоит ли отбросить мысль о том, чтобы собирать такие большие массивы данных? Вопросы такого рода было бы очень полезно обсудить прежде, чем приступать к сбору данных, а временной график работы должен быть составлен заранее самым тщательным образом. После полной проверки всех переменных надо провести переоценку возможностей решения задачи.

Разрешима ли задача в принципе?

Так как изложенная выше процедура отсеивания делает мыслимым исключение многих факторов, то уменьшаются шансы решить задачу вообще. Однако этого, как правило, не происходит. На данной стадии планирования должно быть принято одно из трех возможных решений:

1. Первоначальный замысел следует отбросить.
2. Замысел следует пересмотреть в свете новых знаний, полученных к этому моменту.
3. Замысел представляется реальным и планирование следует продолжить.

Корреляционная матрица и первые прогоны регрессии

Если работе дано «добро», то к ее планированию можно теперь приступить на основе экспериментальных данных, и именно теперь можно проанализировать все трудности, возникающие в связи с поставленной задачей, поскольку надо составить план-график выполнения работы с учетом трудоемкости, бюджета и т. д.

Если это возможно, то следовало бы получить выборку данных, вычислить и распечатать для нее описательные статистики, корреляционную матрицу и матрицу, обратную к корреляционной. Диагональные элементы обратной корреляционной матрицы переменных X — это так называемые *«инфляционные» множители для дисперсии* (ИМД). При их обсуждении в работе, где рассматриваются обобщенные обращения, ридж-регрессия, смещенное линейное оценивание и нелинейное оценивание Д. Маркуардт (M a r q u a r d t D. W. Generalized inverses, ridge regression, biased linear estimation and nonlinear estimation.— Technometrics, 1970, **12**, p. 591—612) указывает (на с. 610) как на желательное, что было бы хорошо иметь ИМД «больше чем 1,0, но, безусловно, не больше чем 10». Если любой из ИМД >10, то соответствующий коэффициент, найденный методом наименьших квадратов, будет скорее всего оцениваться так плохо, что это может послужить указанием на желательность некоторой модификации модели. В этом случае исходные предикторы X коррелируют столь сильно, что с такими данными процесс построения модели не пойдет «как по маслу».

Следующее, что важно сделать — установить для каждого из рассматриваемых откликов корреляции с ним каждой из X-переменных. При этом для каждого отклика хорошо было бы обнаружить одну-две большие корреляции. Если же их нет, то придется вновь проанализировать всю ситуацию. Может случиться так, что диапазон изменения X-переменных окажется слишком мал. Помните, что пока пространство X не будет «достаточно большим», получить хорошее предсказание будет скорее всего нелегко. Подобные отсечения и ветвления на столь ранней стадии планирования позволяют экспериментатору получить разумные оценки ожидаемых затрат времени и средств, а также представление о шансах на создание отличной модели для предсказания.

Установление целей и составление сметы расходов

На этой стадии исследователь и статистик должны установить цели предпринимаемой работы, составить план-график решения конкретных задач и подготовить задания сотрудникам и компьютеру. В плане-графике предусматриваются контрольные точки, после чего вся предварительная работа предъявляется к приемке и утверждению. Ниже приведен пример простых наметок, какие могут представляться на утверждение.

Образец формуляра для намечаемой работы

НАЗВАНИЕ РАБОТЫ: Оценка уравнения для расхода пара на заводе А.

ИНЖЕНЕР-ИСПОЛНИТЕЛЬ: Джо ДОУ.
ПРЕДПОЛАГАЕМЫЙ МЕТОД ИССЛЕДОВАНИЯ: Множественная регрессия.

ЦЕЛИ РАБОТЫ:

1. Окончательное уравнение должно объяснять более 80 % вариации ($R^2 > 0{,}8$).
2. Стандартное отклонение оценки должно составлять менее 5 % от среднего значения количества используемого пара.
3. Число предикторов должно быть _____*[1].
4. Все оценки коэффициентов окончательного уравнения должны быть статистически значимы при $\alpha = 0{,}05$.

---

*[1] Мы полагаем, что, как правило, должно быть порядка десяти полных наборов наблюдений для каждой переменной, которая претендует на место в нашей модели. Это означает, что если мы надеемся, что в окончательной модели будет четыре X-переменных плюс свободный член, то должно быть по меньшей мере сорок наборов наблюдений ($n = 40$).

5. В остатках не должно быть заметных связей.

Смета расходов	Всего долларов
Рабочая сила: 2 человека в месяц	4000
Компьютер: 6 часов	1800
Итого	5800

ПЛАН-ГРАФИК:

1. Сбор данных, прикидочные расчеты и предварительный анализ — 2 недели
2. Модификация и усовершенствование — 4 недели
3. Составление отчета и завершение работы — 2 недели

Приемлемы ли цели и бюджет?

Если работа принята во всем объеме, то она вступает в стадию разработки, а если нет, то осуществляется пересмотр, направленный на уменьшение объема затрат, или же работа просто приостанавливается.

8.2. РАЗРАБОТКА МАТЕМАТИЧЕСКОЙ МОДЕЛИ

Теперь все готово для разработки предсказывающего уравнения. Методы, используемые при построении регрессионной модели, обсуждались в гл. 1—7. В этом параграфе содержатся только упоминания о различных контрольных точках в ходе решения. Возможны многочисленные вариации

Сбор данных, проверка их качества, построение графиков, пробные модели

Тщательное управление процессом сбора данных гарантирует удовлетворение всем ограничениям, введенным на разных этапах планирования. Так, например, если требовалось собрать данные в течение одного дня, то не надо добирать на следующий день то, что не успели собрать накануне. (А такое встречается очень часто) Числа проверяйте настолько тщательно, насколько это только возможно, поскольку запятая, отделяющая целую часть числа от дробной, всегда оказывается не на месте; вообще качество «неупорядоченных» данных часто удручающе неудовлетворительно. Не приступайте к построению модели, не обеспечив должного качества данных! В обследованиях людей очень часто можно встретить мужчин ростом 14 дюймов, женщин весом 1000 фунтов, студентов, у которых «нет» легких, и т. п. Не удивляйтесь, если в данных о работе предприятия вы встретите

10 000 фунтов материала в бочке вместимостью 100 фунтов и многие-многие другие подобные ошибки. Как бы ни планировать и ни обучать персонал, ошибки такого рода устранить нельзя. Существуют «человеческие» ошибки, не свойственные компьютеру, хотя и компьютер часто тоже заслуживает порицания. Наш опыт работы с «неупорядоченными» данными в течение десятилетий свидетельствует: мы почти никогда не встречали большие массивы данных, которые были бы совершенно свободны от подобных качественных недостатков.

После того как соблюдены все предосторожности, можно приступать к самому процессу моделирования. Теперь очередь за такими процедурами, как нанесение данных на графики, подбор модели, исследование остатков, и теми аналитическими подходами, которыми заполнены гл. 1—7.

Экспертный совет для критического разбора

По мере продвижения работ потенциальные модели следует обсуждать для их оценки и анализа со специалистами в данной области.

Графики для новых факторов и построение векторов, которые объясняют остатки

Если (как это обычно бывает) эксперты предлагают для опробования новые факторы, то надо получить требующиеся данные и исследовать их вместе с соответствующими значениями отклика и остатками для полученного уравнения.

Исследование некоторых построенных уравнений регрессии

Выбранные предикторы теперь можно включить в уравнение регрессии. В нем могут участвовать и преобразованные предикторы. Например, по конкретным графикам для остатков можно выяснить, что лучше применить логарифм предиктора, чем сам предиктор. Многие преобразования предлагаются самими исследователями, хорошо знающими изучаемый процесс. Наш опыт говорит, что принцип «давайте пробовать все подряд» редко работает хорошо при преобразованиях исходных входных факторов. Чем лучше мы понимаем суть задачи и применяем это понимание для выбора преобразований X-переменных, тем больше шансов ускорить продвижение к ее решению.

Цели достигнуты?

В итоге, иногда после нескольких попыток, исследователь обнаруживает, что он получил наилучшую модель и должен изучить ее в свете целей, сформулированных на стадии планирования. Если заданные стандарты не достигнуты, то решается вопрос о том, следует ли приостановить реализацию работы или повторить весь цикл. Возможно, потребуется больше денег, а быть может, удастся изменить исходные цели. Это одна из контрольных точек в работе.

8.3. ПРОВЕРКА И ИСПОЛЬЗОВАНИЕ МАТЕМАТИЧЕСКОЙ МОДЕЛИ

После того как уравнение удовлетворит множеству целей, выбранных на стадии планирования, и модель будет признана полезной для предсказания, целесообразно определить процедуры для ее проверки и использования.

Стабильны ли параметры в выборочном пространстве?

При изучении стабильности параметров нам будет полезно различать два вида массивов данных: это данные, собранные в течение длительного отрезка времени («продольные» данные), и данные, собранные за короткое время («поперечные» данные, «мгновенная фотография» объекта).

Данные за длительное время. Если модель была построена по наблюдениям, проведенным в течение длительного времени, то можно проверить устойчивость b-коэффициентов, строя модели для более коротких отрезков времени и сравнивая оценки. Если, например, имеются месячные данные за четыре последовательных года, то можно построить модели за каждый год отдельно и получить четыре множества оценок коэффициентов регрессии. Если оцениваемые коэффициенты проявят определенную тенденцию, то использование уравнения, построенного по всем данным для целей предсказания, будет неразумным.

Данные, полученные за короткое время. Если данные можно рассматривать как информацию, собранную в «одно мгновение», скажем, за одну рабочую смену или на одной партии сырья, то для такого случая есть несколько методов, описанных, например, в работах Мостеллера и Тьюки (Mosteller and Tukey, 1968), Аллена (Allen, 1971), Стоуна (Stone, 1974), Джиссера (Geisser, 1975), Маккарти (McCarthy, 1976) и Сни (Snee, 1977),— все они указаны в библиографии. В области социальных наук есть такие работы, как Финифтер (Finifter, 1972), Новик и др. (Novick et al., 1972) и Киш и Френкель (Kish, Frankel, 1974). Основная идея этих работ заключается в том, чтобы сначала на основе некоторого рационального критерия или критериев разделить имеющийся массив данных на подмножества, а затем использовать одну часть данных для построения «предсказывающего» уравнения, а оставшуюся часть — для «проверки» («экзамена») этого уравнения, т. е. для того, чтобы посмотреть, насколько хорошо оно предсказывает! Как и при выборе переменных, здесь *нет* единственного или наилучшего ответа на все вопросы. Зато есть несколько подходов, среди которых читатель может выбирать.

1. *Подход «выбрасывать по одному наблюдению».* Метод PRESS, предложенный Алленом (1971) и обсуждавшийся в § 6.8, представляет собой процедуру проверки такого типа. После выбрасывания какого-нибудь одного наблюдения, строим заданную модель для тех наблюдений, что остались, предсказываем с ее помощью отброшенное значение и получаем квадрат расхождения между фактическим и предсказанным. Повторяем эту операцию, последовательно отбрасывая каждую из экспериментальных точек; в итоге получим сумму

квадратов расхождений для некоторой данной модели. Исследуя, таким образом, различные модели, как описано в § 6.8, найдем некоторую «наилучшую» модель, которая окажется наиболее «жизнеспособной» при фиксированном наборе данных. Значит, индивидуальные расхождения можно было бы исследовать для отыскания несостоятельных данных. Правда, прежде надо было бы изучить свойства оценок β в таком процессе.

2. *Подход «выбрасывать более, чем по одному наблюдению».* Джиссер (1975) рассматривал метод, аналогичный методу Аллена, в котором исходная идея обобщалась на случай отбрасывания m наблюдений и использования остальных $n—m$ наблюдений для построения модели и испытания ее на «m» отброшенных точках. Стоун (1974) обсуждал методы такого типа и попытался найти «оптимальное» разбиение на подвыборки.

3. *Идеи «делить пополам».* Использование половины данных для построения модели и второй половины для ее проверки — вот что делалось на протяжении многих лет. Сни (1977) рассматривает проблему выбора половины данных для построения модели. Алгоритм ДУПЛЕКС (DUPLEX), который он рассматривает, это прежде всего некий метод, связанный с условием, что свойства определителя матрицы $\mathbf{X}'\mathbf{X}$ должны быть аналогичны для той половины данных, по которым построена модель, и для той, которая пошла на проверку. Использованное Сни правило включает вычисление корня k-й степени из отношения двух определителей, а именно

$$\left\{\frac{|\mathbf{X}'\mathbf{X}| \text{ для оценивающего множества}}{|\mathbf{X}'\mathbf{X}| \text{ для предсказывающего множества}}\right\}^{1/k},$$

где все исходные X-переменные стандартизированы и ортогонализированы, так что определитель $|\mathbf{X}'\mathbf{X}|$ представлен в корреляционной форме, а k — число X-переменных в матрице $\mathbf{X}$. Если исходный массив данных разделить правильно, то эта статистика должна быть приблизительно равна 1. Более того, Сни (1977) рекомендует не делить данные пополам, если не выполняется неравенство n (полный объем выборки) $> 2p + 25$, где p — число параметров, получившееся в окончательной модели. Алгоритм ДУПЛЕКС был предложен Кеннардом (R. W. Kennard), одним из пионеров метода ридж-регрессии,— см. § 6.7.

Проведение какой-либо из указанных проверок — полезная и необходимая часть полного метода множественного регрессионного анализа. Иногда информация, полученная таким путем, может привести к полному пересмотру всей задачи [4].

---

[4] С момента, когда авторы завершили переработку своей книги для второго издания, методы, кратко описанные в этом разделе, претерпели бурное развитие и превратились в настоящий момент (1985 г.) в самостоятельное направление, включающее как упомянутые авторами методы «перепроверки» (cross-validation) и «складного ножа» (jack knife), так и новые подходы, главным среди которых стал предложенный Б. Эфроном в 1979 г. метод «бутстреп» (boot strap). Метод «складного ножа» был предложен Кенуем в 1949 г., усовершенствован Тьюки в 1958 г., обобщен Грейем и Шукаии в 1972 г. и систематизирован Миллером в 1974 г.

Имеется ли систематическая неадекватность?

Даже если параметры построенного уравнения оказываются очень стабильными, некоторые факторы все-таки могут быть пропущены. Всегда нужно исследовать остатки всеми возможными способами, чтобы выявить какие-либо признаки, указывающие на наличие подобных пропусков.

Практика рассмотрения моделей

Приемлемы ли коэффициенты? Этот вопрос может показаться необычным, но следует помнить, что моделью будут пользоваться и те, кто не подозревает о том, что регрессионные МНК-коэффициенты зависят от остальных факторов, входящих в регрессию. Поэтому могут иметь место попытки предсказать отклик, меняя только один фактор и используя соответствующий ему коэффициент, который теперь якобы хорошо заменяет остальные. Если все коэффициенты оцениваются независимо, то вред не будет большим. Однако, когда независимые переменные сильно коррелированы и оцениваемые коэффициенты тоже сильно коррелированы, доверять индивидуальным коэффициентам опасно. Разумно ограничить предсказание областью пространства X, в которой получены исходные данные; полезно также проверить, будут ли индивидуальные коэффициенты безусловно кор-

---

Вот основные работы: 1. Quenouille M. Approximate tests of correlation in time series.— J. Royal Statist. Soc., Ser. B, 1949. **11**, p. 18—84; 2. Quenouille M. Notes on bias in estimation.— Biometrika, 1956, **43**, p. 353—360; 3. Tukey J. Bias and confidene in not quite large samples, abstract.— AMS, 1958, **29**, p. 614; 4. Gray H. L., Shucany W. R. The generalized jackknife statistic.— New York: Marcel Dekker Inc., 1972; 5. Miller R. G. The jackknife — a review.— Biometrika, 1974, **61**, p. 1—17; 6. Miller R. G. An unbalanced jackknife.— Ann. Statist., 1974, **2**, p. 880—891. На русском языке описание этой процедуры и некоторых близких к ней можно найти в работах: Клейнен Дж. Статистические методы в имитационном моделировании/Пер. с англ. Под ред. Ю. П. Адлера, В. Н. Варыгина.— М.: Статистика, 1978, вып. 1 — 221 с.; вып. 2— 335 с. (особо 1-й, с. 178—180) и более подробно: Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия/Пер. с англ. Под ред. Ю. П. Адлера.— М.: Финансы и статистика, 1982, вып. 1, гл. 8, с. 143—170, а также в книге: Поиск зависимостн и оценка погрешности/Под ред. И. Ш. Пинскера.— М.: Наука, 1985.— 148 с.

Метод «бутстреп» был предложен в работе: Efron B. Bootstrap methods: another look at the jackknife.— Ann. Statist., 1979, 7, p. 1—26. Более обстоятельно он описан в книге: Efron B. The jackknife, the bootstrap and other resampling planes.— Philadelphia. Pa.: SIAM, 1982.— p. 92 (в издательстве «Финансы и статистика» готовится русский перевод этой книги и еще нескольких работ ее автора). Есть русский перевод одной из популярных работ на эту тему: Диаконис П., Эфрон Б. Статистические методы с интенсивным использованием ЭВМ.— В мире науки, 1983, № 7, с. 60—73. Еще см.: Адлер Ю. П. Предисловие в кн.: Иванова В. М. Случайные числа и их применение.— М.: Финансы и статистика, 1984.— 111 с. (особо с. 3—10). Появление ЭВМ с новыми возможностями (см.: ЭВМ пятого поколения/Пер. с англ. Под ред. А. А. Рывкина, В. М. Савинкова.— М.: Финансы и статистика, 1984.— 110 с.) гарантирует быстрое развитие подобных методов в ближайшие годы.— *Примеч. пер.*

ректными. Например, если X_1 есть количество продукции, а Y — общий выход, то коэффициент b_1 должен быть положительным.

Правдоподобно ли уравнение? Тщательно ли рассмотрели уравнение эксперты? Подходящи ли включенные в уравнение факторы и нет ли очевидных пропусков?

Пригодно ли уравнение? Окончательная модель может содержать множество переменных, которые полезны для предсказания, но, возможно, ими нельзя воспользоваться для управления. Это показывает следующий пример.

Для некоторого процесса задан набор стандартных условий работы, которые требуют уточнений, и k управляющих факторов. Уравнение для предсказания выхода процесса содержит только p из этих факторов, причем $p < k$. Если попытаться с помощью построенного уравнения найти способ, позволяющий улучшить выход, то будет игнорироваться $k—p$ факторов, не входящих в модель. Это произойдет потому, что при стандартных рабочих условиях $k—p$ факторов так тесно связаны с p факторами из предсказывающей модели, что в предсказании по ним нет необходимости. Однако определение этих $k—p$ факторов так же необходимо для управления данным процессом, как и p факторов, входящих в модель. Так, например, в данных Хальда, приведенных в приложении Б, можно отлично предсказывать выход цемента только по факторам X_1 и X_2, но, чтобы сделать цемент, нужны все-таки все четыре ингредиента.

Использование модели

Если все предшествующие критерии пропустили модель и все контрольные точки были благополучно пройдены, то следует определить процедуру использования модели. Физические условия меняются, и поэтому необходимо определить, когда отклонения фактических наблюдений от предсказанных значений указывают признаки несостоятельности модели. Если статистик располагает множеством контрольных карт для отклонений, то стандартная процедура статистического контроля качества с помощью контрольных карт как раз и призвана служить для проверки адекватности модели. И последний совет по использованию модели: подвергайте модель периодическим проверкам статистическими методами, так как это поможет обнаружить более сложные причины для беспокойства; никогда не оставляйте использование модели полностью на совесть заказчика, будь то химик или инженер.

Выводы

Если исследователь желает применить множественную регрессию как средство, помогающее ему решать задачи, то крайне необходимо, чтобы он следовал в общем приведенным выше ориентирам. Много времени и усилий может быть потрачено зря, если пытаться придавать какой-либо смысл сильно коррелированным данным; для применения методов множественной регрессии необходимо планирование серии контрольных точек, в которых оцениваются произведенные и ожидаемые затраты (стоимости). Наконец, никакой исследователь не может принуждаться к отказу от его научного понимания и прин-

ципов в пользу пристрастия к некоторым вычислительным процедурам статистического отсеивания. Методы множественной регрессии — мощное средство, если только им пользуются с умом и осторожностью [5].

---

[5] Задача построения моделей, несомненно, одна из центральных в науке. Можно думать, что ее роль столь велика, что она перестает принадлежать только науке, но становится феноменом общечеловеческой культуры. Данная глава, конечно, претендует только на охват простейших, чисто технических, аспектов этой проблемы, да и то не полно. Написать примечание, восполняющее этот пробел,— значит написать новую книгу не меньшего объема, чем эта. Поэтому ограничимся здесь только одной ссылкой на работу, в которой предлагается точка зрения, не во всем совпадающая с авторской: Бокс Дж. Е. П. Устойчивость в стратегии построения научных моделей.— В кн.: Устойчивые статистические методы оценки данных/Под ред. Р. Л. Лонера, Г. Н. Уилкинсона/Пер. с англ. Под ред. Н. Г. Волкова.— М.: Машиностроение, 1984, с. 164—188.— *Примеч. пер.*

Глава 9 ● ПРИЛОЖЕНИЕ МНОЖЕСТВЕННОЙ РЕГРЕССИИ К ЗАДАЧАМ ДИСПЕРСИОННОГО АНАЛИЗА

9.0. ВВЕДЕНИЕ

Методы множественного регрессионного анализа рассматривались в гл. 1, 2 и 3. Напомним, что для получения решения $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ нормальных уравнений $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$ необходимо, чтобы матрица $\mathbf{X}'\mathbf{X}$ была неособенной. Практически это означает, что нормальные уравнения должны содержать ровно столько независимых уравнений, сколько параметров подлежит оцениванию. Это наиболее распространенный случай. Однако если данные получаются из спланированных экспериментов, то надо проявить осторожность и проверить, все ли нормальные уравнения независимы друг от друга. Если это не так, то должны быть предприняты дополнительные шаги, чтобы все же получить оценки [1].

Один из наиболее распространенных методов анализа данных, полученных на основе спланированных экспериментов, — это метод дисперсионного анализа. Обычно он трактуется как нечто постороннее и совершенно отличное от регрессионного анализа. В большинстве вычислительных центров есть программы, специально разработанные для дисперсионного анализа. Некоторые специалисты даже не подозревают, что *любой* случай дисперсионного анализа при использовании модели с «фиксированными эффектами» факторов (модель 1) можно обработать с помощью общей регрессионной процедуры, если только модель правильно идентифицирована и предприняты необходимые меры для получения независимых нормальных уравнений. (На самом деле модели, отличные от моделей с фиксированными эффектами факторов, также можно обрабатывать по схеме регрессионного анализа, но мы ограничиваемся здесь только моделью 1.) [2]

---

[1] На самом деле в подобном случае всегда можно получить оценки параметров. Более того, существует бесконечное множество наборов таких оценок, доставляющих одно и то же минимальное значение сумме квадратов отклонений (см.: Горский В. Г., Спивак С. И.//Заводская лаборатория, 1981, **47**, № 10, с. 30—47). Поэтому более точно следовало бы сказать о том, что надо предпринять для получения *единственного* решения задачи оценивания.— *Примеч. пер.*

[2] С дисперсионным анализом и, в частности, с классификацией моделей дисперсионного анализа можно ознакомиться подробнее в книгах: Шеффе Г. Дисперсионный анализ. 2-е изд./Пер. с англ.— М.: Физматгиз, 1960, 626 с.; Хикс Ч. Р. Основные принципы планирования эксперимента/Пер. с англ.— М.: Мяр, 1967.— 406 с.; Хьютсон А. Дисперсионный анализ/Пер. с англ. Под ред. Т. И. Голиковой.— М.: Статистика, 1971, 88 с.; Маркова Е. В., Денисов В. И., Полетаева И. А., Пономарев В. В. Дис-

Мы *не рекомендуем* решать задачи дисперсионного анализа в случае модели с фиксированными эффектами факторов с помощью общих регрессионных методов. Мы лишь показываем, что они *могут* использоваться, если предприняты правильные шаги при решении задачи, и важно понимать это. Чаще всего приходится сталкиваться с тем, что на вопрос: «Какую модель Вы рассматриваете?» следует ответ: «Я не рассматриваю модель, а применяю дисперсионный анализ». Осознание того, что модель подразумевается в любой задаче дисперсионного анализа и что именно она служит основой для построения таблиц дисперсионного анализа, позволяет понять, что дисперсионный анализ практически эквивалентен регрессионному. Ведь известно, что регрессионный анализ без модели немыслим.

Особенности анализа дисперсионных (ANOVA) моделей, вообще говоря, состоят в том, что эти модели перепараметризованы, т. е. они содержат больше параметров, чем это необходимо для представления интересующих нас эффектов. Указанная перепараметризация обычно компенсируется введением определенных ограничений на параметры. Регрессионная обработка задач дисперсионного анализа включает построение фиктивных переменных, и, как это обычно бывает в подобных случаях, имеется неограниченное число способов выполнения этой операции, некоторые из которых более удобны на практике, чем другие.

В следующих параграфах мы обсудим сначала одностороннюю классификацию, а затем двустороннюю классификацию с равным числом наблюдений в ячейках, используя для каждого случая практический пример с реальными данными, чтобы дать читателю некоторое представление о достоинствах и недостатках регрессионного подхода. Будут также теоретически рассмотрены более общие односторонние и двусторонние классификации с одинаковым числом наблюдений в ячейках и будет приведен пример для последней из этих классификаций.

9.1. ОДНОСТОРОННЯЯ КЛАССИФИКАЦИЯ. ПРИМЕР

Считается, что кофеин при приеме его внутрь оказывает возбуждающее действие, результат и разброс которого зависят от величины принятой дозы. Чтобы получить сведения о действии кофеина при выполнении физической деятельности, был проведен следующий простой эксперимент.

1. *Эксперимент.* Использовались три уровня обработки (три дозы): 0, 100 и 200 мг кофеина. Было отобрано и обучено быстро ударять пальцем по клавише 30 здоровых студентов колледжа мужского пола

персионный анализ и синтез планов на ЭВМ.— М.: Наука, 1982.— 195 с.; Ветров А. А., Ломовацкий Г. И. Дисперсионный анализ в экономике.— М.: Статистика, 1975.— 120 с.; Глясс Дж., Стэнли Дж. Статистические методы в педагогике и психологии/Пер. с англ. Под ред. Ю. П. Адлера.— М.: Прогресс, 1976.— 600 с.; Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ/Пер. с нем.— М.: Финансы и статистика, 1985.— 230 с.— *Примеч. пер.*

одинакового возраста с примерно одинаковыми физическими данными. После обучения из них были скомплектованы случайным образом три группы по 10 человек. Попавшие в каждую группу студенты получали одну из трех доз, однако никто из испытуемых и наблюдающих за ними физиологов заранее не знал, какую дозу будет принимать тот или иной человек. Это знал только статистик, который занимался обработкой данных. Через два часа после приема доз каждый из испытуемых должен был выполнить предусмотренное выстукивание пальцем [3]. Было записано число ударов в минуту — см. табл. 9.1.

Таблица 9.1. **Число ударов по клавише в минуту, зарегистрированное у 30 студентов-мужчин, получивших разные дозы кофеина**

Обработки	Наблюдения	Обработка (строка)		
		суммы $T_i=\sum_j Y_{ij}$	средние $\bar{Y}_i$	эффекты $\hat{t}_i=\bar{Y}_i-\bar{Y}$
$i=1$; 0 мг кофеина (плацебо)[4]	242, 245, 244, 248, 247, 248, 242, 244, 246, 242	2448	244,8	—1,7
$i=2$; 100 мг кофенна	248, 246, 245, 247, 248, 250, 247, 246, 243, 244	2464	246,4	—0,1
$i=3$; 200 мг кофеина	246, 248, 250, 252, 248, 250, 246, 248, 245, 250	2483	248,3	1,8
		$\sum_i\sum_j Y_{ij}=7395$	$246,5=\bar{Y}$	

2. *Модель дисперсионного анализа для рассматриваемого эксперимента.* Пусть

[3] Изучаемый в этом эксперименте навык — удары пальцем по клавише — определяет скорость работы на пишущей машинке или на ключе, при передаче сигналов азбуки Морзе. Кофеин — алкалоид, содержащийся в семенах кофейного дерева, листьях чая, орехах кола и др. Оказывает возбуждающее действие на центральную нервную и сердечно-сосудистую системы. Служит стимулятором. Эксперимеит, в котором пациеит не знает, какую дозу (или какой препарат) он получает, называется «слепым». А если этой информацией не располагает и врач, то это уже «дважды слепой опыт». Цель такого умолчания — избежать влияния на результаты психологических эффектов, вольно или невольно проявляющихся как у пациента, так и у врача. Такая организация исследования при изученин новых медикаментов общепринята.— *Примеч. пер.*

[4] Плацебо—лекарство, не отличимое по внешним признакам от изучаемого препарата, но содержащее нейтральные и безвредные вещества. Используется для изучения роли внушения и как контроль прн планировании испытаний лекарств.— *Примеч. пер.*

Y_{ij} — число ударов пальцем в минуту для j-го человека, получившего i-ю дозу,

μ — истинное среднее значение числа ударов для всей популяции мужчин, из которой была сформирована случайная выборка из тридцати человек,

t_i — эффект i-й обработки, т. е. дополнительный эффект i-й обработки выше или ниже среднего μ. Чтобы μ было истинным средним значением для всей выборки, надо предположить, что $t_1 + t_2 + t_3 = 0$,

e_{ij} — случайный эффект, т. е. отклонение фактического числа ударов у j-го студента, получившего i-ю дозу, от величины $\mu + t_i$, обусловленное случайной ошибкой.

С учетом указанных обозначений ANOVA-модель (модель дисперсионного анализа) имеет вид

$$Y_{ij} = \mu + t_i + e_{ij}, \qquad (9.1.1)$$

и мы принимаем обычные предположения о нормально распределенных ошибках.

3. *Стандартные ANOVA-вычисления.* Обычные ANOVA-вычисления представлены в табл. 9.2.

Таблица 9.2. **Таблица дисперсионного анализа для примера с ударами по клавише**

Источник вариации	Степени свободы	SS	MS	F
Между обработкамн	2	61,40	30,70	6,18
Линейный*	1	61,25	61,25	12,32
Квадратичный*	1	0,15	0,15	0,03
Внутрн обработок	27	134,10	$s^2 = 4,97$	
Общий (скорректнрованный)	29	195,50		
Среднее	1	1822867,50		
Общий	30	1823063,00		

* Это расщепление суммы квадратов, обусловленной различиями «между обработками», объясняется в п. 5. Формулировки «между обработками» и просто «обработки» имеют один и тот же смысл. Мы их используем в этой главе как равноценные.

4. *Проверка гипотезы о равенстве эффектов обработки.* Чтобы проверить гипотезу $H_0: t_1 = t_2 = t_3$[5] против альтернативы H_1, что это не так, надо сравнить $F = 6,18$ с $F(2; 27; 0,95) = 3,35$. И таким образом, поскольку эмпирическое F выше табличного, гипотеза H_0 отвергается.

[5] Если учесть ранее введенное условие $t_1 + t_2 + t_3 = 0$, то гипотеза H_0 равноценна предположению $t_1 = t_2 = t_3 = 0$.— *Примеч. пер.*

5. *Линейный и квадратичный контрасты.* Сумма квадратов с двумя степенями свободы, обусловленная различиями «между обработками», может быть расщеплена разными способами. Один из них — стандартный прием, который имеет преимущества при равномерном распределении обработок, состоит в построении ортогональных линейных и квадратичных контрастов. Благодаря ортогональности сумма квадратов двух контрастов равна сумме квадратов «между обработками». При вычислении контрастов используются линейные комбинации сумм T_i, $T_i = \sum_j Y_{ij}$ в виде

Контраст	Кофеин (мг) в обработке		
	0	100	200
Линейный (L)	−1	0	1
Квадратичный (Q)	1	−2	1

Таким образом, контрасты равны:

$$L = -T_1 + T_3 = 35,$$

$$Q = T_1 - 2T_2 + T_3 = 3.$$

Если $i = 1, 2, \ldots, I$ и $j = 1, 2, \ldots, J$ (в нашем примере $I = 3$; $J = 10$), то суммы квадратов, связанные с L и Q, вычисляются по общей формуле и соответственно равны:

$$\mathrm{SS}(L) = \frac{L^2}{J\{(-1)^2 + 0^2 + 1^2\}} = \frac{(35)^2}{10 \cdot 2} = 61{,}25,$$

$$\mathrm{SS}(Q) = \frac{Q^2}{J\{1^2 + (-2)^2 + 1^2\}} = \frac{(3)^2}{10 \cdot 6} = 0{,}15.$$

Эти результаты отражены в табл. 9.2. Заметим, что в знаменателях этих сумм стоит величина J, представляющая собой число наблюдений, связанных с суммами T_i, которые использовались при вычислении контрастов L и Q. Другой сомножитель, стоящий в знаменателе, есть сумма квадратов коэффициентов, на которые умножаются T_i при вычислении контрастов.

Критерии для проверки линейного и квадратичного эффектов обнаруживают значимый линейный эффект (поскольку $12{,}32 > F$ (1; 27; 0,95) = 4,21); квадратичный эффект оказывается незначимым. Таким образом мы заключаем, что в пределах дозы кофеина от 0 до 200 мг истинное число нажатий на клавишу в минуту возрастает (поскольку $\overline{Y}_1 < \overline{Y}_2 < \overline{Y}_3$) линейно с увеличением получаемой дозы.

На этом заканчивается стандартный дисперсионный анализ для данного примера, и теперь мы обсудим регрессионный подход.

9.2. РЕГРЕССИОННЫЙ АНАЛИЗ ДЛЯ ПРИМЕРА С ОДНОСТОРОННЕЙ КЛАССИФИКАЦИЕЙ

Модель, выраженная уравнением (9.1.1), содержит четыре параметра: μ, t_1, t_2, t_3. Но два из них — μ и один из параметров t_i — определяют любое наблюдение Y_{ij}; следовательно, $t_1 + t_2 + t_3 = 0$.

Естественный первый шаг в регрессионном подходе — записать саму модель

$$Y = \mu X_0 + t_1 X_1 + t_2 X_2 + t_3 X_3 + \varepsilon \tag{9.2.1}$$

и затем рассмотреть, какие значения следует придать переменным X_i, чтобы получить уравнение (9.1.1). Элементарный анализ показывает, что для этого достаточно воспользоваться фиктивной переменной $X_0 = 1$, а остальным переменным придать значения:

$$X_i = \begin{cases} 1, & \text{если результат } Y_{ij} \text{ получен в } i\text{-й обработке,} \\ 0, & \text{если это не так,} \end{cases} \tag{9.2.2}$$

для $i = 1, 2, 3$. В таком случае получаем:

$$\underset{(30\times 1)}{\mathbf{Y}} = \begin{bmatrix} 242 \\ 245 \\ \vdots \\ 246 \\ 242 \\ \hline 248 \\ 246 \\ \vdots \\ 243 \\ 244 \\ \hline 246 \\ 248 \\ \vdots \\ 245 \\ 250 \end{bmatrix}, \qquad \underset{(30\times 4)}{\mathbf{X}} = \begin{array}{c} \begin{matrix} \mathbf{X_0} & \mathbf{X_1} & \mathbf{X_2} & \mathbf{X_3} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \hline 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{bmatrix} \end{array},$$

$$\underset{(4\times 1)}{\boldsymbol{\beta}} = \begin{bmatrix} \mu \\ t_1 \\ t_2 \\ t_3 \end{bmatrix}. \tag{9.2.3}$$

Недостаток такого подхода виден сразу. Поскольку вектор-столбцы матрицы $\mathbf{X}$ связаны между собой соотношением

$$\mathbf{X}_0 = \mathbf{X}_1 + \mathbf{X}_2 + \mathbf{X}_3, \tag{9.2.4}$$

матрица $\mathbf{X}'\mathbf{X}$ будет обязательно вырожденной, и потому нормальные уравнения не будут иметь единственного решения. До сих пор мы не принимали во внимание ограничение ANOVA-модели, а именно $t_1 + t_2 + t_3 = 0$. Если им воспользоваться, то решение нормальных уравнений станет единственным. Однако это представляет собой дополнительное усложнение нашего подхода (которое обсуждается далее в § 9.4) и означает, что мы не получаем на самом деле регрессионные выражения в стандартной форме. Как же быть? Имеется много возможностей (в общем их неограниченное число), как это обычно имеет место, когда используются фиктивные переменные. Одна из них — к ее описанию мы и переходим — позволяет воспользоваться линейными и квадратичными контрастами, которые участвуют в разложении, представленном в табл. 9.2. Определим фиктивные переменные Z_1 и Z_2, чтобы заменить ими X_1, X_2 и X_3, используя соотношения

Z_1	Z_2	при	X_1	X_2	X_3
-1	1		1	0	0
0	-2		0	1	0
1	1		0	0	1

(9.2.5)

Тогда модель можно записать в регрессионной форме таким образом:

$$Y = \beta_0 X_0 + \beta_1 Z_1 + \beta_2 Z_2 + \varepsilon, \tag{9.2.6}$$

а данные и параметры, используемые в регрессионном анализе, будут иметь вид

$$
\underset{(30\times 1)}{\mathbf{Y}} = \begin{bmatrix} 242 \\ 245 \\ \cdot \\ \cdot \\ \cdot \\ 242 \\ \hline 248 \\ 246 \\ \cdot \\ \cdot \\ \cdot \\ 244 \\ \hline 246 \\ 248 \\ \cdot \\ \cdot \\ \cdot \\ 250 \end{bmatrix}, \qquad \underset{(30\times 3)}{\mathbf{X}} = \begin{array}{c} \begin{matrix} \mathbf{X}_0 & \mathbf{Z}_1 & \mathbf{Z}_2 \end{matrix} \\ \begin{bmatrix} 1 & -1 & 1 \\ 1 & -1 & 1 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & -1 & 1 \\ \hline 1 & 0 & -2 \\ 1 & 0 & -2 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 0 & -2 \\ \hline 1 & 1 & 1 \\ 1 & 1 & 1 \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ 1 & 1 & 1 \end{bmatrix} \end{array}, \qquad \underset{(3\times 1)}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}. \tag{9.2.7}
$$

Заметим, что векторы $\mathbf{X}_0$, $\mathbf{Z}_1$ и $\mathbf{Z}_2$ взаимно ортогональны. Вследствие этого МНК-решение становится чрезвычайно простым. Мы находим

$$
\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{bmatrix} 1/30 & 0 & 0 \\ 0 & 1/20 & 0 \\ 0 & 0 & 1/60 \end{bmatrix} \begin{bmatrix} 7395 \\ 35 \\ 3 \end{bmatrix} =
$$

$$
= \begin{bmatrix} 246{,}50 \\ 1{,}75 \\ 0{,}05 \end{bmatrix}, \tag{9.2.8}
$$

$$
\mathrm{SS}(b_0,\ b_1,\ b_2) = \mathbf{b}\mathbf{X}'\mathbf{Y} = 1822867{,}50 + 61{,}25 + 0{,}15 = 1822928{,}90 =
$$

$$
= \mathrm{SS}(b_0) + \mathrm{SS}(b_1) + \mathrm{SS}(b_2) \tag{9.2.9}
$$

благодаря ортогональности, упомянутой выше. Результаты расчетов в деталях представлены в табл. 9.2. Подогнанная регрессионная модель имеет вид

$$\hat{Y} = 246{,}50 + 1{,}75\ Z_1 + 0{,}05\ Z_2. \tag{9.2.10}$$

Если мы теперь опустим незначимый член «$+\ 0{,}05\ Z_2$» и заметим, что переменная Z_1 может быть представлена в кодированном виде

$$Z_1 = (C - 100)/100, \tag{9.2.11}$$

где C — количество кофеина в мг, принимаемое внутрь, т. е уровень кофеина, мы увидим, что подогнанное уравнение приобретает вид

$$\hat{Y} = 244{,}75 + 0{,}0175\ C. \tag{9.2.12}$$

Следовательно, в пределах наблюдаемого интервала изменения C предсказываемое число нажатий пальцем на клавишу в течение одной минуты для обученных студентов колледжа можно вычислять в зависимости от C по уравнению (9.2.12).

(Заметим, что при удалении такой переменной, как Z_2, мы обычно должны пересоставлять регрессионное уравнение. Было бы неправильным в общем случае просто удалить какой-либо член. Однако здесь вследствие ортогональности столбца $\mathbf{Z}_2$ к другим столбцам матрицы $\mathbf{X}$ результат будет тем же самым при разных способах построения регрессии — см. § 2.8.)

Предупреждение

Благодаря тщательному выбору уровней фиктивных переменных Z_1 и Z_2, позволившему получить диагональную матрицу $\mathbf{X}'\mathbf{X}$, а также равенству объемов трех выборок приведенный выше пример оказался чрезвычайно простым и кратким. В общем случае на это рассчитывать не приходится. Вернемся снова к выражениям (9.2.3). Предположим, что мы решили вместо замены переменных просто исключить один параметр, скажем, t_3, из β и, следовательно, соответствующий столбец $\mathbf{X}_3$ из матрицы $\mathbf{X}$. Подходящее МНК-решение тогда имело бы вид

$$\begin{bmatrix} \hat{\mu} \\ \hat{t}_1 \\ \hat{t}_2 \end{bmatrix} = \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (10)^{-1} \begin{bmatrix} 3 & 1 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 7395 \\ 2448 \\ 2464 \end{bmatrix} =$$

$$= \frac{1}{10} \begin{bmatrix} 1 & -1 & -1 \\ -1 & 2 & 1 \\ -1 & 1 & 2 \end{bmatrix} \begin{bmatrix} 7395 \\ 2448 \\ 2464 \end{bmatrix} = \begin{bmatrix} 248{,}3 \\ -3{,}5 \\ -1{,}9 \end{bmatrix}, \tag{9.2.13}$$

$$SS(\hat{\mu},\ \hat{t}_1,\ \hat{t}_2) = \mathbf{b}'\mathbf{X}'\mathbf{Y} = (248{,}3;\ -3{,}5;\ -1{,}9)\begin{bmatrix} 7395 \\ 2448 \\ 2464 \end{bmatrix} =$$

$$= 1\,822\,928{,}90. \qquad (9.2.14)$$

Мы получили величину, в точности совпадающую с общей суммой в (9.2.9), как это и должно быть, однако информативного расщепления суммы на три слагаемых здесь не происходит, поскольку отсутствует ортогональность столбцов матрицы $\mathbf{X}$ и поскольку матрица $\mathbf{X}'\mathbf{X}$ здесь недиагональна. Мы можем, конечно, получить дополнительную сумму квадратов для $\hat{t}_1$ и $\hat{t}_2$, вычитая из полученного результата $n\overline{Y}^2$:

$$SS(\hat{t}_1,\ \hat{t}_2 \mid \hat{\mu}) = 1\,822\,928{,}90 - 1\,822\,867{,}50 = 61{,}40, \qquad (9.2.15)$$

однако мы не смогли бы вычислить дополнительную сумму квадратов, указанную в гл. 4, и двигаться дальше. Мораль такова: хотя многие приемы регрессионного анализа и применимы для задач дисперсионного анализа, некоторые из них более информативны, чем другие. Тщательный выбор хорошего описания воздастся сторицей.

Продолжая вычисления с неортогональным описанием дальше, получим модель

$$\hat{Y} = \hat{Y}(X_1,\ X_2) = 248{,}3 - 3{,}5X_1 - 1{,}9X_2, \qquad (9.2.16)$$

которая дает следующие предсказываемые значения:

$\hat{Y}$ (1,0) = 244,8, если принимается 0 мг кофеина,
$\hat{Y}$ (0,1) = 246,4, если принимается 100 мг кофеина,
$\hat{Y}$ (0,0) = 248,3, если принимается 200 мг кофеина.

У неподготовленного читателя появление в уравнении (9.2.16) отрицательных коэффициентов может вызвать удивление, поскольку на первый взгляд противоречит закономерности, в силу которой число ударов по клавише растет с увеличением дозы кофеина. Однако противоречия здесь нет. Фиктивные переменные здесь были выбраны так, чтобы $\hat{Y}$ (0, 0) соответствовало основному уровню. В таком случае для достижения соответствующих величин $\hat{Y}$ (1, 0) и $\hat{Y}$ (0, 1) необходимы *отрицательные* значения коэффициентов регрессии. Мы получили еще одно подтверждение того, как необходима осторожность при использовании регрессионных методов.

В § 9.3—9.5 односторонняя классификация будет рассмотрена более детально, после чего мы перейдем к двусторонней классификации.

9.3. ОДНОСТОРОННЯЯ КЛАССИФИКАЦИЯ

Предположим, что мы имеем данные в виде I групп, $i = 1, 2, \ldots, I$ с J_i наблюдениями в каждой группе, как это указано ниже:

группа 1 $Y_{11}, Y_{12}, \ldots, Y_{1J_1}$ среднее $\overline{Y}_1$,

группа 2 $Y_{21}, Y_{22}, \ldots, Y_{2J_2}$, среднее $\overline{Y}_2$,

. . . . . .

группа I $Y_{I1}, Y_{I2}, \ldots, Y_{IJ_I}$ среднее $\overline{Y}_I$.

Обычная модель дисперсионного анализа с фиксированными эффектами факторов для такой ситуации имеет вид

$$E(Y_{ij}) = \mu + t_i, \quad i = 1, 2, \ldots, I, \quad j = 1, 2, \ldots, J_i, \tag{9.3.1}$$

где $t_1, t_2, \ldots, t_I$ — параметры, удовлетворяющие условию

$$J_1 t_1 + J_2 t_2 + \ldots + J_I t_I = 0. \tag{9.3.2}$$

Ограничение (9.3.2) необходимо, поскольку модель (9.3.1) содержит больше параметров, чем это в действительности необходимо, и приходится рассматривать μ в качестве общего среднего, а t_i — в качестве разности между средним по i-й группе и общим средним. Таким образом, сумма всех разностей между группами и общим уровнем равна нулю, что и отражает уравнение (9.3.2). Обычная таблица дисперсионного анализа имеет в данном случае такой вид

$$\left(\text{здесь } \overline{Y} = \sum_{i=1}^{I} J_i \overline{Y}_i \Big/ \sum_{i=1}^{I} J_i\right):$$

Источник вариации	Степени свободы	SS	MS
Между группами	$I-1$	$\sum_{i=1}^{I} J_i (\overline{Y}_i - \overline{Y})^2$	s_B^2
Внутри групп	$\sum_{i=1}^{I} (J_i - 1) = n - I$	$\sum_{i=1}^{I} \sum_{j=1}^{J_i} (Y_{ij} - \overline{Y}_i)^2$	s_W^2
Среднее	1	$n\overline{Y}^2$	
Общий	$\sum_{i=1}^{I} J_i = n$	$\sum_{i=1}^{I} \sum_{j=1}^{J_i} Y_{ij}^2$	

Как всегда, необходимо проверить гипотезу о том, что нет различия между средними по группам (т. е. H_0: $t_1 = t_2 = \ldots = t_I = 0$), сравнивая отношение $F = s_B^2/s_W^2$ с подходящей процентной точкой F распределения

$$F\left(I - 1;\ \sum_{i=1}^{I} (J_i - 1)\right).$$

9.4. РЕГРЕССИОННАЯ ОБРАБОТКА ОДНОСТОРОННЕЙ КЛАССИФИКАЦИИ С ИСПОЛЬЗОВАНИЕМ ИСХОДНОЙ МОДЕЛИ

Вместо (9.3.1) запишем

$$E(Y) = \mu X_0 + t_1 X_1 + t_2 X_2 + \ldots + t_I X_I. \qquad (9.4.1)$$

Выразим теперь тот факт, что математическое ожидание наблюдения Y_{ij} из i-й группы должно быть равно $\mu + t_i$. Введем обозначения:

$$\mathbf{Y}' = (Y_{11}, Y_{12}, \ldots, Y_{1J_1}; \quad Y_{21}, Y_{22}, \ldots, Y_{2J_2}; \quad \ldots;$$
$$Y_{I1}, Y_{I2}, \ldots, Y_{IJ_I}),$$

$$\mathbf{X} = \begin{array}{c} \begin{array}{ccccccc} \mathbf{X_0} & \mathbf{X_1} & \mathbf{X_2} & \mathbf{X_3} & & & \mathbf{X_I} \end{array} \\ \left[\begin{array}{ccccccc} 1 & 1 & 0 & 0 & \ldots & & 0 \\ 1 & 1 & 0 & 0 & \ldots & & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & & \vdots \\ 1 & 1 & 0 & 0 & \ldots & & 0 \\ \hdashline 1 & 0 & 1 & 0 & \ldots & & 0 \\ 1 & 0 & 1 & 0 & \ldots & & 0 \\ \vdots & \vdots & \vdots & \vdots & \ldots & & \vdots \\ 1 & 0 & 1 & 0 & \ldots & & 0 \\ \hdashline \vdots & \vdots & \vdots & \vdots & \ldots & & \vdots \\ \hdashline 1 & 0 & 0 & 0 & \ldots & & 1 \\ 1 & 0 & 0 & 0 & \ldots & & 1 \\ \vdots & \vdots & \vdots & \vdots & \ldots & & \vdots \\ 1 & 0 & 0 & 0 & \ldots & & 1 \end{array}\right] \end{array},$$

где пунктирные линии разделяют матрицу на подматрицы, содержащие соответственно $J_1, J_2, \ldots, J_I$ строк. Заголовки показывают, каким переменным X соответствуют столбцы. Далее обозначим

$$\boldsymbol{\beta}' = (\mu, t_1, t_2, \ldots, t_I).$$

Тогда соотношение

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$$

представляет собой запись уравнения (9.3.1) в матричной форме.

Теперь

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} n & J_1 & J_2 & \dots & J_I \\ J_1 & J_1 & 0 & \dots & 0 \\ J_2 & 0 & J_2 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ J_I & 0 & 0 & \dots & J_I \end{bmatrix},$$

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} n\overline{Y} \\ J_1\overline{Y}_1 \\ J_2\overline{Y}_2 \\ \dots \\ J_I\overline{Y}_I \end{bmatrix}. \qquad (9.4.2)$$

Если обозначить через b_0 и b_i МНК-оценки параметров μ и t_i, то нормальные уравнения $(\mathbf{X}'\mathbf{X})\,\mathbf{b} = \mathbf{X}'\mathbf{Y}$ можно записать в виде:

$$\begin{aligned} &nb_0 + J_1b_1 + J_2b_2 + \dots + J_Ib_I = n\overline{Y}, \qquad (9.4.3)\\ &J_1b_0 + J_1b_1 \qquad\qquad\qquad\qquad = J_1\overline{Y}_1,\\ &J_2b_0 \qquad + J_2b_2 \qquad\qquad\qquad = J_2\overline{Y}_2,\\ &\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\\ &J_Ib_0 \qquad\qquad\qquad\qquad + J_Ib_I = J_I\overline{Y}_I. \end{aligned}$$

В данном случае обратной матрицы $(\mathbf{X}'\mathbf{X})^{-1}$ не существует, так как $\mathbf{X}'\mathbf{X}$ особенная (уравнения (9.4.3) *не являются* независимыми, ибо первое уравнение есть сумма остальных I уравнений). В нашем распоряжении есть только I уравнений с $I + 1$ неизвестными $b_0, b_1, \dots, b_I$, поскольку исходная модель (9.3.1) содержит больше параметров, чем это фактически необходимо. Эта «особенность» матрицы $\mathbf{X}'\mathbf{X}$ вытекает также и из того факта, что столбец $\mathbf{X}_0$ матрицы $\mathbf{X}$ равен сумме столбцов $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_I$. Имеющаяся зависимость отражается в нормальных уравнениях (9.4.3), что мы уже отмечали. Как же теперь выйти из положения? Мы не принимали пока во внимание условие (9.3.2), которое справедливо как для параметров, так и для их оценок. Следовательно,

$$J_1b_1 + J_2b_2 + \dots + J_Ib_I = 0, \qquad (9.4.4)$$

что дает дополнительное необходимое нам независимое уравнение. Возьмем теперь *любые* I уравнений из системы (9.4.3) вместе с урав-

нением (9.4.4) и образуем из них систему нормальных уравнений. Из (9.4.3) удобнее отбросить первое уравнение, содержащее больше всего членов. Тогда придем к такой системе нормальных уравнений:

$$\begin{aligned}
&J_1 b_1 + J_2 b_2 + \ldots + J_I b_I = 0,\\
&J_1 b_0 + J_1 b_1 = J_1 \overline{Y}_1,\\
&J_2 b_0 + J_2 b_2 = J_2 \overline{Y}_2,\\
&\ldots\ldots\ldots\ldots\ldots\ldots\\
&J_I b_0 + J_I b_I = J_I \overline{Y}_I.
\end{aligned} \tag{9.4.5}$$

Чтобы сохранить симметрию, мы не будем сокращать уравнения, начиная со второго, на соответствующие общие множители. В матричной форме система (9.4.5) имеет вид

$$\begin{bmatrix} 0 & J_1 & J_2 & J_I \\ J_1 & J_1 & 0 & 0 \\ J_2 & 0 & J_2 & 0 \\ \ldots & \ldots & \ldots & \ldots \\ J_I & 0 & 0 & J_I \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \ldots \\ b_I \end{bmatrix} =$$

$$= \begin{bmatrix} 0 \\ J_1 \overline{Y}_1 \\ J_2 \overline{Y}_2 \\ \ldots \\ J_I \overline{Y}_I \end{bmatrix}. \tag{9.4.6}$$

Поскольку мы не можем выразить эти уравнения в виде $(\mathbf{X}'\mathbf{X})\,\mathbf{b} = \mathbf{X}'\mathbf{Y}$, нецелесообразно использовать приемы, ориентированные на такую форму записи. Каждое из уравнений (9.4.5), начиная со второго, позволяет найти

$$b_i = \overline{Y}_i - b_0, \quad i = 1, 2, \ldots, I.$$

После подстановки в первое уравнение получим

$$0 = \sum_{i=1}^{I} J_i b_i = \sum_{i=1}^{I} J_i (\overline{Y}_i - b_0) = \sum_{i=1}^{I} J_i \overline{Y}_i - b_0 \sum_{i=1}^{I} J_i = n\overline{Y} - n b_0.$$

Таким образом,

$$b_0 = \overline{Y},$$

$$b_i = \overline{Y}_i - \overline{Y}.$$

Сумма квадратов, обусловленная вектором оценок параметров $\mathbf{b}$, которые определяются из уравнений $\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$, выражается ве-

личиной $\mathbf{b}'\mathbf{X}'\mathbf{Y}$, даже для вырожденной матрицы $\mathbf{X}'\mathbf{X}$, не имеющей обратной (дополнительные условия $\mathbf{Qb} = 0$ необходимы, чтобы обеспечить единственное решение). Величина $\mathbf{b}'\mathbf{X}'\mathbf{Y}$ есть инвариант по отношению к вектору $\mathbf{Q}$ (в рассматриваемом случае $\mathbf{Q} = (0, 1, 1, \ldots, 1)$. В самом деле, если $\mathbf{b}_1$ и $\mathbf{b}_2$ — два решения, соответствующие различным «дополнительным условиям», то справедливо соотношение

$$\mathbf{b}_1'(\mathbf{X}'\mathbf{Y}) = \mathbf{b}_1'(\mathbf{X}'\mathbf{X}\mathbf{b}_2) = (\mathbf{X}'\mathbf{X}\mathbf{b}_1)'\,\mathbf{b}_2.$$

Осуществив перегруппировку и воспользовавшись известным из теории матриц свойством $(\mathbf{AB})' = \mathbf{B}'\mathbf{A}'$, продолжим цепочку равенств:

$$(\mathbf{X}'\mathbf{X}\mathbf{b}_1)\,\mathbf{b}_2 = (\mathbf{X}'\mathbf{Y})'\,\mathbf{b}_2 = \mathbf{b}_2'\mathbf{X}'\mathbf{Y}.$$

Таким образом, сумма квадратов, обусловленная регрессией, есть

$$\mathbf{b}'\mathbf{X}'\mathbf{Y} = n\bar{Y}^2 + \sum_{i=1}^{I} J_i\bar{Y}_i(\bar{Y}_i - \bar{Y}) = n\bar{Y}^2 + \sum_{i=1}^{I} J_i(\bar{Y}_i - \bar{Y})^2$$

с I степенями свободы, так как дополнительный член

$$\sum_{i=1}^{I}(-\bar{Y})\,J_i(\bar{Y}_i - \bar{Y}),$$

добавляемый в правую часть уравнения, равен нулю по определению средних. Если бы модель содержала только один член μ, то мы имели бы

$$\mathrm{SS}(b_0) = n\bar{Y}^2$$

с одной степенью свободы. Для общего случая

$$\mathrm{SS}(b_1, b_2, \ldots, b_I \mid b_0) = \mathbf{b}'\mathbf{X}'\mathbf{Y} - n\bar{Y}^2 = \sum_{i=1}^{I} J_i(\bar{Y}_i - \bar{Y})^2$$

с $I - 1$ степенями свободы.

Итак, полученная сумма квадратов обусловлена «средним» и рассеянием «между группами», которые указаны в таблице дисперсионного анализа в § 9.3. Сумма квадратов, вызванная рассеянием «внутри групп», находится, как обычно, по разности $\mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y}$, которая совпадает с развернутым выражением из таблицы в § 9.3. Проверка гипотезы H_0: $t_1 = t_2 = \ldots = t_I = 0$ выполняется точно так же, как в дисперсионном анализе.

Мы уже видели, что дисперсионный анализ в случае одностороней классификации может быть выполнен формально, если применить регрессию, используя исходную модель. Однако, чтобы провести вычисления на машине, вероятно, лучше сначала избавиться от вырожденности путем преобразования модели.

(Примечание. Из сказанного выше ясно, как поступать вообще в регрессионной задаче, когда число параметров, подлежащих оцениванию, больше числа независимых нормальных уравнений. Если нет никаких естественных ограничений, как в случае дисперсионного анализа, то надо ввести ограничения произвольного вида. Хотя выбор ограничений и оказывает влияние на фактические значения ко-

эффициентов регрессии, он не влияет на величиину суммы квадратов, обусловленную регрессией. Обычно мы будем выбирать ограничения таким образом, чтобы облегчить решение нормальных уравнений.)

Пример. Предположим, что нормальные уравнения имеют вид:

$$
\begin{array}{rrrrrrl}
22b_1 & +10b_2 & +12b_3 & +5b_4 & +8b_5 & +9b_6 & = 34{,}37 \\
10b_1 & +10b_2 & & +3b_4 & +4b_5 & +3b_6 & = 21{,}21 \\
12b_1 & & +12b_3 & +2b_4 & +4b_5 & +6b_6 & = 13{,}16 \\
 & & & & & & \\
5b_1 & +3b_2 & +2b_3 & +5b_4 & & & = 10{,}28 \\
8b_1 & +4b_2 & +4b_3 & & +8b_5 & & = 14{,}23 \\
9b_1 & +3b_2 & +6b_3 & & & +9b_6 & = 9{,}86
\end{array}
$$

(Эти уравнения взяты из книги: Plackett R. L. Regression analysis.— Oxford: Clarendon Press, 1960, p. 44. Они вытекают из двусторонней классификации с различным числом наблюдений в ячейках. Но такие данные могут возникнуть и в том случае, когда проводится дисперсионный анализ с одинаковым числом наблюдений в ячейках, но некоторые наблюдения теряются. В следующем параграфе будет рассмотрен случай одинакового числа наблюдений в ячейках.)

Только четыре из приведенных шести уравнений независимы, так как второе и третье уравнения в сумме дают первое, а сумма четвертого, пятого и шестого уравнений также совпадает с первым уравнением. Таким образом, чтобы получить шесть уравнений относительно шести неизвестных, нужны еще два дополнительных уравнения. Это должны быть два дополнительных независимых ограничения, связывающие b_1, b_2, . . . , b_6, причем таких, чтобы они не были линейными комбинациями имеющихся уравнений.

Так как фактически имеется только четыре независимых нормальных уравнения, мы можем опустить два зависимых уравнения, например первое и шестое. Оставшиеся четыре уравнения можно записать в матричной форме так:

$$
\begin{array}{c}
\begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\
\begin{bmatrix} 10 & 10 & 0 & 3 & 4 & 3 \\ 12 & 0 & 12 & 2 & 4 & 6 \\ 5 & 3 & 2 & 5 & 0 & 0 \\ 8 & 4 & 4 & 0 & 8 & 0 \end{bmatrix}
\end{array}
\begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{bmatrix}
=
\begin{bmatrix} 21{,}21 \\ 13{,}16 \\ 10{,}28 \\ 14{,}23 \end{bmatrix}
$$

Поскольку исходная матрица была симметричной, указанная выше зависимость строк (или уравнений) выражается также в том, что первый столбец равен как сумме второго и третьего, так и сумме четвертого, пятого и шестого столбцов. При добавлении двух ограничений на коэффициенты b мы должны позаботиться о двух моментах. Добавленные к четырем отобранным уравнениям ограничения приведут к появлению двух новых строк в матрице и двух дополнительных

нулей в записи вектора, стоящего в правой части равенства (обычно выбирают ограничения в форме $\Sigma c_i b_i = 0$). Если мы хотим получить единственное решение, то окончательная матрица должна быть такой, чтобы ее строки и столбцы оказались независимыми. (Более изящный матричный способ представления этих рассуждений дан, например, Плэккеттом (Plackett, 1960), но мы не приводим его в нашем, более элементарном изложении). Так, например, мы *не можем* воспользоваться ограничениями

$$7b_1 + 6b_2 + b_3 + b_4 + b_5 + 5b_6 = 0,$$
$$11b_1 + 9b_2 + 2b_3 + 4b_4 + 4b_5 + 3b_6 = 0,$$

так как в этом случае сохраняется первоначальная зависимость столбцов. Даже если остается только одна связь между столбцами, как, скажем, в случае ограничений

$$3b_4 + 4b_5 + 3b_6 = 0,$$
$$9b_1 + 5b_2 + 4b_3 = 0,$$

при которых первый столбец будет равен сумме второго и третьего, такие ограничения также будут бесполезны. Но соотношения

$$3b_4 + 4b_5 + 3b_6 = 0,$$
$$b_2 + b_3 = 0$$

в качестве ограничений уже приемлемы, поскольку в этом случае не будет никакой зависимости, и мы получим шесть уравнений относительно шести переменных, что и требуется. (Различные ограничения были исследованы Плэккеттом, книгу которого мы рекомендуем для более углубленного ознакомления с возникающими в таких задачах проблемами [6].)

Идея добавления произвольных ограничений может показаться на первый взгляд несколько необычной. Следует, однако, напомнить, что к этому приходится прибегать лишь тогда, когда в модели больше параметров, чем требуется для описания. Подобное несоответствие в какой-то момент должно быть устранено, что и достигается путем добавления ограничений.

9.5. РЕГРЕССИОННАЯ ОБРАБОТКА ДАННЫХ В СЛУЧАЕ ОДНОСТОРОННЕЙ КЛАССИФИКАЦИИ: НЕЗАВИСИМЫЕ НОРМАЛЬНЫЕ УРАВНЕНИЯ

Допустим, что модель дисперсионного анализа имеет вид

$$E(Y_{ij}) = \mu + t_i. \tag{9.5.1}$$

Обозначим

$$\beta_i = \mu + t_i, \qquad i = 1, 2, \ldots, I.$$

[6] Этот вопрос обстоятельно исследован также в книге: Маркова Е. В., Денисов В. И., Полетаев И. А., Пономарев В. В. Дисперсионный анализ и синтез планов на ЭВМ.—М.: Наука, 1982.— 195 с.— *Примеч. пер.*

Теперь, используя обозначения, принятые в регрессионном анализе, мы можем записать

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}, \tag{9.5.2}$$

где, если сравнивать с обозначениями § 9.3, $\mathbf{Y}$ есть тот же самый вектор; $\mathbf{X}$ — матрица, образуемая из прежней матрицы $\mathbf{X}$ путем вычеркивания из нее столбца $\mathbf{X}_0$; $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_I)$. Пусть $\mathbf{b}' = (b_1, b_2, \ldots, b_I)$, тогда

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} J_1 & & & 0 \\ & J_2 & & \\ & & \ddots & \\ 0 & & & J_I \end{bmatrix}, \qquad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} J_1\bar{Y}_1 \\ J_2\bar{Y}_2 \\ \ldots \\ J_I\bar{Y}_I \end{bmatrix} \tag{9.5.3}$$

Так как $\mathbf{X}'\mathbf{X}$ — диагональная матрица с элементами J_i, $i = 1, 2, \ldots, I$, на главной диагонали и остальными нулями, обратная матрица от нее также диагональная с элементами $1/J_i$ на главной диагонали. Отсюда легко видеть, что

$$b_i = \bar{Y}_i. \tag{9.5.4}$$

Сумма квадратов, обусловленная вектором оценок $\mathbf{b}$, равна:

$$\mathbf{b}\,\mathbf{X}'\mathbf{Y} = \sum_{i=1}^{I} \bar{Y}_i (J_i \bar{Y}_i) = \sum_{i=1}^{I} J_i \bar{Y}_i^2. \tag{9.5.5}$$

Остаточная сумма квадратов имеет вид

$$\begin{aligned} \mathbf{Y}'\mathbf{Y} - \mathbf{b}'\mathbf{X}'\mathbf{Y} &= \sum_{i=1}^{I}\sum_{j=1}^{J_i} Y_{ij}^2 - \sum_{i=1}^{I} J_i \bar{Y}_i^2 = \\ &= \sum_{i=1}^{I} \left\{ \sum_{j=1}^{J_i} Y_{ij}^2 - J_i \bar{Y}_i^2 \right\} = \sum_{i=1}^{I}\sum_{j=1}^{J_i} (Y_{ij} - \bar{Y}_i)^2; \end{aligned} \tag{9.5.6}$$

ей соответствует $n-I$ степеней свободы.

Гипотеза H_0: $t_1 = t_2 = \ldots = t_I = 0$ выражается в новых обозначениях так:

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \ldots = \boldsymbol{\beta}_I = \mu.$$

Если бы H_0 была верна, то модель имела бы вид

$$E(Y_{ij}) = \mu$$

или

$$E(\mathbf{Y}) = \mathbf{j}\mu, \tag{9.5.7}$$

где $\mathbf{j}$ есть вектор, составленный из единиц и имеющий ту же размерность, что и $\mathbf{Y}$. Ей соответствует единственное нормальное уравнение

$$n\mu = \mathbf{j}'\mathbf{j}\mu = \mathbf{j}'\mathbf{Y} = \sum_{i=1}^{I}\sum_{j=1}^{J_i} Y_{ij} = n\bar{Y}.$$

Таким образом, оценка параметра μ была бы равна:

$$b_0 = \bar{Y}, \tag{9.5.8}$$

$$\mathrm{SS}(b_0) = n\bar{Y}^2, \tag{9.5.9}$$

что приводит, как это можно показать, к остаточной сумме квадратов

$$\sum_{i=1}^{I}\sum_{j=1}^{J_i} Y_{ij}^2 - n\bar{Y}^2 = \sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y})^2 \tag{9.5.10}$$

с $n-1$ степенями свободы.

Сумма квадратов, обусловленная гипотезой H_0, есть разность между выражениями (9.5.10) и (9.5.6), а именно

$$\mathrm{SS}(H_0) = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\{(Y_{ij}-\bar{Y})^2-(Y_{ij}-\bar{Y}_i)^2\} = \sum_{i=1}^{I} J_i(\bar{Y}_i-\bar{Y})^2$$

с $(n-1)-(n-I) = I-1$ степенями свободы. Статистика для проверки гипотезы H_0 представляет собой, таким образом, величину

$$F = \frac{\mathrm{SS}(H_0)}{I-1}\Bigg/\frac{\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_i)^2}{n-I},$$

в точности совпадающую с выражением, которое мы получаем из дисперсионного анализа. Таким образом, если модель для односторонней классификации записана в виде $E(Y_{ij}) = \beta_i$ и проверяется гипотеза H_0: $\beta_1 = \beta_2 = \ldots = \beta_I = \mu$, то можно воспроизвести дисперсионный анализ с помощью регрессионного анализа при использовании стандартных программ. Оценки параметров t_i могут быть получены как разности $b_i - b_0$.

9.6. ДВУСТОРОННЯЯ КЛАССИФИКАЦИЯ С РАВНЫМ ЧИСЛОМ НАБЛЮДЕНИЙ В ЯЧЕЙКАХ. ПРИМЕР

Принципы, изложенные в § 9.1 и 9.2, становятся даже более важными при переходе к более сложным экспериментальным планам. Приведем пример двусторонней классификации с основными эффектами и взаимодействиями, рассмотренный Смитом в работе: Smith H. The analysis of data from designed experiment.— Journal of Quality Technology, 1969, 1, p. 259—263.

Владелец каталитического завода был обеспокоен состоянием дел с производительностью, которая характеризовалась скоростью производства продукта. После оживленной дискуссии в исследовательском отделе было решено исследовать эффекты 12 различных комбинаций из 4 реагентов и 3 катализаторов. Трудность, с которой столкнулись при экспериментировании, состояла в том, что не удавалось стабильно воспроизводить скорость получения продукта при, казалось бы, идентичных условиях. Чтобы получить оценки эффектов

в условиях такой вариабельности, было решено каждый опыт повторять дважды. Таким образом, весь эксперимент включал 24 опыта, проводившиеся в случайном порядке. Их результаты в закодированном и округленном виде приведены в табл. 9.3.

Таблица 9.3. **Двадцать четыре значения скорости производства продукта (закодированные и округленные) для двенадцати комбинаций реагентов и катализаторов**

Реагент	Катализаторы		
	1	2	3
A	4,6	11,7	5.9
B	6,4	13,15	9,7
C	13,15	15,9	13,13
D	12,12	12,14	7,9

ANOVA-модель для этого эксперимента была принята в виде

$$Y_{ijk} = \mu + R_i + C_j + (RC)_{ij} + \varepsilon_{ijk}, \tag{9.6.1}$$

где μ — генеральное среднее (1 параметр),
R_i — эффект i-го реагента (4 параметра),
C_j — эффект j-го катализатора (3 параметра),
$(RC)_{ij}$ — эффект взаимодействия реагента i и катализатора j (12 параметров);
ε_{ijk} — случайная ошибка внутри (i, j)-ячейки с k наблюдениями; предполагается, что она распределена нормально $N(0, \sigma^2)$ и что ошибки попарно не коррелированы.

Итак,

$$i = 1, 2, \ldots, I \qquad (I = 4),$$
$$j = 1, 2, \ldots, J \qquad (J = 3),$$
$$k = 1, 2, \ldots, K \qquad (K = 2).$$

Заметим, что эта модель включает двадцать параметров и является перепараметризованной. Введем следующие ограничения:

$$\sum_i R_i = \sum_j C_j = 0, \tag{9.6.2}$$

$$\sum_i (RC)_{ij} = 0, \qquad j = 1, 2, \ldots, J, \tag{9.6.3}$$

$$\sum_j (RC)_{ij} = 0, \qquad i = 1, 2, \ldots, I. \tag{9.6.4}$$

Число независимых параметров свелось, таким образом, к $1 + 3 + 2 + 6 = 12$ или в общем случае $1 + (I-1) + (J-1) + (I-1)(J-1) = IJ$, что равно числу ячеек. Обычный анализ приведен в ANOVA-таблице 9.4.

Таблица 9.4 **Дисперсионный анализ для модели (9.6.1)**

ANOVA

Источник	Степени свободы	SS	MS	F
Между реагентами	3	120	40	10,0**
Между катализаторами	2	48	24	6,0*
Взаимодействие (реагенты × катализаторы)	6	84	14	3,5*
«Чистая» ошибка	12	48	4	
Общий (скорректированный)	23	200		

* Значим при уровне $\alpha = 0{,}05$.
** Значим при уровне $\alpha = 0{,}01$.

Проанализируем теперь эти данные, используя регрессионную методологию.

9.7. РЕГРЕССИОННАЯ ОБРАБОТКА ПРИМЕРА С ДВУСТОРОННЕЙ КЛАССИФИКАЦИЕЙ

Как мы уже упоминали, в рассматриваемом примере необходимо найти двенадцать независимых параметров модели. Попытка определения более двенадцати параметров привела бы к вырождению матрицы $\mathbf{X'X}$. Из этих двенадцати параметров один приходится на генеральное среднее, три — на эффекты реагентов, два — на эффекты катализаторов и шесть — на взаимодействия этих эффектов. Как всегда, можно по-разному ввести фиктивные переменные, чтобы построить регрессионную модель. Выберем их так, чтобы матрица $\mathbf{X'X}$ была диагональной. Запишем регрессионную модель в виде

$$\begin{aligned} Y = {} & \beta_0 X_0 + && \text{(постоянный член; } X_0 = 1\text{)} \\ & + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + && \text{(для реагентов)} \\ & + \beta_4 X_4 + \beta_5 X_5 + && \text{(для катализаторов)} \\ & + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9 + \\ & + \beta_{10} X_{10} + \beta_{11} X_{11} + && \text{(для } R \times C \text{ взаимодействий)} \\ & + \varepsilon && \text{(ошибка)} \end{aligned} \qquad (9.7.1)$$

и определим фиктивные переменные с помощью следующих соотношений:

Используемый реагент	Соответствующие значения фиктивных переменных		
	X_1	X_2	X_3
A	−1	0	−1
B	1	0	−1
C	0	−1	1
D	0	1	1

Заметим, что три независимых столбца по существу позволяют провести сравнения между реагентами:

$$A \text{ и } B \text{ с помощью } X_1,$$
$$C \text{ и } D \text{ с помощью } X_2, \qquad (9.7.2)$$
$$(A+B) \text{ и } (C+D) \text{ с помощью } X_3$$

и что эти независимые сравнения (или контрасты) «содержат» 3 степени свободы. Эти столбцы позволяют генерировать некоторые другие типы сравнений. Так, например, чтобы сравнить A и D мы можем использовать столбец (зависимый, конечно, от столбцов X_1, X_2 и X_3), который получен суммированием данных трех столбцов, что дает $(-2, 0, 0, 2)'$. Кроме того, в силу симметрии плана и матрицы $\mathbf{X}$ столбцы, связанные с X_1, X_2 и X_3, взаимно ортогональны. При условии, что они также ортогональны к другим $\mathbf{X}$-столбцам, как это и есть на самом деле, индивидуальные дополнительные суммы квадратов могут быть соотнесены с каждым из трех контрастов, которые генерируются с помощью столбцов X_1, X_2 и X_3. Введем теперь фиктивные переменные, связанные с различными катализаторами.

Катализатор	Соответствующие значения фиктивных переменных	
	X_4	X_5
1	−1	1
2	0	−2
3	1	1

Эти два столбца являются независимыми, они позволяют провести сравнение между катализаторами

$$1 \text{ и } 3 \text{ с помощью } X_4,$$
$$(1+3) \text{ и } 2 \text{ с помощью } X_5. \qquad (9.7.3)$$

Благодаря тому, что этот двусторонний план полностью сбалансирован, а в каждой ячейке содержится одинаковое число наблюдений, можно сконструировать фиктивные переменные, связанные с взаимодействиями, перемножая соответствующие элементы других фиктивных переменных, в виде шести комбинаций: три фиктивные переменные, связанные с реагентами, умножаются на две фиктивные переменные, связанные с катализаторами. Таким образом,

$$X_6 = X_1X_4, \qquad X_7 = X_1X_5,$$
$$X_8 = X_2X_4, \qquad X_9 = X_2X_5, \qquad (9.7.4)$$
$$X_{10} = X_3X_4, \qquad X_{11} = X_3X_5.$$

Запишем теперь вектор $\mathbf{Y}$ и матрицу $\mathbf{X}$:

$$\mathbf{Y}=\begin{bmatrix}4\\6\\11\\7\\5\\9\\6\\4\\13\\15\\9\\7\\13\\15\\15\\9\\13\\13\\12\\12\\12\\14\\7\\9\end{bmatrix},\quad \mathbf{X}=\begin{array}{c}\begin{matrix}X_0 & X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 & X_9 & X_{10} & X_{11}\end{matrix}\\ \begin{bmatrix}1&-1&0&-1&-1&1&1&-1&0&0&1&-1\\1&-1&0&-1&-1&1&1&-1&0&0&1&-1\\1&-1&0&-1&0&-2&0&2&0&0&0&2\\1&-1&0&-1&0&-2&0&2&0&0&0&2\\1&-1&0&-1&1&1&-1&-1&0&0&-1&-1\\1&-1&0&-1&1&1&-1&-1&0&0&-1&-1\\1&1&0&-1&-1&1&-1&1&0&0&1&-1\\1&1&0&-1&-1&1&-1&1&0&0&1&-1\\1&1&0&-1&0&-2&0&-2&0&0&0&2\\1&1&0&-1&0&-2&0&-2&0&0&0&2\\1&1&0&-1&1&1&1&1&0&0&-1&-1\\1&1&0&-1&1&1&1&1&0&0&-1&-1\\1&0&-1&1&-1&1&0&0&1&-1&-1&1\\1&0&-1&1&-1&1&0&0&1&-1&-1&1\\1&0&-1&1&0&-2&0&0&0&2&0&-2\\1&0&-1&1&0&-2&0&0&0&2&0&-2\\1&0&-1&1&1&1&0&0&-1&-1&1&1\\1&0&-1&1&1&1&0&0&-1&-1&1&1\\1&0&1&1&-1&1&0&0&-1&1&-1&1\\1&0&1&1&-1&1&0&0&-1&1&-1&1\\1&0&1&1&0&-2&0&0&0&-2&0&-2\\1&0&1&1&0&-2&0&0&0&-2&0&-2\\1&0&1&1&1&1&0&0&1&1&1&1\\1&0&1&1&1&1&0&0&1&1&1&1\end{bmatrix}\end{array}$$

Исследование показывает, что все столбцы матрицы $\mathbf{X}$ взаимно ортогональны, так что $\mathbf{X}'\mathbf{X}$ — диагональная матрица с элементами

$$\mathbf{X}'\mathbf{X}=\operatorname{diag}\{24,\ 12,\ 12,\ 24,\ 16,\ 48,\ 8,\ 24,\ 8,\ 24,\ 16,\ 48\} \tag{9.7.6}$$

и матрица $(\mathbf{X}'\mathbf{X})^{-1}$ также диагональная с элементами $\{1/24,\ 1/12,\ \ldots,\ 1/16,\ 1/48\}$. Затем

$$\mathbf{X}'\mathbf{Y}=(240,\ 12,\ -12,\ 48,\ 0,\ -48,\ 2,\ -18,\ -6,\ -18,\ -20,\ 36)'. \tag{9.7.7}$$

Отсюда следует, что $\mathbf{b}=(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ имеет значения

$$\mathbf{b}=(10,\ 1,\ -1,\ 2,\ 0,\ -1,\ 1/4,\ -3/4,\ -3/4,\ -3/4,\ -5/4,\ 3/4)'. \tag{9.7.8}$$

Сумма квадратов, обусловленная регрессией, $\mathbf{b'X'Y}$, включает следующие двенадцать независимых составляющих, собранных для удобства в группы, каждый элемент которых представляет собой сумму квадратов, обусловленную своим b-коэффициентом:

$$\mathbf{b'X'Y} = 2400 + \{12 + 12 + 96\} + \{0 + 48\} + \left\{1/2 + 13\frac{1}{2} + 4\frac{1}{2} + \right.$$

$$\left. + 13\frac{1}{2} + 25 + 27\right\} = 2400 + 120 + 48 + 84 = 2652. \qquad (9.7.9)$$

Эти результаты позволили нам составить ANOVA-таблицу для рассматриваемого примера (см. табл. 9.5).

Таблица 9.5. ANOVA-таблица для примера с двусторонней классификацией, обрабатываемого с помощью регрессионного анализа

Источник вариации	Степени свободы	SS	MS	F
Между реагентами	3	120	40	10**
A против B	1	12	12	2
C против D	1	12	12	3
$A + B$ против $C + D$	1	96	96	24**
Между катализаторами	2	48	24	6*
1 против 3	1	0***	0	0
(1 + 3) против 2	1	48	48	12**
Реагенты × катализаторы	6	84****	14	3,5*
«Чистая» ошибка	12	48	$s^2 = 4$	
Общий (скорректированный)	23	300		
b_0	1	2400		
Общий	24	2700		

* Значим при уровне $\alpha = 0{,}05$.
** Значим при уровне $\alpha = 0{,}01$.
*** Сумма квадратов может быть равной нулю в точности при использовании реальных данных очень редко, и это часто означает, что данных скомпилированы. В данном случае это произошло потому, что исходные данные были закодированы и числа были округлены, чтобы упростить вычисления.
**** Сумму квадратов можно расщепить, как это видно из уравнения (9.7.9), на отдельные суммы квадратов. Значимый вклад вносят b_{10} и b_{11}, что указывает на существование «действительных» взаимодействий между и X_3 и X_4 а также между X_3 и X_5.

Параметры (9.7.8) позволяют получить модель

$$\hat{Y} = 10 + X_1 - X_2 + 2X_3 + 0X_4 - X_5 + \frac{1}{4}X_6 - \frac{3}{4}X_7 -$$

$$- \frac{3}{4}X_8 - \frac{3}{4}X_9 - \frac{5}{4}X_{10} + \frac{3}{4}X_{11}, \qquad (9.7.10)$$

и это уравнение можно использовать для предсказания скорости производства продукта при различных условиях и для определения остатков. Например, для реагента C и катализатора 2 мы видим, что

$$X_1=0, \quad X_4=0, \quad X_6=0, \quad X_7=0,$$
$$X_2=-1, \quad X_5=-2, \quad X_8=0, \quad X_9=2,$$
$$X_3=1, \quad X_{10}=0, \quad X_{11}=-2,$$

так что

$$\hat{Y}=10+1+2+2-\frac{3}{4}(2)+\frac{3}{4}(-2)=12.$$

В этой ячейке фактические наблюдения равны 15 и 9, так что соответствующие остатки, которые должны давать в сумме нуль, равны 3 и —3. Сумма остатков в каждой ячейке должна равняться нулю, так как модель по существу без остатков подогнана к средним значениям по ячейкам. Это происходит во всех регрессионных ситуациях, когда в каждой ячейке содержится одинаковое число повторных опытов [7].

Дисперсии оцениваемых параметров и предсказываемых величин отклика могут быть получены по обычным формулам регрессионного анализа. Дальнейший анализ данных этого примера читателю полезно провести самостоятельно.

В § 9.8 приведены стандартные выкладки дисперсионного анализа для модели с двусторонней классификацией при равном числе наблюдений в ячейке, а в § 9.9 даны два альтернативных варианта регрессионной обработки. Один из них опирается на принцип дополнительной суммы квадратов, другой — на «несимметричное исключение параметров», что позволяет получить невырожденную матрицу $\mathbf{X}'\mathbf{X}$. Первый из этих методов иллюстрируется на примере в § 9.10.

9.8. ДВУСТОРОННЯЯ КЛАССИФИКАЦИЯ С РАВНЫМ ЧИСЛОМ НАБЛЮДЕНИЙ В ЯЧЕЙКАХ

Предположим, что в нашем распоряжении есть двусторонняя классификация с I строками, J столбцами и K наблюдениями Y_{ijk}, $k=1, 2, \ldots, K$, в ячейках. Обычная модель дисперсионного анализа с фиксированными эффектами факторов есть

$$E(Y_{ijk})=\mu+\alpha_i+\beta_j+\gamma_{ij}, \qquad i=1, 2, \ldots, I,$$
$$j=1, 2, \ldots, J, \qquad k=1, 2, \ldots, K, \tag{9.8.1}$$

с ограничениями

$$\sum_{i=1}^{I}\alpha_i=\sum_{j=1}^{J}\beta_j=\sum_{i=1}^{I}\gamma_{ij}\ (\text{для всех } j)=\sum_{j=1}^{J}\gamma_{ij}\ (\text{для всех } i)=0. \tag{9.8.2}$$

[7] Это связано с тем, что число параметров модели в точности равно числу ячеек.— *Примеч. пер.*

Эти ограничения позволяют рассматривать μ как генеральное среднее, в то время как α, β_j и γ_{ij} есть соответственно разности между эффектом некоторой строки и генеральным средним, эффектом некоторого столбца и генеральным средним, а также между эффектом ячейки и объединенным эффектом строки и столбца. Обычная таблица дисперсионного анализа имеет вид

ANOVA

Источник	Степени свободы	SS	MS
Строки	$I-1$	$JK\sum_{i=1}^{I}(\overline{Y}_{i..}-\overline{Y})^2$	s_r^2
Столбцы	$J-1$	$IK\sum_{j=1}^{J}(\overline{Y}_{.j.}-\overline{Y})^2$	s_c^2
Ячейки (взаимодействие)	$(I-1)(J-1)$	$K\sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{Y}_{ij}-\overline{Y}_{i..}-\overline{Y}_{.j.}+\overline{Y})^2$	s_{rc}^2
Остаток	$IJ(K-1)$	Находится вычитанием	s^2
Среднее	1	$IJK\overline{Y}^2$	
Общий	IJK	$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{k}Y_{ijk}^2$	

где $\overline{Y}_{i..}$ — среднее по всем наблюдениям в i-й строке,
$\overline{Y}_{.j.}$ — среднее по всем наблюдениям в j-м столбце,
$\overline{Y}_{ij}$ — среднее по всем наблюдениям в ячейке (i, j),
$\overline{Y}$ — среднее по всем наблюдениям, содержащимся в эксперименте.

Обычные тесты для проверки гипотез имеют вид:

H_0: все $\alpha_i=0$ $F=S_r^2/s^2$ сравнивается с $F[(I-1),\ IJ(K-1)]$,

H_0: все $\beta_j=0$ $F=s_c^2/s^2$ сравнивается с $F[J-1),\ IJ(K-1)]$,

H_0: все $\gamma_{ij}=0$ $F=s_{rc}^2/s^2$ сравнивается с $F[(I-1)(J-1),\ IJ(K-1)]$.

9.9. РЕГРЕССИОННАЯ ОБРАБОТКА ДВУСТОРОННЕЙ КЛАССИФИКАЦИИ С РАВНЫМ ЧИСЛОМ НАБЛЮДЕНИЙ В ЯЧЕЙКАХ

При желании мы могли бы решить эту задачу подобно тому, как это сделано в § 9.4, записывая *зависимые* нормальные уравнения относительно параметров μ, α_i, β_j и γ_{ij}, добавляя к ним уравнения, задаваемые ограничениями (9.8.2), и решая систему из $(I+1)(J+1)$ отобранных независимых уравнений. Однако мы будем решать данную

задачу другим способом, при котором в вычислениях участвует невырожденная матрица $\mathbf{X}'\mathbf{X}$.

Вообще здесь имеется $1 + I + J + IJ = (I + 1)(J + 1)$ параметров, однако они зависимы благодаря существованию $1 + 1 + J + I - 1 = I + J + 1$ ограничений, определяемых уравнениями (9.8.2). Мы должны исключить одно ограничение из общего числа очевидных (на первый взгляд) ограничений, чтобы сделать поправку, связанную с тем, что если все суммы параметров γ_{ij} *по строкам* равны нулю, то общая сумма всех γ_{ij} также равна нулю. Значит, достаточно указать только, что $J-1$ сумм параметров γ_{ij} по столбцам равны нулю, чтобы обеспечить равенство нулю суммы параметров по последнему столбцу. Таким образом, в действительности необходимо только IJ параметров для описания модели, и мы можем определить последнюю так:

$$\delta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \quad i = 1, 2, \ldots, I, \quad j = 1, 2, \ldots, J.$$

Рассмотрим следующие модели:

а) $E(Y_{ijk}) = \delta_{ij}$,

б) $E(Y_{ijk}) = \delta_{.j}$ независимо от i,

в) $E(Y_{ijk}) = \delta_{i.}$ независимо от j,

г) $E(Y_{ijk}) = \delta$ независимо от i и j.

Мы можем выразить все модели в матричной форме. Пусть

$$\mathbf{Y} = (Y_{111}, Y_{112}, \ldots, Y_{11K}; \quad Y_{121}, Y_{122}, \ldots, Y_{12K};$$

$$\ldots; Y_{IJ1}, Y_{IJ2}, \ldots, Y_{IJK})',$$

где выдержана следующая нумерация ячеек:

$$(11), (12), \ldots, (1J); \quad (21), (22), \ldots, (2J); \quad \ldots;$$

$$(I1), (I2), \ldots, (IJ);$$

третий индекс используется для обозначения порядкового номера наблюдения внутри ячейки. Тогда с помощью матриц, которые мы укажем в дальнейшем, можем записать модели а), б), в) и г) в форме $E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$.

Модель а)

$$\mathbf{X} = \begin{array}{c} \begin{matrix} \delta_{11} & \delta_{12} \ldots & \delta_{1J} & \delta_{21} \ldots & \delta_{2J} & \ldots & \delta_{I1} & \delta_{I2} \ldots & \delta_{IJ} \end{matrix} \\ \left[\begin{array}{ccccccccc} 1 & 0 \ldots 0 & 0 \ldots 0 & & & 0 \ldots 0 & 0 \ldots 0 \\ 1 & 0 \ldots 0 & 0 \ldots 0 & & & 0 \ldots 0 & 0 \ldots 0 \\ \ldots & \ldots & \ldots & & & \ldots & \ldots \\ 1 & 0 \ldots 0 & 0 \ldots 0 & & & 0 \ldots 0 & 0 \ldots 0 \\ \hline 0 & 1 \ldots 0 & 0 \ldots 0 & & & 0 \ldots 0 & 0 \ldots 0 \\ 0 & 1 \ldots 0 & 0 \ldots 0 & & & 0 \ldots 0 & 0 \ldots 0 \\ \ldots & \ldots & \ldots & & & \ldots & \ldots \\ 0 & 1 \ldots 0 & 0 \ldots 0 & & & 0 \ldots 0 & 0 \ldots 0 \\ \hline \ldots & \ldots & \ldots & & & \ldots & \ldots \\ \hline 0 & 0 \ldots 0 & 0 \ldots 0 & & & 0 \ldots 0 & 1 \ldots 1 \\ 0 & 0 \ldots 0 & 0 \ldots 0 & & & 0 \ldots 0 & 0 \ldots 1 \\ \ldots & \ldots & \ldots & & & \ldots & \ldots \\ 0 & 0 \ldots 0 & 0 \ldots 0 & & & 0 \ldots 0 & 0 \ldots 1 \end{array}\right] \end{array},$$

где каждый отрезок столбца содержит K элементов,

$$\boldsymbol{\beta}' = (\delta_{11}, \delta_{12}, \ldots, \delta_{1J}; \delta_{21}, \delta_{22}, \ldots, \delta_{2J}; \ldots; \delta_{I1}, \delta_{I2}, \ldots, \delta_{IJ}).$$

Модель б)

$$\mathbf{X} = \begin{array}{c} \begin{matrix} \delta_{\cdot 1} & \delta_{\cdot 2} & \delta_{\cdot 3} \ldots \delta_{\cdot J} \end{matrix} \\ \left[\begin{array}{ccc} \mathbf{j} & \mathbf{0} & \mathbf{0} \ldots \mathbf{0} \\ \mathbf{0} & \mathbf{j} & \mathbf{0} \ldots \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{j} \ldots \mathbf{0} \\ \ldots & \ldots & \ldots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \ldots \mathbf{j} \\ \hline \text{Другие } (I-1) \text{ блоков точно} \\ \text{такие же, как и выше} \end{array}\right] \end{array},$$

где $\mathbf{j}$ означает $(K \times 1)$-вектор, составленный из единиц

$$\beta' = (\delta_{.1}, \delta_{.2}, \ldots, \delta_{.J}).$$

Модель в)

$$\mathbf{X} = \begin{array}{c} \begin{matrix} \delta_1. & \delta_2. & \delta_3. & \ldots & \delta_J. \end{matrix} \\ \begin{bmatrix} \mathbf{j} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{j} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{j} & \ldots & \mathbf{0} \\ . & . & . & \ldots & . \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \ldots & \mathbf{j} \end{bmatrix} \end{array},$$

где $\mathbf{j}$, как и ранее, означает вектор, размерности $JK \times 1$, составленный из единиц:

$$\beta' = (\delta_{1.}, \delta_{2.}, \ldots, \delta_{I.}).$$

Модель г)

$$\mathbf{X} = \mathbf{j},$$

причем $\mathbf{j}$ есть $(IJK \times 1)$-вектор из единиц, а $\beta = \delta$ является скаляром.

Можно составить стандартную таблицу дисперсионного анализа, используя регрессионный анализ, следующим образом. Обозначим через S_1, S_2, S_3 и S_4 суммы квадратов, обусловленные регрессиями, указанными выше. И пусть $S = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} Y_{ijk}^2$. Тогда, применяя принцип «дополнительной суммы квадратов», рассмотренный в § 2.7, построим следующую таблицу:

ANOVA

Источник	Степени свободы	MS
Строки	$I-1$	$S_3 - S_4$
Столбцы	$J-1$	$S_2 - S_4$
Взаимодействия	$(I-1)(J-1)$	$S_1 - S_2 - S_3 + S_4$
Остаток	$IJ\ (K-1)$	$S - S_1$
Среднее	1	S_4
Общий	IJK	S

(Сумма квадратов, обусловленная взаимодействиями, в действительности получается из соотношения

$$S_1 - (S_2 - S_4) - (S_3 - S_4) - S_4,$$

которое сводится к выражению, указанному в таблице.)

Эквивалентность этих сумм квадратов и тех, которые можно получить с помощью процедуры дисперсионного анализа, может быть

легко доказана математически, но мы на этом не будем здесь останавливаться.

Обычно нас интересуют оценки m, a_i, b_j и c_{ij} истинных параметров μ, α_i, β_j и γ_{ij} дисперсионной модели. Эти оценки можно найти, исходя из оценок d_{ij}, $d_{.j}$, $d_{i.}$ и d регрессионных коэффициентов δ_{ij}, $\delta_{.j}$, $\delta_{i.}$ и δ четырех моделей. А именно:

$$m = d,$$
$$a_i = d_{i.} - d,$$
$$b_j = d_{.j} - d,$$
$$c_{ij} = d_{ij} - d_{i.} - d_{.j} + d.$$

Альтернативный метод

Предложенный выше метод рассмотрения двусторонней классификации дисперсионного анализа включал четыре симметричные процедуры регрессионного анализа и опирался на принцип дополнительной суммы квадратов. Чтобы ограничиться одной процедурой регрессионного анализа, нужно записать несимметричную модель, не содержащую некоторые из зависимых параметров стандартной модели регрессионного анализа. Проиллюстрируем это на примере. Рассмотрим двустороннюю классификацию, при которой в каждой ячейке содержится по два наблюдения.

	Столбец $j = 1$	Столбец $j = 2$
Строка $i = 1$	Y_1, Y_2	Y_3, Y_4
Строка $i = 2$	Y_5, Y_6	Y_7, Y_8
Строка $i = 3$	Y_9, Y_{10}	Y_{11}, Y_{12}

Стандартная модель дисперсионного анализа имеет вид

$$E(Y_{ij}) = \mu + \alpha_i + \beta_j + \gamma_{ij},$$

где

$$\alpha_1 + \alpha_2 + \alpha_3 = 0, \qquad \gamma_{11} + \gamma_{12} = 0,$$
$$\beta_1 + \beta_2 = 0, \qquad \gamma_{21} + \gamma_{22} = 0,$$
$$\gamma_{31} + \gamma_{32} = 0,$$
$$\gamma_{11} + \gamma_{21} + \gamma_{31} = 0,$$
$$\gamma_{12} + \gamma_{22} + \gamma_{32} = 0.$$

Следовательно, если, например, параметры μ, α_1, α_2, β_1, γ_{11} и γ_{21} известны или найдены их оценки, то все другие параметры, точнее их оценки, можно найти из ограничений. Таким образом, можно

записать регрессионную модель

$$E(Y_{ij}) = \mu + \alpha_1 X_1 + \alpha_2 X_2 + \beta_1 X_3 + \gamma_{11} X_4 + \gamma_{21} X_5,$$

или

$$E(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta},$$

где

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \\ Y_9 \\ Y_{10} \\ Y_{11} \\ Y_{12} \end{bmatrix}, \quad \mathbf{X} = \begin{array}{c} \begin{matrix} \mu & \alpha_1 & \alpha_2 & \beta_1 & \gamma_{11} & \gamma_{21} \end{matrix} \\ \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 1 & 0 & -1 & -1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & 0 & 1 & -1 & 0 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & -1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & -1 & 1 & 1 \end{bmatrix} \end{array}$$

$$\boldsymbol{\beta}' = (\mu,\ \alpha_1,\ \alpha_2,\ \beta_1,\ \gamma_{11},\ \gamma_{21}).$$

(Примечание. Элементы столбцов, соответствующих γ_{ij}, получаются в результате перемножения соответствующих элементов столбцов, отвечающих α_i и β_j.)

Для получения такой модели могут использоваться любые независимые подмножества параметров, и поэтому можно построить много разных форм модели. Для оценивания β здесь применимы обычные регрессионные методы. Вследствие ортогональности некоторых столбцов матрицы **X** удается получить раздельные, ортогональные суммы квадратов для оценок 1) μ, 2) α_1 и α_2, 3) β_1, 4) γ_{11} и γ_{21}. Это будут обычные суммы квадратов для 1) среднего, 2) строк, 3) столбцов, 4) взаимодействия в стандартном дисперсионном анализе.

9.10. ПРИМЕР. ДВУСТОРОННЯЯ КЛАССИФИКАЦИЯ

Приводимые ниже данные для случая двусторонней классификации взяты из книги Браунли: Brownlee K. A. Statistical Theory and Methodology in Science and Engineering, second edition.— New York: Wiley, 1965, p. 475. Описательные детали не приводятся.

	Столбец 1	Столбец 2	Столбец 3
Строка 1	17, 21, 49, 54	64, 48, 34, 63	62, 72, 61, 91
Строка 2	33, 37, 40, 16	41, 64, 34, 64	56, 62, 57, 72

Следуя процедуре, изложенной в § 9.9, можно вычислить указанные ниже величины, пользуясь регрессионными методами.

а) $S_1 = 65863$; $\begin{pmatrix} d_{11} & d_{12} & d_{13} \\ d_{21} & d_{22} & d_{23} \end{pmatrix} = \begin{pmatrix} 35{,}25 & 52{,}25 & 71{,}50 \\ 31{,}50 & 50{,}75 & 61{,}75 \end{pmatrix}$,

б) $S_2 = 65\,640{,}25$; $(d_{\cdot 1},\ d_{\cdot 2},\ d_{\cdot 3}) = (33{,}375;\ 51{,}5;\ 66{,}625)$,

в) $S_3 = 61\,356$; $\begin{pmatrix} d_{1\cdot} \\ d_{2\cdot} \end{pmatrix} = \begin{pmatrix} 53 \\ 48 \end{pmatrix}$,

г) $S_4 = 61\,206$; $d = 50{,}5$.

ANOVA

Источник	Степени свободы	SS
Строки	1	150,00
Столбцы	2	4 434,25
Взаимодействия	2	72,75
Остаток	18	3 495,00
Среднее	1	61 206,00
Общий (нескорректированный)	24	69 358,00

Точно такая же таблица была получена Браунли (1965) с помощью обычной процедуры дисперсионного анализа [8]. Оценки параметров в случае обычного дисперсионного анализа равны:

$$m = 50{,}5,$$

$$a_1 = 53 - 50{,}5 = 2{,}5, \qquad a_2 = 48 - 50{,}5 = -2{,}5.$$

(Примечание. $a_1 + a_2 = 0$, как и должно быть.)

$$b_1 = 33{,}375 - 50{,}5 = -17{,}125, \qquad b_2 = 51{,}5 - 50{,}5 = 1{,}0,$$

$$b_3 = 66{,}625 - 50{,}5 = 16{,}125.$$

(Примечание. $b_1 + b_2 + b_3 = 0$, как и должно быть.)

$$c_{11} = 35{,}25 - 53 - 33{,}375 + 50{,}5 = -0{,}625,$$

[8] Есть русский перевод: Браунли К. Статистическая теория и методология в науке и технике/Пер. с англ. Под ред. Л. Н. Большева.— М.: Наука, 1977.— 408 с. Однако раздел, из которого заимствован данный пример, в русском переводе опущен.— *Примеч. пер.*

$$c_{12} = 52{,}25 - 53 - 51{,}5 + 50{,}5 = -1{,}750,$$
$$c_{13} = 71{,}50 - 53 - 66{,}625 + 50{,}5 = 2{,}375,$$
$$c_{21} = 31{,}50 - 48 - 33{,}375 + 50{,}5 = 0{,}625,$$
$$c_{22} = 50{,}75 - 48 - 51{,}5 + 50{,}5 = 1{,}750,$$
$$c_{23} = 61{,}75 - 48 - 66{,}625 + 50{,}5 = -2{,}375.$$

(Примечание. $\sum_{i=1}^{2} c_{ij} = \sum_{j=1}^{3} c_{ij} = 0$, как и следовало ожидать.)

Остатки от этой модели дисперсионного анализа выражаются соотношением

$$e_{ijk} = Y_{ijk} - m - a_i - b_j - c_{ij} = Y_{ijk} - d - (d_{i\cdot} - d) - (d_{\cdot j} - d) - (d_{ij} - d_{i\cdot} - d_{\cdot j} + d) = Y_{ijk} - d_{ij};$$

они также будут остатками и в случае регрессионного анализа с использованием модели а). При желании их можно исследовать методами, описанными в гл. 3. Могут быть также рассмотрены графики остатков для каждой строки и каждого столбца [9].

9.11. КОММЕНТАРИИ

На ряде специально рассмотренных примеров мы видели, что дисперсионный анализ при необходимости может быть выполнен с использованием стандартных регрессионных процедур. Если модель исследуется тщательно и надлежащим образом репараметризована, то таблицу дисперсионного анализа получают подобным же образом и для других моделей. Надлежащий выбор фиктивных переменных является решающим для удачного представления результатов и позволяет заметно упростить вычисления. Однако многие репараметризации представляют самостоятельную ценность при работе с фиктивными переменными; при этом в определенных ситуациях некоторые репараметризации оказываются проще других. Чем более сложный план, тем более сложной будет регрессионная обработка и тем больше размеры матрицы **X**. Могут потребоваться значительные усилия, но если план стандартный и все данные имеются в наличии (т. е. нет пропущенных наблюдений), они, как правило, невелики. На практике лучше прибегать к соответствующей вычислительной процедуре дисперсионного анализа или к машинной программе, если она есть. Тем не менее полезно установить связь между этими двумя методами анализа по следующим причинам:

1. Надо акцентировать внимание на том, что в задачах дисперсионного анализа непременно присутствует модель.

2. Остатки в случае модели дисперсионного анализа играют ту же роль, что и остатки в регрессионном анализе, и они должны ис-

[9] Об анализе остатков для моделей дисперсионного анализа см., например: Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия/Пер. с англ. Под ред. Ю. П. Адлера.— М.: Финансы и статистика, 1982.— *Примеч. пер.*

следоваться, чтобы извлечь содержащуюся в них информацию о возможной неадекватности модели. (По-видимому, в большинстве случаев дисперсионного анализа молчаливо предполагается, что модель выбрана правильно.)

3. Если некоторые из наблюдений, необходимых для дисперсионного анализа, отсутствуют, то зачастую их удается «оценить» с помощью стандартных формул. Если же это неудобно или отсутствует слишком много наблюдений, то обычно данные можно анализировать с помощью регрессионной процедуры, как это было показано выше, но при условии вычеркивания из матрицы X строк, для которых нет результатов наблюдений.

(Термин «оценены» взят в кавычки, поскольку никакое оценивание в действительности не проводится. «Оценки» — это просто включаемые в рассмотрение числа (исходя из вычислительных целей), чтобы получить те же оценки параметров, которые могут быть получены на основе неполных даиных с помощью несимметричного и зачастую трудного анализа.)

Упражнения

1. Для каждой из двух двусторонних классификаций а) и б), указанных ниже, выполните следующие процедуры:

1) Проанализируйте эти даиные, используя любой из методов, описаниых в гл. 9.

2) Вычислите предсказанные значения и исследуйте остатки всеми возможиыми методами. Зафиксируйте любые обнаруженные вами аномалии.

3) Убедитесь, что альтернативные регрессионные методы, рассмотренные в гл. 9, приводят к одинаковым результатам.

а) Проводился эксперимент для выявлення влияния давлення пара и времени продувки на процентное содержаиие посторонней примеси, остающейся в фильтруемом материале. Результаты эксперимента представлены в таблице.

Давленне пара (фунты)	Процентное содержанне посторонней примеси		
	Время продувкн (часы)		
	1	2	3
10	45,2; 46,0	40,0; 39,0	35,9; 34,1
20	41,8; 20,6	27,8; 19,0	22,5; 17,7
30	23,5; 33,1	44,6; 52,2	42,7; 48,6

б) Для устаиовлеиия влияиии скорости предварительного перемешивании (начальной скорости) и конечной скорости смесителя (миксера) иа высоту центральиой части кекса проведеи эксперимент. Для каждой из двух перемеиных было выбрано по три значеиия скорости [10]. Получеиы следующие данные:

[10] В процессе выпечки кекса наблюдается характерный подъем центральной части, косвенно характеризующий качество. При прочих равных условиях этот подъем, обусловленный припеком, зависит от качества перемешивания теста в миксере. При этом число оборотов миксера изменяется по ходу перемешиваиия. Характеризовать его можно двуми значениями — начальным и коиечным. В обсуждаемом эксперименте исследуютси различные эффекты комбинаций скоростей перемешивании теста.— *Примеч. пер.*

(Начальная скорость — 5)	(Конечная скорость — 3,5) × 2	(Высота центральной части — 2) × 100
X_1	X_2	Y
−1	−1	4; −3
−1	0	3; 2
−1	1	−1; −5
0	−1	3; 10
0	0	2; 2
0	1	0; 0
1	−1	−1; −10
1	0	1; 2
1	1	7; 9

2. Проведен химический эксперимент по исследованию влияния температуры экструзии X_1 и температуры охлаждения X_2 на упругость получаемого полимерного изделия. На основании имеющихся сведений о процессе было принято предположение, что наблюдаемые вариации могут быть вполне удовлетворительно описаны моделью следующего вида:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 + \varepsilon.$$

Были выбраны два уровня температуры экструзии, два уровня температуры охлаждения и реализованы все четыре комбинации. Каждый из этих опытов повторялся четырежды. Получены следующие данные:

ANOVA

Источник рассеяния	Степени свободы	SS	MS
Общий	16	921,0000	
Обусловленный регрессией		881,2500	
b_0	1	798,0625	
b_1	1	18,0625	
b_2			
b_{12}		5,0625	
Остаток			

1) а) Заполните приведенную выше таблицу.
б) Приняв $\alpha = 0{,}05$, ответьте на следующие вопросы: является ли регрессионное уравнение в целом статистически значимым? все ли b-коэффициенты значимы?
в) Вычислите квадрат множественного коэффициента корреляции R^2.
2) При наличии следующей дополнительной информации:

$$\Sigma x_{0i} y_i = 113, \qquad \Sigma x_{1i} y_i = 17, \qquad \Sigma x_{2i} y_i = 31 \quad \text{и} \quad \Sigma x_{1i} x_{2i} y_i = -9,$$

где

$$x_{ji} = X_{ji} - \overline{X}_j, \; j = 0, 1, 2 \quad \text{и} \quad y_i = Y_i - \overline{Y},$$

выполните следующие операции:
а) Определите b_0, b_1, b_2 и b_{12} и выпишите полученное уравнение регрессии.
б) Предсказываемая величина $\hat{Y}$ при $X_1 = 70°$ и при $X_2 = 150°$ имеет значение 54. Дисперсия этой величины равна 0,6875. Какова дисперсия предсказываемого значения отклика в точке $X_1 = 70°$, $X_2 = 150°$?

в) Определите 95 %-ные доверительные пределы для истинного среднего значения величины Y в точке $X_1 = 70°$, $X_2 = 150°$.

3) К каким выводам можно прийти по результатам проведенного анализа?

3. Проанализируйте данные, приведенные на с. 183, используя альтернативный метод, изложенный на с. 181—182.

4. (Источник: Keeping E. S. Introduction to Statistical Inference.— Princeton: Van Nostrand, NJ, 1962, p. 216.

Примените регрессионный метод либо из § 9.4, либо из § 9.5 к двусторонней классификации, описанной ниже. Выполните полностью все этапы вычислений в соответствии с выбранным методом вместо того, чтобы сразу приступать к анализу дисперсионной таблицы.

Были получены данные для проверки предположения о влиянии малых количеств угольной пыли, добавляемой в песок, на качество бетона, изготавливаемого из такого песка. Несколько партий были замешаны при практически одинаковых условиях за исключением содержания угольной пыли. Из каждой партии было изготовлено по 4 цилиндрических образца, которые были испытаны затем на разрушающую нагрузку (фунты/дюйм^2). Один цилиндр в третьей партии оказался дефектным, так что эта партия характеризуется тремя результатами.

Номер партии	1	2	3	4	5
Содержание угля, %	0	0,05	0,1	0,5	1,0
Разрушающая нагрузка	1690 1580 1745 1685	1550 1445 1645 1545	1625 1450 1510	1725 1550 1430 1445	1530 1545 1565 1520

(Подсказка. Работайте с $Y_{ij} - 1430$, это легче. Ответьте для себя на вопрос: как это отразится на анализе?)

5. Важной операцией в производстве бумаги является удаление из нее воды. Применительно к некоторому конкретному процессу предполагалось, что на содержание воды, остающейся в бумаге, могут влиять четыре фактора. Чтобы исследовать это, было решено провести полный факторный эксперимент 2^4. Факторы и по два уровня для каждого из них представлены ниже.

Фактор	Обозначение	Нижний уровень	Верхний уровень
Вакуум в первом барабане	A	0	18
Вакуум во втором барабане	B	0	19
Вес бумаги	C	10,0	13,7
Скорость движения	D	1700	2000

Полученные данные приведены в таблице. Обозначение *bcd*, например, означает, что в этом опыте фактор A поддерживается на нижнем уровне, в то

время как факторы B, C и D — на их верхних уровнях. Обозначение «1» относится к опыту, в котором все факторы имеют нижние уровни.

Номер опыта	Комбинация факторов	Количество удаленной воды, % Y	Номер опыта	Комбинация факторов	Количество удаленной воды, % Y
1	*bcd*	39,7	9	*bc*	38,9
2	*abcd*	41,1	10	*ac*	40,0
3	*cd*	40,6	11	*abc*	41,0
4	*acd*	40,4	12	*c*	42,9
5	*ad*	41,0	13	*a*	40,2
6	*bd*	37,6	14	*b*	35,4
7	*d*	38,7	15	*ab*	39,4
8	*abd*	39,0	16	1	39,0

1) Проанализируйте эти данные, используя регрессионный подход.

2) Покажите, что обычная обработка полного факторного эксперимента и использование модели дисперсионного анализа для этих данных приводят в точности к одним и тем же результатам и/или выводам.

Ответы к упражнениям

1. 1) Оба метода анализа приводят к следующей таблице дисперсионного анализа:

ANOVA

Источник	Степени свободы	SS	MS	F	$F_{0,95}$
Общий	18	24 403,750			
Среднее	1	22 352,027			
Общий (скорректированный)	17	2 051,723			
Давление пара	2	963,721	481,861	11,728*	4,26
Время продувки	2	37,481	18,741	0,456	4,26
Взаимодействие	4	680,756	170,189	4,142*	3,63
«Чистая» ошибка	9	369,765	41,085		

Следовательно, эффекты давления пара и взаимодействия давления пара и времени продувки статистически значимы. Регрессионное уравнение, полученное для этой задачи, имеет вид

$$\hat{Y} = 35{,}239 + 4{,}794\ X_1 - 10{,}339\ X_2 - 0{,}206\ X_3 + 1{,}861\ X_4 + \\ + 5{,}773\ X_1X_3 - 2{,}394\ X_1X_4 + 6{,}506\ X_2X_3 - 3{,}361\ X_2X_4,$$

где X_1, X_2 — фиктивные переменные, связанные с давлением пара следующим образом:

X_1	X_2	
1	0	давление пара 10 фунтов,
0	1	давление пара 20 фунтов,
—1	—1	давление пара 30 фунтов.

а X_3, X_4 — фиктивные переменные, связанные со временем продувки:

X_3	X_4	
1	0	время продувки 1 час,
0	1	время продувки 2 часа,
—1	—1	время продувки 3 часа.

Анализ остатков показывает, что отклики имеют наименьшие дисперсии, когда давление пара находится на нижнем уровне. Поскольку в каждой ячейке только по два повторных опыта, анализ остатков не обязательно будет несостоятельным, однако он, несомненно, указывает на необходимость дальнейшего исследования. Чтобы оценить эффект взаимодействия, надо исследовать средние значения из таблицы или вычертить графики средних значений отклика в зависимости от времени продувки при каждом фиксированном значении давления пара.

2) ANOVA

Источник	Степени свободы	SS	MS	F	$F_{0,95}$
Общий	18	417,000			
Среднее	1	34,722			
Общий (скорректированный)	17	382,278			
Начальная скорость смесителя	2	24,111	12,055	1,080	4,26
Конечная скорость смесителя	2	7,444	3,722	0,333	4,26
Взаимодействие	4	250,223	62,556	5,602*	3,63
«Чистая» ошибка	9	100,500	11,167		

Следовательно, в этом эксперименте статистически значим только член взаимодействия. Полученное уравнение регрессии имеет вид:

$$\hat{Y} = 1{,}39 - 1{,}39X_1 + 1{,}44\ X_2 - 0{,}89\ X_3 + 0{,}61\ X_4 + 1{,}39\ X_1X_2 + \\ + 1{,}89\ X_1X_4 + 4{,}56\ X_2X_3 - 1{,}44\ X_2X_4,$$

где X_1, X_2 — фиктивные переменные, связанные с начальной скоростью, а X_3, X_4 — фиктивные переменные, связанные с конечной скоростью:

X_1	X_2		X_3	X_4	
1	0	начальная скорость 1	1	0	конечная скорость 1
0	1	начальная скорость 2	0	1	конечная скорость 2
—1	—1	начальная скорость 3	—1	—1	конечная скорость 3

Анализ остатков показывает, что наибольшая вариабельность откликов имеет место на нижнем уровне конечной скорости смесителя. Это надо исследовать.

Значимость взаимодействия наиболее легко прослеживается с помощью следующей таблицы:

Таблица средних значений отклика

		X_2		
		—1	0	+1
X_1	—1	0,5	2,5	—3
	0	6,5	2,0	0
	+1	—5,5	1,5	8

2.
1а)

ANOVA

Источник	Степени свободы	SS	MS	F
Общий	16	921,0000		
Регрессия	4	881,2500	220,3125	66,51*
b_0	1	798,0625	798,0625	
b_1	1	18,0625	18,0625	5,45*
b_2	1	60,0625	60,0625	18,13*
b_{12}	1	5,0625	5,0625	1,53НЗ
Остаток	12	39,7500	3,3125	

б) Регрессиоиное уравнение значимо.
Все коэффнциенты значнмы, за нсключением b_{12}.
в) $R^2 = 95{,}68\ \%$.
2а)

$$b_0 = \frac{798{,}0625}{113} = 7{,}0625,$$

$$b_1 = \frac{18{,}0625}{17} = 1{,}0625,$$

$$b_2 = \frac{60{,}0625}{31} = 1{,}9375,$$

$$b_{12} = \frac{5{,}0625}{-9} = -0{,}5625;$$

$$\hat{Y} = 7{,}0625 + 1{,}0625X_1 + 1{,}9375X_2 - 0{,}5625X_1X_2.$$

б)

$$s^2(\mathbf{X'CX}) = 0{,}6875,$$

$$(3{,}3125)(\mathbf{X'CX}) = 0{,}6875,$$

$$\mathbf{X'CX} = \frac{0{,}6875}{3{,}3125} = 0{,}207547.$$

Дисперсия отдельного наблюдения равна:

$$s^2(1 + \mathbf{X'CX}) = (3{,}3125)(1 + 0{,}207547) = 4{,}0000.$$

в) $\hat{Y}$ есть 54 при $X_1 = 70°$ и $X_2 = 150°$.

$$V(\hat{Y}) = 0{,}6875.$$

Доверительные пределы для истинного среднего значения величины Y равны:

$$\hat{Y} \pm (11;\ 0{,}95)\, s(\hat{Y}) = 54 \pm (2{,}201)\sqrt{0{,}6875} = 54 \pm (2{,}201)(0{,}8292) =$$
$$= 54 \pm 1{,}8251.$$

3а) На основании этого аиализа можно прийти к уравнению

$$\hat{Y} = 7{,}0625 + 1{,}0625X_1 + 1{,}9375X_2 - 0{,}5625X_1X_2.$$

б) Член взаимодействия, т. е. переменная X_1X_2, статистически незначим при уровне $\alpha = 0{,}05$. Следовательно, есть основание сомневаться в правильности постулированной модели. Однако такие сомнения базируются на малом числе наблюдений, $n = 16$, а первоначальная модель была основана на знаниях химика. Прежде чем исключить эффект взаимодействия X_1X_2, надо провести дополнительную экспериментальную работу. В подобном случае следует руководствоваться утверждением: «Даже если вы предполагаете, что переменная статистически незначима, не следует считать, что она не оказывает никакого влияния на результат эксперимента».

3. Решение не приводится.

4. Наше решение опирается на метод из § 9.5. Преобразованные данные:

Номера партий	1	2	3	4	5
$Y_{lj} - 1430$	260 150 315 255	120 15 215 115	195 20 80	295 120 0 15	100 115 135 90
$J\bar{Y}_l$	980	465	295	430	440
J_l	4	4	3	4	4
$\bar{Y}_l = b_l$	245	116,25	98,33	107,5	110

На основании уравнений (9.5.5) и (9.5.6) остаточная сумма квадратов SS равна: 524 450—417 788,6 = 106 661,4. Чтобы проверить гипотезу о равенстве эффектов групп, исходя из уравнений (9.5.10) и (9.5.6), получим SS, обусловленную гипотезой H_0 : (524 450—358 531,6) — 106 661,4 = 59 257.

$$F = \frac{\frac{1}{4}\,59\,257}{\frac{1}{14}\,106\,661{,}4} = 1{,}94 < F(4;\ 14;\ 0{,}95) = 3{,}11.$$

Гипотеза H_0 не отвергается. Отсюда мы заключаем, что полученные данные не дают основания считать, что добавки угольной пыли значимо влияют на прочность бетона.

Вычисления с использованием $Y_{lj} - 1430$ отражаются только на полной сумме квадратов и на корректирующем факторе. Они не сказываются на SS, обусловленных вариациями между группами и внутри групп.

5. Решение не приводится.

Глава 10 ● ВВЕДЕНИЕ В НЕЛИНЕЙНОЕ ОЦЕНИВАНИЕ

10.0. ВВЕДЕНИЕ

Эта глава представляет собой краткое введение в проблемы нелинейного оценивания. Поскольку представления о геометрии метода наименьших квадратов позволяют глубже понять проблемы нелинейного оценивания, в § 10.5 и 10.6, кратко обсуждается эта геометрия. В конце книги приведена библиография, которая содержит список многих важных публикаций по нелинейному оцениванию.

В предыдущих главах метод наименьших квадратов применялся для построения моделей, которые были *линейными относительно параметров* и имели вид

$$Y = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \ldots + \beta_p Z_p + \varepsilon, \qquad (10.0.1)$$

где Z_i — некоторые функции от основных предикторных переменных $X_1, X_2, \ldots, X_k$. Хотя уравнение (10.0.1) может описывать весьма широкий класс разнообразных задач (см. гл. 5), существует много ситуаций, в которых модель такого вида непригодна; например при наличии определенной информации о форме связи отклика с предикторными переменными. Такая информация может включать непосредственные знания о действительной форме истинной модели или может быть выражена в виде системы дифференциальных уравнений, которым должна удовлетворять модель. Иногда информация приводит к нескольким альтернативным моделям; в подобных случаях представляют интерес методы дискриминации моделей. Если мы приходим к заключению, что модель имеет нелинейную форму, то мы должны воспользоваться для описания именно такой моделью, а не более простой, но, возможно, менее реалистической линейной моделью.

В отличие от линейных моделей вида уравнения (10.0.1) в *нелинейных моделях* зависимость отклика от коэффициентов при предикторах оказывается нелинейной, такие модели называют также нелинейными *по параметрам*. Вот два примера таких моделей:

$$Y = \exp(\theta_1 + \theta_2 t^2 + \varepsilon), \qquad (10.0.2)$$

$$Y = \frac{\theta_1}{\theta_1 - \theta_2}\left[e^{-\theta_2 t} - e^{-\theta_1 t}\right] + \varepsilon. \qquad (10.0.3)$$

В этих примерах параметры, подлежащие оцениванию, обозначаются буквой θ, а не β, которая использовалась прежде; t — единственная предикторная переменная, а ε — случайная ошибка, удовлетворяющая обычным предположениям: $E(\varepsilon) = 0$, $V(\varepsilon) = \sigma^2$. (Мы можем

также записать эти модели без ε и заменить Y на η. Тогда модели будут отражать истинную зависимость величины отклика η от t. Здесь мы хотим, однако, обратить внимание на то, каким образом входит ошибка в каждую конкретную модель.)

Обе модели (10.0.2) и (10.0.3) нелинейны в том смысле, что параметры θ_1 и θ_2 входят в них нелинейно, но существенно отличаются одна от другой. Уравнение (10.0.2) может быть приведено путем логарифмирования по основанию e к форме

$$\ln Y = \theta_1 + \theta_2 t^2 + \varepsilon, \tag{10.0.4}$$

которая имеет вид (10.0.1) и *линейна* относительно параметров. Мы можем, таким образом, сказать, что модель (10.0.2) *внутренне линейна*, так как она может быть преобразована к линейному виду. (Некоторые авторы в отличие от нас называют такие модели *внешне нелинейными*.)

Однако уравнение (10.0.3) невозможно преобразовать к форме, линейной по параметрам. Такую модель называют *внутренне нелинейной*. Между тем иногда может оказаться полезным преобразовать модель этого типа таким образом, чтобы получить модель более удобную для подгонки, хотя она и будет сохранять нелинейную форму. Если это не будет оговариваться специально, все модели, упоминаемые в данной главе, будут внутренне нелинейными [1].

(Примечание. Среди моделей с аддитивными ошибками внутренне линейной моделью называют такую, которая может быть приведена к линейному виду путем преобразования параметров. К такому типу относится, например, модель $Y = e^{\theta}X + \varepsilon$, поскольку, используя подстановку $\beta = e^{\theta}$, можно записать эту модель в виде $Y = \beta X + \varepsilon$. Некоторые авторы используют термин *внутренне линейная* только для моделей такого типа.)

10.1. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ В НЕЛИНЕЙНОМ СЛУЧАЕ

Стандартные обозначения в задачах нелинейного МНК-оценивания отличаются от обозначений, используемых в задачах с использованием линейного МНК. Поначалу это может смущать читателя, но такие обозначения прочно утвердились в литературе. Различия между обозначениями отражены ниже, в табл. 10.1.

---

[1] Нелинейным по параметрам моделям пока посвящено относительно мало публикаций на русском языке. Среди них надо отметить следующие издания: Бард Й. Нелинейное оценивание параметров/Пер. с англ. Под ред. В. Г. Горского.— М.: Статистика, 1979.— 351 с.; Демиденко Е. З. Линейная и нелинейная регрессии.— М.: Финансы и статистика, 1981.— 304 с.; Горский В. Г. Планирование кинетических экспериментов.— М.: Наука, 1984.— 241 с.; Горский В. Г., Сарылов В. Н., Адлер Ю. П.— Математический анализ и планирование эксперимента при исследовании кинетики химических реакций.— М.: Рукопись деп. в ВИНИТИ, 1975, № 3129—75— Деп.— 50 с. Проблема оценивания нелинейных моделей относится к числу наиболее актуальных и быстро развивающихся в настоящее время.— *Примеч. пер.*

Таблица 10.1 **Стандартные обозначения для линейного и нелинейного методов наименьших квадратов**

Линейный	Нелинейный
Отклик Y	Y
Порядковый номер, индекс наблюдения $i = 1, 2, \ldots, n$	$u = 1, 2, \ldots, n$
Предикторные переменные $X_1, X_2, X_3, \ldots, X_k$	$\xi_1, \xi_2, \ldots, \xi_k$ (иногда t — время, T — температура и т. д., иной раз даже $X_1, X_2, \ldots, X_k$)
Параметры $\beta_0, \beta_1 \ldots, \beta_p$	$\theta_1, \theta_2, \ldots, \theta_p$ (иногда $\alpha, \beta, \ldots, \psi$ и т. д.)

Предположим, что постулированная модель имеет форму

$$Y = f(\xi_1, \xi_2, \ldots, \xi_k; \theta_1, \theta_2, \ldots \theta_p) + \varepsilon. \qquad (10.1.1)$$

Если ввести обозначения

$$\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_k)', \qquad \boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)',$$

то уравнение (10.1.1) можно записать в виде

$$Y = f(\boldsymbol{\xi}, \boldsymbol{\theta}) + \varepsilon$$

или

$$E(Y) = f(\boldsymbol{\xi}, \boldsymbol{\theta}), \qquad (10.1.2)$$

если мы предполагаем, что $E(\varepsilon) = 0$. Относительно ошибок принимаются обычные предположения, что они не коррелированы, что $V(\varepsilon) = \sigma^2$ и, как обычно, что $\varepsilon \sim N(0, \sigma^2)$, а значит, они и независимы.

Если имеется n наблюдений в виде

$$Y_u; \ \xi_{1u}, \xi_{2u}, \ldots, \xi_{ku}$$

для $u = 1, 2, \ldots, n$, то можно записать модель в другой форме:

$$Y_u = f(\xi_{1u}, \xi_{2u}, \ldots, \xi_{ku}; \theta_1, \theta_2, \ldots, \theta_p) + \varepsilon_u, \qquad (10.1.3)$$

где ε_u есть ошибка в опыте с номером u, $u = 1, 2, \ldots, n$.

Это выражение можно записать более кратко:

$$Y_u = f(\boldsymbol{\xi}_u, \boldsymbol{\theta}) + \varepsilon_u, \qquad (10.1.4)$$

где $\boldsymbol{\xi}_u = (\xi_{1u}, \xi_{2u}, \ldots, \xi_{ku})'$.

Предположение о нормальности и независимости ошибок может быть выражено так: $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$, где $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)'$ и, как обычно, $\mathbf{0}$ — нулевой вектор, а $\mathbf{I}$ — единичная матрица соответствующей размерности.

Запишем теперь *сумму квадратов ошибок* для нелинейной модели:

$$S(\boldsymbol{\theta}) = \sum_{u=1}^{n} \{Y_u - f(\boldsymbol{\xi}_u, \boldsymbol{\theta})\}^2. \qquad (10.1.5)$$

Поскольку Y_u и $\boldsymbol{\xi}_u$ фиксированы, сумма квадратов есть функция от $\boldsymbol{\theta}$. Будем обозначать буквой $\hat{\boldsymbol{\theta}}$ *МНК-оценку* вектора $\boldsymbol{\theta}$, т. е. такую ве-

личину $\boldsymbol{\theta}$, которая минимизирует $S(\boldsymbol{\theta})$. (Можно показать, что если $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$, то МНК-оценка $\boldsymbol{\theta}$ становится также оценкой максимального правдоподобия для данного вектора: функцию правдоподобия для такой задачи можно записать в виде

$$l(\boldsymbol{\theta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} e^{-S(\boldsymbol{\theta})/2\sigma^2},$$

так что если σ^2 известна, то максимизация $l(\boldsymbol{\theta}, \sigma^2)$ по отношению к $\boldsymbol{\theta}$ эквивалентна минимизации $S(\boldsymbol{\theta})$ по отношению к $(\boldsymbol{\theta})$.

Чтобы найти МНК-оценку $\hat{\boldsymbol{\theta}}$, мы должны продифференцировать (10.1.5) по $\boldsymbol{\theta}$. Это дает *p нормальных уравнений* относительно $\hat{\boldsymbol{\theta}}$:

$$\sum_{u=1}^{n} \{Y_u - f(\boldsymbol{\xi}_u, \hat{\boldsymbol{\theta}})\} \left[\frac{\partial f(\boldsymbol{\xi}_u, \boldsymbol{\theta})}{\partial \theta_i}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0, \quad i = 1, 2, \ldots, p, \tag{10.1.6}$$

где величина, заключенная в квадратные скобки, есть производная от $f(\boldsymbol{\xi}_u, \boldsymbol{\theta})$ по θ_i при условии, что в этом выражении вектор $\boldsymbol{\theta}$ заменен на вектор $\hat{\boldsymbol{\theta}}$. Напомним, что если бы функция $f(\boldsymbol{\xi}_u, \boldsymbol{\theta})$ была линейной (по параметрам), то эта производная была бы функцией только от $\boldsymbol{\xi}_u$ и вовсе не зависела бы от $\boldsymbol{\theta}$. Так, например, если

$$f(\boldsymbol{\xi}_u, \boldsymbol{\theta}) = \theta_1 \xi_{1u} + \theta_2 \xi_{2u} + \ldots + \theta_p \xi_{pu},$$

то

$$\partial f / \partial \theta_i = \xi_{iu}, \qquad i = 1, 2, \ldots, p,$$

т. е. производные не зависят от вектора $\boldsymbol{\theta}$. Это приводит, как мы уже видели в предыдущих главах, к линейным нормальным уравнениям относительно $\theta_1, \theta_2, \ldots, \theta_p$. Для нелинейной по параметрам модели также будут получены нормальные уравнения. Проиллюстрируем это на сравнительно простом примере, когда оценивается только один параметр.

Пример. Пусть нужно получить нормальное уравнение для отыскания МНК-оценки $\hat{\theta}$ параметра θ модели $Y = f(\theta, t) + \varepsilon$, где $f(\theta, t) = e^{-\theta t}$. Допустим, что имеется n пар наблюдений: (Y_1, t_1), $(Y_2, t_2), \ldots, (Y_n, t_n)$. Находим

$$\frac{\partial f}{\partial \theta} = -te^{-\theta t}.$$

Применяя соотношение (10.1.6), приходим к единственному нормальному уравнению

$$\sum_{u=1}^{n} [Y_u - e^{-\hat{\theta} t_u}][-t_u e^{-\hat{\theta} t_u}] = 0,$$

или

$$\sum_{u=1}^{n} Y_u t_u e^{-\hat{\theta} t_u} - \sum_{u=1}^{n} t_u e^{-2\hat{\theta} t_u} = 0.$$

Мы видим, что даже в случае одного параметра и сравнительно простой нелинейной модели отыскание оценки $\hat{\theta}$ путем решения всего

лишь одного нормального уравнения оказывается вовсе не такой уж легкой задачей. Если же параметров несколько и модель становится более сложной, то решение нормальных уравнений может превратиться в чрезвычайно трудную задачу. Для ее решения, как правило, приходится применять итеративные методы. Эти трудности усугубляются еще и тем, что может существовать множество решений, соответствующих множеству стационарных значений функции $S(\hat{\boldsymbol{\theta}})$. Теперь мы обсудим методы, которые используются для оценивания параметров в нелинейных системах.

10.2. ОЦЕНИВАНИЕ ПАРАМЕТРОВ НЕЛИНЕЙНЫХ СИСТЕМ

В некоторых нелинейных задачах более удобно непосредственно записать нормальные уравнения (10.1.6) и применить для их решения итеративные методы. Успешность этих методов зависит от формы уравнений и конкретных особенностей используемого итеративного метода. Для решения задачи с помощью такого подхода имеется несколько широко распространенных приемов, пригодных для получения оценок параметров с помощью стандартных вычислений на ЭВМ. Мы рассмотрим здесь три из них: 1) метод линеаризации, 2) метод наискорейшего спуска и 3) компромиссный метод Маркуардта.

Метод линеаризации

Метод линеаризации (или метод разложения в ряд Тейлора) состоит в многократном использовании результатов линейного МНК. Пусть постулированная модель имеет вид (10.1.4). Пусть, далее, $\theta_{10}, \theta_{20}, \ldots, \theta_{p0}$ есть исходные значения оценок параметров $\theta_1, \theta_2, \ldots, \theta_p$. Эти начальные значения могут быть разумно предугаданы или предварительно оценены на основе любой имеющейся информации. (Можно взять, например, значения, полученные при подгонке аналогичных уравнений в другой лаборатории или представляющиеся экспериментатору правдоподобными на основе знаний и интуиции.) Можно надеяться, что исходные оценки будут улучшаться в описанном ниже процессе последовательных итераций.

Если разложить функцию $f(\boldsymbol{\xi}_u, \boldsymbol{\theta})$ в ряд Тейлора в окрестности точки $\boldsymbol{\theta}_0 = (\theta_{10}, \theta_{20}, \ldots, \theta_{p0})'$ и ограничиться производными первого порядка, то можно утверждать, что для близких друг к другу векторов $\boldsymbol{\theta}$ и $\boldsymbol{\theta}_0$ приблизительно верно соотношение

$$f(\boldsymbol{\xi}_u, \boldsymbol{\theta}) = f(\boldsymbol{\xi}_u, \boldsymbol{\theta}_0) + \sum_{i=1}^{p} \left[\frac{\partial f(\boldsymbol{\xi}_u, \boldsymbol{\theta})}{\partial \theta_i}\right]_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} (\theta_i - \theta_{i0}). \qquad (10.2.1)$$

Введя обозначения

$$f_u^0 = f(\boldsymbol{\xi}_u, \boldsymbol{\theta}_0), \quad \beta_i^0 = \theta_i - \theta_{i0}, \quad Z_{iu}^0 = [\partial f(\boldsymbol{\xi}_u, \boldsymbol{\theta})/\partial \theta_i]\boldsymbol{\theta} = \boldsymbol{\theta}_0, \qquad (10.2.2)$$

можно приближенно записать модель (10.1.4) в виде

$$Y_u - f_u^0 = \sum_{i=1}^{p} \beta_i^0 Z_{iu}^0 + \varepsilon_u. \qquad (10.2.3)$$

Таким образом, получена линейная форма этой модели вида (10.0.1), удовлетворяющая выбранной степени аппроксимации. Параметры β_j^0, $j = 1, 2, \ldots, p$, можно теперь оценить, применяя теорию линейного МНК. Пусть

$$\mathbf{Z}_0 = \begin{vmatrix} Z_{11}^0 & Z_{21}^0 & \ldots & Z_{p1}^0 \\ Z_{12}^0 & Z_{22}^0 & \ldots & Z_{p2}^0 \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ Z_{1u}^0 & Z_{2u}^0 & \ldots & Z_{pu}^0 \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ Z_{1n}^0 & Z_{2n}^0 & \ldots & Z_{pn}^0 \end{vmatrix} = \{Z_{iu}^0\}, \quad n \times p, \qquad (10.2.4)$$

$$\mathbf{b}_0 = (b_1^0, b_2^0, \ldots, b_p^0)' \text{ и } \mathbf{y}_0 = (Y_1 - f_1^0, Y_2 - f_2^0, \ldots, Y_n - f_n)' = \mathbf{Y} - \mathbf{f}^0. \qquad (10.2.5)$$

Тогда оценка вектора $\boldsymbol{\beta}_0 = (\beta_1^0, \beta_2^0, \ldots, \beta_p^0)'$ выражается формулой

$$\mathbf{b}_0 = (\mathbf{Z}_0'\mathbf{Z}_0)^{-1}\mathbf{Z}_0'(\mathbf{Y} - \mathbf{f}^0). \qquad (10.2.6)$$

Следовательно, вектор $\mathbf{b}_0$ будет минимизировать сумму квадратов

$$SS(\theta) = \sum_{u=1}^{n} \left\{ Y_u - f(\xi_u, \boldsymbol{\theta}_0) - \sum_{i=1}^{p} \beta_i^0 Z_{iu}^0 \right\}^2 \qquad (10.2.7)$$

по отношению к β_i^0, $i = 1, 2, \ldots, p$. Таким образом, величины $b_i^0 = \theta_{i1} - \theta_{i0}$, $i = 1, 2, \ldots, p$, можно считать улучшенными оценками элементов вектора $\boldsymbol{\beta} = \boldsymbol{\theta} - \boldsymbol{\theta}_0$.

Отметим, кстати, несовпадение между суммой квадратов $S(\boldsymbol{\theta})$ в (10.1.5), где использовалась соответствующая *нелинейная* модель, и суммой квадратов (10.2.7), где фигурирует *линейная аппроксимация* (линейное разложение) модели.

Мы можем теперь считать, что величины θ_{i1} — уточненные оценки параметров — служат исходными для последующего уточнения, т. е. они играют ту же роль, что прежде величины θ_{i0}. Затем можно продолжить уточнение параметров точно тем способом, который был описан выше, используя уравнения (10.2.1) и (10.2.7), в которых все нулевые индексы должны быть заменены на единичные. Это приведет к другому набору уточненных оценок θ_{i2} и т. д. Обобщая прежние

обозначения, можно записать в векторной форме

$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j + \mathbf{b}_j,$$

$$\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j + (\mathbf{Z}_j'\mathbf{Z}_j)^{-1}\,\mathbf{Z}_j'\,(\mathbf{Y} - \mathbf{f}^j), \qquad (10.2.8)$$

где

$$\mathbf{Z}_j = \{Z_{iu}^j\}, \quad \mathbf{f}^j = (f_1^j, f_2^j, \ldots, f_n^j)',$$

$$\boldsymbol{\theta}_j = (\theta_{1j}, \theta_{2j}, \ldots, \theta_{pj})'. \qquad (10.2.9)$$

Этот итеративный процесс продолжается до тех пор, пока он не сойдется, т. е. пока последовательные итерации j, $j+1$ не будут удовлетворять условию $|\{\theta_{i(j+1)} - \theta_{ij}\}/\theta_{ij}| < \delta$, $i = 1, 2, \ldots, p$, где δ есть некоторая наперед заданная малая величина (например, 0,000001). Вычисляя для каждой итерации сумму $S(\boldsymbol{\theta}_j)$, проверяем, происходит ли ее уменьшение.

Процедура линеаризации имеет ряд недостатков, существенных при решении некоторых конкретных задач.

1. Она может сходиться очень медленно, т. е. может потребоваться много итераций, прежде чем решение стабилизируется, хотя сумма квадратов $S(\boldsymbol{\theta}_j)$ может и уменьшаться последовательно, с увеличением номера итерации j. Такое поведение не является типичным, но может встретиться.

2. Могут возникнуть сильные колебания, т. е. периодические увеличения и уменьшения суммы квадратов. Тем не менее в конце концов решение может стабилизироваться.

3. Процедура может вообще не сходиться и даже расходиться, так что сумма квадратов от итерации к итерации будет увеличиваться.

Для преодоления этих недостатков в программе, написанной в 1958 г. Бузом (Booth G. W.) и Питерсоном (Peterson T. I.) под руководством Бокса (Box G. E. P.) — см.: Non-linear estimation, IBM SHARE Program Pa, N 687 (WLNLI), предусматривается возможность изменения длины корректирующего вектора $\mathbf{b}_j$ в уравнении (10.2.8), а именно уменьшение длины вдвое, если справедливо неравенство $S(\boldsymbol{\theta}_{j+1}) > S(\boldsymbol{\theta}_j)$, и, наоборот, увеличение длины вдвое при соблюдении альтернативного неравенства $S(\boldsymbol{\theta}_{j+1}) < S(\boldsymbol{\theta}_j)$. Такое уменьшение или увеличение длины корректирующего вектора продолжается до тех пор, пока не найдутся три точки между $\boldsymbol{\theta}_j$ и $\boldsymbol{\theta}_{j+1}$, между которыми заключен локальный минимум $S(\boldsymbol{\theta})$. Чтобы локализовать минимум, используется квадратичная интерполяция, и итеративный цикл начинается снова.

Хотя теоретически этот метод сходится всегда (см.: Hartley H. O. The modified Gauss—Newton method for fitting of non-linear regression functions by least squares.— Technometrics, 1961, 3, p. 269—280), на практике могут возникнуть трудности. Так, например, в случае, рассмотренном Смитом (см.: Smith N. H. Transient operation of continuous stirred tank reactors.— University of Wisconsin, Ph. D. Thesis, 1963), нелинейность модели приводила к огромным «выбросам» и, несмотря на то что длина корректирующего вектора десять раз уменьшалась вдвое, никакого уменьшения суммы $S(\boldsymbol{\theta})$

по сравнению с первоначальной величиной достигнуто не было. Хотя первоначальная величина составляла меньше 10 ед., вычисления привели к такой большой величине суммы $S(\theta)$ (более 10^{308}), что она «переполнила» память машины. Безуспешная попытка получить оценки данным методом не была, между прочим, описана в этих материалах. Вместо этого был использован метод случайного поиска. Мы упоминаем данный пример лишь для того, чтобы подчеркнуть, что подобные трудности существуют. Метод линеаризации полезен и позволяет успешно решать многие нелинейные задачи. В тех случаях, когда это не так, надо рассматривать такие альтернативы, как репараметризация модели (см. § 10.4) или компромиссный метод Маркуардта [2].

Замечание о производных. Многие вычислительные программы, в которых используется метод линеаризации, предполагают знание значений производных от функции отклика по параметрам в некоторых точках. Вместо них обычно вычисляются отношения вида

$$\{f(\xi_u, \theta_{10}, \theta_{20}, \ldots, \theta_{i0}+h_i, \ldots, \theta_{p0}) - f(\xi_u, \theta_{10}, \theta_{20}, \ldots, \theta_{p0})\}/h_i, \quad i=1, 2, \ldots, p,$$

где h_i — заранее выбранное малое приращение. Отношение, указанное выше, служит приближенным выражением производной

$$\left[\frac{\partial f(\xi_u, \theta)}{\partial \theta_i}\right]_{\theta=\theta_0},$$

так как если h_i стремится к 0, то предел этого отношения есть производная (по определению).

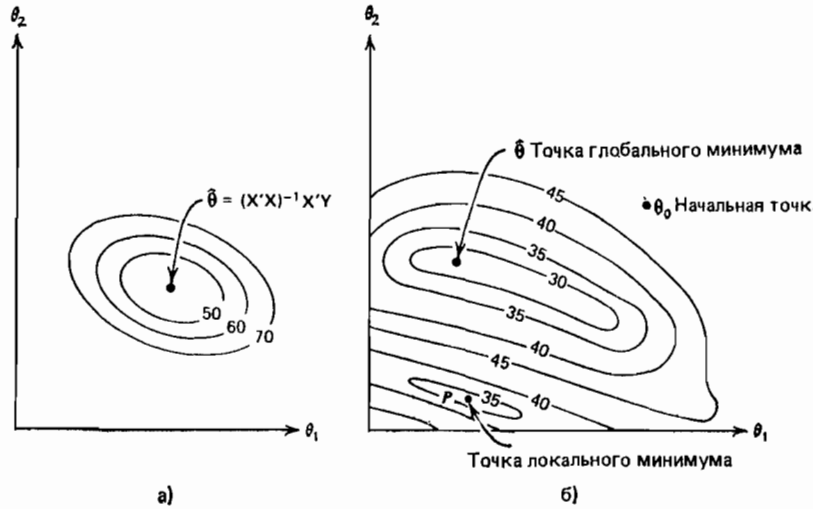
Геометрическая интерпретация линеаризации

Сумма квадратов $S(\theta)$ есть функция только от параметров θ. Экспериментальные данные определяют лишь некоторые численные коэффициенты в выражении $S(\theta)$, и они не изменяются в любой конкретной задаче нелинейного оценивания. В *параметрическом пространстве*, т. е. в p-мерном геометрическом пространстве величин $\theta_1, \theta_2, \ldots, \theta_p$ функцию $S(\theta)$ можно представить в виде контуров поверхности [3]. Для линейной по параметрам модели контуры имеют эллипсоидную форму и существует точка $\hat{\theta}$, в которой достигается единственный локальный (а также и глобальный) минимум, равный $S(\hat{\theta})$. Для нелинейной по параметрам модели контуры не будут эллипсоидными, а будут иметь нерегулярную, зачастую «банановидную» форму. При этом может быть несколько точек локального

[2] Или другие эффективные методы поиска.— *Примеч. пер.*

[3] Под контурами поверхности $S(\theta)$ здесь понимаются фигуры, которые могут быть построены исходя из уравнения $S(\theta) = \alpha$, где α есть некоторое заранее выбранное положительное число. Такие контуры по существу представляют собой сечения поверхности $S(\theta)$ гиперплоскостями, параллельными координатным осям $\theta_1, \theta_2, \ldots, \theta_p$. Размерность таких контуров на единицу меньше числа параметров в модели.— *Примеч. пер.*

минимума и возможно более одной точки глобального минимума, т. е. одно и тоже наименьшее значение $S(\boldsymbol{\theta})$ может достигаться в нескольких точках параметрического пространства. На рис. 10.1 приведены примеры для случая $p = 2$. На рис. 10.1, а изображены эллиптические контуры поверхности $S(\boldsymbol{\theta})$ для линейной модели, а на рис. 10.1, б — нерегулярные контуры поверхности $S(\boldsymbol{\theta})$ для нелинейной модели.



а) б)

Рис. 10.1. а) Эллиптические контуры поверхности $S(\boldsymbol{\theta})$ для линейной модели $Y = \theta_1 X_1 + \theta_2 X_2 + \varepsilon$; эти эллипсы имеют единственную точку минимума $\hat{\boldsymbol{\theta}}$. б) Нерегулярные контуры поверхности $S(\boldsymbol{\theta})$ для нелинейной модели с двумя точками локального минимума. Искомой точкой является $\hat{\boldsymbol{\theta}}$, но итерации могут привести в точку P. Итеративные процедуры для проверки надо осуществлять, стартуя из нескольких случайно разбросанных начальных точек

Точная форма и ориентация контуров $S(\boldsymbol{\theta})$ зависит от модели и от данных. Если контуры, окружающие точку $\hat{\boldsymbol{\theta}}$, сильно вытянуты и имеется много точек $\boldsymbol{\theta}$, почти таких же «хороших», как и $\hat{\boldsymbol{\theta}}$, в том смысле, что соответствующие им значения $S(\boldsymbol{\theta})$ близки к $S(\hat{\boldsymbol{\theta}})$, то говорят, что задача плохо обусловлена, оценки $\hat{\boldsymbol{\theta}}$ могут быть вычислены с трудом. Плохая обусловленность может указывать на то, что модель перепараметризована, т. е. что она содержит больше параметров, чем это необходимо, или на то, что данных недостаточно, и это не позволяет нам оценить постулированные параметры. Поскольку это две стороны одной медали, выбор той или иной причины зависит от априорных знаний о практической задаче и от определенной точки зрения [4]. Рассмотрим, например, функцию $f(t; \theta_1, \theta_2)$, входящую в уравнение (10.0.3). Этой функции соответствует кривая, которая начинается при $t = 0$ и «заканчивается при $t = \infty$. Ее ордината

[4] В настоящее время разработаны методы исследования идентифицируемости параметров нелинейных моделей, позволяющие обнаружить одну из

в начале и в конце равна нулю, и в некоторой точке на этом интервале она достигает максимума. Наклон кривой [5] в начале координат, при $t = 0$, равен θ_1, а максимум достигается при

$$t_{\text{реак}} = \ln(\theta_1/\theta_2)/(\theta_1 - \theta_2).$$

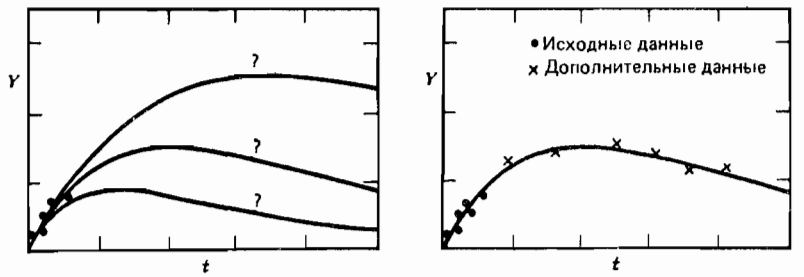


Рис. 10.2. а) Эти данные позволяют оценить начальный наклон кривой (начальную скорость) θ_1, но не дают возможности найти максимум на кривой, который зависит также и от θ_2. Поверхность $S(\theta)$, соответствующая уравнению (10.0.3), плохо обусловлена.
б) Дополнительные данные, показанные на рисунке, позволяют оценить оба параметра θ_1 и θ_2, входящие в уравнение (10.0.3). Поверхность $S(\theta)$ теперь оказывается сравнительно хорошо обусловленной

Отсюда следует, что, если имеющиеся данные охватывают только начальный участок кривой (см. рис. 10.2, а), существует возможность оценить только θ_1, но не θ_2. Чтобы оценить второй параметр, должна быть получена информация с того участка кривой, на котором находится точка максимума, как показано на рис. 10.2, б. Однопараметрическая модель $Y = \theta t + \varepsilon$ была бы адекватна данным, указанным на рис. 10.2, а, однако этих данных недостаточно для оценивания двухпараметрической модели, выражаемой уравнением (10.0.3).

Метод линеаризации сводит задачу отыскания минимума суммы $S(\theta)$ для нелинейной модели, начиная с некоторой исходной точки θ_0, к последовательности задач с линейными моделями. Начальная

рассматриваемых ситуаций, а именно выявить тот случай, когда модель перепараметризована. Этому посвящены работы: Спивак С. И., Горский В. Г.//Докл. АН СССР, 1981, т. 257, № 2, р. 412—415; Горский В. Г., Спивак С. И.//Заводская лаборатория, 1981, т. 47, № 10, с. 30—47; Горский В. Г. Теоретико-групповой анализ идентифицируемости параметров нелинейных моделей.— В кн.: Планирование эксперимента/ Под ред. Ю. П. Адлера, Ю. В. Грановского.— М.: МДНТП им. Дзержинского, 1985, с. 4—8; Горский В. Г., Храименков М. И. Геометрическая природа неидентифицируемости параметров и симметрия нелинейно параметризованных моделей неполного ранга.— М.: Рукопись деп. в ВИНИТИ, 1985, № 5570—85.— 51 с.

Если повинна сама модель, т. е. она перепараметризована, то, какими бы «хорошими» ни были экспериментальные данные, невозможно однозначно определить оценки ее параметров.— *Примеч. пер.*

[5] Точнее, тангенс угла наклона, который в данном случае совпадает с начальной скоростью изменения концентрации промежуточного вещества, участвующего в последовательной химической реакции.— *Примеч. пер.*

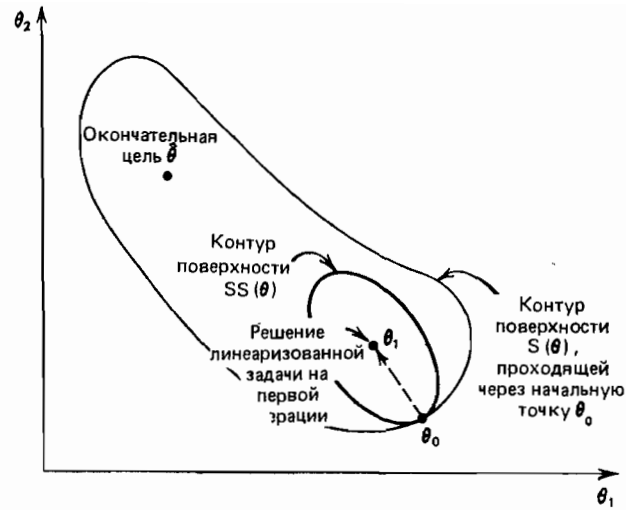


Рис. 10.3. Первая итерация процедуры линеаризации. Эллиптический контур поверхности $SS(\theta)$ «близко» примыкает к контуру суммы $S(\theta)$ при $\theta = \theta_0$, в том смысле, как указано в тексте. Линеаризованная задача дает решение θ_1, и процедура затем повторяется начиная с этой точки

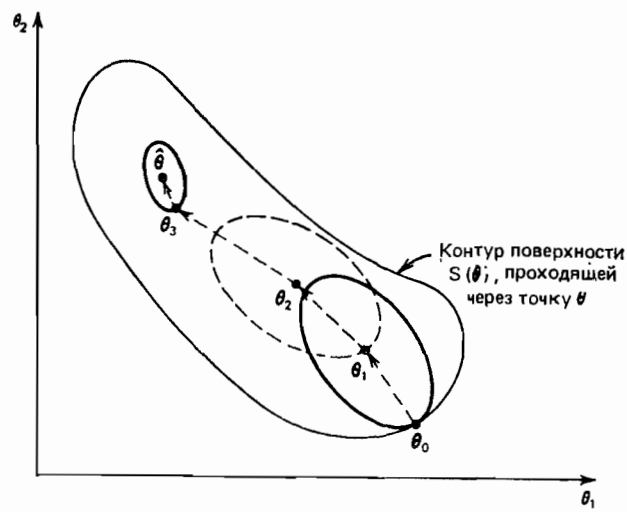


Рис. 10.4. Последовательные итерации процедуры линеаризации, показывающие сходимость к точке $\hat{\theta}$, где достигается минимум $S(\theta)$. Сходимость гарантируется теоретически, на практике же процесс может разойтись

линеаризация функции $f(\xi, \theta)$ в окрестности точки θ_0 в соответствии с выражением (10.2.1) приводит к замене нерегулярного контура $S(\theta)$ на эллипсоидный контур $SS(\theta)$, форма которого «выглядит правильной», т. е. для него характерны те же самые производные от соответствующей функции по параметрам при $\theta = \theta_0$. Как мы увидим далее, в примере § 10.3, такой способ аппроксимации истинных контуров $S(\theta)$ может быть плохим или хорошим в зависимости от следующих обстоятельств: самой постулированной модели, имеющихся данных и относительного расположения точек θ_0 и $\hat{\theta}$ в параметрическом пространстве. В любом случае мы решаем «линеаризованную при θ_0» задачу, продвигаясь к точке минимума θ_1 для контуров линеаризованной модели при θ_0 (с помощью довольно простых МНК-вычислений), как это показано на рис. 10.3. Затем мы повторяем весь «линеаризованный» процесс при θ_1. Мы надеемся, что последовательные итерации приведут нас к точке $\hat{\theta}$, как показано на рис. 10.4, и процесс не разойдется. Обычно линеаризация проходит успешно, если стартовая точка близка к $\hat{\theta}$, поскольку в этом случае истинные контуры поверхности суммы квадратов будут хорошо аппроксимироваться приближенными контурами, соответствующими линеаризованной модели.

Более подробно геометрия линейного и нелинейного методов наименьших квадратов рассматривается в § 10.5 и 10.6.

Метод наискорейшего спуска

Метод наискорейшего спуска использует выражение (10.1.5) для суммы квадратов, а также итеративный процесс нахождения минимума этой функции. Основная идея состоит в том, чтобы двигаться из исходной точки θ_0 в направлении вектора с компонентами

$$\left[-\frac{\partial S(\theta)}{\partial \theta_1}, \quad -\frac{\partial S(\theta)}{\partial \theta_2}, \ldots, -\frac{\partial S(\theta)}{\partial \theta_p}\right],$$

величины которых непрерывно изменяются вдоль траектории. Один из путей реализации этого движения без использования точных функциональных выражений производных состоит в оценивании составляющих указанного выше вектора антиградиента в различных точках пространства параметров путем аппроксимации поверхности $S(\theta)$ плоскостью. Этот метод имеет большое значение в экспериментальных исследованиях для нахождения стационарной точки [6]. Полное опи-

[6] Стационарной точкой поверхности называют точку, в которой равны нулю составляющие вектора-градиента этой поверхности по независимым переменным. Исследования, о которых здесь идет речь,— это планирование эксперимента при изучении поверхности отклика методом Бокса—Уилсона. Кроме ссылки в тексте см. еще книгу: Адлер Ю. П., Маркова Е. В., Грановский Ю. В. Планирование эксперимента при поиске оптимальных условий.— Изд. 2-е.— М.: Наука, 1976.— *Примеч. пер.*

сание метода дается в книге: D a v i e s O. L. Design and analysis of Industrial Experiments.— Edinburgh: Oliver and Boyd, 1954; здесь он будет рассмотрен совсем кратко.

Процедура состоит в следующем. Начиная из некоторой области пространства θ, или *параметрического пространства*, как мы его будем чаще именовать, делают несколько шагов путем выбора n комбинаций уровней параметров и производят вычисления величины $S(\theta)$ для этих комбинаций. Шаги обычно выбирают по схеме факторного эксперимента на двух уровнях [7]. Используя вычисленные значения $S(\theta)$ как наблюдения зависимой переменной и комбинации уровней параметров как соответствующие значения факторов, мы получаем модель

$$\text{«НАБЛЮДАЕМ } S(\theta)\text{»} = \beta_0 + \sum_{i=1}^{p} \beta_i (\theta_i - \bar{\theta}_i)/s_i + \varepsilon,$$

которую мы оцениваем, пользуясь стандартным МНК. Здесь $\bar{\theta}_i$ есть среднее значение уровней θ_{iu}, $u = 1, 2, \ldots, n$, переменных θ_i, используемых в шагах, а s_i — масштабный фактор, который выбирается таким образом, чтобы выполнялось условие

$$\sum_{u=1}^{n} (\theta_{iu} - \bar{\theta}_i)/s_i^2 = \text{const}.$$

Мы полагаем, что истинная поверхность, определяемая функцией $S(\theta)$, в той области пространства параметров, в которой мы делаем шаги, может быть аппроксимирована плоскостью. Оценки коэффициентов

$$b_1, b_2, \ldots, b_p$$

указывают направление наискорейшего возрастания функции [8], а отрицательные компоненты

$$-b_1, -b_2, \ldots, -b_p$$

задают направление наискорейшего спуска. Это означает, что до тех пор, пока справедлива линейная аппроксимация, максимальное уменьшение величины $S(\theta)$ будет получаться при движении вдоль линии, которая содержит такие точки, как $(\theta_i - \bar{\theta}_i)/s_i \propto -b_i$. Если обозначить коэффициент пропорциональности через λ, то линия наискорейшего

[7] Факторное планирование на двух уровнях — краеугольный камень планирования эксперимента. Оно описано во множестве публикаций, см. например: Н а л и м о в В. В., Ч е р н о в а Н. А. Статистические методы планирования экстремальных экспериментов.— М.: Наука, 1965.— 360 с.; Б р о д с к и й В. З. Введение в факторное планирование эксперимента.— М.: Наука, 1974.— 223 с.; Таблицы планов эксперимента для факторных и полиномиальных моделей: Справочное издание/Под ред. В. В. Налимова.— М.: Металлургия, 1982.— 752 с.; R a c t o e B. L., H e d a y a t A., F e d e r e r W. T. Factorial Designs.— New York: J. Wiley, 1981, p. 209. — *Примеч. пер.*

[8] Направление градиента.— *Примеч. пер.*

спуска содержит точки $(\theta_1, \theta_2, \ldots, \theta_p)$, удовлетворяющие уравнению

$$\frac{\theta_i - \bar{\theta}_i}{s_i} = -\lambda b_i,$$

где $\lambda > 0$, или

$$\theta_i = \bar{\theta}_i - \lambda b_i s_i.$$

Придавая различные значения λ, будем двигаться по траектории наискорейшего спуска. Значения λ выбираются таким образом и движение вдоль траектории наискорейшего спуска производится до тех пор, пока величина $S(\theta)$ убывает. Если этого не происходит, то переходят к другому экспериментальному плану, и процесс продолжают до тех пор, пока он не сходится к величине $\hat{\theta}$, которая минимизирует $S(\theta)$.

Хотя теоретически метод наискорейшего спуска должен сходиться, на практике могут встретиться такие ситуации, когда после довольно быстрого первоначального продвижения происходит резкое замедление. В частности, медленная сходимость, вероятно, имеет место тогда, когда контуры поверхности $S(\theta)$ оказываются узкими и имеют форму банана (как зачастую это и бывает на практике), а также когда траектория наискорейшего спуска имеет зигзагообразную форму, с медленным перемещением вдоль узкого оврага. Каждая итерация здесь приводит лишь к незначительному уменьшению величины $S(\theta)$. (Это не столь существенно для лабораторных исследований, где вмешательство экспериментатора позволяет на каждом этапе вычислений пересматривать экспериментальные планы, а также менять масштаб независимых переменных и т. д.). Подобные трудности привели к появлению ряда модификаций основного метода наискорейшего спуска, в которых используется нелинейная аппроксимация (см. например: Spang H. A. A review of minimization techniques of nonlinear functions.— Society for Industrial and Applied Mathematics Review, 1962, **4**, p. 343—365). Одна из возможных модификаций, в частности, состоит в использовании аппроксимации второго порядка вместо первого порядка. Хотя это и приводит к лучшему описанию действительной поверхности $S(\theta)$, однако требует дополнительных вычислений при проведении итерации [9].

Другой недостаток метода наискорейшего спуска состоит в том, что он неинвариантен по отношению к изменению масштаба. Это означает, что если изменить масштабные коэффициенты s_i, то определяемое направление движения изменится неодинаково для всех переменных. Метод наискорейшего спуска, вообще говоря, несколько

[9] Альтернативный подход состоит в использовании последовательного симплексного метода и его многочисленных модификаций. Об этом см. например, работы: Горский В. Г., Адлер Ю. П. Планирование промышленных экспериментов. Модели статики.— М.: Металлургия, 1974.— 264 с.; Дамбраускас А. П. Симплексный поиск.— М.: Энергия, 1979.— 175 с.— *Примеч. пер.*

менее предпочтителен, чем метод линеаризации, но является вполне удовлетворительным для многих нелинейных ситуаций, особенно если используются некоторые модификации основного метода.

В общем, этот метод работает хорошо в той области параметрического пространства, которая находится вдали от искомой точки $\hat{\theta}$, что имеет обычно место на ранних итерациях. По мере приближения к точке $\hat{\theta}$ с помощью этой процедуры происходит зигзагообразное движение, тогда как метод линеаризации работает лучше. Процедура Маркуардта, основанная на работе: Levenberg K. A method for the solution of certain nonlinear problems in least squares.— Quarterly of Applied Mathematics, 1944, 2, p. 164—168, учитывает эти обстоятельства.

Компромиссный метод Маркуардта

Метод, развитый Маркуардтом (см.: Marquardt D. W. An algorithm for least squares estimation of nonlinear parameters.—Journal of the Society for Industrial and Applied Mathematics, 1963, 2, p. 431—441), по-видимому, существенно расширяет число практических задач, в которых может применяться нелинейное оценивание. Метод Маркуардта представляет собой компромисс между методом линеаризации (или методом с разложением модели в ряд Тейлора) и методом наискорейшего спуска. Он, вероятно, сочетает в себе наилучшие черты обоих методов, устраняя в то же время их наиболее серьезные недостатки. Он хорош тем, что почти всегда сходится и не приводит к замедлению, как это часто бывает при использовании метода наискорейшего спуска. Однако мы снова подчеркиваем, что во многих практических задачах будут вполне хорошо работать различные методы, если только не нарушены их ограничения. (Вообще мы должны иметь в виду, что если предлагается какой-то особый метод, то обычно можно сконструировать такую задачу, которая покажет его полную несостоятельность. И наоборот, если имеется данная частная задача и предложен какой-то метод для данного случая, то он может оказаться более эффективным с точки зрения скорости сходимости, чем другие методы. Метод Маркуардта — это такой метод, который, по-видимому, хорошо работает в самых различных задачах, и это является существенным аргументом при выборе данного метода во многих практических ситуациях. По причинам, которые были сформулированы выше, не существует, однако, такого метода, который можно было бы назвать «наилучшим» для всех нелинейных задач.)

Идею метода Маркуардта можно пояснить кратко следующим образом. Предположим, что мы начинаем двигаться из некоторой точки пространства параметров θ. Если применяется метод наискорейшего спуска, то мы получим некоторый вектор δ_g (где индекс g означает, что вектор направлен вдоль градиента), определяющий направление движения из начальной точки. Из-за большой вытянутости контуров поверхности $S(\theta)$ это может быть лучшим *локальным* направлением, при движении по которому получается наименьшее значение $S(\theta)$,

но это может и не быть наилучшим *глобальным* направлением. Однако угол между наилучшим направлением и направлением вектора δ_g не должен превышать 90°, иначе функция $S(\theta)$ станет увеличиваться. Метод линеаризации приводит к другому корректирующему вектору δ, который задается формулой, подобной (10.2.6). Маркуардт нашел, что в тех практических задачах, которые он изучал, угол, который обозначим буквой φ, между δ_g и δ лежит в пределах $80° < \varphi < 90°$. Другими словами, оба направления почти всегда составляют прямой угол! Алгоритм Маркуардта приводит к интерполяции между векторами δ_g и δ и позволяет получать также подходящие размеры шагов.

Мы не будем в деталях рассматривать этот метод. Основной алгоритм описан в работах, приведенных в библиографии; обсуждение метода содержится в работе: Meeter D. A. Problems in the analysis of nonlinear models by least squares.— University of Wisconsin, Ph. D. Thesis, 1964. С программами вычислений можно ознакомиться в работах: Marquardt D. W. Least squares estimation of nonlinear parameters, a computer program in Fortran IV language.— IBM SHARE Library, Distribution N 309401, August 1966. (Successor to Distribution Numbers 1428 and 3094); Marquardt D. W., Stanley R. M. NLIN 2 — Least squares estimation of nonlinear parametrs, supplement to S.D.A. — 3093 (NLIN), 1964. Mimeo manuscript.

Примечание. Программы непрерывно совершенствуются, поэтому и сейчас могут появляться свежие, более интересные варианты [10].

Доверительные контуры

Некоторые дополнительные представления о нелинейности модели [11] можно получить при дальнейшем ее изучении после нахождения оценки вектора θ в результате построения эллипсоидных до-

---

[10] Алгоритмическому и программному обеспечению задач нелинейного оценивания посвящено большое количество публикаций. Много интересных алгоритмов и программ содержнтся в более поздних изданиях, см. например: Химмельблау Д. Анализ процессов статистическими методами/Пер. с англ. Под ред. В. Г. Горского.— М.: Мир, 1973.— 959 с. (особо см. с. 434—452); Химмельблау Д. Прикладное нелинейное программирование.— М.: Мир, 1975.— 536 с.; Численные методы условной оптимизации/Под ред. Ф. Гилла, У. Мюррея; Пер. с англ. Под ред. А. А. Петрова.— М.: Мир, 1977.— 292 с.; Бард Й. Нелинейное оценивание параметров/Пер. с англ. Под ред. В. Г. Горского.— М.: Статистика, 1979.— 351 с.; Демиденко Е. З. Линейная и нелинейная регрессия. Фортран-IV.— М.: ИМЭМО, 1979.— 82 с.; Демиденко Е. З. Гребневая регрессия (Препринт).— М.: ИМЭМО, 1982.— 127 с.; Демиденко Е. З. Нелинейная регрессия. Ч. 1. Алгоритмы (Препринт).— М.: ИМЭМО, 1984.— 74 с.; Демиденко Е. З. Нелинейная регрессия. Ч. 2. Программы (Препринт).— М.: ИМЭМО, 1984.— 72 с. Список программ нелинейного оценивания приведен в приложении к книге: Планирование эксперимента в задачах нелинейного оценивания и распознавания образов/Круг Г. К., Кабанов В. А., Фомин Г. А., Фомина Е. С.— М.: Наука, 1981.— 172 с. (особо с. 161—163).— *Примеч. пер.*

[11] Количественные методы оценки нелинейности модели начали развиваться после работы Била (Beale E. M. L.— J. Roy. Statist. Soc. B, 1960, **22**,

верительных областей, исходя из линеаризованной формы модели. Эллипсоидная доверительная область задается неравенством

$$(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})'\hat{\mathbf{Z}}'\hat{\mathbf{Z}}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \leqslant ps^2F(p,\ n-p,\ 1-\alpha),$$

где $\hat{\mathbf{Z}}$ означает матрицу, приведенную в уравнении (10.2.4), но с подстановкой вектора $\hat{\boldsymbol{\theta}}$ вместо $\boldsymbol{\theta}_0$ и где также

$$s^2 = S(\hat{\boldsymbol{\theta}})/(n-p).$$

Заметим, что если разница между последовательными величинами $\boldsymbol{\theta}_{l+1}$ и $\boldsymbol{\theta}_l$ будет достаточно малой и процедура линеаризации закончится при значении $\boldsymbol{\theta}_{l+1} = \hat{\boldsymbol{\theta}}$, то $S(\hat{\boldsymbol{\theta}})$ станет минимальным значением суммы $S(\boldsymbol{\theta})$ согласно уравнению (10.1.5) в пределах точности, связанной с выбранным правилом останова. Это можно увидеть, исследуя уравнение (10.2.7), в котором величины $\boldsymbol{\theta}_0$, β_i^0 и Z_{iu}^0 заменены соответственно величинами $\hat{\boldsymbol{\theta}}$, β_i^{l+1} и Z_{iu}^{l+1}, и вспоминая, что в соответствии с правилом останова точность определяется соотношением $\mathbf{b}_{l+1} = \mathbf{0}$. Эллипсоид, описанный выше, не будет представлять истинную доверительную область, если модель окажется нелинейной. Можно, однако, определить конечные точки на главных осях этого эллипсоида путем канонического преобразования уравнения (см., например: Davies O. L. Design an Analysis of Industrial Experiments.— Edinburgh, Scotland: Oliver and Boyd, 1954) [12]. Для этих точек могут быть вычислены и сопоставлены между собой *истинные* значения суммы $S(\boldsymbol{\theta})$. В случае линейной параметризации все они должны быть одинаковыми.

Границы точной доверительной области определяются выражением $S(\boldsymbol{\theta}) = \text{const}$, но так как неизвестно действительное распределение случайных величин $S(\hat{\boldsymbol{\theta}})$ в общем нелинейном случае, то не удается найти соответствующую доверительную вероятность (уровень

N1, p. 41—75). Вот некоторые этапные публикации: Gutman I., Meeter D. A.— Technometrics, 1965, **7**, N4, p. 623—637; Heuts H.— Statist. H., 1974, **15**, N4, p. 234—255; Linssen H. N.— Statist. Neer, 1975, **29**, N3, p. 93—99; Bates D. M., Watts D. G.—Ann. Statist., 1981, **9**, N6, p. 1152—1167. На русском языке эти вопросы нашли отражение в книгах: Деми-денко Е. З. Линейная и нелинейная регрессии.— М.: Финансы и статистика, 1981.— 304 с. (особо с. 278—280); Планирование эксперимента в задачах нелинейного оценивания и распознавания образов. — М.; Наука, 1981, — 172 с. (особо с. 44—46).— *Примеч. пер.*

[12] Приведение поверхности эллипсоида к каноническим осям — это стандартная процедура, рассматриваемая в курсе аналитической геометрии или линейной алгебры (см., например: Выгодский М. Я. Аналитическая геометрия.— М.: Физматгиз, 1968.— 528 с.; Александров П. С. Лекции по аналитической геометрии.— М.: Наука, 1968.— 911 с.; Мишина А. П., Проскурякова И. В. Высшая алгебра.— М.: Физматгиз, 1962.— 300 с.). Применительно к планированию эксперимента это преобразование изложено, в частности, в книге: Рузинов Л. П., Слободчикова Р. И. Планирование эксперимента в химии и химической технологии.— М.: Химия, 1980.— 280 с.— *Примеч. пер.*

значимости). Однако можно, например, выбрать контуры так, чтобы выполнялось соотношение

$$S(\theta) = S(\hat{\theta})\left\{1 + \frac{p}{n-p} F(p, n-p, 1-\alpha)\right\}.$$

В таком случае, если модель линейна, они дают *точную* 100 (1—α) %-ную эллипсоидную границу (см. 2.6 и 10.5) и представляют собой также границу приближенной 100 (1—α) %-ной доверительной области для нелинейного случая. Заметим, что определяемые таким образом контуры *будут соответствовать правильным доверительным контурам в этом случае* (по форме они, в общем, не будут эллипсоидными), *но доверительная вероятность будет приближенной.* Если модель содержит лишь два параметра, то доверительные контуры можно вычертить. В случае большого числа параметров при желании можно построить соответствующие сечения.

В общем случае, если используется линеаризованная форма представления нелинейной модели, могут применяться все обычные формулы и аналитические процедуры линейной регрессии. Любые получаемые результаты, однако, имеют силу лишь постольку, поскольку линеаризованная форма модели дает хорошую аппроксимацию истинной модели.

Сетки и графики

Зачастую из виду упускают два очевидных способа исследования суммы квадратов $S(\theta)$. Они могут быть полезны в особенности тогда, когда итеративная процедура начиная с некоторой исходной точки не дает удовлетворительной сходимости.

В первом из них выбирается сетка из точек и «проводится факторный эксперимент» в пространстве параметров $(\theta_1, \theta_2, \ldots, \theta_p)$, в каждой точке сетки вычисляют (в общем случае — с помощью ЭВМ) соответствующую ей сумму квадратов. Эти значения дают некоторое представление о форме поверхности суммы квадратов отклонений и могут позволить, например, обнаружить, что существует несколько минимумов. Во всяком случае, точка сетки, в которой получено наименьшее значение суммы квадратов, может использоваться как исходная в итеративной процедуре оценивания параметров или для построения более густой сетки и более детального исследования окрестности этой точки, чтобы получить лучшую исходную точку. Простейший тип сетки — это сетка, в которой выбираются два уровня для каждого параметра. В этом случае узлы сетки становятся точками факторного эксперимента 2^p и можно использовать стандартные методы для вычисления факторных эффектов и взаимодействий и, таким образом, получить информацию об эффектах изменения параметров функции $S(\theta)$.

Вторая возможность заключается в том, чтобы вычертить контуры суммы квадратов в какой-либо отдельной области пространства параметров, где имеет место плохая сходимость или для которой нужна дополнительная информация. Это обычно имеет смысл делать тогда,

когда модель содержит один или два параметра. Если параметров больше, то мы можем получить контуры на плоскости, зафиксировав значения всех параметров, кроме двух, и тогда можно воспроизвести сравнительно сложную картину.

Важность выбора хорошей исходной точки

Все итеративные процедуры требуют знания исходных (начальных) значений $\theta_{10}, \theta_{20}, \ldots, \theta_{p0}$ параметров $\theta_1, \theta_2, \ldots, \theta_p$. Чтобы сделать их наиболее подходящими, следует использовать всю имеющуюся информацию. Хорошие исходные значения параметров зачастую обеспечивают сходимость итеративной процедуры к решению намного быстрее, чем это возможно в других случаях. Итак, если существует множество минимумов или имеется несколько локальных минимумов в дополнение к глобальному минимуму, то плохие исходные значения могут в результате привести к сходимости к нежелательной стационарной точке поверхности суммы квадратов. Эта нежелательная точка может соответствовать таким значениям параметров, которые физически нереализуемы или не дают истинной минимальной величины $S(\boldsymbol{\theta})$. Как уже говорилось ранее, очень полезно провести предварительные вычисления значений суммы квадратов $S(\boldsymbol{\theta})$ в ряде точек сетки пространства параметров.

Получение исходных оценок θ_0

Изобрести стандартный, пригодный для любой задачи нелинейного оценивания метод отыскания начальных оценок не удается. На практике в зависимости от особенностей задачи, пользуются одним из следующих приемов *[1].

1. Если постулированная модель содержит p оцениваемых параметров, в нее подставляют p наборов имеющихся данных $(Y_u, \boldsymbol{\xi}_u)$, *считая ошибки эксперимента* равными нулю. Решают полученные p уравнений относительно параметров (если это возможно). При этом выбор наиболее различающихся величин $\boldsymbol{\xi}_u$ нередко дает лучшие результаты.

2. Применяя первый прием или рассматривая в качестве альтернативы поведение функции отклика при условии, что ξ_i стремится к нулю или бесконечности, в полученные выражения подставляют результаты наблюдений, которые более всего отвечают этим условиям. Затем, если это возможно, решают систему полученных уравнений.

3. Исследуют форму модели, полагая аддитивную ошибку равной нулю, с тем чтобы (если это возможно) хотя бы приближенно преобразовать модель к более простому выражению. Так, например, если модель имеет вид $Y = \theta_1 e^{-\theta_2 t} + \varepsilon$, график зависимости $\ln Y$ от t обычно дает хорошие начальные оценки параметров $\ln \theta_1$ и θ_2, первый из которых — отрезок, отсекаемый на оси ординат, а второй равен тангенсу угла наклона прямой к оси абсцисс. (Если бы модель

*[1] Для ознакомления с вычислениями в некоторых примерах см. с. 247.

включала ошибку мультипликативно, т. е. имела бы вид $Y = \theta_1 e^{-\theta_2 t} \cdot \varepsilon$, то преобразование $\ln Y = \ln \theta_1 - \theta_2 t + \ln \varepsilon$ было бы строго обоснованным.)

4. Более сложный пример применения метода 3 — модель $\ln W = -\theta_1^{-1} \ln(\theta_2 + \theta_3 X^{\theta_4}) + \varepsilon$, используемая при описании процессов роста (см. Mead R. Plant density and crop yield.—Applied Statistics, 1970, **19**, p. 64—81). Когда $\theta_1 = \theta_4 = 1$, график зависимости величины W^{-1} от X представляет собой прямую линию с параметрами: θ_2 — отрезок, отсекаемый на оси ординат, θ_3 — тангенс угла наклона прямой к оси абсцисс. Если после вычерчивания такого графика для конкретной задачи обнаружится, что он имеет вид кривой линии, то методом «проб и ошибок» можно подобрать такие значения параметров θ_1 и θ_4, чтобы график зависимости $W^{-\theta_1}$ от X^{θ_4} превратился в достаточно «хорошую» прямую линию. Как только это достигнуто, соответствующие значения величин θ могут использоваться в качестве исходных оценок. Заметим, что в моделях, подобных данной, при плохих начальных оценках величина $\theta_2 + \theta_3 X^{\theta_4}$ может оказаться отрицательной, что приведет к вычислительным трудностям [13]. Хорошие начальные оценки зачастую позволяют избежать таких неприятностей.

5. Если все попытки тщетны, можно использовать сетки и графики — см. с. 209—210.

(Примечание. Когда в качестве исходных значений параметров вначале выбраны малые величины, ожидается, что в итеративной процедуре некоторые параметры будут иметь малые значения, надо позаботиться о том, чтобы интервалы h_i при вычислении частных производных имели соответственно малые значения. Если этого не сделать, то некоторые приемы определения оценок параметров окажутся непригодными.)

10.3. ПРИМЕР

Рассматриваемый пример взят из исследования, выполненного Проктером и Гемблом и описанного в работе: Smith H., Dubey S. D. Some reliability problems in chemical industry.— Industrial Quality Control, 1964, **21** (2), p. 64—70. Проиллюстрируем на этом примере, как можно получить решение задачи нелинейного оценивания решая системы нормальных уравнений непосредственно или с помощью метода линеаризации. Мы не приведем пример использования метода наискорейшего спуска; при желании с таким примером можно познакомиться в работе: Box G. E. P., Coutie G. A. Application of digital computers in the exploration relationships.— Proceedings of the Institution of Electrical Engineers, 1956, **103**, Part B, Supplement N1, p. 100—107. Не будем иллюстрировать примером и компромиссный метод Маркуардта.

---

[13] Поскольку тогда придется находить логарифмы отрицательного числа, которые не существуют.— *Примеч. пер.*

Исследуется качество продукта А, который должен содержать в момент производства не менее 50 % активного хлора. Содержание активного хлора в продукте со временем понижается. Известно, также, что к восьмой неделе хранения происходит снижение содержания активного хлора до 49 %. Поскольку изучаемый процесс подвержен воздействию многих неконтролируемых факторов (таких, как, скажем, условия хранения на складе и др.), теоретическое предсказание ожидаемой доли активного хлора в продукте к заданному сроку хранения оказывается ненадежным. Руководству нужно решить: 1) когда материал на складе должен уничтожаться, 2) когда следует заменять запасы сырья. Чтобы обосновать такое решение, ящики с продуктом анализировались на содержание активного хлора в течение некоторого времени. Полученные данные представлены в табл. 10.2

Таблица 10.2 **Процент активного хлора в продукте**

Номер недели года, отвечающей выпуску партии продукта X	Содержание активного хлора Y	Среднее значение содержания активного хлора $\bar{Y}$	Значение, предсказанное по модели $\hat{Y}$
8	0,49; 0,49	0,490	0,490
10	0,48; 0,47; 0,48; 0,47	0,475	0,472
12	0,46; 0,46; 0,45; 0,43	0,450	0,457
14	0,45; 0,43; 0,43	0,437	0,445
16	0,44; 0,43; 0,43	0,433	0,435
18	0,46; 0,45	0,455	0,427
20	0,42; 0,42; 0,43	0,423	0,420
22	0,41; 0,41; 0,40	0,407	0,415
24	0,42; 0,40; 0,40	0,407	0,410
26	0,41; 0,40; 0,41	0,407	0,407
28	0,41; 0,40	0,405	0,404
30	0,40; 0,40; 0,38	0,393	0,401
32	0,41; 0,40	0,405	0,399
34	0,40	0,400	0,397
36	0,41; 0,38	0,395	0,396
38	0,40; 0,40	0,400	0,395
40	0,39	0,390	0,394
42	0,39	0,390	0,393

(Заметим, что продукт выпускается партиями раз в две недели, а кодированные данные соответствуют номеру недели года. Предсказанные значения, приведенные в таблице, вычислены с помощью аппроксимирующего уравнения, к получению которого мы теперь приступим.) Было постулировано, что нелинейная модель имеет вид

$$Y = \alpha + (0{,}49 - \alpha)\, e^{-\beta (X-8)} + \varepsilon \qquad (10.3.1)$$

и что она пригодна для объяснения вариаций, наблюдаемых в экспериментальных данных при $X \geqslant 8$. Эта модель без учета ошибки дает истинное значение $\eta = 0{,}49$, когда $X = 8$. Кроме того, она учитывает также соответствующий характер колебаний концентрации актив-

ного хлора. Дополнительная информация, которая согласуется с представлениями химика, заключается в том, что при достижении равновесия концентрация активного хлора доходит до 30 %. Задача сводится к нахождению оценок параметров α и β нелинейной модели (10.3.1) на основе данных из таблицы. Сумма квадратов отклонений для этой модели может быть выражена формулой

$$S(\alpha,\ \beta) = \Sigma_u^n [Y_u - \alpha - (0{,}49 - \alpha) e^{-\beta(X_u - 8)}]^2, \qquad (10.3.2)$$

где $(X_u,\ Y_u)$, $u = 1, 2, \ldots, 44$,— соответствующие пары наблюдений из таблицы (например, $X_1 = 8$, $Y_1 = 0{,}49, \ldots, X_{44} = 42$, $Y_{44} = 0{,}39$).

Решение с помощью нормальных уравнений

Продифференцируем уравнения (10.3.2) сначала по α, а затем по β и приравняем соответствующие выражения к нулю. Получим два нормальных уравнения. После исключения сомножителя 2 из первого уравнения и сомножителя 2 $(0{,}49 - \alpha)$ из второго уравнения и выполнения некоторых преобразований получим систему:

$$\begin{cases} \alpha = \dfrac{\Sigma Y_u - \Sigma Y_u e^{-\beta t_u} - 0{,}49 \Sigma e^{-\beta t_u} + 0{,}49 \Sigma e^{-2\beta t_u}}{n - 2\Sigma e^{-\beta t_u} + \Sigma e^{-2\beta t_u}}, & (10.3.3) \\ \alpha = \dfrac{0{,}49 \Sigma t_u e^{-2\beta t_u} - \Sigma Y_u t_u e^{-\beta t_u}}{\Sigma t_u e^{-2\beta t_u} - \Sigma t_u e^{-\beta t_u}}, & (10.3.4) \end{cases}$$

где суммирование производится от $u = 1$ до $u = 44$ и использовано обозначение $t_u = X_u - 8$. Эта система нормальных уравнений имеет особую структуру, благодаря которой параметр α может быть исключен из нее вычитанием второго уравнения из первого. Тогда получим единственное нелинейное уравнение вида $f(\beta) = 0$ относительно β. Это уравнение можно решить с помощью метода Ньютона—Рафсона. Обозначим исходную оценку параметра β через β_0 и будем затем «корректировать» ее, используя величину h_0, которую получим следующим образом. Если корень уравнения $f(\beta) = 0$ есть $\beta_0 + h_0$, тогда приближенно

$$0 = f(\beta_0 + h_0) = f(\beta_0) + h_0 \left[\frac{df(\beta)}{d\beta}\right]_{\beta=\beta_0}, \qquad (10.3.5)$$

откуда и получаем

$$h_0 = -f(\beta_0) \Big/ \left[\frac{df(\beta)}{d\beta}\right]_{\beta=\beta_0}. \qquad (10.3.6)$$

Теперь мы можем вместо β_0 использовать величину $\beta_1 = \beta_0 + h_0$ и повторить процедуру коррекции, определяя поправку h_1 и записывая $\beta_2 = \beta_1 + h_1$. Этот процесс может продолжаться до тех пор, пока он не сойдется к величине $\hat{\beta}$, которая и будет МНК-оценкой параметра β. Величину $\hat{\alpha}$, т. е. МНК-оценку параметра α, можно получить

из уравнений (10.3.3) и (10.3.4) путем подстановки $\hat{\beta}$ в их правые части. Для проверки полезен тот факт, что оба уравнения должны приводить к одному и тому же значению α.

Исходное значение β_0 можно найти, зная, например, что для $X_{44} = 42$ отклик равен $Y_{44} = 0{,}39$. Если считать, что величина $Y_{44} = 0{,}39$ определяется без ошибки, то можно записать

$$0{,}39 = \alpha + (0{,}49 - \alpha)\, e^{-34\beta_0}.$$

В то же время мы знаем (на основании априорной информации от химиков), что величина Y имеет тенденцию стремиться к 0,30, когда X стремится к бесконечности. Поэтому мы можем принять исходное значение для α на уровне $\alpha_0 = 0{,}30$. Отсюда следует

$$0{,}39 = 0{,}30 + (0{,}49 - 0{,}30)\, e^{-34\beta_0},$$

$$e^{-34\beta_0} = \frac{0{,}09}{0{,}19}.$$

Следовательно, приближенно имеем

$$\beta_0 = \frac{-[\ln (0{,}09/0{,}19)]}{34} = 0{,}02.$$

(Заметим, что необходимо $\beta > 0$, иначе наблюдаемое падение концентрации активного хлора нельзя представить с помощью данной функции.)

Если для уравнений (10.3.3) и (10.3.4) воспользоваться соответственно записью

$$\alpha = f_1(\beta), \qquad \alpha = f_2(\beta), \tag{10.3.7}$$

то

$$f(\beta) \equiv f_1(\beta) - f_2(\beta) = 0 \tag{10.3.8}$$

и

$$\frac{\partial f(\beta)}{\partial \beta} \equiv \frac{\partial f_1(\beta)}{\partial (\beta)} - \frac{\partial f_2(\beta)}{\partial \beta}. \tag{10.3.9}$$

Вместо того чтобы выписывать весьма длинное выражение, которое получается при дифференцировании функций $f_1(\beta)$ и $f_2(\beta)$, применим более простой прием для отыскания $\hat{\alpha}$, $\hat{\beta}$. Он состоит в вычерчивании графиков функций (10.3.7) в разумных пределах изменения параметров β и в отыскании точки пересечения этих двух кривых. В итоге будут сразу получены искомые значения $\hat{\alpha}$, $\hat{\beta}$. Некоторые значения функций $f_1(\beta)$ и $f_2(\beta)$ для ряда значений аргументов β приведены в табл. 10.3, а результирующий график — на рис. 10.5.

Оценки, которые можно получить с помощью миллиметровой бумаги, получаются достаточно точными и равны: $\hat{\alpha} = 0{,}39$; $\hat{\beta} = 0{,}10$. Из рисунка видно, что кривые проходят довольно близко друг к другу в сравнительно большом диапазоне изменения параметра β и при

сравнительно малом изменении параметра α. Это указывает на то, что параметр β оценивается несколько хуже, чем параметр α. Например, величина $|f_1(\beta) - f_2(\beta)| < 0{,}0025$ достигается в интервале значений β от 0,07 до 0,12 и в интервале значений α между 0,37 и 0,40. Некоторые пары значений (β, α), такие, как, например, (0,09; 0,385), (0,11; 0,393), лежат там, где кривые на рисунке практически совпадают. В свете имеющихся экспериментальных данных эти наборы нельзя считать неразумными оценками параметров β и α, хотя они и не минимизируют по-настоящему $S(\alpha, \beta)$. Мы увидим далее, что

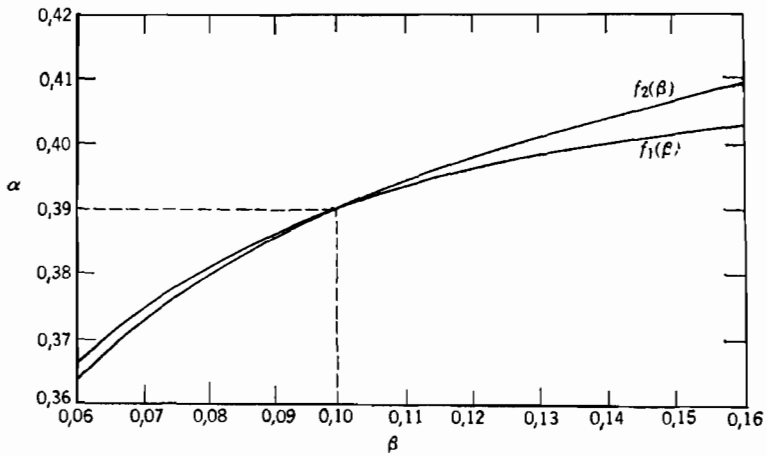


Рис. 10.5. Нахождение оценок $\hat{\beta}$, $\hat{\alpha}$ путем пересечения кривых $f_1(\beta)$ и $f_2(\beta)$.

эти соображения подтверждаются также с помощью доверительной области для истинных значений β и α, которая может быть построена в данном случае (см. рис. 10.8). Подогнанное уравнение теперь приобретает вид

$$\hat{Y} = 0{,}39 + 0{,}10e^{-0{,}10(X-8)}. \qquad (10.3.10)$$

Подставляя в (10.3.10) наблюдаемое значение X, можно получить величины Y, приведенные в табл. 10.2. Аппроксимирующая кривая и результаты наблюдений показаны на рис. 10.6.

Таблица 10.3. Точки кривых $f_i(\beta)$

β	$f_1(\beta)$	$f_2(\beta)$	β	$f_1(\beta)$	$f_2(\beta)$
0,06	0,3656	0,3627	0,12	0,3953	0,3970
0,07	0,3743	0,3720	0,13	0,3975	0,4002
0,08	0,3808	0,3791	0,14	0,3993	0,4031
0,09	0,3857	0,3847	0,15	0,4009	0,4057
0,10	0,3896	0,3894	0,16	0,4023	0,4082
0,11	0,3927	0,3935			

Теперь в соответствии с гл. 3 можно проделать обычный анализ остатков. (Наибольший остаток при $X = 18$ сразу бросается в глаза. Однако никаких особых причин для объяснения этого выброса авторы не нашли.)

Решение с использованием методов линеаризации

Чтобы линеаризовать модель вида (10.2.1), вычислим первые производные от функции

$$f(\xi_u,\ \theta) = f(X_u;\ \alpha,\ \beta) = \alpha + (0{,}49 - \alpha)\, e^{-\beta(X_u - 8)}, \qquad (10.3.11)$$

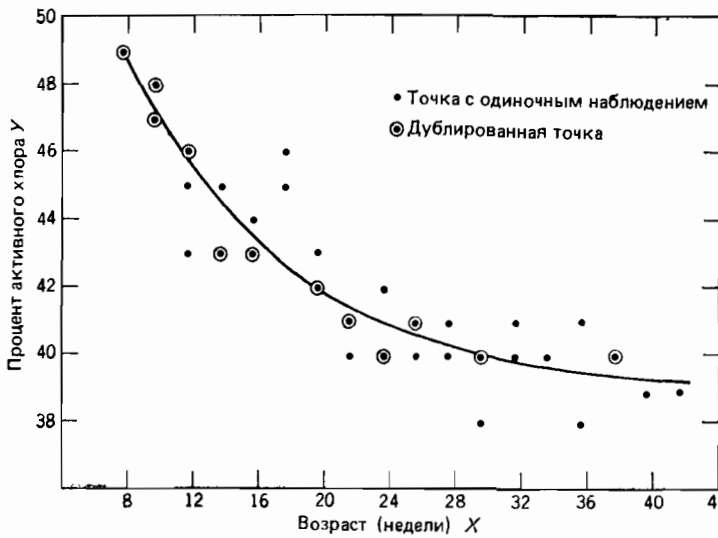


Рис. 10.6. Расчетная кривая и наблюдения

а именно:

$$\frac{\partial f}{\partial \alpha} = 1 - e^{-\beta(X_u - 8)},$$

$$\frac{\partial f}{\partial \beta} = -(0{,}49 - \alpha)(X_u - 8)\, e^{-\beta(X_u - 8)}. \qquad (10.3.12)$$

Таким образом, если $\alpha = \alpha_j$, $\beta = \beta_j$ — величины, вводимые на j-м шаге, как описано в § 10.2, то при обозначениях, используемых в том же параграфе, модель на j-м шаге итерации имеет вид

$$Y_u - f_u^j = \left[1 - e^{-\beta_j(X_u - 8)}\right](\alpha - \alpha_j) + \left[-(0{,}49 - \alpha_j)(X_u - 8)\, e^{-\beta_j(X_u - 8)}\right] \times (\beta - \beta_j) + \varepsilon_u$$

или в матричной форме

$$\mathbf{Y} - \mathbf{f}^j = \mathbf{Z}_j \begin{bmatrix} \alpha - \alpha_j \\ \beta - \beta_j \end{bmatrix} + \boldsymbol{\varepsilon},$$

где

$$f_u^j = \alpha_j + (0{,}49 - \alpha_j)\, e^{-\beta_j (X_u - 8)} \tag{10.3.13}$$

и

$$\mathbf{Z}_j = \begin{bmatrix} 1 - e^{-\beta_j (X_1 - 8)} & -(0{,}49 - \alpha_j)(X_1 - 8)\, e^{-\beta_j (X_1 - 8)} \\ \vdots & \vdots \\ 1 - e^{-\beta_j (X_u - 8)} & -(0{,}49 - \alpha_j)(X_u - 8)\, e^{-\beta_j (X_u - 8)} \\ \vdots & \vdots \\ 1 - e^{-\beta_j (X_n - 8)} & -(0{,}49 - \alpha_j)(X_n - 8)\, e^{-\beta_j (X_n - 8)} \end{bmatrix} \tag{10.3.14}$$

Нужно найти вектор

$$\begin{bmatrix} \alpha - \alpha_j \\ \beta - \beta_j \end{bmatrix}; \tag{10.3.15}$$

его оценки задаются уравнением

$$\begin{bmatrix} \alpha_{j+1} - \alpha_j \\ \beta_{j+1} - \beta_j \end{bmatrix} = (\mathbf{Z}_j' \mathbf{Z}_j)^{-1} \mathbf{Z}_j' \begin{bmatrix} Y_1 - f_1^j \\ Y_2 - f_2^j \\ \vdots \\ Y_n - f_n^j \end{bmatrix}. \tag{10.3.16}$$

Если начать итерации с исходных величин $\alpha_0 = 0{,}30$ и $\beta_0 = 0{,}02$, как указано выше, то, применяя последовательно уравнение (10.3.16), получим следующие оценки:

Итерация	α_j	β_j	$S(\alpha_j, \beta_j)$
0	0,30	0,02	0,0263
1	0,8416	0,1007	4,4881
2	0,3901	0,1004	0,0050
3	0,3901	0,1016	0,0050
4	0,3901	0,1016	0,0050

(Примечание. Эти данные округлены по сравнению с найденными при вычислениях на ЭВМ и содержавшими больше значащих цифр. Естественно, они несколько отличаются от тех чисел, которые получились бы при проведении вычислений на микрокалькуляторе.)

Процесс сходится к некоторым МНК-оценкам, указанным выше, которые и входят в итоговое уравнение (10.3.10): $\widehat{\alpha} = 0{,}39$, $\widehat{\beta} = 0{,}10$.

Заметим, что это происходит несмотря на тот тревожный факт, что после первого шага итерации сумма становится равной $S(\alpha_1, \beta_1) = 4{,}4881$, что примерно в 170 раз больше первоначального значения суммы $S(\alpha_0, \beta_0) = 0{,}0263$. Уменьшение суммы на следующем шаге итерации получилось большим и практически исчерпывающим, ибо дальнейшие итерации приводят к незначительному уменьшению суммы $S(\alpha, \beta)$, которое сказывается лишь в шестом знаке после запятой и не получило отражения в таблице. В некоторых нелинейных задачах не происходит вообще никакого улучшения, и процесс расходится, давая все большие и большие значения $S(\theta)$. (Причины такого поведения описаны, в частности, в § 10.6.)

На рис. 10.7а изображены контуры суммы квадратов для нелинейной модели (10.3.1) в области $0 \leqslant \beta \leqslant 0{,}20$; $0{,}28 \leqslant \alpha \leqslant 0{,}90$ вместе с иллюстрацией перемещения из начальной точки (β_0, α_0) в точку (β_2, α_2). Обоснование такого «пути» отражено на рис. 10.7б и 10.7в. Контуры суммы SS (θ), изображенные на рис. 10.7б, представляют довольно плохо обусловленную задачу и потому приводят к точке минимума суммы $S(\alpha, \beta)$, а именно к точке (β_1, α_1), которая весьма далеко удалена от фактической точки минимума суммы $S(\alpha, \beta)$. На следующей итерации контуры суммы квадратов для линеаризованной модели представляют уже относительно хорошо обусловленную задачу (рис. 10.7в), и их центр также, как мы видим, близок к точке минимума поверхности, описываемой суммой квадратов для нелинейной модели. Заметим, что на обоих рисунках 10.7б и 10.7в направление наискорейшего спуска, перпендикулярное к контурам суммы $S(\alpha, \beta)$ в стартовой точке, привело бы нас к траектории, которая отличается от траектории метода линеаризации.

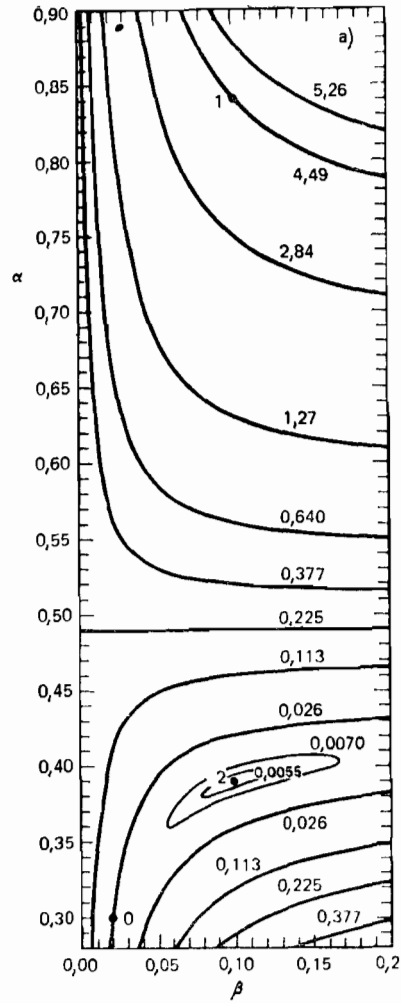


Рис. 10.7а. Контуры суммы квадратов $S(\alpha, \beta)$ для нелинейной модели. Точки 0, 1 и 2 соответствуют значениям параметров (β, α) в начальной точке и на первой и второй итерациях

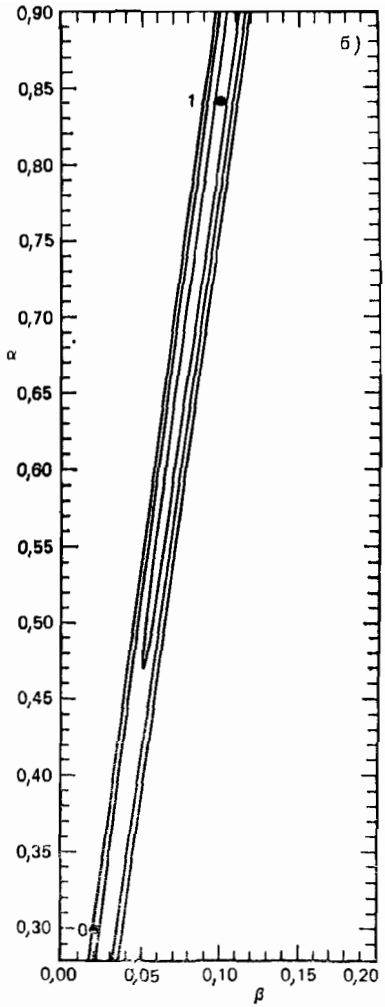


Рис. 10.7б. Контуры суммы квадратов S (α, β) для линеаризованной модели в точке (β, α) = (β_0, α_0) = (0,02; 0,30). Центр системы эллипсов имеет координаты (β_1, α_1) = (0,1007; 0,8416) и следующая итерация начинается с этой точки

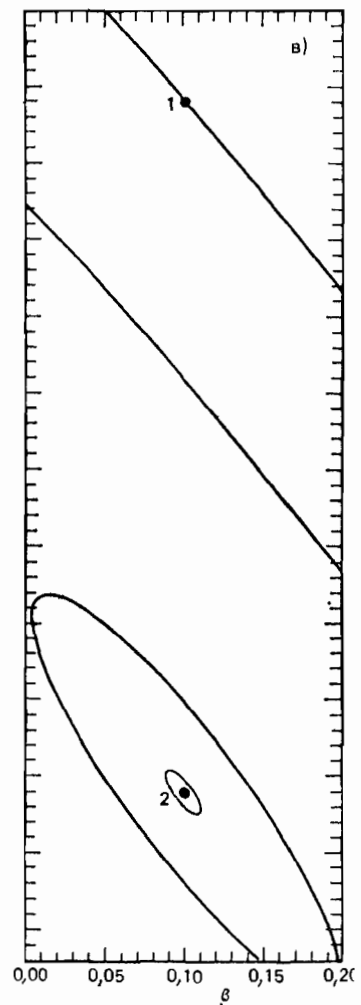


Рис. 10.7в. Контуры суммы квадратов SS (α, β) для линеаризованной модели в точке (β, α) = (β_1, α_1) = = (0,1007; 0,8416). Центр систем эллипсов имеет координаты (β_2, α_2) = = (0,1004; 0,3901) и следующая итерация начинается с этой точки

Дальнейший анализ

Обычные критерии, пригодные для линейных моделей, часто вообще *не* годятся в нелинейном случае. В качестве практической процедуры здесь можно рекомендовать, например, сопоставлять необъяс-

ненную вариацию с оценкой величины $V(Y_u) = \sigma^2$. Однако не удается воспользоваться F-статистикой, чтобы прийти к какому-либо заключению при определенном уровне значимости. Мерой необъясненной вариации служит, например, величина $S(\hat{\alpha}, \hat{\beta}) = 0{,}0050$. При отсутствии точных результатов для нелинейного случая можно считать, что сумма квадратов базируется примерно на $44-2 = 42$ степенях свободы (поскольку оцениваются два параметра). В нелинейном случае это не приводит обычно к несмещенной оценке величины σ^2, как в линейном случае, даже если модель и правильна.

Оценка дисперсии σ^2, получаемая на основе «чистых» ошибок (см. § 1.5), может быть найдена из параллельных наблюдений. Это дает сумму квадратов $S_{pe} = 0{,}0024$ с 26 степенями свободы.

Некоторое приближенное представление о возможной неадекватности модели можно получить путем вычисления величины $S(\hat{\alpha}, \hat{\beta}) - S_{pe} = 0{,}0026$ с $42-26 = 16$ степенями свободы и сравнения средних квадратов:

$$\frac{S(\hat{\alpha}, \hat{\beta}) - S_{pe}}{16} = 0{,}00016,$$

$$\frac{S_{pe}}{26} = 0{,}00009.$$

F-критерий здесь *неприменим*, но мы можем воспользоваться величиной $F(16; 26; 0{,}95) = 2{,}08$ как мерой сравнения. Поскольку $16/9 = 1{,}8$, то это позволяет сказать, что модель, вероятно, неплохо описывает экспериментальные данные.

Доверительные области

Можно вычислить приближенные $100(1-\alpha)$ %-ные доверительные контуры (описанные в § 10.2) путем определения (α, β), удовлетворяющих соотношению

$$S(\alpha, \beta) = S(\hat{\alpha}, \hat{\beta})\left[1 + \frac{p}{n-p} F(p, n-p, 1-q)\right] =$$

$$= 0{,}0050\,[1 + F(2; 42; 1-q)/21] = S_q.$$

Из уравнения (10.3.2) следует, что

$$\sum_{u=1}^{n} \{(Y_u - 0{,}49e^{-\beta(x_u-8)} + \alpha(e^{-\beta(x_u-8)} - 1)\}^2 = S_q.$$

или

$$A\alpha^2 + 2B\alpha + C - S_q = 0,$$

где

$$A = \sum_{u=1}^{n} (e^{-\beta(x_u-8)} - 1)^2,$$

$$B = \sum_{u=1}^{n} \left(Y_u - 0{,}49e^{-\beta(X_u-8)}\right)\left(e^{-\beta(X_u-8)} - 1\right),$$

$$C = \sum_{u=1}^{n} \left(Y_u - 0{,}49e^{-\beta(X_u-8)}\right)^2$$

есть функции только от β. Таким образом, выбор значения q позволяет вычислить

$$\alpha = \frac{\{-B \pm [B^2 - A(C - S_q)]^{1/2}\}}{A}$$

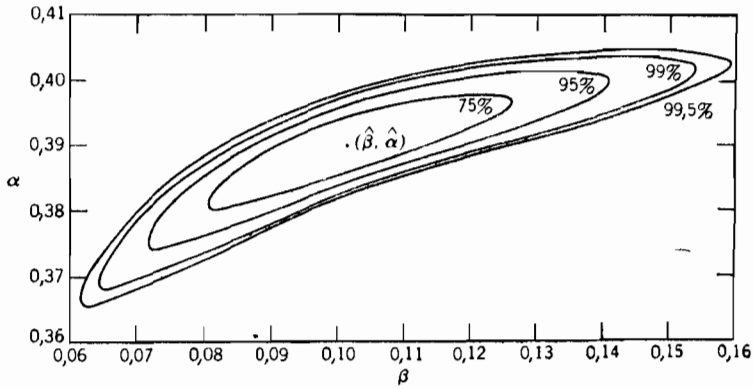


Рис. 10.8. Доверительные области для параметров (β, α). Области являются точными, доверительные вероятности — приближенные

для некоторого интервала значений β, чтобы получить границы доверительной области, отвечающие примерно 100 $(1-q)$ %-ной доверительной вероятности. На рис. 10.8 показаны 75, 95, 99 и 99,5 %-ные области, полученные при $q = 0{,}25$; 0,05; 0,01 и 0,005 соответственно. Точка на рисунке имеет координаты $(\hat{\beta}, \hat{\alpha})$.

Не будет нереалистичным считать, что среди точек, лежащих внутри контура, помеченного, скажем, числом 95 % (отвечающего 95 %-ной доверительной вероятности), находится точка, соответствующая истинным значениям параметров β и α. Ориентация и форма контуров указывают на то, что параметр β определен хуже, чем $\hat{\alpha}$. (Применительно к линейному случаю этот вопрос обсуждается в § 10.5; см. также § 10.4.) [14]

---

[14] В последнее время появился альтернативный подход к построению доверительных областей для нелинейных моделей. Он основан на так называемой процедуре «бутстреп», изложенной, например, в книге: Efron B. The Jackknife, the Bootstrap and Other Resampling Plans.—Philadelphia, Pa: SIAM, 1982.—92 с. Издательством «Финансы и статистика» намечен перевод этой книги на русский язык.—*Примеч. пер.*

Некоторые характерные особенности распечаток при использовании типовых программ нелинейного регрессионного анализа

В разных программах нелинейного МНК можно увидеть отличающиеся друг от друга распечатки. Однако многие из них содержат некоторые повторяющиеся особенности. Опишем и проиллюстрируем их кратко на примере, в котором применяется метод Маркуардта.

1. *Сингулярные числа* [15] — это величины, равные корню квадратному из соответствующих собственных чисел (см. § 6.9 и 6.10) матриц $\mathbf{Z}_j'\mathbf{Z}_j$, где матрица $\mathbf{Z}_j$ получается на основе (10.2.4), но с использованием текущего вектора $\boldsymbol{\theta}_j$, а не $\boldsymbol{\theta}_0$. Сильно различающиеся между собой сингулярные числа указывают на плохую обусловленность задачи. Соответствующие сингулярные векторы определяют ориентацию осей контуров суммы квадратов для линеаризованной модели по отношению к θ осям:

Пример 1. Для $\hat{\theta}$ имеют место следующие результаты:

Сингулярные числа	4,964	0,773
α:	−0,946	0,326
β:	0,326	0,946

Отношение наибольшего и наименьшего сингулярных чисел (в данном примере их всего два, но в общем случае их число равно p, т. е. числу параметров) равно 6,4. Эта величина указывает на довольно хорошую обусловленность задачи. (В плохо обусловленных задачах сингулярные числа могут отличаться в тысячи раз. Если их отношение близко к единице, то контуры суммы квадратов для линеаризованной модели близки к круговым.) Чтобы получить одинаковое изменение суммы SS (θ), следовало бы сдвинуться примерно на 6,4 ед. вдоль оси β и лишь на одну единицу вдоль оси α. На основании приведенных выше результатов мы видим, что первое сингулярное направление

$$\Phi_1 = -0{,}946\alpha + 0{,}326\beta$$

почти совпадает с обратным направлением оси α, в то время как

$$\Phi_2 = 0{,}326\alpha + 0{,}946\beta,$$

а это означает, что второе сингулярное направление почти совпадает с направлением оси β. Рис. 10.8 подтверждает сказанное. Заметим,

---

[15] Сингулярное разложение используется для представления прямоугольных матриц в виде произведения трех матриц. При этом первый и третий сомножители составлены из ортонормированных векторов, а второй содержит на главной диагонали сингулярные числа. Такое представление матриц описано в ряде книг. См. например: Алберт А. Регрессия, псевдоинверсия и рекуррентное оценивание/Пер. с англ. Под ред. Я. З. Цыпкина.— М.: Наука, 1977.— 224 с.; Форсайт Дж., Малькольм М., Моулер К. Машинные методы математических вычислений/Пер. с англ.— М.: Мир, 1980.— 280 с.; Стренг Г. Линейная алгебра и ее применения/Пер. с англ. Под ред. Г. И. Марчука.— М.: Мир, 1980.— 456 с.; Беклемишев Д. В. Дополнительные главы линейной алгебры.— М.: Наука, 1983.— 336 с.— *Примеч. пер.*

что сумма квадратов коэффициентов, входящих в приведенные уравнения (т. е. величин — 0,946 и + 0,326), равна 1 независимо от ошибок округления, и это верно для всех таких уравнений.

2. *Нормализирующие элементы* — величины, равные корню квадратному из диагональных элементов матрицы $\mathbf{Q} = ((q_{il})) = (\mathbf{Z}_j'\mathbf{Z}_j)^{-1}$. Приближенная *корреляционная матрица* $\mathbf{C} = ((c_{il}))$ для $\hat{\boldsymbol{\theta}}$ получается путем деления каждой строки и столбца матрицы на соответствующие нормализирующие элементы, а именно $c_{il} = q_{il}/(q_{ii}q_{ll})^{1/2}$.

Пример 2.

Параметры	α	β
Нормализирующие элементы	0,462	1,226
Приближенная корреляционная матрица α	1,000	0,888
β	0,888	1,000

Коэффициент корреляции между $\hat{\alpha}$ и $\hat{\beta}$ положителен. Это указывает на то, что сравнительно небольшое *одновременное* увеличение или *одновременное* уменьшение параметров α и β по сравнению с $\hat{\alpha}$ и $\hat{\beta}$ почти не сказывается на величине суммы квадратов SS (θ). Из рис. 10.8 видно, как это происходит. (Напомним, что эллипсоидные контуры суммы квадратов SS (θ) с центром в точке $\hat{\theta}$ лишь имитируют истинные контуры суммы квадратов S (θ), но не совпадают с ними в точности.) Хотя коэффициент корреляции, равный 0,888, и достаточно высок, но не настолько, чтобы считать модель перепараметризованной. (Более детально об этом в следующем разделе.)

3. *Доверительные пределы* для истинных параметров θ могут быть вычислены на основе линейной аппроксимации при $\theta = \hat{\theta}$. Обычно их выражают в виде [16] $\hat{\theta}_i \pm 2\text{s.e.}(\hat{\theta}_i)$, где $\text{s.e.}(\hat{\theta}_i)$ — соответствующий диагональный элемент матрицы $[(\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{-1}s^2]^{1/2}$, а $\hat{\mathbf{Z}}$ — матрица $\mathbf{Z}$, составленная согласно (10.2.4), в которой вектор $\boldsymbol{\theta}_0$ заменен на $\hat{\boldsymbol{\theta}}$ и $s^2 = S(\hat{\boldsymbol{\theta}})/(n-p)$. В общем случае этим пределам соответствует доверительная вероятность, составляющая примерно 95 %.

Пример 3.

	Нижний предел	Окончательная оценка	Верхний предел
α	0,380	0,390	0,400
β	0,075	0,102	0,128

Окончательная оценка для α равна $\hat{\alpha} = 0{,}390$, а приближенный, 95 %-ный доверительный интервал для α есть (0,380; 0,400). Окончательная оценка для β равна $\hat{\beta} = 0{,}102$; приближенный 95 %-ный до-

[16] Здесь авторы вопреки принятому ранее соглашению используют обозначение s. e. вместо s. d.— *Примеч. пер.*

верительный интервал для β составляет (0,075; 0,128). Это согласуется с рис. 10.8. (При необходимости читатель может вернуться к с. 131 *, где обсуждаются вопросы доверительного оценивания. При этом надо иметь в виду, во-первых, что в данном случае имеет силу предостережение, высказанное на с. 131 *, во-вторых, что высказанные выше утверждения о доверительном оценивании параметров нелинейной модели делались на основе ее линейной аппроксимации и потому не слишком строги.) Оба интервала не включают нуль, и это указывает на то, что величины α и β, по-видимому, отличны от нуля.

Перепараметризация

Ситуации, в которых модель содержит больше параметров, чем это необходимо для представления данных, вообще говоря, проявляются в структуре корреляционной матрицы оцениваемых коэффициентов, составленной на основе линеаризованной модели. Эта матрица получается из матрицы $(\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{-1}$, если поделить каждый ее элемент, содержащийся в i-й строке и j-м столбце, на

$$\{[i\text{-диагональный элемент } (\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{-1}) \times$$
$$\times [j\text{-диагональный элемент } (\hat{\mathbf{Z}}'\hat{\mathbf{Z}})^{-1}]\}^{1/2}.$$

В результате такой нормализации на главной диагонали окажутся единицы, а недиагональные элементы будут равны коэффициентам корреляции оценок параметров. Если некоторые из этих коэффициентов близки по абсолютному значению к единице, то это указывает на то, что модель содержит слишком много параметров или, более точно, что в этом случае репараметризованная модель, содержащая меньшее число параметров, может работать столь же хорошо, как и исходная. При этом не обязательно исходная модель будет неподходящей для исследуемой физической ситуации. Просто получено указание на то, что имеющиеся в наличии экспериментальные данные могут быть неадекватны задаче оценивания всех параметров, содержащихся в модели. (Так, например, если для линейной модели установлено, что некоторый параметр β не кажется отличным от нуля, то отсюда еще не следует неэффективность соответствующей предикторной переменной X; такой эффект может быть связан с тем, что *в данном конкретном наборе данных* эта предикторная переменная изменялась недостаточно, чтобы ее эффект был различимым. Подобная, но более сложная картина может возникать и в нелинейном случае.)

В случае перепараметризованных моделей контуры суммы квадратов зачастую сильно вытянуты в некоторых направлениях параметрического пространства, и в подобных ситуациях исследование размерности контуров суммы квадратов или их сечений может быть очень полезным.

* См. книгу 1. — *Примеч. пер.*

10.4. НЕКОТОРЫЕ ЗАМЕЧАНИЯ О РЕПАРАМЕТРИЗАЦИИ МОДЕЛИ

Если поверхность суммы квадратов, определяемая уравнением (10.1.5), содержит длинные вытянутые овраги, то, вероятно, при любой итеративной процедуре оценивания будет иметь место медленная сходимость. В качестве простого примера рассмотрим линейную модель вида $Y_u = \theta_0 + \theta_1 X_u + \varepsilon_u$ и предположим, что имеются три наблюдения Y_1, Y_2 и Y_3 при $X = 9$, 10 и 11. Тогда

$$S(\theta) = (Y_1 - \theta_0 - 9\theta_1)^2 + (Y_2 - \theta_0 - 10\theta_1)^2 + (Y_3 - \theta_0 - 11\theta_1)^2 =$$

$$= \sum_{u=1}^{3} Y_u^2 - 2\theta_0 \sum_{u=1}^{3} Y_u - 2\theta_1 (9Y_1 + 10Y_2 + 11Y_3) + 3\theta_0^2 + 302\theta_1^2 + 60\theta_0\theta_1.$$

В координатах (θ_0, θ_1) контуры $S(\theta) = \text{const}$ получаются тонкими вытянутыми эллипсами. Такая поверхность суммы квадратов может быть названа *плохо обусловленной*. Однако, если переписать модель в виде

$$Y_u = (\theta_0 + \theta_1 \overline{X}) + \theta_1 (X_u - \overline{X}) + \varepsilon_u = \Phi_0 + \Phi_1 x_u + \varepsilon_u,$$

где $\Phi_0 = \theta_0 + \theta_1 \overline{X}$, $\Phi_1 = \theta_1$, $x_u = X_u - \overline{X} = (-1; 0; 1$ для $u = 1, 2, 3)$, то получим снова сумму квадратов с параметрами $\Phi = (\Phi_1, \Phi_2)'$, которая выражается уравнением

$$S(\Phi) = \Sigma (Y_u - \Phi_0 - \Phi_1 x_u)^2 = (Y_1 - \Phi_0 + \Phi_1)^2 + (Y_2 - \Phi_0)^2 + (Y_3 -$$

$$- \Phi_0 - \Phi_1)^2 = \sum_{u=1}^{3} Y_u^2 - 2\Phi_0 \sum_{u=1}^{3} Y_u + 2\Phi_1 (Y_1 - Y_3) + 3\Phi_0^2 + 2\Phi_1^2.$$

В координатах (Φ_0, Φ_1) контуры этой поверхности станут хорошо округленными эллипсами. Такую поверхность называют *хорошо обусловленной*.

Случай плохой обусловленности может иметь место при использовании нелинейной модели вида $Y_u = \theta_0 e^{\theta_1 X_u} + \varepsilon_u$, если среднее значение аргумента X_u, равное $\overline{X}$, не близко к нулю. Когда имеешь дело с выражениями такого типа, лучше представлять модель в альтернативной форме:

$$Y_u = \left(\theta_0 e^{\theta_1 \overline{X}}\right)\left(e^{\theta_1 (X_u - \overline{X})}\right) + \varepsilon_u = \Phi_0 e^{\Phi_1 x_u} + \varepsilon_u,$$

где

$$\Phi_0 = \theta_0 e^{\theta_1 \overline{X}}, \quad \Phi_1 = \theta_1 \quad \text{и} \quad x_u = X_u - \overline{X}.$$

(Примечание. В примере из § 10.3 такое преобразование не проводилось, поскольку оно привело бы к усложнению модели, так как параметр α в ней содержится дважды. Замены одного параметра другим в таком случае недостаточно.)

Подходящая репараметризация модели, позволяющая улучшить обусловленность поверхности суммы квадратов, в общем случае не всегда очевидна. Простые преобразования, обеспечивающие «центрирование» некоторых переменных, как это имело место в рассмотренном примере, часто могут оказаться полезными. В худшем случае

они не принесут пользы, но и не нанесут вреда. Дополнительные комментарии по репараметризации моделей содержатся в работе [17]: Box G. E. P. Some notes on nonlinear estimation.— Madison, Wisconsin: Statistics Department Technical Report, N 25, University of Wisconsin, 1964.

10.5. ГЕОМЕТРИЯ ЛИНЕЙНОГО МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ

Чтобы понять, почему итеративные методы при использовании их в нелинейных задачах не всегда приводят к успеху, полезно рассмотреть в первую очередь геометрическую интерпретацию линейного метода наименьших квадратов. В линейном случае, используя обозначения данной главы, можно записать модель так:

$$Y = f(\boldsymbol{\xi}, \boldsymbol{\theta}) + \boldsymbol{\varepsilon} = \theta_1 X_1 + \theta_2 X_2 + \ldots + \theta_p X_p + \varepsilon,$$

где X_i есть функции от $\boldsymbol{\xi}$. Если наблюдения Y_u, содержащие ошибки ε_u, соответствуют значениям $X_{1u}, X_{2u}, \ldots, X_{pu}$ переменных X_i при $u = 1, 2, \ldots, n$, то можно записать модель в альтернативной форме:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon},$$

где

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} X_{11} & X_{21} & \ldots & X_{p1} \\ X_{12} & X_{22} & \ldots & X_{p2} \\ & \cdot & & \cdot \\ & \cdot & & \cdot \\ & \cdot & & \cdot \\ X_{1n} & X_{2n} & \ldots & X_{pn} \end{bmatrix},$$

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \cdot \\ \cdot \\ \cdot \\ \theta_p \end{bmatrix}, \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}.$$

(Заметим, что можно получить «член β_0» в этой модели, полагая, $X_{1u} = 1$ для $u = 1, 2, \ldots, n$.) Поверхность суммы квадратов, соответствующую уравнению (10.1.5), можно записать так:

$$S(\boldsymbol{\theta}) = \sum_{u=1}^{n} \left[Y_u - \sum_{i=1}^{p} \theta_i X_{iu} \right]^2 = (\mathbf{Y} - \mathbf{X}\boldsymbol{\theta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\theta}) = \mathbf{Y}'\mathbf{Y} - 2\boldsymbol{\theta}\,\mathbf{X}'\mathbf{Y} + \boldsymbol{\theta}'\mathbf{X}'\mathbf{X}\boldsymbol{\theta}.$$

---

[17] Репараметризация нелинейных моделей с целью улучшения их обусловленности наиболее полно рассмотрена в книге: Бард Й. Нелинейное оценивание параметров/Пер. с англ. Под ред. В. Г. Горского.— М.: Статистика, 1979.— 351 с. Эта процедура сродни преобразованиям моделей, которые подробно обсуждаются в гл. 5 данной книги.— *Примеч. пер.*

Продифференцируем это выражение по $\boldsymbol{\theta}$, приравняем результаты к нулю и заменим затем $\boldsymbol{\theta}$ на $\hat{\boldsymbol{\theta}}$. Получим нормальные уравнения

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{X}'\mathbf{Y}.$$

Если $\mathbf{X}'\mathbf{X}$ — невырожденная матрица, то решение данной системы уравнений выражается формулой

$$\hat{\boldsymbol{\theta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.$$

Напомним, что сумма квадратов, обусловленная регрессией, есть $\hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{Y}$, а остаточная сумма квадратов выражается формулой $\mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{Y}$. Далее,

$$S(\hat{\boldsymbol{\theta}}) = \mathbf{Y}'\mathbf{Y} - 2\hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}} = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{Y} - \hat{\boldsymbol{\theta}}' \times$$
$$\times (\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}}) = \mathbf{Y}'\mathbf{Y} - \hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{Y},$$

так как $\hat{\boldsymbol{\theta}}$ удовлетворяет нормальным уравнениям. Таким образом, $S(\hat{\boldsymbol{\theta}})$ есть наименьшее значение величины $S(\boldsymbol{\theta})$, и она равна остаточной сумме квадратов в таблице дисперсионного анализа. Мы можем, следовательно, записать

$$S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}\,\mathbf{X}'\mathbf{X}\boldsymbol{\theta} - 2\boldsymbol{\theta}\,\mathbf{X}'\mathbf{Y} + \hat{\boldsymbol{\theta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}} = (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\,\mathbf{X}'\mathbf{X}\,(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}).$$

Когда ошибки ε_u независимы и каждая из них подчиняется распределению $N(0, \sigma^2)$, т. е. $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2)$, можно показать, что если модель верна, то верны следующие результаты:

1. $\hat{\boldsymbol{\theta}} \sim N[\boldsymbol{\theta}, (\mathbf{X}'\mathbf{X})^{-1}\sigma^2]$
2. $S(\hat{\boldsymbol{\theta}}) \sim \sigma^2\chi^2_{n-p}$.
3. $S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}}) \sim \sigma^2\chi^2_p$.

4. $S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})$ и $S(\hat{\boldsymbol{\theta}})$ распределены независимо, так что их отношение подчиняется распределению Фишера:

$$\frac{[S(\boldsymbol{\theta}) - S(\hat{\boldsymbol{\theta}})]/p}{S(\hat{\boldsymbol{\theta}})/(n-p)} \sim F(p,\ n-p).$$

Имеются два различных, но связанных между собой способа исследования контуров, определяемых соотношением $S(\boldsymbol{\theta}) = \text{const}$. Их можно исследовать в *выборочном пространстве* (в котором существо метода наименьших квадратов объясняется наилучшим образом), а также в *пространстве параметров* (где ограничиваются указаниями только контуров поверхности суммы квадратов $S(\boldsymbol{\theta})$). Обсудим оба эти способа.

Выборочное пространство

Выборочное пространство — это n-мерное пространство. Вектор наблюдений $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)'$ означает вектор $\overrightarrow{OY}$, начало которого совпадает с точкой O, а конец — с точкой Y, имеющей коорди-

наты $\mathbf{Y}=(Y_1, Y_2, \ldots, Y_n)$. Матрица $\mathbf{X}$ имеет p столбцов, каждый из которых содержит n элементов. Элементы j-го столбца определяют координаты $(X_{j1}, X_{j2}, \ldots, X_{jn})$ точки X_j в выборочном пространстве, и j-й вектор-столбец матрицы $\mathbf{X}$ определяет вектор $\overrightarrow{OX}$ в выборочном пространстве. Множество из p векторов $\overrightarrow{OX}_1, \overrightarrow{OX}_2, \ldots, \overrightarrow{OX}_p$ определяет [18] подпространство размерности p, называемое *пространством оценок* (estimation space), которое содержится в выборочном пространстве. Любая точка этого подпространства может быть представлена с помощью вектора, являющегося линейной комбинацией векторов, порождающих это пространство, т. е. столбцов матрицы $\mathbf{X}$. Так, например, $\mathbf{X\theta}$, где $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$, представляется $(p \times 1)$-вектором в этом пространстве. Предположим, что вектор $\mathbf{X\theta}$ определяет точку T. Тогда квадрат расстояния YT^2 задается выражением

$$(\mathbf{Y}-\mathbf{X\theta})'(\mathbf{Y}-\mathbf{X\theta}) = S(\boldsymbol{\theta}),$$

как было показано ранее. Таким образом, сумма квадратов $S(\boldsymbol{\theta})$ в выборочном пространстве есть квадрат расстояния от точки Y до произвольной точки T в пространстве оценок. Минимизация $S(\boldsymbol{\theta})$ по $\boldsymbol{\theta}$ как раз и предполагает отыскание такой величины $\boldsymbol{\theta}$ (обозначаемой буквой $\hat{\boldsymbol{\theta}}$), которая дает точку P (определяемую вектором $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\theta}}$) в пространстве оценок, наиболее близкую к точке Y. Следовательно, точка P должна быть основанием перпендикуляра к пространству оценок, проведенного через точку Y, т. е. принадлежать прямой ортогональной ко всем вектор-столбцам матрицы $\mathbf{X}$. С помощью векторов, выходящих из начала координат, мы можем записать

$$\mathbf{Y} = \hat{\mathbf{Y}} + (\mathbf{Y}-\hat{\mathbf{Y}}) = \hat{\mathbf{Y}} + \mathbf{e},$$

где $\mathbf{e}$ есть *вектор остатков*. Вектор $\mathbf{Y}$, таким образом, раскладывается на две ортогональные компоненты: 1) вектор $\hat{\mathbf{Y}}$, принадлежащий пространству оценок, и 2) вектор $\mathbf{Y}-\hat{\mathbf{Y}} = \mathbf{e}$ — вектор остатков, принадлежащий пространству, именуемому *пространством ошибок*. Пространство ошибок определяется как $(n-p)$-мерное подпространство, которое содержится в полном n-мерном пространстве. Оно представляет собой ортогональное дополнение пространства оценок. Можно показать алгебраически, что $\hat{\mathbf{Y}}$ и $\mathbf{e}$ ортогональны:

$$\mathbf{Y}'\mathbf{e} = (\mathbf{X}\hat{\boldsymbol{\theta}})'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}}'\mathbf{X}'(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\theta}}) = \hat{\boldsymbol{\theta}}'(\mathbf{X}'\mathbf{Y}-\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\theta}}) = \mathbf{0}.$$

Поскольку $\hat{\boldsymbol{\theta}}$ удовлетворяет нормальным уравнениям, никакие комментарии к этим выкладкам не требуются. Вектор $\mathbf{e}$ есть вектор $\overrightarrow{OR}$,

---

[18] При условии, конечно, что все они линейно независимы.— *Примеч. пер.*

выходящий из начала координат O. Он имеет длину $OR = YP$, и при этом данный вектор параллелен вектору $\overrightarrow{PY}$ [19].

Если T — произвольная точка в пространстве оценок, а вектор $\overrightarrow{YP}$ ортогонален к этому пространству, то

$$YT^2 = YP^2 + PT^2$$

или

$$S(\theta) = S(\hat{\theta}) + PT^2.$$

Таким образом, контуры, для которых соблюдается условие $S(\theta) = \text{const}$, должны удовлетворять соотношению

$$PT^2 = S(\theta) - S(\hat{\theta}) = \text{const}.$$

Следовательно, контуры, определяемые из условия $S(\theta) = \text{const}$, содержат все точки T в выборочном пространстве, для которых $PT^2 = \text{const}$, т. е. точки вида $\mathbf{X}\theta$, лежащие на p-мерной сфере в пространстве оценок с центром в точке P, задаваемой вектором $\mathbf{X}\hat{\theta}$. Радиус этой сферы равен $[S(\theta) - S(\hat{\theta})]^{1/2}$. Используя приведенное ранее соотношение

$$\frac{[S(\theta) - S(\hat{\theta})]/p}{S(\hat{\theta})/(n-p)} \sim F(p,\ n-p),$$

можно определить границы 100 $(1-\alpha)$ %-ной доверительной области для точки $\mathbf{X}\theta$ с помощью равенства

$$\frac{[S(\theta) - S(\hat{\theta})]/p}{S(\hat{\theta})/(n-p)} = F(p,\ n-p,\ 1-\alpha);$$

откуда получаем

$$S(\theta) = S(\hat{\theta})\left[1 + \frac{p}{n-p} F(p,\ n-p,\ 1-\alpha)\right].$$

Это выражение можно представить в форме $S(\hat{\theta})(1+q^2)$, которая показывает, насколько величина $S(\theta)$ больше, чем минимальная величина суммы квадратов, т. е. $S(\hat{\theta})$. Доверительная область будет, таким образом, включать внутренность сферы в пространстве оценок с центром в точке P и радиусом, равным

$$[S(\theta) - S(\hat{\theta})]^{1/2} = \left[S(\hat{\theta})\frac{p}{n-p} F(p,\ n-p,\ 1-\alpha)\right]^{1/2}.$$

---

[19] Приведенная выше геометрическая трактовка метода наименьших квадратов была впервые дана в известной работе: Колмогоров А. Н. К обоснованию метода наименьших квадратов.— Успехи матем. наук, 1946, вып. 1, с. 57—70.— *Примеч. пер.*

Выборочное пространство при $n=3$, $p=2$

Чтобы проиллюстрировать ранее высказанные соображения, будем исходить из предположения, что $n = 3$. Если $n > 3$, то возникнет более сложная ситуация, которая не изображается графически. Однако мысленное обобщение результатов, полученных при $n = 3$, на случай большей размерности труда не составит.

На рис. 10.9 изображено выборочное пространство для случая, когда $n = 3$, $p = 2$, координатные оси обозначены индексами 1, 2

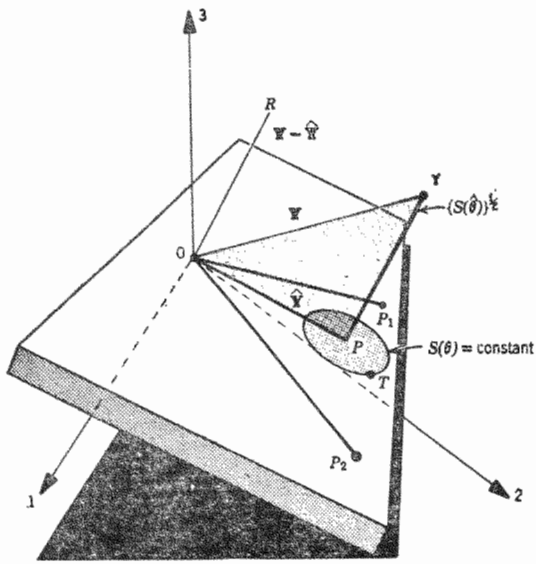


Рис. 10.9. Выборочное пространство при $n = 3$, $p = 2$

и 3, что соответствует трем компонентам (Y_1, Y_2, Y_3) вектора $\mathbf{Y}'$. Мы будем предполагать далее, что имеется $p = 2$ параметров θ_1 и θ_2 и $\mathbf{X}$ есть (3×2)-матрица вида

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{21} \\ X_{12} & X_{22} \\ X_{13} & X_{23} \end{bmatrix}.$$

Столбцы матрицы $\mathbf{X}$ определяют две точки P_1 и P_2 с координатами (X_{11}, X_{12}, X_{13}) и (X_{21}, X_{22}, X_{23}) соответственно, а векторы $\overrightarrow{OP}_1$ и $\overrightarrow{OP}_2$ определяют плоскость, задающую двумерное пространство оценок, в котором должен лежать вектор $\hat{\mathbf{Y}} = \mathbf{X}\hat{\theta}$. Точка Y лежит вне этой плоскости, и перпендикуляр YP из точки Y на плоскость OP_1P_2 пересекает плоскость в точке P. Таким образом, YP есть наикрат-

чайшее расстояние от Y до любой точки в пространстве оценок; точка P определяется с помощью соотношений $\hat{\mathbf{Y}} = \mathbf{X}\hat{\theta}$ и $S(\hat{\theta}) = YP^2$. Поскольку $OY^2 = \mathbf{Y}'\mathbf{Y}$, стандартное разложение, используемое в дисперсионном анализе, можно записать в виде

$$\mathbf{Y}'\mathbf{Y} = \hat{\theta}'\mathbf{X}'\mathbf{Y} + (\mathbf{Y}'\mathbf{Y} - \hat{\theta}'\mathbf{X}'\mathbf{Y})$$

или

$$\mathbf{Y}'\mathbf{Y} = \hat{\theta}'\mathbf{X}'\mathbf{Y} + S(\hat{\theta}),$$

что эквивалентно теореме Пифагора

$$OY^2 = OP^2 + YP^2.$$

Если мы теперь построим отрезок OR, проходящий через точку O, равный по длине (так как $OR^2 = S(\hat{\theta})$) и параллельный вектору $\overrightarrow{PY}$, то $\overrightarrow{OR}$ будет вектором остатков $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$. Вектор $\overrightarrow{OP}$ есть вектор $\hat{\mathbf{Y}}$, так что мы имеем векторное уравнение

$$\overrightarrow{OY} = \overrightarrow{OP} + \overrightarrow{OR}$$

или

$$\mathbf{Y} = \hat{\mathbf{Y}} + (\mathbf{Y} - \hat{\mathbf{Y}}).$$

Теперь мы можем сказать, что контуры постоянных значений $S(\theta)$ вообще представляются в пространстве оценок сферами. Однако на плоскости OP_1P_2 контуры представляют собой окружности. Это легко видеть, ибо если T есть произвольная точка $\mathbf{X}\theta$ на плоскости, то, $S(\theta) = \text{const}$ означает, что $YT^2 = \text{const}$, так что $PT^2 = YT^2 - YP^2 = \text{const}$. Таким образом, мы получаем окружность с центром в точке P. Одна такая окружность показана на рисунке. Окружность, которая дает 100 $(1-\alpha)$ %-ную доверительную область для точки $\mathbf{X}\theta$, соответствующей теоретическим значениям параметров, имеет радиус

$$[2S(\hat{\theta})\, F(2;\ 1;\ 1-\alpha)]^{1/2}.$$

Это соотношение получается путем подстановки $n = 3$, $p = 2$ в общую формулу.

Геометрия выборочного пространства, когда модель ошибочна

Предположим, что $\mathbf{Y} = \mathbf{X}\theta + \varepsilon$ есть постулируемая линейная модель, содержащая p параметров, а истинная линейная модель выражается уравнением

$$\mathbf{Y} = \mathbf{X}\theta + \mathbf{X}_2\theta_2 + \varepsilon$$

и включает дополнительные слагаемые $X_2\theta_2$, не учитываемые в постулируемой модели. Поскольку пространство оценок содержит лишь точки вида $\mathbf{X}\theta$, истинная точка $\eta = \mathbf{X}\theta + \mathbf{X}_2\theta_2$ не может лежать в пространстве оценок. В этом случае перпендикуляр YP из Y на пространство оценок (основание которого P задается с помощью соотношения $\hat{\mathbf{Y}} = \mathbf{X}\hat{\theta}$) будет длиннее, чем он был бы, если бы исполь-

зовалась правильная модель и соответствующее ей пространство оценок. Чтобы проиллюстрировать это, на рис. 10.10 приведено изображение для случая $n = 3$, $p = 1$, когда истинная модель включает два параметра θ и θ_2. Эта модель имеет вид

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} X_{11} \\ X_{12} \\ X_{13} \end{bmatrix} \theta + \begin{bmatrix} X_{21} \\ X_{22} \\ X_{23} \end{bmatrix} \theta_2 + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \end{bmatrix} = \mathbf{X}\theta + \mathbf{X}_2\theta_2 + \boldsymbol{\varepsilon},$$

а постулируемая модель получается из нее при $\theta_2 = 0$. Единственный столбец матрицы $\mathbf{X}$ определяет точку P_1 и линию OP_1, представляющую собой пространство оценок для постулируемой модели. Отрезок YP есть перпендикуляр из точки Y на OP_1, а P есть точка $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\theta}}$. Следовательно, наименьший квадрат расстояния $S(\hat{\boldsymbol{\theta}})$ среди всех возможных квадратов расстояний $S(\boldsymbol{\theta})$ от точки Y до точки $\mathbf{X}\boldsymbol{\theta}$ на линии OP_1 представляет собой квадрат длины отрезка YP. Истинная величина θ определяет неизвестную точку $\mathbf{X}\theta$ на линии OP_1. Доверительный интервал для истинной величины $\mathbf{X}\theta$ может быть построен на OP_1, и это есть окрестность точки P.

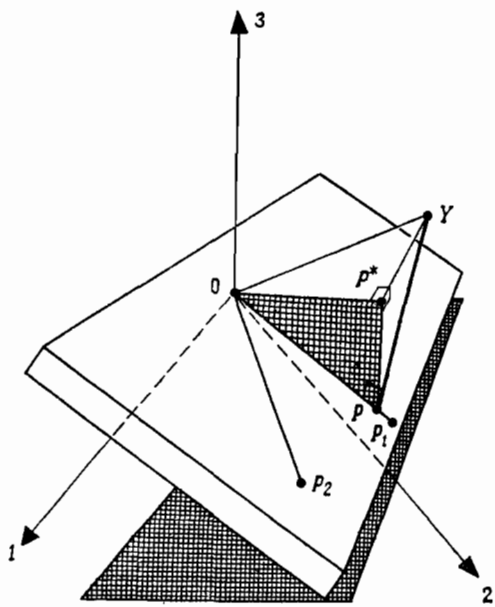


Рис. 10.10. Выборочное пространство при $n = 3$, $p = 1$; модель ошибочна

Теперь второй вектор $\mathbf{X}_2$ в истинной модели определяет линию OP_2, а линии OP_1 и OP_2 определяют плоскость, в которой лежит истинная точка $\mathbf{X}\theta + \mathbf{X}_2\theta_2$. Предположим, что YP^* — есть перпендикуляр из точки Y на *истинное пространство оценок*, задаваемое плоскостью OP_1P_2. Тогда P^* представляет собой точку, которая получилась бы при правильном определении величины $\hat{Y}^*$, *если бы* использовалась правильная модель. Этот перпендикуляр всегда имеет длину, меньшую или равную длине отрезка YP, так как «перпендикуляр к пространству» (в данном случае к плоскости OP_1P_2) не может быть длиннее, чем перпендикуляр к подпространству (в нашем случае — к линии OP_1). Следовательно, если модель неправильна, то $S(\hat{\boldsymbol{\theta}}) = YP^2$ будет во всяком случае слишком большой. (Заметим,

что P и P^* могут совпасть, и тогда величина $S(\hat{\theta})$ будет минимальной, какая бы модель ни использовалась. Но такое совпадение может быть лишь случайным.)

Если постулируемая модель правильна, то в общем случае величина $S(\hat{\theta})$ имеет математическое ожидание, равное $(n-p)\sigma^2$. Зная некоторую априорную оценку дисперсии σ^2 илн получив ее на основе «чистых» ошибок, можно определить приблизительно величину YP^2. Одиако если постулируемая модель не адекватна, то величина YP будет, вероятно, слишком большой. Стандартиая проверка адекватности, выполняемая с использованием отношения (2.6.12), предназначается, таким образом, для ответа на вопрос: будет ли квадрат длины отрезка YP бо̀льшим, чем это следует ожидать на основании имеющейся информации о величиие случайной ошибки эксперимента? Какое именно значение YP^2 следует считать *слишком большим*, определяется на основании характеристик распределения, как это делалось прежде.

Геометрическая интерпретация «чистой» ошибки

Геометрическая интерпретация «чистой» ошибки показана на рис. 10.11. В выборочном пространстве точки O соответствует началу координат, Y есть конец вектора наблюдений $\mathbf{Y}$, а P — основание

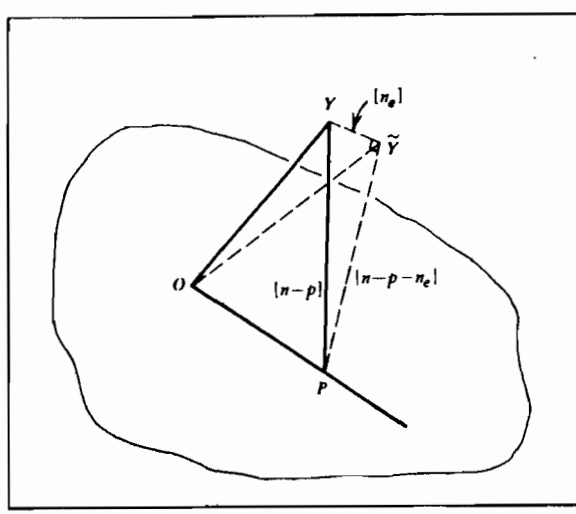


Рис. 10.11. Геометрическая интерпретация чистой ошибки. Символ $[n_e]$ означает, что вектор $\mathbf{Y}-\tilde{\mathbf{Y}}$, который параллелен линии $Y\tilde{Y}$, лежит в подпространстве выборочного пространства размерностью n_e

перпендикуляра, опущенного из точки Y на пространство оценок, порождаемое вектор-столбцами матрицы $\mathbf{X}$. Следовательно, $\overrightarrow{OP}$ — вектор, выражаемый формулой $\hat{\mathbf{Y}} = \mathbf{X}\hat{\theta}$. Точка $\tilde{Y}$ представляет собой конец вектора $\tilde{\mathbf{Y}}$, i-й элемент которого, $i = 1, 2, \ldots, n$, есть

$\tilde{Y}_i$= (среднее значение отклика по группе повторных опытов, к которым принадлежит Y_i)$=\bar{Y}_{i0}$.
Если группа содержит всего один опыт и повторения отсутствуют, то $\bar{Y}_{i0} = Y_i$. Легко видеть, что векторы $\mathbf{Y}-\tilde{\mathbf{Y}}$ и $\hat{\mathbf{Y}}-\tilde{\mathbf{Y}}$ ортогональны, так что отрезки $Y\tilde{Y}$ и $\tilde{Y}P$ взаимно перпендикулярны. В соответствии с теоремой Пифагора $Y\tilde{Y}^2 + \tilde{Y}P^2 = YP^2$. Эти результаты (в несколько иных обозначениях можно обнаружить в ответе к упражнению 8 гл. 1). Таким образом,

YP^2 — квадрат длины вектора остатков $\mathbf{Y}-\mathbf{X}\hat{\theta}$, т. е. сумма квадратов остатков для подгоняемой модели;

$Y\tilde{Y}^2$ — квадрат длины вектора $\mathbf{Y}-\tilde{\mathbf{Y}}$, т. е. сумма квадратов, обусловленная «чистой» ошибкой;

$\tilde{Y}P^2$ — квадрат длины вектора $\hat{\mathbf{Y}}-\tilde{\mathbf{Y}}$, т. е. сумма квадратов, определяющая неадекватность модели.

F — критерий для проверки гипотезы об адекватности модели основан на сопоставлении величины $\tilde{Y}P^2/(n-p-n_e)$ с $Y\tilde{Y}^2/n_e$, где n_e — число степеней свободы для суммы, обусловленной «чистой» ошибкой, $(n-p-n_e)$ — число степеней свободы для суммы, обусловленной неадекватностью модели. Как видно, квадраты длин векторов делятся на размерности подпространств выборочного пространства, в которых лежат эти векторы. Свойства F-распределения опираются на обычное предположение о нормальности случайных ошибок.

Параметрическое пространство

Пространство параметров представляет собой p-мерное пространство, точка которого определяется множеством значений параметров $(\theta_1, \theta_2, \ldots, \theta_p)$. Минимальное значение суммы $S(\boldsymbol{\theta})$ достигается в точке $\hat{\boldsymbol{\theta}} = (\hat{\theta_1}, \hat{\theta_2}, \ldots, \hat{\theta_p})'$. Мы напомним, что

$$S(\theta) - S(\hat{\theta}) = (\theta - \hat{\theta})' \mathbf{X}'\mathbf{X} (\theta - \hat{\theta}).$$

Все величины θ, удовлетворяющие условию $S(\theta) = \text{const} = K$, задаются выражением

$$(\theta - \hat{\theta})' \mathbf{X}'\mathbf{X} (\theta - \hat{\theta}) = K - S(\hat{\theta}),$$

и можно показать, что это уравнение эллипсоидного контура с центром в точке $\hat{\theta}$. Если $K_1 > K_2$, то контур $S(\theta) = K_1$ полностью охватывает контур $S(\theta) = K_2$, а точка $\hat{\theta}$ лежит в центре этой последовательности p-мерных эллипсоидов 100 $(1-\alpha)$ %-ная доверительная область для вектора θ истинных, но неизвестных параметров заключена в контуре, который задается уравнением

$$\frac{[S(\theta) - S(\hat{\theta})]/p}{S(\hat{\theta})/(n-p)} = F(p,\ n-p,\ 1-\alpha)$$

при условии, что ошибки распределены нормально, т. е. $\varepsilon \sim N(0, \mathbf{I}\sigma^2)$. Приведенное выражение можно переписать так.

$$S(\boldsymbol{\theta}) = S(\hat{\boldsymbol{\theta}})\left\{1 + \frac{p}{n-p} F(p,\ n-p,\ 1-\alpha)\right\},$$

где в правой части стоит величина, определяющая размеры контура.

Параметрическое пространство при $p = 2$

Проиллюстрируем рассматриваемую ситуацию при $p = 2$. На рис. 10.12 показаны некоторые возможные контуры, отвечающие общему уравнению $S(\theta) = \text{const}$, для трех значений правой части при $p = 2$. Внешний контур — это 100 $(1-\alpha)$ %-ный доверительный контур, определенный выше. В двумерном пространстве (θ_1, θ_2) контуры представляют собой концентрические эллипсы с центром в точке $(\hat{\theta}_1, \hat{\theta}_2)$. Заметим, что контуры такого типа получаются независимо от того, каким может быть число наблюдений n, поскольку размерность пространства параметров зависит только от p.

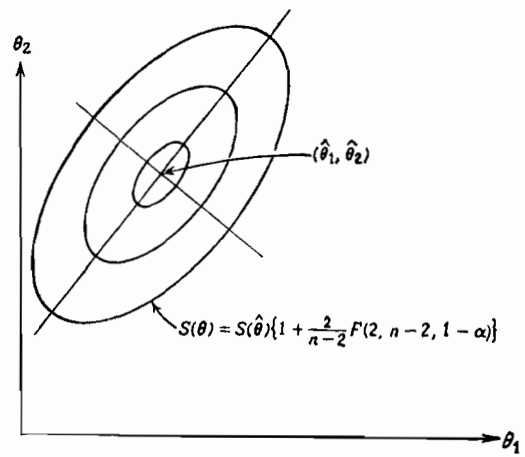


Рис. 10.12. Контуры поверхности $S(\theta)$ в двумерном параметрическом пространстве

Вообще говоря, важны и ориентация и форма эллипсов. Если оси эллипсов параллельны осям θ_1 и θ_2, то величина $\hat{\theta}_1$, минимизирующая $S(\theta_1, \theta_2)$, не зависит от θ_2, т. е. если зафиксировать θ_2 на некотором произвольном уровне, то будет получено одно и то же значение $\theta_1 = \hat{\theta}_1$, обращающее сумму $S(\theta_1, \theta_2)$ в минимум независимо от значения θ_2. Это означает, что определенная информация о величине θ_2, используемая при фиксировании значения этой величины, не оказывает влияния на МНК-оценку θ_1. Такая ситуация имеет место, если выражение для $S(\theta_1, \theta_2)$ можно записать так, чтобы оно не содержало произведения $\theta_1\theta_2$. При $p = 2$ получаем модель

$$Y_u = \theta_1 X_{1u} + \theta_2 X_{2u} + \varepsilon_u, \quad u = 1,\ 2,\ \ldots,\ n.$$

Таким образом,

$$S(\boldsymbol{\theta}) = S(\theta_1, \theta_2) = \sum_{u=1}^{n} (Y_u - \theta_1 X_{1u} - \theta_2 X_{2u})^2 = \Sigma Y_u^2 - \theta_1 2\Sigma X_{1u} Y_u -$$

$$- \theta_2 2\Sigma X_{2u} Y_u + \theta_1^2 \Sigma X_{1u}^2 + \theta_2^2 \Sigma X_{2u}^2 + \theta_1\theta_2 2\Sigma X_{1u} X_{2u},$$

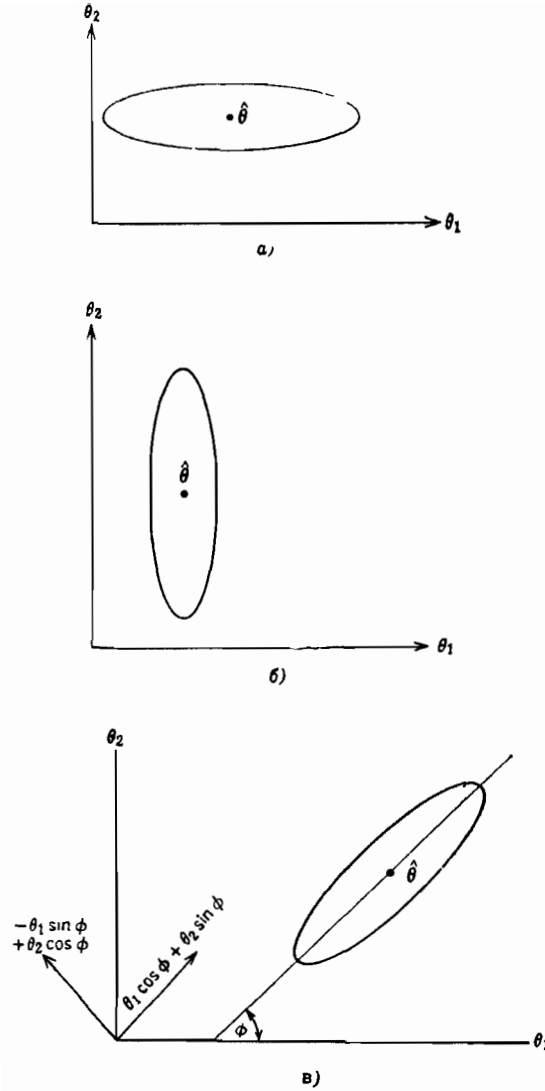


Рис. 10.13. Интерпретация некоторых возможных 95 %-ных доверительных областей для параметров (θ_1, θ_2):

а) θ_1 определяется плохо, θ_2 идентифицируется хорошо, оценки $\hat{\theta}_1$ и $\hat{\theta}_2$ независимы.

б) θ_1 идентифицируется хорошо, θ_2 — плохо, оценки $\hat{\theta}_1$ и $\hat{\theta}_2$ независимы.

в) параметрическая функция $\theta_1 \cos\Phi + \theta_2 \sin\Phi$ идентифицируется плохо, функция $\theta_1 \sin\Phi + \theta_2 \cos\Phi$ идентифицируется хорошо, $\hat{\theta}_1$ и $\hat{\theta}_2$ взаимозависимы

где суммирование ведется по индексу $u = 1, 2, \ldots, n$. Отсюда ясно, что величина θ_1, которая удовлетворяет условию $\partial S(\boldsymbol{\theta})/\partial\theta_1 = 0$ и обозначается как θ_1, не зависит от θ_2 (и наоборот), если в этом выражении отсутствует член с сомножителем $\theta_1\theta_2$, т. е. если $\Sigma X_{1u} X_{2u} = 0$, или, что то же самое, если столбцы матрицы $\mathbf{X}$ ортогональны. Если же столбцы матрицы $\mathbf{X}$ неортогональны и в сумму квадратов входит слагаемое с сомножителем $\theta_1\theta_2$, то эллипсы ориентированы наклонно по отношению к осям θ_1 и θ_2.

Форма контуров поверхности $S(\theta_1, \theta_2)$ характеризует относительную точность оценок θ_1 и $\hat{\theta}_2$[20]. На рис. 10.13 проиллюстрированы некоторые варианты. Единственный контур, который там показан, предназначен для того, чтобы изобразить границы 95 %-ной доверительной области и точку $\hat{\boldsymbol{\theta}}$ с координатами $(\hat{\theta}_1, \hat{\theta}_2)$, представляющую собой в каждом случае МНК-оценку параметров.

[20] Квадрат площади эллипса пропорционален обобщенной дисперсии, равной определителю дисперсионной матрицы оценок параметров, $(\mathbf{X}'\mathbf{X})^{-1}$. Эта связь лежит в основе концепции D-оптимальности в планировании экспери-

10.6. ГЕОМЕТРИЯ НЕЛИНЕЙНОГО МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ

Выборочное пространство

Для нелинейной модели не удается построить матрицу $\mathbf{X}$, которой мы располагали в линейном случае. К тому же пространство оценок нельзя определить с помощью набора векторов. Оно может быть очень сложным. Пространство оценок, называемое также *геометрическим местом точек решения*, состоит из всех точек, координаты которых имеют вид

$$\{f(\xi_1, \theta), f(\xi_2, \theta), \ldots, f(\xi_n, \theta)\}.$$

Поскольку сумма квадратов $S(\theta)$ все же представляет собой квадрат расстояния от точки $(Y_1, Y_2, \ldots, Y_n)$ до точки пространства оценок, минимизация функции $S(\theta)$ по θ соответствует геометрически нахождению в пространстве оценок такой точки P, которая ближе всего расположена к Y. Выборочное пространство для очень простого нелинейного случая с $n = 2$ наблюдениями Y_1 и Y_2, полученными при $\xi = \xi_1$ и $\xi = \xi_2$ соответственно, и единственным параметром θ показано на рис. 10.14. Пространство оценок представляет собой кривую, которая содержит точки

$$\{f(\xi_1, \theta), f(\xi_2, \theta)\},$$

Рис. 10.14. Выборочное пространство при $n = 2$, $f(\xi, \theta)$ нелинейная

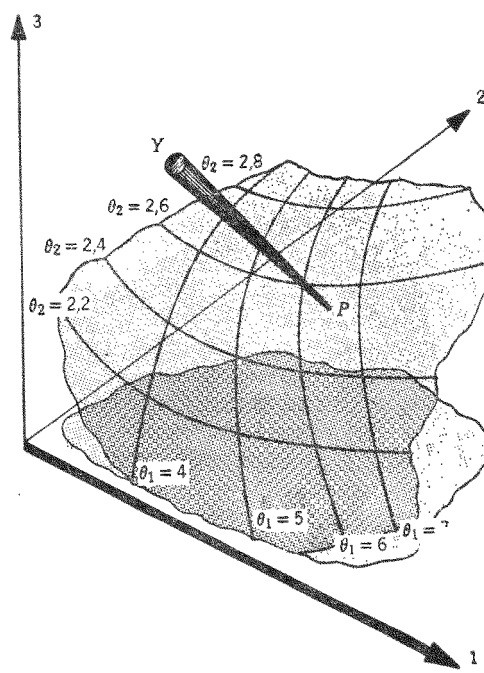


Рис. 10.15. Выборочное пространство при $n = 3$ и $p = 2$, $f(\xi, \theta)$ нелинейная

мента (см.: Федоров В. В. Теория оптимального эксперимента. — М.: Наука, 1971.— 312 с.).— *Примеч. пер.*

где θ меняется, а ξ_1 и ξ_2 фиксированы. Точка Y имеет координаты (Y_1, Y_2), а P есть точка пространства оценок, лежащая ближе всего к Y.

На рис. 10.15 для примера показано выборочное пространство с $n = 3$ наблюдениями Y_1, Y_2, Y_3, полученными при $\xi = \xi_1$, ξ_2 и $\xi = \xi_3$ соответственно, при условии, что модель содержит два параметра θ_1 и θ_2. Кривые линии определяют систему координат для параметров в пространстве оценок или в геометрическом месте точек решения, которое включает все точки вида

$$\{f(\xi_1, \theta_1, \theta_2),\ f(\xi_2, \theta_1, \theta_2),\ f(\xi_3, \theta_1, \theta_2)\},$$

когда θ_1 и θ_2 меняются, а ξ_1, ξ_2 и ξ_3 фиксированы. Точка Y имеет координаты (Y_1, Y_2, Y_3), а P есть точка в пространстве оценок, которая ближе всего к точке Y. Применяя метод линеаризации к нелинейным задачам, выбирают точку в пространстве оценок, скажем, точку $\boldsymbol{\theta}_0$, как новое начало, определяющее линеаризованное пространство оценок параметров в форме касательного пространства, проходящего через точку $\boldsymbol{\theta}_0$, и решают линейную задачу, используя метод наименьших квадратов, как было описано ранее. Это решение (выраженное в единицах скорости изменения величины $\boldsymbol{\theta}$ и поэтому пригодное лишь в окрестности $\boldsymbol{\theta}_0$) служит для построения новой начальной точки $\boldsymbol{\theta}_н$. Она соотносится снова с нелинейной задачей, *где решение может быть* и неправильным. Затем проводится следующая итерация. Для нелинейной задачи, включающей только два наблюдения для модели с одним параметром, эффект линеаризации изображен на рис. 10.16.

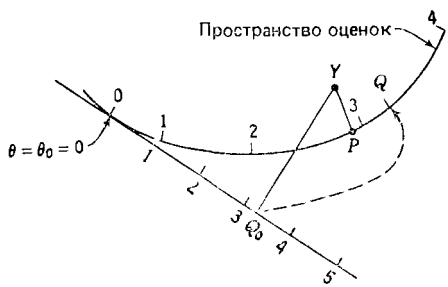


Рис. 10.16. Геометрическая интерпретация метода линеаризации ($n = 2$, $p = 1$)

На рис. 10.16 представлено пространство оценок или геометрическое место точек решения с нанесенными на кривой единицами масштаба для параметра θ. Здесь предполагается, что $\theta_0 = 0$ и что точка, обозначенная через $\theta = 1$, есть точка пространства оценок, соответствующая значению параметра $\theta = 1$, и т. д. Заметим, что на кривой масштаб для θ *неравномерен* в силу нелинейности модели и неравномерности координатной системы. На рисунке показана также касательная к кривой, т. е. к пространству оценок при $\theta = \theta_0 = 0$, на которой изменение θ отражено уже в равномерном масштабе и соответствует единицам ($\theta = 0, 1, 2, \ldots$) скорости изменения, найденной при θ_0. Мы видим теперь, что МНК-оценка θ основана на линейном подходе.

Геометрически это означает отыскание такой точки Q_0, что отрезок YQ_0 перпендикулярен к касательной линии. Мы видим, что в линеаризованных единицах величина θ, отвечающая точке Q (см. рис. 10.16), равна примерно 3,2. На следующей итерации метода линеаризации мы, таким образом, используем касательную к *простран-*

ству оценок в точке, где $\theta = 3, 2$, т. е. в точке Q_0. Отсюда легко видеть причину, вследствие которой процедура линеаризации иногда приводит к неверным результатам. Если скорость изменения величины $f(\xi, \theta)$ в точке θ_0 мала, но быстро возрастает, то шкала масштаба на касательной линии может оказаться довольно нереалистичной. Подобная ситуация отражена на рис. 10.17. Скорость изменения в точке $\theta_0 = 0$ мала, и потому мала единица линеаризованного масштаба для θ. А настоящие единицы масштаба (на кривой) увеличиваются при этом быстро. Если мы начинаем теперь следующую итерацию, используя найденную величину θ, которая равна примерно 26 в точке Q_0, то наша исходная точка на кривой, в которой θ равно этому значению, будет дальше от лучшей точки P, чем точка, отвечающая начальной оценке $\theta = \theta_0 = 0$. Такая ситуация в некоторых случаях может быть скорректирована в ходе последующих итераций, но иногда это не удается. (Хотя мы использовали начальную оценку $\theta_0 = 0$ и единицы масштаба 1, 2, . . . для простоты, предыдущее замечание справедливо и в общем случае, какой бы ни была исходная величина оценки θ_0 и какой бы ни была шкала масштаба поблизости от точки θ_0).

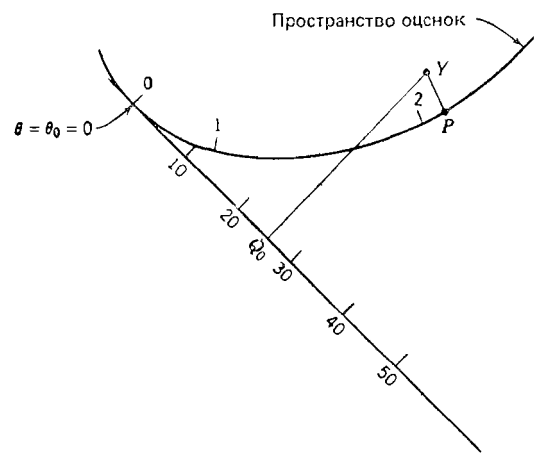


Рис. 10.17. Влияние большой неравномерности шкалы на метод линеаризации ($n = 2$, $p = 1$)

Если имеется более двух наблюдений и более одного параметра, то в общем случае сохраняется то же положение, но ситуация становится более сложной и ее трудно или даже невозможно изобразить графически.

Если модель линейная, то контуры постоянных значений $S(\theta)$ в выборочном пространстве представляют собой сферы. В нелинейных случаях это уже несправедливо и могут возникнуть довольно нерегулярные контуры, включающие все точки пространства оценок эквидистантные с точкой $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$.

Пространство параметров

В случае линейной модели контуры постоянных значений $S(\theta)$ в параметрическом пространстве есть концентрические эллипсоиды. Если модель нелинейна, то контуры иногда имеют вид банановидных фигур, зачастую довольно вытянутых. В некоторых случаях контуры оказываются бесконечно вытянутыми и незамкнутыми или

могут иметь множество петель, окружающих ряд стационарных точек. Если существует несколько стационарных точек, им могут соответствовать разные значения суммы квадратов, вследствие чего процедура оценивания может приводить к разным результатам.

Рассмотрим, например, модель

$$Y = \frac{(\theta_1 e^{-\theta_2 t} - \theta_2 e^{-\theta_1 t})}{(\theta_2 - \theta_1)} + \varepsilon.$$

Эта модель инвариантна относительно перестаиовки параметров θ_1 и θ_2. Таким образом, если минимум $S(\boldsymbol{\theta})$ достигается при $(\theta_1, \theta_2) = (\hat{\theta}_1, \hat{\theta}_2)$, то в точности такая же минимальная величина этой суммы достигается и при $(\theta_1, \theta_2) = (\hat{\theta}_2, \hat{\theta}_1)$, т. е. существуют два решения. О существовании множества решений часто нелегко узнать вообще [21]. Примеры банаиоподобных контуров приведены в § 10.3.

Доверительные контуры в нелинейном случае

Для нелинейной модели некоторые результаты, справедливые для линейного случая, неприменимы. Если предположить, что ошибка ε в нелинейной модели (10.1.1) распределена нормально, то оценки параметров $\hat{\boldsymbol{\theta}}$ совсем не обязательно будут подчиняться нормальному распределению. Оценка $s^2 = S(\hat{\theta})/(n-p)$ не служит несмещенной оценкой дисперсии σ^2, и вообще здесь нет никакой матрицы дисперсий и ковариаций оценок параметров [22] вида $(\mathbf{X}'\mathbf{X})^{-1}\sigma^2$.

Хотя доверительные области все же могут быть *определены* с помощью выражения

$$S(\boldsymbol{\theta}) = S(\hat{\boldsymbol{\theta}})\left\{1 + \frac{p}{n-p} F(p,\ n-p,\ 1-\alpha)\right\},$$

которое в линейном случае дает 100 $(1-\alpha)$ %-ную доверительную область, но даже при нормально распределенных ошибках доверительная вероятность в случае нелинейной параметризации уже не будет равна $1-\alpha$. Мы не знаем вообще, какой она будет, и потому можем

---

[21] Здесь авторы затрагивают одну из фундаментальных проблем — проблему идентифицируемости параметров нелинейных моделей. Приведенный выше пример иллюстрирует так называемую глобальную неидентифицируемость параметров моделей. См.: Спивак С. И., Горский В. Г.//Доклады АН СССР, 1981.— 257, № 2, с. 412—417; Спивак С. И., Горский В. Г.//Химическая физика, 1982, 1, № 2.— с. 237—243; Горский В. Г. Планирование кинетических экспериментов.— М.: Наука, 1984.— 241 с. (особо с. 128—135).— *Примеч. пер.*

[22] Построение матрицы дисперсий-ковариаций оценок параметров нелинейных моделей представляет собой сложную задачу, точное решение которой неизвестно. Для приближенного решения могут быть использованы асимптотические разложения моментов оценок параметров. См.: Иванов А. В.— Украинский матем. журнал, 1982, 34, № 2, с. 164—170; Математическая теория планирования эксперимента/Под ред. С. М. Ермакова.— М.: Наука, 1983.— 392 с. (особо с. 41—44).— *Примеч. пер.*

назвать такие области *приблизительно 100 (1—α) %-ными доверительными областями* для θ[23].

Банановидные контуры в примере § 10.3 были получены именно таким путем. И хотя подходящие сравнения средних квадратов и остаются все же наглядными, использование F-критерия Фишера для проверки гипотез о параметрах нелинейной регрессионной модели и проверки ее адекватности нельзя считать корректным.

Мера нелинейности

Было сделано несколько предложений по поводу того, как измерять степень нелинейности модели в нелинейных задачах. Такого рода мера должна нам помочь, например, ответить, на вопрос: позволяет ли линеаризация модели получить приемлемую аппроксимацию? Обсуждение различных аспектов этой проблемы и подходящие ссылки на литературу можно найти в работе: Bates D. M., Watts D. G. Relative curvature measures of nonlinearity.— Jornal of the Royal Statistical Society, 1980, **B—42**, p. 1—16, discussion 16—25 [24].

10.7. НЕЛИНЕЙНЫЕ МОДЕЛИ РОСТА

В этом параграфе мы приводим некоторые примеры нелинейных моделей, которые используются для описания процессов роста, развертывающихся во времени. Модели роста находят применение во многих областях науки и техники: в биологии, ботанике, лесном деле, зоологии и экологии; с их помощью описывается рост организмов, растений, деревьев и кустарников, животных и людей. В химии и химической технологии с их помощью описывают результаты химических реакций. В экономике и политических науках эти модели используют для описаиия изменений, происходящих с организациями, ресурсами продовольствия и материалов, странами и др. [25].

Типы моделей

Те или иные типы моделей, необходимые в определенных областях и в определенных задачах, зависят от специфических особенностей рассматриваемых объектов. Вообще модели роста оказываются скорее механистическими [26], чем эмпирическими. Механистическая модель обычно возникает как результат принятия определенных гипотез относительно типа роста. Гипотезы выписывают в виде дифференциаль-

---

[23] Наиболее полно этот вопрос рассмотрен на русском языке в книге: Демиденко Е. З. Линейная и нелинейная регрессии.— М.: Финансы и статистика, 1981.— 302 с.— *Примеч. пер.*

[24] См. примечание 11 к этой главе.— *Примеч. пер.*

[25] Добавим еще, что моделями этого типа описываются также, например, характеристики развития науки (см.: Налимов В. В., Мульченко З. М. Наукометрия.— М.: Наука, 1969), переходные процессы в автоматических системах и многие другие процессы.— *Примеч. пер.*

[26] Под этим термином подразумеваются модели, основанные на теоретических представлениях об изучаемом объекте.— *Примеч. пер.*

ных или разностных уравнений, которые решают и получают в результате модель роста. (Эмпирическую модель, напротив, выбирают интуитивно, чтобы аппроксимировать неизвестную механистическую модель. Часто эмпирическая модель имеет вид полинома соответствующего порядка.)

Пример механистической модели роста

Рассмотрим процесс роста, в котором предполагается, что *скорость* роста в каждый момент времени прямо пропорциональна разности между некоторым предельным (максимально возможным) уровнем, обозначаемым α, и текущим уровнем ω, соответствующим моменту времени t. Тогда

$$\frac{d\omega}{dt} = k(\alpha - \omega), \tag{10.7.1}$$

где k есть *константа скорости* роста. Интегрируя это уравнение, получим

$$\omega = \alpha(1 - \beta e^{-kt}) \tag{10.7.2}$$

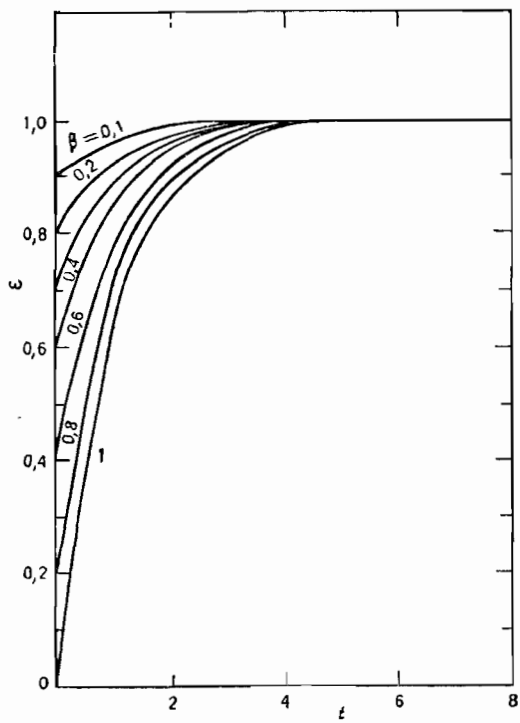


Рис. 10.18. Теоретические кривые зависимости $\omega = 1 - \beta e^{-t}$ при различных значениях β

Это выражение известно также под названием мономолекулярной функции роста [27]. Она описывает кривую, монотонно возрастающую от значения $\alpha \times (1-\beta)$ при $t = 0$ до предельного значения α. На интервале $(0, \infty)$ эта функция не имеет точки перегиба (знак второй производной $d^2\omega/dt^2$ не меняется) и всегда остается возрастающей, хотя сама скорость, как видно из уравнения (10.7.1), все время уменьшается. Эта модель использовалась например, в работе:

[27] Данное выражение заимствовано из химической кинетики, где под мономолекулярной реакцией понимают одностадийную реакцию превращения вещества, которая может быть описана линейным дифференциальным уравнением (такую реакцию также называют реакцией первого порядка.).— *Примеч. пер.*

Gregory F. G. Studies in the energy relation of plants, II.— Annals of Botany, 1928, **42**, p. 469—507 *[2].

Составим теперь представление о том, как выглядит кривая, описываемая уравнением (10.7.2). Поскольку α — просто масштабный множитель, можно положить $\alpha = 1$; кроме того, поскольку параметр k и предикторная переменная t входят в модель только в виде произведения kt, можно положить $k = 1$. Тогда, варьируя параметр β, получим кривые, подобные тем, которые показаны на рис. 10.18. При разных значениях α изменяется только шкала кривой по вертикали, а выбор разных значений k приводит только к растяжению или сжатию кривой по горизонтали. Каждая кривая начинается при $t = 0$ со значения $\omega = \alpha(1-\beta)$, которое равно $1-\beta$ при $\alpha = 1$, как показано на рис. 10.18.

Кривые, изображенные на рис. 10.18, получаются, конечно, теоретическими, поскольку они базируются на теоретической функции (10.7.2). Если предположить, что фактически наблюдаемые значения ω в моменты времени $t_1, t_2, \ldots, t_n$ есть w_i, $i = 1, 2, \ldots, n$, то можно постулировать модель в виде

$$w_i = \alpha\left(1-\beta e^{-kt_i}\right)+\varepsilon_i \qquad (10.7.3)$$

где ε — случайная ошибка, имеющая, например, такие характеристики: $E(\varepsilon_i)=0$, $V\varepsilon_i=\sigma_i^2$. Примем что все ошибки ε_i имеют равные дисперсии и не коррелированы. В таком случае для подгонки модели к имеющимся данным резонно воспользоваться методом наименьших квадратов. Если не оговаривается иное, во всех случаях предполагается, что ошибки входят в модель аддитивно. Поэтому, в частности, выражение «подогнать модель, заданную уравнением (10.7.2)» означает, что речь идет об использовании уравнения (10.7.3). То же самое относится и к другим моделям.

Некоторые настораживающие моменты использования МНК

Описывающие процесс роста данные далеко не всегда удовлетворяют «обычным предположениям метода наименьших квадратов». Так, например, если все наблюдения w_i производятся на одних и тех же растениях, животных или организмах, не резонно ожидать, что полученные данные будут некоррелированными. В идеале экспериментатор должен бы использовать разные независимые объекты для каждого наблюдения, но это не всегда разумно и возможно в силу имеющихся ограничений на экспериментальный материал или оборудование. Также не всегда выполняется условие постоянства дисперсии отклика $\sigma_i^2 = V(\varepsilon_i)$, которая может зависеть, например, от достигнутого к данному моменту уровня или от некоторой его функции. Если есть конкретная информация на этот счет, то ею можно воспользоваться и применить, например, взвешенный МНК. Другой путь такой: применить стандартную процедуру МНК для подгонки модели,

*[2] Ссылки в этом параграфе и связанные с ними соответствующие работы отражены в библиографии к гл. 10, раздел *B* «Модели роста»

а затем исследовать остатки и попытаться выявить причину несостоятельности исходных предположений [28]. Эти соображения могут быть затем использованы для повторения процедуры улучшения оценивания и/или для изменения модели.

Логистическая модель

Предположим, что скорость роста пропорциональна произведению величины, характеризующей текущий уровень на разность между предельным (максимальным) значением этого уровня и его текущим значением. В таком случае модель может быть выражена уравнением

$$\frac{d\omega}{dt} = \frac{k\omega(\alpha - \omega)}{\alpha}, \qquad (10.7.4)$$

где k — коэффициент пропорциональности, $k > 0$; α — предельный уровень. Если сравнить это уравнение с (10.7.1), то можно видеть, что оно отражает динамику, при которой темп прироста характеристики уровня линейно падает с увеличением ω. Интегрируя (10.7.4), получаем

$$\omega = \alpha/\{1 + \beta e^{-kt}\}; \qquad (10.7.5)$$

эта функция известна под названием логистической (или автокаталитической) функции роста [29]. Она описывает S-образную кривую. Заметим, что при $t=0$ величина ω равна: $\omega = \frac{\alpha}{1+\beta}$ — это начальный уровень развития; при $t \to \infty$ она стремится к предельному уровню, равному $\omega = \alpha$. Отсюда также следует, что $\beta > 0$. Ясно также, что и $k > 0$. Из (10.7.4) видно, что тангенс угла наклона касательной к кривой в точке $t = 0$ по отношению к оси абсцисс всегда положителен, а вторая производная выражается формулой

$$\frac{d^2\omega}{dt^2} = \frac{k}{\alpha}(\alpha - 2\omega). \qquad (10.7.6)$$

Из (10.7.6) вытекает, что вторая производная положительна при $\omega > \frac{1}{2}\alpha$, равна нулю в точке перегиба, где $\omega = \omega_1 = \frac{1}{2}\alpha$, и отрицательна при $\omega > \frac{1}{2}\alpha$. Из (10.7.5) находим, что точка перегиба имеет

---

[28] Связанные с этим вопросы рассматриваются, например, в книге: Вучков И. Н., Бояджиева Л. Н., Солаков Е. Б. Прикладной линейный регрессионный анализ. Это книга болгарских авторов, которую предполагается издать в русском переводе в издательстве «Финансы и статистика» в 1987 г.— *Примеч. пер.*

[29] Данный термин также заимствован из теории химической кинетики. С помощью такой функции описывается автокаталитическая реакция первого порядка. Скорость такой реакции зависит от концентраций исходного и образующегося веществ.— *Примеч. пер.*

координату $t_I = (\ln \beta)/k$. Если ввести обозначение $t = t_I + u$, то можно представить выражение (10.7.5) в виде $\omega = \alpha/(1+e^{-ku})$; тогда можно записать

$$\omega - \omega_I = \frac{\alpha}{2}\left\{\frac{1-e^{-ku}}{1+e^{-ku}}\right\} = g(u). \qquad (10.7.7)$$

Из этого соотношения ясно, что данная кривая симметрична относительно точки перегиба, поскольку $g(-u) = g(u)$. Конечно, кривая простирается влево только до $t = 0$, т. е. $u = -t_I$, тогда как вправо — до $t = \infty$, т. е. $u = \infty$. Укажем еще, что если $0 < \beta < 1$, кривая начинается с точки (при $t = 0$), которая выше точки перегиба [30], хотя

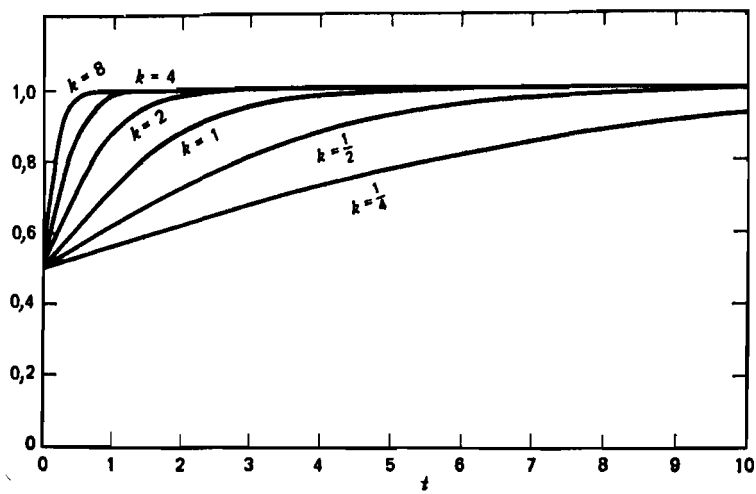


Рис. 10.19. Теоретические кривые зависимости $1/\{1 + e^{-kt}\}$ при различных значениях k

если β положительно и имеет большое значение, то точка перегиба имеет большую координату t, которая может оказаться за пределами рабочего интервала времени. Конечно, $\beta > 0$, поскольку $\alpha > \alpha/(1 + \beta)$.

Чтобы получить некоторое представление о характере кривых (10.7.5), можно без потери общности положить $\alpha = 1$ и вычертить графики, варьируя β и k. Некоторые иллюстративные кривые показаны на рис. 10.19 и 10.20. При изменении коэффициента β меняется положение начальной кривой на оси ординат при $t = 0$. Изменение k приводит к изменению крутизны кривой. Поскольку в выражение модели (10.7.5) величины k и t входят только в виде произведения, изменение величины k может быть компенсировано за счет изменения масштаба по оси t, например

[30] Иными словами, в этом случае кривая не содержит точки перегиба.— *Примеч. пер.*

$$kt = \left(\frac{1}{2}k\right)(2t) = KT,$$

где

$$K = \frac{1}{2}k \quad \text{и} \quad T = 2t.$$

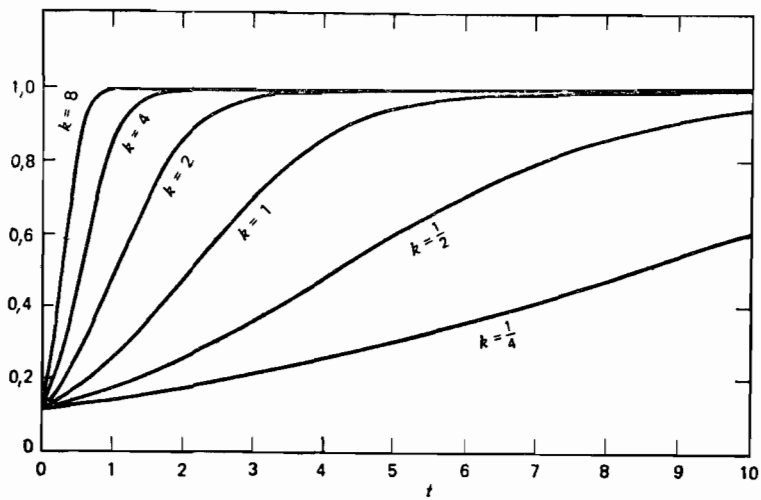


Рис. 10.20. Теоретические кривые зависимости $1/\{1 + 8e^{-kt}\}$ при различных значениях k

Другая форма логистической модели

Рассмотрим модель

$$\eta = \delta - \ln(1 + \beta e^{-kt}), \tag{10.7.8}$$

которая получается из (10.7.5) после логарифмирования и введения обозначений $\eta = \ln \omega$ и $\delta = \ln \alpha$. По существу вид соответствующей кривой аналогичен виду кривых, изображенных на рис. 10.19 и 10.20 с точностью до изменения масштаба по вертикальной оси за счет логарифмирования. При подгонке уравнений (10.7.5) и (10.7.8) к экспериментальным данным с помощью метода наименьших квадратов приходится исходить из совершенно разных представлений о характере отклонений данных от этих моделей. Согласно работе Недлера (N e d l e r J. A. The fitting of a generalization of the logistic curve.— Biometrics, 1961, **17**, p. 89—110) предположение о том, что измеряемые отклики $y_i = \ln \omega_i$ имеют постоянную дисперсию, обычно обосновано только при описании роста растений. На практике для подгонки модели к экспериментальным данным можно опираться на любую из форм (10.7.5) или (10.7.8); проверка остатков покажет, какая из этих моделей лучше подходит для экспериментатора.

Как мы получаем начальные оценки параметров?

Как уже отмечалось, процедура нелинейного оценивания требует знания исходных (начальных) оценок параметров, и, чем лучше эти оценки, тем скорее процедура сходится к искомым оценкам. Практический опыт работы с моделями роста показывает, что, если исходные оценки плохи, процедура может сойтись к ошибочным конечным оценкам.

Нет общих методов получения начальных оценок. Для этих целей используют любую информацию, которая имеется в наличии. Так, например, для логистической модели (10.7.8) мы рекомендуем следующие приемы.

1. Если $t = \infty$, то $\eta = \delta$. Так что примите $\delta_0 = y_{\max}$.

2. Для любых двух других наблюдений, скажем i-го, и j-го, используя выражение (10.7.8) и игнорируя при этом ошибки наблюдения, можно записать:

$$y_i = \delta_0 - \ln(1 + \beta_0 e^{-k_0 t_i}),$$

$$y_j = \delta_0 - \ln(1 + \beta_0 e^{-k_0 t_j}).$$

Затем из приведенных уравнений находим:

$$\exp(\delta_0 - y_i) - 1 = \beta_0 \exp(-k_0 t_i),$$

$$\exp(\delta_0 - y_j) - 1 = \beta_0 \exp(-k_0 t_j),$$

откуда путем деления, логарифмирования и несложных преобразований получаем

$$k_0 = \frac{1}{t_j - t_i} \ln\left\{\frac{\exp(\delta_0 - y_i) - 1}{\exp(\delta_0 - y_j) - 1}\right\}.$$

Чем сильнее между собой различаются наблюдения по времени, тем более устойчивыми оказываются оценки.

3. Из i-го уравнения, приведенного выше, можно вычислить

$$\beta_0 = \exp(k_0 t_i)\{\exp(\delta_0 - y_i) - 1\}.$$

(Безразлично, какое уравнение использовать, i-е или j-е.)

4. Подстановка значения $\delta_0 = y_{\max}$ в два предыдущих уравнения позволяет определить значения k_0 и β_0.

В качестве альтернативы на первом шаге следует принять величину δ_0 равной несколько большему значению, чем y, скажем, 1,1 $y_{\max}$, или какому-либо другому значению, подсказанному опытом. Обычно это дает несколько лучшие начальные оценки. В некоторых задачах можно принять η при $t = 0$ равной $y_{\min}$ или 0,9 $y_{\min}$ и т. д. Вообще для получения исходных оценок следует применять любой метод, который приводит к простым уравнениям.

Исходные оценки для подгонки модели (10.7.5) могут быть получены аналогичным образом с помощью формул:

$$\alpha_0 = y_{\max},$$

$$\beta_0 = \{(\alpha_0 - w_i)/\alpha_0\} \exp(k_0 t_i),$$

$$k_0 = \frac{1}{t_i - t} \ln\left\{\frac{\alpha_0 - w_j}{\alpha_0 - w_i}\right\}.$$

Модель Гомпертца

Если скорость роста подчиняется дифференциальному уравнению

$$\frac{d\omega}{dt} = k\omega \ln(\alpha/\omega), \tag{10.7.9}$$

то в результате интегрирования получим из него модель Гомпертца

$$\omega = \alpha \exp\{-\beta e^{-kt}\}. \tag{10.7.10}$$

Хотя эта кривая и является S-образной подобию логистической кривой, однако она не симметрична относительно точки перегиба. Точка перегиба имеет место, когда $\frac{d^2\omega}{dt^2} = 0$ и соответственно $\omega_I = \alpha/e = 0{,}368\alpha$. Координата этой точки $t_I = (\ln \beta)/k$. Из уравнений (10.7.9) и (10.7.10) вытекают соотношения

$$\frac{d\omega/dt}{\omega} = k(\ln \alpha - \ln \omega), \tag{10.7.11}$$

$$\frac{d\omega/dt}{\omega} = k\beta e^{-kt}. \tag{10.7.12}$$

Последнее может быть переписано в виде

$$\ln\left\{\frac{d\omega/dt}{\omega}\right\} = \ln(k\beta) - kt. \tag{10.7.13}$$

Таким образом, из уравнения (10.7.11) следует, что относительная скорость изменения ω (темп прироста) и $\ln \omega$ связаны между собой линейным соотношением, а уравнение (10.7.13) показывает, что имеется линейная связь между логарифмом относительной скорости и временем. Как утверждается в работе: Richards F. J. A flexible growth function for empirical use.— Journal of Experimental Botany, 1959, **10**, p. 290—300, последнее уравнение лучше пригодно для описания развития популяций и роста животных, нежели для приложений в ботанике. Оно использовалось, например, в работе П. Медовара (Medawar P. B. Growth, growth energy, and ageing of the chicken's heart.— Proceeding of the Royal Society of London, 1940, **B—129**, p. 332—355) при изучении развития сердца у цыплят. Тем не менее оно применялось при исследовании роста растения Пеларгонии зональной [31] (см.: Amer F. A., Williams W. T. Leaf-area growth in Pelargonium zonale.— Annals of Botany N. S., 1957, **21**, p. 339—342).

Когда $t \to \infty$, характеристика уровня роста стремится к предельному значению α. При $t = 0$ получаем начальный уровень $\omega = \alpha e^{-\beta}$.

Если без потери общности рассуждений, положить $\alpha = 1$, то для $k = 1$ и выбранных значений β можно построить кривые, показанные

[31] Пеларгония зональная (Pelargonium zonale) — наиболее известный вид комнатной герани (семейства гераниевых — Geraniaceae, насчитывающего около 250 видов), одного из древнейших комнатных растений в истории человечества. Родина — Южная Африка (Капская область).— *Примеч. пер.*

на рис. 10.21. При каждом фиксированном значении β вариации параметра k дают наборы кривых, выходящих из одной и той же исходной точки, подобно тому, как это показано на рис. 10.19 и 10.20 для логистической модели.

Модель Берталанфи

Эта четырехпараметрическая модель имеет форму

$$\omega = \{\alpha^{1-m} - \theta e^{-kt}\}^{1/(1-m)}, \qquad (10.7.14)$$

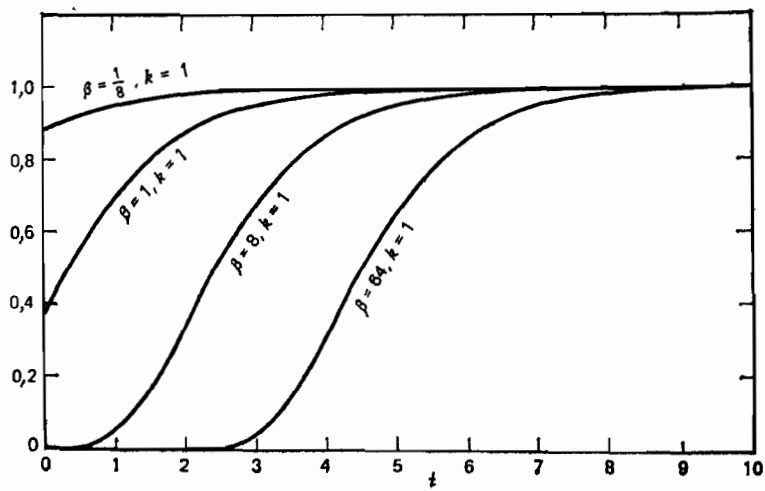


Рис. 10.21. Отдельные теоретические кривые модели Гомпертца $\omega = \alpha \exp(-\beta e^{-kt})$ для $\alpha = 1$

где α, θ, k и m — параметры, подлежащие оцениванию. В исходных работах: von Bertalanffy L. Stoffwechseltypen und Wachstumstypen.— Biol Zentralbl., 1941, **61**, p. 510—532; Quantitative laws in metabolism and growth.— Quarterly Review of Biology, 1957, **32**, p. 218—231, где эта модель была предложена, на параметр m были наложены ограничения. Однако в дальнейшем Ричардс (Richards F. J. A flexible growth function for empirical use.— Journal of Experimental Botany, 1959, **10**, p. 290—300) показал, что целесообразно этот параметр использовать в других пределах. Особый интерес представляют следующие факты.

1. При $m = 0$ получаем мономолекулярную функцию, если записать $\theta = \alpha\beta$.

2. При $m = 2$ получаем логистическую функцию, если записать $\theta = \beta/\alpha$.

3. При $m \to 1$ кривая стремится к виду модели Гомпертца, как можно показать, исследуя предельное поведение скорости роста; непосредственная подстановка не удается. Следовательно, близость оценки m к единице показывает, что в данном случае полезна кривая Гомпертца.

4. При $m > 1$ величина θ отрицательна; при $m < 1$ она положительна.

Подгонка этой модели может вызвать затруднения, если текущие значения параметров приводят к отрицательным значениям функции. Чтобы избежать этого, надо использовать дробные показатели степени и вводить ограничения. В силу этого модель Берталанфи может оказаться менее удобной для подгонки, чем другие модели, описанные выше.

10.8. НЕЛИНЕЙНЫЕ МОДЕЛИ: ДРУГИЕ РАБОТЫ

После ознакомления с методами оценивания параметров нелинейной модели можно уделить внимание другим проблемам. Обсудим кратко некоторые важные вопросы и укажем источники, содержащие более подробную информацию.

Планирование экспериментов в нелинейном случае

Источник. Box G. E. P., Lucas H. L. Design of experiments in nonlinear situations.— Biometrika, 1959, **46**, p. 77—90.

Если метод линеаризации применяется для оценивания параметров нелинейной модели, то мы приходим к итеративной формуле

$$\mathbf{b}_j = (\mathbf{Z}_j'\mathbf{Z}_j)^{-1}\,\mathbf{Z}_j'(\mathbf{Y} - \mathbf{f}^j),$$

позволяющей получить $\boldsymbol{\theta}_{j+1} = \boldsymbol{\theta}_j + \mathbf{b}_j$. Можно показать, что приближенная доверительная область (см. с. 207) для $\boldsymbol{\theta}$ на этой итерации имеет объем, пропорциональный $|(\mathbf{Z}_j'\mathbf{Z})^{-1}|$. Следовательно, если рассмотреть наилучший план, представляющий собой наилучший набор опытов, подлежащих выполнению, то это будет план, который минимизирует объем доверительной области или максимизирует определитель матрицы $\mathbf{Z}_j'\mathbf{Z}_j$.

Как реализовать эту идею на практике?

Если никакие опыты еще не проводились, то выбираем такой набор из n опытов (предполагается, что n задано), который максимизирует $|\mathbf{Z}_0'\mathbf{Z}_0|$ (когда имеется несколько планов на выбор, выбираем из них тот, которому отвечает наибольшее значение $|\mathbf{Z}_0'\mathbf{Z}_0|$).

Если уже выполнено n опытов и предстоит выбрать $(n+1)$-й опыт, то можно записать $|\mathbf{Z}_j'\mathbf{Z}_j|$ как функцию условий $(n+1)$-го опыта и максимизировать $|\mathbf{Z}_j'\mathbf{Z}_j|$ по отношению к условиям $(n+1)$-го опыта.

В общем случае эта максимизация выполняется численно на ЭВМ. Аналитическое решение возможно лишь в простых случаях.

Пример. Даны исходные значения параметров $\boldsymbol{\theta}_0' = (\theta_{10}, \theta_{20}) = (0{,}7;\ 0{,}2)$ и модель

$$Y = \frac{\theta_1}{\theta_1 - \theta_2}\{e^{-\theta_2\xi} - e^{-\theta_1\xi}\} + \varepsilon = f(\theta_1,\ \theta_2,\ \xi) + \varepsilon.$$

Надо выбрать два опыта, отвечающие значениям ξ_1 и ξ_2, при которых $|\mathbf{Z}_0'\mathbf{Z}_0|$ достигает максимума.

Дифференцируя $f(\theta_1, \theta_2, \xi)$ по θ_1 и θ_2 и подставляя в полученные выражения $\theta = \theta_0$, получим:

$$Z^0_{1u} = (0{,}8 + 1{,}4\xi_u)\, e^{-0{,}7\xi_u} - 0{,}8e^{-0{,}2\xi_u},$$

$$Z^0_{2u} = -2{,}8e^{-0{,}7\xi_u} + (2{,}8 - 1{,}4\xi_u)\, e^{-0{,}2\xi_u}.$$

Следовательно,

$$\mathbf{Z}_0 = \begin{bmatrix} Z^0_{11} & Z^0_{21} \\ Z^0_{12} & Z^0_{22} \end{bmatrix}.$$

Затем, поскольку здесь $n = p = 2$, можно записать $|\mathbf{Z}_0'\mathbf{Z}_0| = |\mathbf{Z}_0'| \times |\mathbf{Z}_0| = |\mathbf{Z}_0|^2$. Таким образом, в данном конкретном случае задача свелась к максимизации абсолютного значения величины

$$|\mathbf{Z}_0| = \mathbf{Z}^0_{11}\mathbf{Z}^0_{22} - \mathbf{Z}^0_{12}\mathbf{Z}^0_{21}.$$

Можно показать, что максимум достигается при $\xi_1 = 1{,}23$ и $\xi_2 = 6{,}86$; это и есть наш план. Теперь можно найти значения Y при этих двух значениях предикторов и перейти к оцениванию θ_1 и θ_2, начиная с исходных оценок (0,7; 0,2). Примеры, связанные с планированием $(n + 1)$-го опыта при уже выполненных n опытах, можно найти в работе: Box G. E. P., Hunter W. G. Sequential design of experiment for nonlinear models.— Proceedings of the IBM Scientific Computing Symposium on Statistics, 1963, October 21—23, p. 113—137.

Когда уже имеется n экспериментов, экспериментатор обращается к компьютеру для проведения операций с моделью, данными и текущими оценками параметров.

Компьютер позволяет получать:

1. Новые МНК-оценки параметров.
2. Наилучшие условия для проведения следующего эксперимента.
3. Информацию об устойчивости наилучших условий.
4. Другие величины, представляющие определенный интерес, например, дисперсии и ковариации оценок параметров [32].

---

[32] Оптимальному планированию эксперимента при нелинейной параметризации посвящен целый ряд книг на русском языке. Среди них укажем следующие: Федоров В. В. Теория оптимального эксперимента. — М.: Наука. 1971.— 312 с.; Успенский А. Б., Федоров В. В. Вычислительные аспекты метода наименьших квадратов при анализе и планировании регрессионных экспериментов.— М.: Изд-во МГУ, 1975.— 168 с.; Денисов В. И. Математическое обеспечение системы ЭВМ — экспериментатор.— М.: Наука, 1977.— 252 с.; Писаренко В. Н., Зиятдинов А. Ш., Кафаров В. В. Планирование эксперимента и кинетика промышленных органических реакций.— М.: Научный совет по кибернетике АН СССР, 1977.— 25 с.; Планирование эксперимента в исследовании технологических процессов/ Хартман К., Лецкий Э., Шефер В. и др./Пер. с нем. Под ред. Э. К. Лецкого.—М.: Мир, 1977.—552 с. (особо гл. 10); Горский В. Г., Адлер Ю. П., Талалай А. М. Планирование промышленных экспериментов (Модели динамики).— М.: Металлургия, 1978.— 112 с.; Планирование эксперимента в задачах нелинейного оценивания и распознавания образов/Круг Г. К., Кабанов В. А., Фомин Г. А., Фомина Е. С.— М.: Наука, 1981.— 172 с.; Горский В. Г. Планирование кинетических экспериментов.— М: Наука, 1984.— 241 с. Упомянем еще интересный сборник: Kinetic data analysis: Design and analysis of enzyme and pharmacokinetic experiments/Ed. L. Endrenyi.— Toronto, Canada: Plenum, 1981 — p. 438. В издательстве «Медицина» предполагается русский перевод.— *Примеч. пер.*

Полезная методика построения моделей

Источник. Box G. E. P., Hunter W. G. A useful method of model building.— Technometrics, 1962, 4, p. 301—318. Предположим, что нужно подогнать модель

$$Y=f(\theta_1,\ \theta_2,\ \ldots,\ \theta_p;\quad W_1,\ W_2,\ \ldots,\ W_l)+\varepsilon=f(\boldsymbol{\theta},\ \mathbf{W})+\varepsilon.$$

Пусть $X_1,\ X_2,\ \ldots,\ X_k$ — набор предикторных переменных, которые *не входят* в данную модель. Обозначим некоторый j-й набор значений этих переменных через $X_{1j},\ X_{2j},\ \ldots,\ X_{kj},\ j=1,\ 2,\ \ldots,\ n$. Предположим, что при каждом из этих наборов выполнено несколько различных опытов с вариацией значений W, причем такой, что можно найти оценки параметров для каждого из n наборов величин X.

Составим следующую таблицу:

						столбцы $\hat{\theta}_i$			
X_{11}	X_{21}	X_{31}	. . .	X_{k1}	$\hat{\theta}_{11}$	$\hat{\theta}_{21}$	. . .	$\hat{\theta}_{i1}$	. . . $\hat{\theta}_{p1}$
X_{12}	X_{22}	X_{32}	. . .	X_{k2}	$\hat{\theta}_{12}$	$\hat{\theta}_{22}$	. . .	$\hat{\theta}_{i2}$	. . . $\hat{\theta}_{p2}$
.					.				
.					.				
.					.				
X_{1n}	X_{2n}	X_{3n}	. . .	X_{kn}	$\hat{\theta}_{1n}$	$\hat{\theta}_{2n}$	. . .	$\hat{\theta}_{ln}$	. . . $\hat{\theta}_{pn}$

Можно ожидать, что каждый столбец $\hat{\boldsymbol{\theta}}_i$ будет «устойчивым», т. е. его элементы окажутся практически постоянными, если переменные X никак не влияют на отклик. Отсюда вытекает способ проверки того, зависит ли функция отклика от переменных X. Для этого надо построить регрессию столбцов $\hat{\boldsymbol{\theta}}_i$ по отношению к наборам X-ов и проверить, получились ли какие-либо коэффициенты регрессии значимыми, т. е. нужно подогнать модель

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\ \ldots\ +\beta_kX_k+\varepsilon,$$

или

$$\mathbf{Y}=\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon},$$

где $\mathbf{Y}=\hat{\boldsymbol{\theta}}_i$ и $\mathbf{X}$ — полный блок переменных X, указанный выше. Если оцениваемый коэффициент при X_q значимо отличается от нуля, мы заключаем, что параметр θ_i зависит от переменной X_q и, следовательно, переменная X_q должна присутствовать в исходной модели. (Так должно быть и с любой другой переменной X, имеющей значимые коэффициенты.) Другими словами, исходная модель $Y=f(\boldsymbol{\theta},\ \mathbf{W})+\boldsymbol{\varepsilon}$ неадекватна, и необходимо ее пересмотреть (см. рис. 10.22).

Примеры применения подобной методики, когда переменные X варьировались по схеме дробного факторного эксперимента типа 2^{k-p} ($p\neq 0$) или полного факторного эксперимента, приводятся в работах:

1. Hunter W. G., Mezaki R. A model-building technique for chemical engineering kinetics.— Am. Inst. Chem. Eng. J., 1964, **10**, p. 315—322 (обратите внимание на ошибку в тексте под табл. 4 на с. 320; если бы шестой остаток был положительным (а он как раз отрицателен), то картина была бы очевидной, здесь же требуется дальнейшее исследование).

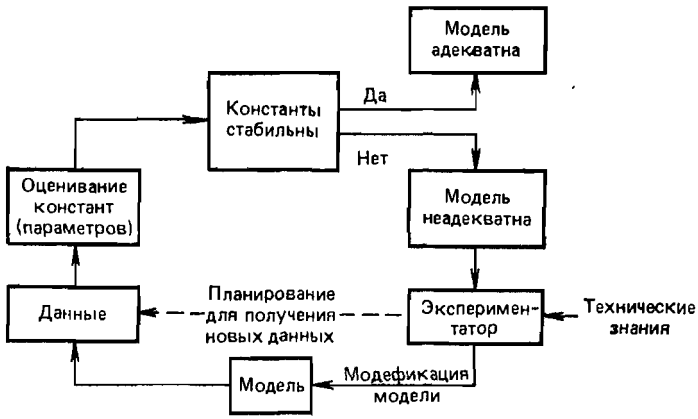


Рис. 10.22. Диаграмма адаптивной процедуры построения модели. Переработана на основе статьи: Box C. E. P., Hunter W. G. — Technometrics, 1962, **4**, p. 302.

2. Box G. E. P., Hunter W. G. A useful method of model building.— Technometrics, 1962, **4**, p. 301—318.

Многомерные (векторные) отклики

Источники: Box G. E. P., Draper N. R. The Bayesian estimation of common parameters from several responses.— Biometrika, 1965, **52**, p. 355—361; Erjavec J., Box G. E. P., Hunter W. G., Mac Gregor J. F. Some problems associated with the analysis of multiresponse data.— Technometrics, 1973, **15**, p. 33—51. В некоторых ситуациях может одновременно наблюдаться несколько переменных откликов, а модели, соответствующие этим откликам, могут содержать некоторую часть или все одинаковые параметры. Хорошим примером для такого случая служит многооткликовая модель (в данном случае — трехоткликовая), описывающая последовательную мономолекулярную реакцию, в которой вещество A превращается в B, а вещество B превращается в C:

$$\eta_1 = \exp(-\phi_1 t),$$
$$\eta_2 = \{\exp(-\phi_1 t) - \exp(-\phi_2 t)\}\,\phi_1/(\phi_2 - \phi_1),$$
$$\eta_3 = 1 + \{-\phi_2 \exp(-\phi_1 t) + \phi_1 \exp(-\phi_2 t)\}/(\phi_2 - \phi_1).$$

Как видно, все отклики зависят от одной предикторной переменной — времени t; отклик η_1 зависит от одного, а отклики η_2 и η_3 — от двух параметров ϕ_1, ϕ_2. Заметим, что здесь $\eta_1 + \eta_2 + \eta_3 = 1$, однако соответствующие наблюдения (y_{1u}, y_{2u}, y_{3u}), $u = 1, 2, \ldots, n$, для (η_1, η_2, η_3) могут быть и независимыми, т. е. их сумма $y_{1u} + y_{2u} + y_{3u}$ не обязательно равна 1 из-за случайных ошибок наблюдения. В таком случае подходящий способ оценивания параметров состоит в минимизации детерминанта матрицы $[v_{ij}]$, где

$$v_{ij} = \sum_{u=1}^{n} (y_{iu} - \eta_{iu})(y_{ju} - \eta_{ju}),$$

по отношению к этим параметрам. Элементами указанной матрицы, как видно, служат суммы квадратов и суммы смешанных произведений отклонений соответствующих экспериментальных значений откликов y_{iu} от величин η_{iu}, определяемых с помощью моделей.

Использование этого критерия для оценивания может привести к трудностям, если один или несколько откликов определяются арифметически, исходя из других измеренных непосредственно откликов. В нашем примере это может иметь место, если, например, y_{1u} и y_{2u} являются фактически измеряемыми откликами, тогда как отклик y_{3u} находится как разность, т. е. $y_{3u} = 1 - y_{1u} - y_{2u}$. Способы обнаружения подобных ситуаций и рекомендации по тому, что надо в таких случаях делать, указаны во второй из приведенных выше работ [33].

Упражнения

1. Оцените параметр θ, входящий в нелинейную модель

$$Y = e^{-\theta t} + \varepsilon,$$

исходя из следующих наблюдений:

t	Y
1	0,80
4	0,45
16	0,04

Постройте приближенный 95 %-й доверительный интервал для θ.

2. Оцените параметр θ нелинейной модели

$$Y = e^{-\theta t} + \varepsilon$$

[33] Надо, однако, иметь в виду, что линейные связи между математическими ожиданиями откликов-концентраций могут возникать в связи с наличием параллельных стадий в сложной химической реакции (см.: Горский В. Г. Планирование кинетических экспериментов.— М.: Наука, 1984.— 241 с. (особо с. 43—47)). Заметим еще, что для задач химической кинетики и многих других задач характерно использование многооткликовых моделей (см., например: Аренс Х., Лёйтер Ю. Многомерный дисперсионный анализ./Пер. с нем.— М.: Финансы и статистика, 1985.— 232 с.). — *Примеч. пер.*

исходя из следующих наблюдений:

t	Y
0,5	0,96; 0,91
1	0,86; 0,79
2	0,63; 0,62
4	0,48; 0,42
8	0,17; 0,21
16	0,03; 0,05

Постройте приближенный 95 %-ный доверительный интервал для θ.

3. Оцените параметры α, β нелинейной модели

$$Y = \alpha + (0{,}49 - \alpha)\, e^{-\beta\,(X-8)} + \varepsilon,$$

используя следующие наблюдения:

X	Y
10	0,48
20	0,42
30	0,40
40	0,39

Постройте приближенную 95 %-ную доверительную область для (α, β).

4. Связь между урожаем некоторой культуры Y и количеством вносимых удобрений X выражается формулой $Y = \alpha - \beta\rho^X + \varepsilon$, где $0 < \rho < 1$. Имеем

X	Y
0	44,4
1	54,6
2	63,8
3	65,7
4	68,9

Получите оценки α, β и ρ. Затем постройте приближенную 95 %-ную доверительную область для (α, β, ρ).

5. Связь между давлением и температурой насыщенного пара можно представить в виде

$$Y = \alpha(10)^{\beta t/(\gamma + t)} + \varepsilon,$$

где Y — давление,
t — температура,
α, β, γ — неизвестные константы.

Были собраны следующие данные:

t (°C)	Y (давление)	t (°C)	Y (давление)
0	4,14	70	224,74
10	8,52	80	341,35
20	16,31	85	423,36
30	32,18	90	522,78
40	64,62	95	674,32
50	98,76	100	782,04
60	151,13	105	920,01

Получите оценки параметров α, β и γ. Кроме того, постройте приближенную 95 %-ную доверительную область для (α, β, γ).

6. Рассмотрите модель

$$Y = \theta + \alpha X_1 X_3 + \beta X_2 X_3 + \alpha\gamma X_1 + \beta\gamma X_2 + \varepsilon.$$

Будет ли эта модель нелинейной? Можно ли получить оценки параметров, пользуясь только методом линейной регрессии, если данные (Y_u, X_{1u}, X_{2u}, X_{3u}) имеются [34]?

7. (Источник: Exploring the Atmosphere's First Mile/H. H. Lettau, B. Davidson, eds.— New York: Pergamon Press, 1957, vol. 1, p. 332—336.) При адиабатических условиях скорость ветра выражается нелинейной моделью

$$Y = \theta_1 \log(\theta_2 X + \theta_3) + \varepsilon,$$

где θ_1 — скорость трения (скорость холостого хода) в см/с; $\theta_2 = 1 +$ (сдвиг нулевой точки); $\theta_3 =$ (длина неровности)$^{-1}$ в см$^{-1}$; X — номинальная высота анемометра (ветрометра). Оцените параметры θ_1, θ_2 и θ_3 исходя из нижеследующих данных. Постройте приближенную 95 %-ную доверительную область для (θ_1, θ_2, θ_3).

X	Y
40	490,2
80	585,3
160	673,7
320	759,2
640	837,5

8. (Источник. Hunter W. G., Atkinson A. C.— Technical Report, N59, Statistics Dept., Wisconsin: University of Wisconsin, Madison, December 1965. Краткое содержание этой работы содержится в статье: Statistical designs for pilot-plant and laboratory experiments Part 11.— Chemical Engineering, 1966, June 6, p. 159—164. Исходные данные были опубликованы в работе: Srinivasan R., Levi A. A. Kinetics of the thermal isomerization of bicyclo [2.1.1] hexane. — Journal of the American Chemical Society, 1963; November 5, p. 3363—3364. Данные на с. 3364, где величина давления в мм была меньше единицы, были опущены).

Некоторая химическая реакция описывается с помощью нелинейной, модели

$$Y = \exp\left\{-\theta_1 X_1 \exp\left[-\theta_2\left(\frac{1}{X_2} - \frac{1}{620}\right)\right]\right\} + \varepsilon,$$

где θ_1 и θ_2 — параметры, подлежащие оцениванию;

Y — доля остающегося непревращенным исходного материала;

X_1 — время реакции в минутах,

X_2 — температура в градусах Кельвина.

Используя приведенные ниже данные, полученные в результате пассивного эксперимента, оцените θ_1 и θ_2, а также постройте 95 %-ную доверительную

---

[34] Так ставить вопрос не совсем корректно. Ведь данные могут быть разными. При одних данных оценки параметров регрессии можно найти однозначно, тогда как при других данных это может и не получиться.— *Примеч. пер.*

область для точки (θ_1, θ_2). При желании можете воспользоваться предварительными оценками $\theta_0' = (\theta_{10}, \theta_{20}) = (0{,}01155;\ 5000)$.

Номер опыта	X_1	X_2	Y	Номер опыта	X_1	X_2	Y
1	120,0	600	0,900	20	60,0	620	0,802
2	60,0	600	0,949	21	60,0	620	0,802
3	60,0	612	0,886	22	60,0	620	0,804
4	120,0	612	0,785	23	60,0	620	0,794
5	120,0	612	0,791	24	60,0	620	0,804
6	60,0	612	0,890	25	60,0	620	0,799
7	60,0	620	0,787	26	30,0	631	0,764
8	30,0	620	0,877	27	45,1	631	0,688
9	15,0	620	0,938	28	40,0	631	0,717
10	60,0	620	0,782	29	30,0	631	0,802
11	45,1	620	0,827	30	45,0	631	0,695
12	90,0	620	0,696	31	15,0	639	0,808
13	150,0	620	0,582	32	30,0	639	0,655
14	60,0	620	0,795	33	90,0	639	0,309
15	60,0	620	0,800	34	25,0	639	0,689
16	60,0	620	0,790	35	60,1	639	0,437
17	30,0	620	0,883	36	60,0	639	0,425
18	90,0	620	0,712	37	30,0	639	0,638
19	150,0	620	0,576	38	30,0	639	0,659

9. (Источник. Hunter, Atkinson, см. ссылку в упражнении 3.) Используя данные, приведенные ниже, которые получены с помощью вычислительного эксперимента, выполненного по плану, оцените параметры θ_1 и θ_2 в модели, предложенной в упражнении 3, и постройте доверительную область с доверительной вероятностью 95 % для (θ_1, θ_2). (Интересно отметить, что 8 опытов, указанных ниже, которые были получены методом последовательного планирования, дают доверительную область для (θ_1, θ_2), несколько меньшей площади, чем 38 неспланированных опытов из упражнения 3. Этот пример служит демонстрацией поразительных преимуществ, которые могут дать специально спланированные опыты. Заметим, в частности, что пределы изменения переменной X_1 теперь намного больше, чем раньше. Это само по себе могло бы привести к уменьшению размеров доверительной области, если бы использовались 38 неплаиируемых опытов. Планирование экспериментов в добавление к этому позволяет использовать даже намного меньше опытов при сохранении той же точности).

Номер опыта	X_1	X_2	Y	Номер опыта	X_1	X_2	Y
1	109	600	0,912	5	1270	600	0,342
2	65	640	0,382	6	69	640	0,358
3	1180	600	0,397	7	1230	600	0,348
4	66	640	0,376	8	68	640	0,376

10. (Источник. Box G. E. P., Hunter W. G. Sequental design of experiments for nonlinear models.— Proceedings of the IBM Scientific Computing Symposium on Statistics, 1963, October 21—23, published in 1965, p. 113—137.)

Некоторая химическая реакция может быть описана с помощью нелинейной модели

$$Y = \theta_1\theta_3 X_1/(1 + \theta_1 X_1 + \theta_2 X_2) + \varepsilon,$$

где Y — скорость реакции,

X_1 и X_2 — парциальные давления реагента и продукта соответственно,

θ_1 и θ_2 — константы адсорбционного равновесия (адсорбционные коэффициенты) для реагента и продукта соответственно,

θ_3 — эффективная константа скорости.

Используя данные, приведенные ниже, оцените θ_1, θ_2 и θ_3, а также постройте 95 %-ную доверительную область для $(\theta_1, \theta_2, \theta_3)$. Вы можете воспользоваться предварительными оценками параметров $\theta'_0 = (\theta_{10}, \theta_{20}, \theta_{30}) = (2{,}9;\ 12{,}2;\ 0{,}69)$, если желаете. Эти данные воспроизводятся с разрешения фирмы IBM.

Номер опыта	X_1	X_2	Y	Номер опыта	X_1	X_2	Y
1	1,0	1,0	0,126	8	3,0	0,0	0,614
2	2,0	1,0	0,219	9	0,3	0,0	0,318
3	1,0	2,0	0,076	10	3,0	0,8	0,298
4	2,0	2,0	0,126	11	3,0	0,0	0,509
5	0,1	0,0	0,186	12	0,2	0,0	0,247
6	3,0	0,0	0,606	13	3,0	0,8	0,319
7	0,2	0,0	0,268				

11. (Источник. Hald A. Statistical Theory with Engineering Application.— New York: J. Wiley, 1960, p. 564.) [35]

Постройте по приведенным ниже даиным нелинейную модель

$$Y = \theta_1 X^{\theta_2} + \varepsilon,$$

найдите 95 %-ную доверительную область для параметров (θ_1, θ_2). Большой объем данных приведен в работе, упомянутой выше.

Скорость автомобиля, X	Тормозной путь, Y
4	5
10	20
17	45
22	66
25	85

12. (Источник. Marske D. Biomedical oxygen demand data interpretation using the sum of squares surface.— M. S. thesis in Civil Engineering.— Wisconsin: University of Wisconsin, Madison, 1967.)

Для каждого набора данных, приведенных ниже, постройте нелинейную модель

$$Y = \theta_1 (1 - e^{-\theta_2 t}) + \varepsilon$$

и приближенную 95 %-ную область для (θ_1, θ_2).

[35] Есть русский перевод: Хальд А. Математическая статистика с техническими приложениями/Пер. с англ. Под ред. Ю. В. Линника.— М.: ИЛ, 1956.— 664 с.— *Примеч. пер.*

Набор 1

t	Y
1	82
2	112
3	153
4	163
5	176
6	192
7	200

Набор 2

t	Y
1	0,47
2	0,74
3	1,17
4	1,42
5	1,60
7	1,84
9	2,19
11	2,17

Набор 3

t	Y
1	168
2	336
3	468
5	660
6	708
7	696

Набор 4

t	Y
1	9
2	9
3	16
4	20
5	21
7	22

Набор 5

t	Y
1	4,3
2	8,2
3	9,5
4	10,4
5	12,1
7	13,1

Набор 6

t	Y
1	6,8
2	12,7
3	14,8
4	15,4
5	17,0
7	19,9

Набор 7

t	Y
1	109
2	149
3	149
5	191
7	213
10	224

Набор 8

t	Y
1	8,3
2	10,3
3	19
4	16
5	15,6
7	19,8

Набор 9

t	Y
1	4710
2	7080
3	8460
4	9580

13. Используя данные о кристаллах льда, приведенные в упражнении 22 гл. 5, постройте нелинейную модель $M = \alpha T^{\beta} + \varepsilon$. Исследуйте остатки и сформулируйте вывод.

14. Приведенные ниже данные были получены при наблюдении за ростом пятн апельсиновых деревьев в Риверсайде, Калифорния, в период 1969—1973 гг. Отклик w в таблице есть диаметр ствола в миллиметрах, а предикторная переменная t означает время, выраженное в днях, причем первый день соответствует 1 января 1969 г. Подгоните к этим данным модели, выражаемые уравнениями (10.7.2), (10.7.5), (10.7.8) и (10.7.10). Основываясь на визуальном исследовании построенных моделей, ответьте, какая из моделей представляется вам наиболее полезной.

Отклик w для каждого из пяти деревьев

t	1	2	3	4	5
118	30	33	30	32	30
484	58	69	51	62	49
664	87	111	75	112	81
1004	115	156	108	167	125
1231	120	172	115	179	142
1372	142	203	139	209	174
1582	145	203	140	214	177

15. Оцените параметры α, β нелинейной модели $Y = \alpha + X^{\beta} + \varepsilon$, используя данные $(X, Y) = (0; -1{,}1)$, $(1; 0)$, $(2; 2{,}9)$, $(3; 8{,}1)$. (Это можно сделать по-разному. Во-первых, с помощью нелинейного МНК. И во-вторых, зафиксировав β и полагая, что $\hat{\alpha}(\beta) = \{\text{среднее из } (Y_u - X_u^{\beta})\}$, можно с помощью линейного МНК найти подходящие оценки α для фиксированных β; после этого построить график зависимости $S(\alpha, \beta) = \sum_{u=1}^{n} \{Y_u - \hat{\alpha}(\beta) - X_u^{\beta}\}^2$ от β и найти оценку $\hat{\beta}$, как ту величину β, которая минимизирует $S(\alpha, \beta)$. Подходящее значение $\hat{\alpha}$ есть тогда та самая величина $\hat{\alpha}(\hat{\beta})$, которая соответствует $\hat{\beta}$.)

16. Напряженное состояние ковкого чугуна [36] можно охарактеризовать тремя показателями:

x — удлинение, %,

y — предел прочности при растяжении, кг на дюйм ^{2},

z — предел текучести, кг на дюйм ^{2}.

Было высказано предположение (это было сделано Лопером и Котши (Loper C. R., Kotschi R. M.) из Висконсинского университета, которым мы выражаем благодарность за этот пример), что модели имеют форму

$$Y = \alpha + \beta/x^{\gamma} + \varepsilon,$$

$$Z = \delta + \theta/x^{\phi} + \varepsilon$$

и что их параметры могут рассматриваться как некоторые подходящие характеристики качества чугуна.

Приведенные ниже в таблице данные представляют собой минимальные значения показателей качества, содержащиеся в промышленном стандарте ASTM A 53 6-70, на основании которого производится выпуск ковкого чугуна. Подгоните приведенные выше модели к этим данным и таким образом получите оценки параметров.

(Подсказка. Обе модели линейны относительно преобразованных предикторных переменных $x^{-\gamma}$ и $x^{-\phi}$. Для первой модели предположите на некоторое время, что γ есть фиксированная величина. Тогда можно решить нормальные МНК-уравнения и определить оценки $\hat{\alpha}$ и $\hat{\beta}$, которые зависят от γ. После этого можно записать выраженне для суммы квадратов, зависящее только от γ. Варьируя γ и вычерчивая график или распечатывая таблицу зависимости $S(\gamma)$ от γ, мы найдем $\hat{\gamma}$, минимизирующее $S(\gamma)$. Это значение $\hat{\gamma}$ и соответствующие величины $\hat{\alpha}$ и $\hat{\beta}$ как раз и будут подходящими МНК-оценками параметров,

[36] Ковкий чугун — разновидность чугуна, получаемая отжигом белого чугуна. Используется в фасонном литье.— *Примеч. пер.*

которые можно использовать. Аналогичный подход можно применнть, чтобы получить оценки $\hat{\delta}$, $\hat{\theta}$ и $\hat{\phi}$.)

Удлинение, % x	Предел прочности, кг/дюйм^2 y	Выходное напряжение, кг/дюйм^2 z
2	120	90
3	100	70
6	80	55
12	65	45
18	60	40

17. Подгоните модель к данным упражнения 26 гл. 5; модель имеет форму

$$Y = \alpha X_1^{\beta} X_2^{\gamma} + \varepsilon.$$

18. (Источники: Ledger H. P., Sayers A. R. The utilization of dietary energy by steers during periods of restricted food intake and subsequent realimentation, Part 1.— Journal of Agricultural Science, Cambrige, 1977, 88, p. 11—26; Ledger H. P., Part 1, p. 27—33 of the same journal issue. Адаптировано с разрешения авторов.)

Таблица к упражнению 10.16. **Групповые средние ежедневного потребления сухого корма в % от живого веса**

Условное обозначение группы животных		Недели содержания							
		3	6	9	12	15	18	21	24
В 185 кг	%	1,835	1,255	1,037	1,045	0,900	0,901	0,836	0,896
	ст. откл.	0,412	0,315	0,692	0,712	0,149	0,197	0,156	0,234
В 275 кг	%	1,702	1,313	1,037	0,952	0,863	0,879	0,897	0,800
	ст. откл.	0,292	0,262	0,245	0,187	0,161	0,145	0,160	0,199
В × Н 275 кг	%	1,545	1,134	0,941	0,840	0,791	0,822	0,855	0,773
	ст. откл.	0,206	0,205	0,215	0,211	0,293	0,095	0,072	0,124
$\frac{3}{4}$ В 450 кг	%	0,990	0,803	0,790	0,797	0,734	0,687	0,687	0,716
	ст. откл.	0,287	0,191	0,168	0,093	0,127	0,145	0,162	0,224
$\frac{3}{4}$ Н 450 кг	%	0,818	0,753	0,737	0,713	0,664	0,660	0,706	0,670
	ст. откл.	0,249	0,185	0,133	0,140	0,196	0,122	0,122	0,114

В таблице показаны значения отклика Y (групповое среднее ежедневного потребления сухого корма в процентах от живого веса бычков), измеренные в течение восьми недель, равномерно отстоящих друг от друга; X — недели содержания животных. В таблице даны также стандартные отклонения откликов. Приведено 5 групп таких данных, коды которых указаны в левом столбце. Каждой группе отвечает, следовательно, свой отклик. Так, например, в четвертой группе шестое наблюдение при $X = 18$ (18 недель содержания) равно $Y = 0{,}687$. Этому отклику соответствует стандартное отклонение 0,145. К каждой группе данных в отдельности, используя взвешенный МНК, подгоните модель вида

$$Y = \beta_0 + \beta_1 \theta^X + \varepsilon$$

и выполните обычный анализ.

(Подсказка. Если в вашем распоряжении нет программы взвешенного МНК, можно использовать следующие подходы:

1. Можно воспользоваться взвешенным линейным МНК, фиксируя различные значения параметра θ, скажем, в интервале $0 \leqslant \theta \leqslant 1$ и определяя при этом оценки параметров β_0 и β_1 и остаточную сумму квадратов, которую они минимизируют. Зная значения остаточной суммы для разных θ, можно определить соответствующее минимуму суммы и выписать отвечающие этой величине оценки $\hat{\beta}_0$ и $\hat{\beta}_1$.

2. Записать модель в виде

$$Y/\text{ст.откл.} = (1/\text{ст.откл.})\,\beta_0 + (\theta^X/\text{ст.откл.})\,\beta_1 + \text{ошибка}$$

и использовать затем обычный невзвешенный нелинейный МНК, чтобы оценить β_0, β_1 и θ.

3. Можно записать модель в преобразованной форме, как указано в предыдущем подходе, а затем воспользоваться обычным (невзвешенным) линейным МНК, поступая как в первом подходе. Иначе говоря, поочередно фиксируя выбранные, скажем, из интервала $0 \leqslant \theta \leqslant 1$, различные значения параметра θ, определять на каждом таком шаге оценки параметров β_0 и β_1 и величины остаточной суммы квадратов. Зная последние, можно найти ту величину $\hat{\theta}$, которая соответствует наименьшей из сумм. Затем остается лишь выписать, связанные с этой суммой оценки $\hat{\beta}_0$ и $\hat{\beta}_1$.

Заметим, что $\mathbf{V}(\mathbf{Y})$ в данной задаче есть диагональная матрица.)

19. (Источник. House C. C. U. S. Department of Agriculture.— Washington: D. C., 20250.) Приведенные далее табличные данные представляют собой 8 выборок, полученных при изучении скорости роста кукурузы. Эти данные содержат:

Y — средний по четырем растениям вес сухого зерна,

t — среднее время с момента образования пестиков в початках («шелкования»)[37] по четырем растениям.

1. Для каждой из приведенных выборок постройте нелинейную модель вида

$$Y = \delta - \ln(1 + \beta e^{-kt}) + \varepsilon,$$

где δ, β и k — параметры, подлежащие оцениванию.

2. Нанесите данные каждой выборки на график и проведите там же полученную кривую. Прокомментируйте, что вы увидите как на графиках, так и в распечатках с ЭВМ.

3. Постройте точечную диаграмму по всем восьми значениям $\hat{\delta}$, затем то же сделайте для величин $\hat{\beta}$ и $\hat{k}$. Прокомментируйте, что вы увидите на этих диаграммах. Что должны были бы вы видеть, если бы не было никакой разницы между выборками?

4. Постройте указанную в п. 1 нелинейную модель сразу по всем данным и добавьте полученные оценки параметров к тем, которые нанесены на точечные диаграммы в п. 3. Каково Ваше заключение?

5. Отклики, приведенные в таблице,— это средние из четырех наблюдений. Если бы были известны результаты индивидуальных наблюдений, стали бы вы их использовать в процедуре подгонки вместо средних значений? Объясните

---

[37] «Шелком» называется множество пестиков в початке кукурузы. При производстве кукурузы на зерно пестики удаляют. Эту операцию называют «шелкованием». Существует предположение, что скорость роста кукурузы зависит от момента проведения этой операции, что и изучается в данном примере.— *Примеч. пер.*

свой ответ. Укажите преимущества и недостатки обоих приемов.

Выборка	Y	t	Выборка	Y	t
8	11,44	13,625	24	37,84	18,625
	29,51	19,750		67,34	29,125
	69,05	28,625		157,10	49,250
	98,79	41,750			
	138,01	49,625			
	162,82	69,625			
			32	7,84	13,375
				15,12	16,875
14	35,88	19,750		73,97	28,250
	106,89	30,250		110,58	43,125
	168,58	35,625		114,36	47,500
	136,84	43,375		185,87	58,000
	164,50	49,500		115,18	60,625
16	3,52	6,500	52	10,60	8,750
	10,56	14,875		15,13	14,875
	41,55	24,625		38,74	22,500
	94,55	35,875		120,19	37,625
	122,52	49,875		126,32	41,750
	130,19	56,875		171,75	51,625
				156,67	63,875
22	14,26	17,125			
	50,51	25,625	54	11,92	13,625
	60,83	29,625		66,82	24,750
	104,78	39,625		28,29	24,875
	96,46	46,375		106,92	39,000
	97,02	54,250		129,83	53,875
	172,41	62,125		143,26	60,875

20. (Источник. Bates D. M., Watts D. G. Relative curvature measures of nonlinearity. — Journal of the Royal Statistical Society, 1980, **B—42**, p. 1—16; discussion 16—25.) С помощью метода наименьших квадратов подгоните нелинейную модель

$$Y = \theta_1 X/(\theta_2 + X) + \varepsilon$$

к имитированным данным, приведенным в таблице. Выполните полный анализ результатов, включая составление графика исходных данных и полученной кривой. Предположим, что вас попросили бы запланировать один дополнительный опыт. При каком значении X вы предложили бы провести этот опыт? А если бы потребовалось запланировать два дополнительных опыта, то при каких наилучших значениях X вы рекомендовали бы провести такие опыты?

X	Y	X	Y	X	Y	X	Y
2,000	0,0615	0,667	0,0258	0,286	0,0129	0,222	0,0169
2,000	0,0527	0,400	0,0138	0,286	0,0183	0,200	0,0129
0,667	0,0334	0,400	0,0258	0,222	0,0083	0,200	0,0087

21. Вернитесь к модели упражнения 20. Более общая модель этого типа имеет вид

$$Y = (\theta_4 + \theta_1 X^{\theta_3})/(\theta_2 + X^{\theta_3}) + \varepsilon.$$

Обратите внимание на то, что если $\theta_4 = 0$, то эта регрессионная функция проходит через начало координат; в таком виде это модель Хилла. Если же $\theta_4 = 0$ и $\theta_3 = 1$, то получим так называемую модель Михаэлиса-Ментена, рассмотренную в упражнении 20 (см. статью: Morgan P. H., Mercer L. P., Flodin N. W. General model for nutritional responses of Higher organisms.— Proceedings of the National Academy of Sciences, USA, 1975, **72**, November, p. 4327—4331). Дополнительные примеры использования этой модели содержатся в статье: Mercer L. P., Flodin N. W., Morgan P. H. New methods for comparing the biological efficiency of alternate nutrient sources.— Journal of Nutrition, 1978, **108**, August, p. 1244—1249.

Прежде всего подгоните эту более общую модель к данным упражнениям 18 и затем найдите «дополнительную сумму квадратов» для параметра θ_4 при наличии в модели θ_1, θ_2 и θ_3; затем — для параметра θ_3 при наличии θ_1, θ_2, полагая при этом, что $\theta_4 = 0$, и, наконец, для параметров θ_3, θ_4 при наличии θ_1, θ_2. (Получите эти результаты путем вычитания сумм квадратов, отвечающих соответствующим моделям.) Сравните эти «дополнительные SS» с величиной среднего квадрата, отвечающего «чистой» ошибке s_e^2. К какому заключению вы пришли? Какую модель вы использовали бы для представления данных и почему?

22. Подгоните модель (10.0.3) к каждому из двух наборов данных, приведенных ниже, и проведите стандартный анализ. Какова корреляция между $\widehat{\theta}_1$ и $\widehat{\theta}_2$ в каждом случае? Как вы могли бы охарактеризовать эти два набора данных?

Набор данных 1		Набор данных 2	
t	Y	t	Y
0,2	0,142	1	0,445
0,4	0,240	2	0,585
0,6	0,329	3	0,601
0,8	0,381	4	0,532
1,0	0,455	5	0,470

Ответы к упражнениям

1. $\widehat{\theta} = 0{,}20345$, $S(\widehat{\theta} = 0{,}00030$; $0{,}179 \leqslant \theta \leqslant 0{,}231$.

2. $\widehat{\theta} = 0{,}20691$, $S(\widehat{\theta}) = 0{,}01202$; $0{,}190 \leqslant \theta \leqslant 0{,}225$.

3. $(\widehat{\alpha}, \widehat{\beta}) = (0{,}38073;\ 0{,}07949)$, $S(\widehat{\alpha}, \widehat{\beta}) = 0{,}00005$; $S(\alpha, \beta) = 0{,}001$ (или более точно 0,0009).

4. $(\widehat{\alpha}, \widehat{\beta}, \widehat{\rho}) = (72{,}4326;\ 28{,}2519;\ 0{,}5968)$, $S(\widehat{\alpha}, \widehat{\beta}, \widehat{\rho}) = 3{,}5688$; $S(\alpha, \beta, \rho) = 106{,}14$.

5. $(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}) = (5{,}2673;\ 8{,}5651;\ 294{,}9931)$, $S(\widehat{\alpha}, \widehat{\beta}, \widehat{\gamma}) = 1718{,}2108$: $S(\alpha, \beta, \gamma) = 3400$.

6. Запишите модель в виде

$$Y = \theta + \alpha(X_1X_3 + \gamma X_1) + \beta(X_2X_3 + \gamma X_2) + \varepsilon.$$

Зафиксируйте γ и найдите оценки $\widehat{\theta}$, $\widehat{\alpha}$ и $\widehat{\beta}$. Повторите эту операцию многократно и найдите значение $\widehat{\gamma}$, при котором сумма $S(\theta, \alpha, \beta, \gamma)$ является наименьшей среди всех сумм, отвечающих разным γ.

7. $(115{,}2;\ 2{,}310;\ -22{,}022)$, $S(\widehat{\theta}) = 7{,}0133$. $S(\widehat{\theta}) = 209{,}0$.

8. $(0{,}00376;\ 27{,}539)$, $S(\widehat{\theta}) = 0{,}00429326$. $S(\theta) = 0{,}00507559$.

9. (0,00366; 27,627), $S(\hat{\theta}) = 0{,}000754$, $S(\theta) = 0{,}00204$.

10. (3,57; 12,77; 0,63), $S(\hat{\theta}) = 0{,}00788$, $S(\theta) = 0{,}01665$.

11. (0,480; 1,603), $S(\hat{\theta}) = 7{,}301$, $S(\theta) = 53{,}8$.

Вытянутый тонкий контур указывает на то, что имеется большое число пар значений параметров, которые почти такие же подходящие, как и фактические МНК-оцеики параметров.

12.

№	$\hat{\theta}_1$	$\hat{\theta}_2$	$S(\hat{\theta})$	$S(\theta)$
1	205,25	0,431	252	835,8
2	2,498	0,202	0,0262	0,0712
3	892,67	0,245	3 376,5	15 093
4	25,475	0,323	17,004	76,007
5	13,809	0,398	0,866	3,871
6	19,903	0,441	3,716	16,61
7	213,82	0,547	1 168	5 221
8	19,142	0,531	25,99	116,18
9	10 525	0,569	68349	1 366 980

13. Решение не приводится.

14. (Часть решения.)

Дерево № 1

$$(10.7.2)\ \hat{\alpha} = 268{,}3 \quad \hat{\beta} = 0{,}9478 \quad \hat{k} \cdot 10^4 = 4{,}740$$
$$(10.7.5)\ \hat{\alpha} = 154{,}1 \quad \hat{\beta} = 5{,}643 \quad \hat{k} \cdot 10^3 = 2{,}759$$
$$(10.7.8)\ \hat{\alpha} = 5{,}032 \quad \hat{\beta} = 5{,}792 \quad \hat{k} \cdot 10^3 = 2{,}814$$
$$(10.7.10)\ \hat{\alpha} = 172{,}2 \quad \hat{\beta} = 2{,}813 \quad \hat{k} \cdot 10^3 = 1{,}626$$

Визуально модель (10.7.8) представляется лучшей, хотя это может быть связано с преобразованием отклика с помощью логарифмирования. Первая модель (10.7.2) не передает хорошо S-образное поведение данных. Однако набор данных, содержащий только семь наблюдений, слишком мал, чтобы можно было сделать определенные выводы.

Дерево № 2

$$(10.7.2)\ \hat{\alpha} = 519{,}3 \quad \hat{\beta} = 0{,}9820 \quad \hat{k} \cdot 10^4 = 3{,}208$$
$$(10.7.5)\ \hat{\alpha} = 218{,}9 \quad \hat{\beta} = 8{,}225 \quad \hat{k} \cdot 10^3 = 3{,}010$$
$$(10.7.8)\ \hat{\alpha} = 5{,}398 \quad \hat{\beta} = 8{,}228 \quad \hat{k} \cdot 10^3 = 2{,}962$$
$$(10.7.10)\ \hat{\alpha} = 248{,}4 \quad \hat{\beta} = 2{,}645 \quad \hat{k} \cdot 10^3 = 1{,}703$$

15—17. Решение не приводится.

18.

Группа	Уравнение $\hat{Y} =$
В 185 кг	$0{,}8822 + 2{,}2290\ (0{,}7512)X$
В 275 кг	$0{,}8444 + 1{,}7083\ (0{,}796)X$
В×Н 275 кг	$0{,}8079 + 1{,}7701\ (0{,}7483)X$
0,75 В 450 кг	$0{,}7066 + 0{,}5700\ (0{,}7905)X$
0,75 Н 450 кг	$0{,}6699 + 0{,}2448\ (0{,}8458)X$

19. (Основные детали). Оценки параметров равны:

Выборка	$\hat{\delta}$	$\hat{\beta}$	$\hat{k}$
8	4,96	78,76	0,15
14	5,08	222,36	0,21
16	4,88	103,70	0,15
22	4,79	128,24	0,17
24	5,71	23,84	0,07
32	4,88	217,93	0,20
52	5,18	50,66	0,12
54	4,95	49,17	0,13
Все	4,96	89,32	0,15

Хотя выборки 14, 24 и 32 и кажутся резко выделяющимися, различные графики обнаруживают полную последовательность этих данных.

Индивидуальные наблюдения позволили бы найти оценку дисперсии, основанную на «чистой» ошибке, и позволили бы проверить ее (предполагаемое) постоянство; поскольку каждое среднее определяется по одинаковому числу наблюдений, фактические оценки параметров не изменятся.

20. (Основные результаты.)

$$\hat{\theta} = (0{,}10579;\ 1{,}7007)',$$

S, обусловленная чистой ошибкой $= 1{,}998 \times 10^{-4}$ (6 степеней свободы)
S, обусловленная неадекватностью $= 1{,}08 \times 10^{-6}$ (4 степени свободы).

Модель, по-видимому, удовлетворительная. Единственный наилучший опыт есть тот, в котором предикторная переменная X имеет предельно возможное наибольшее значение; наилучшие два опыта — это опыты, которые следует проводить при одних и тех же условиях, указанных выше.

21. Решение не приводится.

22. Первый набор данных имеет характеристики, показанные на рис. 10.2,а. Приближенный 95 %-ный доверительный контур сильно вытянут, и оценки $\hat{\theta}_1$ и $\hat{\theta}_2$ сильно коррелированы. Второй набор данных похож на дополнительные данные, приведенные на рис. 10.2, б. Он дает намного лучшее оценивание, приближенный 95 %-ный контур охватывает значительно меньшую площадь. В качестве дополнительного упражнения произведите оценивание сразу по всем девяти опытам (опуская повторный опыт при $t = 1$) и проведите сравнительный анализ всех трех вариантов.

Нормальное распределение (одностороннее)

Доля (A) всей площади, лежащей справа от точки $x = \mu + z\sigma$, $[z = (x - \mu)/\sigma]$.

(z)		0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09	Приставка (z)	
0,0	0,5	000	960	920	880	840	801	761	721	681	641	0,4	0,0
0,1	0,4	602	562	522	483	443	404	364	325	286	247	0,4	0,1
0,2	0,4	207	168	129	090	052	013	974	936	897	859	0,3	0,2
0,3	0,3	821	783	745	707	669	632	594	557	520	483		0,3
0,4		446	409	372	336	300	264	228	192	156	121	0,3	0,4
0,5	0,3	085	050	015	981	946	912	877	843	810	776	0,2	0,5
0,6	0,2	743	709	676	643	611	578	546	514	483	451		0,6
0,7		420	389	358	327	296	266	236	206	177	148	0,2	0,7
0,8	0,2	119	090	061	033	005	977	949	922	894	867	0,1	0,8
0,9	0,1	841	814	788	762	736	711	685	660	635	611		0,9
1,0		587	562	539	515	492	469	446	423	401	379		1,0
1,1		357	335	314	292	271	251	230	210	190	170	0,1	1,1
1,2	0,1	151	131	112	093	075	056	038	020	003	985	0,0	1,2
1,3	0,0	968	951	934	918	901	885	869	853	838	823		1,3
1,4		808	793	778	764	749	735	721	708	694	681		1,4
1,5		668	655	643	630	618	606	594	582	571	559		1,5
1,6		548	537	526	516	505	495	485	475	465	455		1,6
1,7		446	436	427	418	409	401	392	384	375	367		1,7
1,8		359	351	344	336	329	322	314	307	301	294		1,8
1,9		287	281	274	268	262	256	250	244	239	233		1,9
2,0		228	222	217	212	207	202	197	192	188	183		2,0
2,1		179	174	170	166	162	158	154	150	146	143		2,1
2,2		139	136	132	129	125	122	119	116	113	110	0,0	2,2
2,3	0,0	107	104	102	990	964	939	914	889	866	842	0,00	2,3
2,4	0,00	820	798	776	755	734	714	695	676	657	639		2,4
2,5		621	604	587	570	554	539	523	508	494	480		2,5
2,6		466	453	440	427	415	402	391	379	368	357		2,6
2,7		347	336	326	317	307	298	289	280	272	264		2,7
2,8		256	248	240	233	226	219	212	205	199	193		2,8
2,9	0,00	187	181	175	169	164	159	154	149	144	139	0,00	2,9

Заимствовано из книги: Davies O. L., ed. The Design and Analysis of Industrial Experiments, 2nd ed.— Edinburg: Oliver and Boyed, 1956 (книга переиздавалась в 1978 г. издательством Longman Group в Нью-Йорке); сжатое изложение и адаптация из книги: Pearson E. S., Hartly H. O. Biometrika Tables for Statisticians.— New York: Cambridge University Press, 1954 vol. 1.

Нормальное распределение

[расширение на большие значения отклонений]

Отклонение z	Доля площади A	Отклонение z	Доля площади A	Отклонение z	Доля площади A	Отклонение z	Доля площади A
3,0	0,00135	3,5	0,000233	4,0	$0,0^{4}317$	4,5	$0,0^{5}340$
3,1	0,000968	3,6	0,000159	4,1	$0,0^{4}207$	4,6	$0,0^{5}211$
3,2	0,000687	3,7	0,000108	4,2	$0,0^{4}133$	4,7	$0,0^{5}130$
3,3	0,000483	3,8	$0,0^{4}723$	4,3	$0,0^{6}854$	4,8	$0,0^{6}793$
3,4	0,000337	3,9	$0,0^{4}481$	4,4	$0,0^{5}541$	4,9	$0,0^{6}479$

Адаптировано из книги: D a v i e s O. L., ed. The Design and Analysis of Industrial Experiments, 2nd ed.— Edinburgh: Oliver and Boyd, 1956; сжатое изложение и адаптация из книги Пирсона и Хартли: P e a r s o n E. S., H a r t l e y H. O. Biometrika Tables for Statisticians.— New York, Cambridge University Press, 1954, vol. 1.

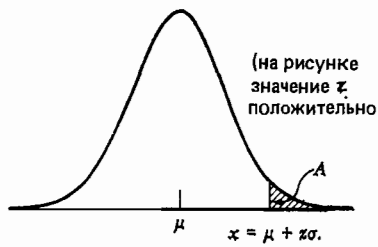


Иллюстрация показывает нормальную кривую. Масштаб кривой выбран так, чтобы вся площадь под кривой была равна единице. Заштрихованная часть площади под кривой равна значениям A, указанным выше в таблице. Эти значения соответствуют положительным значениям аргумента z. Отрицательным значениям соответствуют дополнения, которые можно найти, вычитая из единицы значения A, отвечающие положительным значениям данного аргумента.

П р и м е р ы. *Пусть* $z = +1{,}96$. Приставка $= 0{,}0$. Табличное значение $= 0{,}0250$, откуда вытекает, что площадь справа $= 0{,}0250$. Площадь слева (от точки, отвечающей аргументу $z = 1{,}96$) составляет $1-0{,}0250 = 0{,}975$.

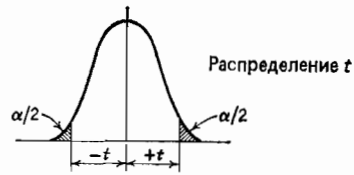
Пусть $z = -3{,}00$. Табличное значение $= 0{,}00135$. Поскольку z отрицательное, табличное значение соответствует площади *слева*. Площадь справа $= 1-0{,}00135 = 0{,}99865$.

Пусть $z = +4{,}50$. Таблнчное значение $= 0{,}00000340$. Площадь слева $= 0{,}99999660$.

Чтобы найти величину z, соответствующую заданному значению A, можно воспользоваться таблицей «в обратном порядке».

Пусть, например, известна *площадь справа*, т. е. $A = 0{,}10$. Два ближайших к этой величине табличных значения: $A = 0{,}1103$ для $z = 1{,}28$ и $A = 0{,}0985$ для $z = 1{,}29$. Теперь надо выполнить линейную интерполяцию, чтобы получить требуемое значение z. Выполняя такую операцию, получаем $z = 1{,}28 + 3 \cdot 0{,}01/18 = 1{,}2817$.

Нормальное распределение (двустороннее)



Вероятность = площадь двух хвостов распределения
за пределами величин ± t из таблицы

Степени свободы	Вероятность									
	0,9	0,7	0,5	0,3	0,2	0,1	0,05	0,02	0,01	0,001
1	0,158	0,510	1,000	1,963	3,078	6,314	12,706	31,821	63,657	636,619
2	0,142	0,445	0,816	1,386	1,886	2,920	4,303	6,965	9,925	31,598
3	0,137	0,424	0,765	1,250	1,638	2,353	3,182	4,541	5,841	12,924
4	0,134	0,414	0,741	1,190	1,533	2,132	2,776	3,747	4,604	8,610
5	0,132	0,408	0,727	1,156	1,476	2,015	2,571	3,365	4,032	6,869
6	0,131	0,404	0,718	1,134	1,440	1,943	2,447	3,143	3,707	5,959
7	0,130	0,402	0,711	1,119	1,415	1,895	2,365	2,998	3,499	5,408
8	0,130	0,399	0,706	1,108	1,397	1,860	2,306	2,896	3,355	5,041
9	0,129	0,398	0,703	1,100	1,383	1,833	2,262	2,821	3,250	4,781
10	0,129	0,397	0,700	1,093	1,372	1,812	2,228	2,764	3,169	4,587
11	0,129	0,396	0,697	1,088	1,363	1,796	2,201	2,718	3,106	4,437
12	0,128	0,395	0,695	1,083	1,356	1,782	2,179	2,681	3,055	4,318
13	0,128	0,394	0,694	1,079	1,350	1,771	2,160	2,650	3,012	4,221
14	0,128	0,393	0,692	1,076	1,345	1,761	2,145	2,624	2,977	4,140
15	0,128	0,393	0,691	1,074	1,341	1,753	2,131	2,602	2,947	4,073
16	0,128	0,392	0,690	1,071	1,337	1,746	2,120	2,583	2,921	4,015
17	0,128	0,392	0,689	1,069	1,333	1,740	2,110	2,567	2,898	3,965
18	0,127	0,392	0,688	1,067	1,330	1,734	2,101	2,552	2,878	3,922
19	0,127	0,391	0,688	1,066	1,328	1,729	2,093	2,539	2,861	3,883
20	0,127	0,391	0,687	1,064	1,325	1,725	2,086	2,528	2,845	3,850
21	0,127	0,391	0,686	1,063	1,323	1,721	2,080	2,518	2,831	3,819
22	0,127	0,390	0,686	1,061	1,321	1,717	2,074	2,508	2,819	3,792
23	0,127	0,390	0,685	1,060	1,319	1,714	2,069	2,500	2,807	3,767
24	0,127	0,390	0,685	1,059	1,318	1,711	2,064	2,492	2,797	3,745
25	0,127	0,390	0,684	1,058	1,316	1,708	2,060	2,485	2,787	3,725
26	0,127	0,390	0,684	1,058	1,315	1,706	2,056	2,479	2,779	3,707
27	0,127	0,389	0,684	1,057	1,314	1,703	2,052	2,473	2,771	3,690
28	0,127	0,389	0,683	1,056	1,313	1,701	2,048	2,467	2,763	3,674
29	0,127	0,389	0,683	1,055	1,311	1,699	2,045	2,462	2,756	3,659
30	0,127	0,389	0,683	1,055	1,310	1,697	2,042	2,457	2,750	3,646
40	0,126	0,388	0,681	1,050	1,303	1,684	2,021	2,423	2,704	3,551
60	0,126	0,387	0,679	1,046	1,296	1,671	2,000	2,390	2,660	3,460
120	0,126	0,386	0,677	1,041	1,289	1,658	1,980	2,358	2,617	3,373
∞	0,126	0,385	0,674	1,036	1,282	1,645	1,960	2,326	2,576	3,291

F-распределение. Верхние 10 % -ные точки F (υ_1, υ_2, 0,90)
Степени свободы для числителя

10%

υ_2 \ υ_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19	60,71	61,22	61,74	62,00	62,26	62,53	62,79	63,06	63,33
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39	9,41	9,42	9,44	9,45	9,46	9 47	9,47	9,48	9,49
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23	5,22	5,20	5,18	5,18	5,17	5,16	5,15	5,14	5,13
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92	3,90	3,87	3,84	3,83	3,82	3,80	3,79	3,78	3,76
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30	3,27	3,24	3,21	3,19	3,17	3,16	3,14	3,12	3,10
6	3,78	3,46	3,29	3,18	3 11	3,05	3,01	2,98	2,96	2,94	2,90	2,87	2,84	2,82	2,80	2,78	2,76	2,74	2,72
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70	2,67	2,63	2,59	2,58	2,56	2,54	2,51	2,49	2,47
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54	2,50	2,46	2,42	2,40	2,38	2,36	2,34	2,32	2,29
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42	2,38	2,34	2,30	2,28	2,25	2,23	2,21	2,18	2,16
10	3,29	2,92	2 73	2,61	2,52	2,46	2,41	2,38	2,35	2,32	2,28	2,24	2,20	2,18	2,16	2,13	2,11	2,08	2,06
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25	2,21	2,17	2,12	2,10	2,08	2,05	2,03	2,00	1,97
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19	2,15	2,10	2,06	2,04	2,01	1,99	1,96	1,93	1,90
13	3,14	2,76	2,56	2,43	2,35	2 28	2,23	2,20	2,16	2,14	2,10	2,05	2,01	1,98	1,96	1,93	1,90	1,88	1,85
14	3,10	2,73	2,52	2,39	2,31	2 24	2,19	2,15	2,12	2,10	2,05	2,01	1,96	1,94	1,91	1,89	1,86	1,83	1,80
15	3,07	2,70	2,49	2,36	2 27	2,21	2,16	2,12	2,09	2,06	2,02	1,97	1,92	1,90	1,87	1,85	1,82	1,79	1,76
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03	1,99	1,94	1,89	1,87	1,84	1,81	1,78	1,75	1,72
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00	1,96	1,91	1,86	1,84	1,81	1,78	1,75	1,72	1,69
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98	1,93	1,89	1,84	1,81	1,78	1,75	1,72	1,69	1,66
19	2,99	2,61	2,40	2,27	2,17	2,11	2,06	2,02	1,98	1,96	1,91	1,86	1,81	1,79	1,76	1,73	1,70	1,67	1,63
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94	1,89	1,84	1,79	1,77	1,74	1,71	1,68	1,64	1,61
21	2,96	2,57	2,36	2,23	2,14	2,08	2,02	1,98	1,95	1,92	1,87	1,83	1,78	1,75	1,72	1,69	1,66	1,62	1,59
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90	1,86	1,81	1,76	1,73	1,70	1,67	1,64	1,60	1,57
23	2,94	2,55	2,34	2,21	2,11	2,05	1,99	1,95	1,92	1,89	1,84	1,80	1,74	1,72	1,69	1,66	1,62	1,59	1,55
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88	1,83	1,78	1,73	1,70	1,67	1,64	1,61	1,57	1,53
25	2,92	2,53	2,32	2,18	2,09	2,02	1,97	1,93	1,89	1,87	1,82	1,77	1,72	1,69	1,66	1,63	1,59	1,56	1,52
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86	1,81	1,76	1,71	1,68	1,65	1,61	1,58	1,54	1,50
27	2,90	2,51	2,30	2,17	2,07	2,00	1,95	1,91	1,87	1,85	1,80	1,75	1,70	1,67	1,64	1,60	1,57	1,53	1,49
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84	1,79	1,74	1,69	1,66	1,63	1,59	1,56	1,52	1,48
29	2,89	2,50	2,28	2,15	2,06	1,99	1,93	1,89	1,86	1,83	1,78	1,73	1,68	1,65	1,62	1,58	1,55	1,51	1,47
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82	1,77	1,72	1,67	1,64	1,61	1,57	1,54	1,50	1,46
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76	1,71	1,66	1,61	1,57	1,54	1,51	1,47	1,42	1,38
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71	1,66	1,60	1,54	1,51	1,48	1,44	1,40	1,35	1,29
120	2,75	2,35	2,13	1,99	1,90	1,82	1,77	1,72	1,68	1,65	1,60	1,55	1,48	1,45	1,41	1,37	1,32	1,26	1,19
∞	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60	1,55	1,49	1,42	1,38	1,34	1,30	1,24	1,17	1,00

Перепечатано из книги: Pearson E. S., Hartley H. O. Biometrika Tables for Statisticians,— New York: Cambridge University Press, 1954, vol. 1.

F-распределение. Верхние 5 %-ные точки $F(\nu_1, \nu_2, 0{,}95)$

Степени свободы для числителя

5%

ν_2 \ ν_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1	252,2	253,3	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00

Pearson E. S., Hartley H. O. Biometrika Tables for Statisticians.— New York: Cambridge University Press, 1954, vol. I.

F-распределение. Верхние 1 % -ные точки $F(v_1, v_2, 0{,}99)$
Степени свободы для числителя

1%

v_2 \ v_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052	4999,5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

Pearson E. S., Hartley H. O. Biometrika Tables for Statisticians. — New York: Cambridge University 1954, vol. 1.

ПРИЛОЖЕНИЯ

ПРИЛОЖЕНИЕ А

ОБЪЯСНЕНИЕ ДАННЫХ И СИМВОЛОВ В ПРИЛОЖЕНИЯХ

Определение переменных

Каждое приложение начинается с названия регрессионной задачи, за которым следует определенне используемых переменных.

Первоначальные и/или преобразованные данные

Затем перечисляются входные данные. Каждая строка представляет совокупность одновременных наблюдений над всеми независимыми н зависимой переменными (опыт). Номеру опыта соответствует номер строки слева от матрицы исходных данных.

Средние преобразованных переменных

Строка средних значенний каждого столбца.

Стандартные отклонения преобразованных переменных

Каждый элемент этой строкн является стандартным отклонением столбца исходных данных.

Корреляционная матрица

Матрнца вычисленных коэффициентов корреляцни r_{ij}. Диагональные элементы должны быть в точностн единицамн. Нарушение этого условия объясняется ошибками округления, возникающими при машинном счете.

Информация для управления

Информация для управления включает:

а) Число наблюдений — равно числу наблюдений отклика в данной задаче.

б) Номер, соответствующий переменной-отклику. Номер обозначает, какой столбец матрицы исходных данных следует рассматривать как отклик. Например, № 4 будет указывать, что данные столбца 4 следует рассматривать как отклик или переменную Y.

в) Уровень риска β для доверительного интервала. Эта строка обозначает вероятность α совершения ошибки первого рода. Используется для расчета доверительных интервалов уровня $1-\alpha$ коэффициентов регрессии.

г) Перечень исключенных переменных. Это номера, определяющие, какие векторы не должны рассматриваться при подборе регрессионной модели.

д) Включаемые переменные. Независимые переменные, которые выбраны для введения в регрессию на определенном шаге.

е) Последовательный F-критерий. F-критерий служит для проверки того, существенно ли влияет последняя переменная, включенная в регрессию, на понижение величины необъясненной вариации. Детально обсуждается в гл. 2 и 4.

ж) Доля объясненной вариацни в % — R^2. Это квадрат коэффициента множественной корреляцни, R^2. Объяснение этой статистики дается в гл. 1, 2 и 4.

з) Стандартное отклонение остатков. Корень квадратный нз среднего квадрата ошибкн в таблице дисперсионного анализа.

и) Средний отклнк. Арифметическое среднее всех наблюденных значений отклика.

к) Стандартное отклонение в процентах от среднего отклнка. Мера велнчины стандартного отклонения остатков относительно среднего отклика рассчитывается как отношение стандартного отклонения остатков к среднему отклику.

л) Степени свободы. Число степеней свободы, используемое для расчета стандартного отклонения остатков.

м) Значение определителя. Значение определителя корреляционной матрицы всех переменных в регрессии на каждом шаге машинного счета.

Дисперсионный анализ (ANOVA)

Это таблица дисперсионного анализа для регрессионных задач.

Источник. Под заголовком «Источник» перечисляются все источники рассеяния в данной задаче.

Степени свободы. Столбец показывает число степеней свободы для каждого из источников рассеяния.

SS — сумма квадратов. В столбце с таким заголовком указываются суммы квадратов для каждого источника рассеяния.

MS — средний квадрат. Столбец «средний квадрат» получается при делении каждой суммы квадратов на соответствующее ей число степеней свободы.

Общий *F*. *F*-статистика служит для определения статистической значимости регрессионной модели, рассматриваемой на каждом этапе. Рассчитывается следующим образом:

$$F = \frac{\text{Средний квадрат, обусловленный регрессией}}{\text{Средний квадрат, обусловленный остатком}}.$$

β-коэффициенты и доверительные пределы.

Номер переменной. Указывается номер каждой независимой переменной в регрессионной модели.

Среднее. Среднее значение всех наблюдений для независимой переменной.

Натуральный *B*-коэффициент. Если при вводе данных в вычислительную машину производилось любое кодирование, программа декодирует *b*-коэффициент в первоначальные единицы и напечатает его.

Пределы (верхний/нижний). Имеются в виду 95 %-ные доверительные пределы истинного коэффициента регрессии, β. Расчетные формулы для этих пределов приведены в гл. 1 и 4.

Стандартная ошибка. Стандартная ошибка коэффициента *b*. Формула для ее вычисления дается в гл. 1 и 4.

Частный *F*-критерий. *F*-критерий для каждой переменной в предположении, что она была последней переменной, включенной в регрессию. Обсуждение этой статистики дается в гл. 5 и 6.

Свободный член в предсказывающем уравнении. МНК-оценка коэффициента β_0.

Квадраты частных коэффициентов корреляции переменных, не введенных в уравнение регрессии

Частный коэффициент корреляции между каждой переменной, не введенной в регрессию, и откликом вычисляется и возводится в квадрат. Обсуждение этого статистического показателя дано в гл. 5 и 6.

Анализ остатков

Под этим заголовком записываются результаты подбора модели для «сглаживания» всех данных.

Наблюдаемый *Y*. Наблюдение отклика *Y* для каждой данной точки, показанной в матрице исходных данных.

Предсказываемый *Y*. Значение предсказанного *Y*, или $\hat{Y}$, полученное при использовании подобранной модели.

Остаток. Разность между наблюдением и предсказанным значениями *Y* или $Y-\hat{Y}$.

Нормальное отклонение. Разность $Y-\hat{Y}$, деленная на стандартное отклонение остатков *s*, или $\dfrac{Y-\bar{Y}}{s}$.

Исходные и [нли] преобразованные данные

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	10,980	5,2000	0,61000	7,4000	31,000	20,000	22,000	35,300	54,800	4,000
2	11,130	5,1200	0,64000	8,0000	29,000	20,000	25,000	29,700	64,000	5,000
3	12,510	6,1900	0,78000	7,4000	31,000	23,000	17,000	30,800	54,800	4,000
4	8,400	3,8900	0,49000	7,5000	30,000	20,000	22,000	58,800	56,300	4,000
5	9,270	6,2800	0,84000	5,5000	31,000	21,000	0,000	61,400	30,300	5,000
6	8,730	5,7600	0,74000	8,9000	30,000	22,000	0,000	71,300	79,200	4,000
7	6,360	3,4500	0,42000	4,1000	31,000	11,000	0,000	74,400	16,800	2,000
8	8,500	6,5700	0,87000	4,1000	31,000	23,000	0,000	76,700	16,800	5,000
9	7,820	5,6900	0,75000	4,1000	30,000	21,000	0,000	70,700	16,800	4,000
10	9,140	6,1400	0,76000	4,5000	31,000	20,000	0,000	57,500	20,300	5,000
11	8,240	4,8400	0,65000	10,3000	30,000	20,000	11,000	46,400	106,100	4,000
12	12,190	4,8800	0,62000	6,9000	31,000	21,000	12,000	28,900	47,600	4,000
13	11,880	6,0300	0,79000	6,6000	31,000	21,000	25,000	28,100	43,600	5,000
14	9,570	4,5500	0,60000	7,3000	28,000	19,000	18,000	39,100	53,300	5,000
15	10,940	5,7100	0,70000	8,1000	31,000	23,000	5,000	46,800	95,600	4,000
16	9,580	5,6700	0,74000	8,4000	30,000	20,000	7,000	48,500	70,600	4,000
17	10,090	6,7200	0,85000	6,1000	31,000	22,000	0,000	59,300	37,200	6,000
18	8,110	4,9500	0,67000	4,9000	30,000	22,000	0,000	70,000	24,000	4,000
19	6,830	4,6200	0,45000	4,6000	31,000	11,000	0,000	70,000	21,200	3,000
20	8,800	6,6000	0,95000	3,7000	31,000	23,000	0,000	74,500	13,700	4,000
21	7,680	5,0100	0,64000	4,7000	30,000	20,000	0,000	72,100	22,100	4,000
22	8,470	5,6800	0,75000	5,3000	31,000	21,000	1,000	58,100	28,100	6,000
23	8,860	5,2800	0,70000	6,2000	30,000	20,000	14,000	44,600	38,400	4,000
24	10,360	5,3600	0,67000	6,8000	31,000	20,000	22,000	33,400	46,200	4,000
25	11,080	5,8700	0,70000	7,5000	31,000	22,000	22,000	28,600	56,300	5,000
					Средние преобразованных переменных					
1	9,424	5,4424	0,69520	6,3560	30,480	20,240	9,160	52,600	43,364	4,3200
					Стандартные отклонения преобразованных переменных					
1	1,630	0,8169	0,12586	1,7540	0,770	3,017	10,282	17,265	23,198	0,8524

Корреляционная матрица

1	1,00000	0,38318	0,30555	0,47431	0,13674	0,53612	0,64065	—0,84524	0,39454	0,38212
2	0,38318	1,00000	0,94364	—0,12609	0,38213	0,68509	—0,19112	—0,00188	—0,13135	0,61630
3	0,30555	0,94364	1,00000	—0,14367	0,24823	0,76446	—0,22636	—0,06774	—0,13419	0,60130
4	0,47431	—0,12609	—0,14367	1,00000	—0,31677	0,23114	0,55810	—0,61634	0,98996	0,07390
5	0,13674	0,38213	0,24823	—0,31677	1,00000	0,02008	—0,20475	0,07738	—0,32100	—0,05330
6	0,53612	0,68509	0,76446	0,23114	0,02008	1,00000	0,11688	—0,20976	0,21248	0,60059
7	0,64065	—0,19112	—0,22636	0,55810	—0,20475	0,11688	1,00000	—0,85761	0,49151	0,11751
8	— 0,84524	—0,00188	0,06774	—0,61634	0,07738	—0,20976	—0,85761	1,00000	—0,54142	—0,23695
9	0,39454	—0,13135	—0,13419	0,98996	—0,32100	0,21248	0,49151	—0,54142	1,00000	0,02842
10	—0,38212	0,61630	0,60130	0,07390	—0,05330	0,60059	0,11751	—0,23695	0,02842	1,00000

Пример регрессии — модель выпарного аппарата

Распечатки, демонстрирующие процесс построения трехфакторной регрессии

Исходные данные:

Столбец 1. Вектор отклика— колнчество пара, используемого ежемесячно, в фунтах.

Столбец 2. Количество жирной кислоты в хранилище в месяц в фунтах.

Столбец 3. Количество произведенного глицерина-сырца в фунтах.

Столбец 4. Средняя скорость ветра в милях в час.

Столбец 5. Число календарных дней в месяце.

Столбец 6. Чнсло рабочнх дней в месяце.

Столбец 7. Число дней в месяце, когда температура опускалась ниже 32° F.

Столбец 8. Средняя атмосферная температура в градусах по Фаренгейту.

Столбец 9. (Средняя скорость ветра)^2.

Столбец 10. Число запусков.

Простая линейная регрессия

$$\hat{X}_1 = f(X_8).$$

Это соответствует зависимости $\hat{Y} = f(X_8)$, или количеству пара как линейной функции от средней атмосферной температуры в °F.

Информация для управления

Число наблюдений	25
Номер отклика	1
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	2, 3, 4, 5, 6, 7, 9, 10
Включаемая переменная	8
Последовательный F-критерий	57,5427930
Доля объясненной вариации R^2в %	71,4437600
Стандартное отклонение остатков	0,8901244
Средний отклик	9,4240000
Стандартное отклонение в % от среднего отклика	9,445
Степени свободы	23
Значение определителя	0,9999998

Дисперсионный анализ (ANOVA)

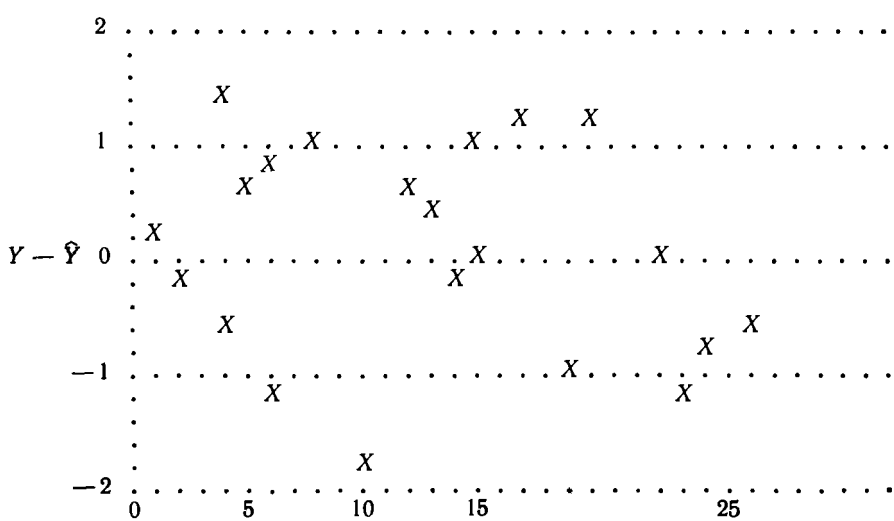
Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	24	63,8158000		
Регрессия	1	45,5924060	45,5924060	57,5428050
Остаток	23	18,2233950	0,7923215	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
8	52,6000000	—0,0798287	—0,0580554 —0,1016020	0,0105236	57,5427970

Свободный член в предсказывающем уравнении равен 13,622 9890.

Анализ остатков для $\hat{X}_1$ или $\hat{Y} = f(X_8)$



Анализ остатков для $\hat{X}_1$ или $\hat{Y} = f(X_8)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	10,9800000	10,8050370	0,1749630	0,1965602
2	11,1300000	11,2520770	—0,1220770	—0,1371460
3	12,5100000	11,1642660	1,3457340	1,5118492
4	8,4000000	8,9290620	—0,5920620	—0,5943686
5	9,2700000	8,7215080	0,5484920	0,6161970
6	8,7300000	7,9312040	0,7987960	0,8973981
7	6,3600000	7,6837350	—1,3237350	—1,4871347
8	8,5000000	7,5001290	0,9998710	1,1232935
9	7,8200000	7,9791010	—0,1591010	—0,1787402
10	9,1400000	9,0328400	0,1071600	0,1203877
11	8,2400000	9,9189380	—1,6789380	—1,8861834
12	12,1900000	11,3159400	0,8740600	0,9819526
13	11,8800000	11,3798030	0,5001970	0,5619405
14	9,5700000	10,5016880	—0,9316880	—1,0466941
15	10,9400000	9,8870070	1,0529930	1,1829727
16	9,5800000	9,7512980	—0,1712980	—0,1924428
17	10,0900000	8,8891480	1,2008520	1,3490832
18	8,1100000	8,0349810	0,0750190	0,0842792
19	6,8300000	8,0349810	—1,2049810	—1,3537219
20	8,8800000	7,6757520	1,2042480	1,3528984
21	7,6800000	7,8673410	—0,1873410	—0,2104661
22	8,4700000	8,9849420	—0,5149420	—0,5785056
23	8,8600000	10,0626300	—1,2026300	—1,3510807
24	10,3600000	10,9567110	—0,5967110	—0,6703681
25	11,0800000	11,3398890	—0,2598890	—0,2919693

Простая линейная регрессия $\hat{X}_6 = f(X_8)$.

или количество рабочих дней в месяц как линейная функция средней атмосферной температуры в °F.

Информация для управления

Число наблюдений	25
Номер отклика	6
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	1, 2, 3, 4, 5, 7, 9, 10
Включаемая переменная	8
Последовательный F-критерий	1,0585742
Доля объясненной вариации R^2 в %	4,4000000
Стандартное отклонение остатков	3,0140493
Средний отклик	20,2400000
Стандартное отклонение в % от среднего отклика	14,892
Степени свободы	23
Значение определителя	0,9999998

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	24	218,5599900		
Регрессия	1	9,6166395	9,6166395	1,0585773
Остаток	23	208,9433500	9,0844934	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный *B*-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный *F*-критерий
8	52,6000000	—0,0366626	0,0370639 —0,1103892	0,0356339	1,0585740

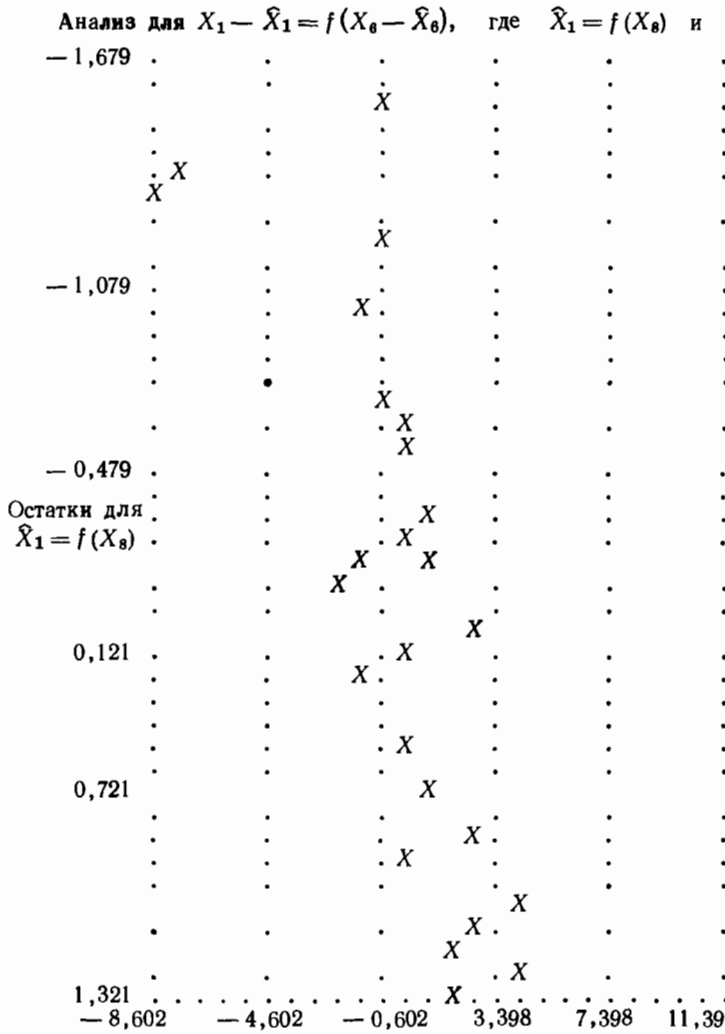
Свободный член в предсказывающем уравнении равен 22,1684550.

Анализ остатков для $\hat{X}_6 = f(X_8)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	20,0000000	20,8742640	—0,8742640	—0,2900629
2	20,0000000	21,0795750	—1,0795750	—0,3581309
3	23,0000000	21,0392460	1,9607540	0,6505381
4	20,0000000	20,0126920	—0,0126920	—0,0042109
5	21,0000000	19,9173690	1,0826310	0,3591949
6	22,0000000	19,5544090	2,4455910	0,8113971
7	11,0000000	19,4407550	—8,4407550	—2,8004701
8	23,0000000	19,3564310	3,6435690	1,2088617
9	21,0000000	19,5764060	1,4235940	0,4723194
10	20,0000000	20,0603530	—0,0603530	—0,0200239
11	20,0000000	20,4673090	—0,4673090	—0,1550436
12	21,0000000	21,1089050	—0,1089050	—0,0361325
13	21,0000000	21,1382350	—0,1382350	—0,0458635
14	19,0000000	20,7349460	—1,7349460	—0,5756196
15	23,0000000	20,4526440	2,5473560	0,8451607
16	20,0000000	20,3903170	—0,3903170	—0,1294992
17	22,0000000	19,9943610	2,0056390	0,6654301
18	22,0000000	19,6020700	2,3979300	0,7955842
19	11,0000000	19,6020700	—8,6020700	—2,8539911
20	23,0000000	19,4370880	3,5629120	1,1821014
21	20,0000000	19,5250790	0,4749210	0,1575691
22	21,0000000	20,0383560	0,9616440	0,3190538
23	20,0000000	20,5333010	—0,5333010	—0,1769384
24	20,0000000	20,9439230	—0,9439230	—0,3131744
25	22,0000000	21,1199040	0,8800960	0,2919979

Анализ остатков

Остатки $X_1 - \hat{X}_1$, где $\hat{X}_1 = f(X_8)$, вычерчены в зависимости от $X_6 - \hat{X}_6$, где $\hat{X}_6 = f(X_8)$. Затем строится регрессионная зависимость $X_1 - \hat{X}_1$ от $X_6 - \hat{X}_6$. Наконец, вычисляются остатки.



Информация для управления

Число наблюдений	25
Номер отклика	$X_1 - \hat{X}_1$
Уровень риска для доверительного интервала β	5 %
Включаемая переменная	$X_6 - \hat{X}_6$
Последовательный F-критерий	20,5300700
Доля объясненной вариации R^2 в %	47,1629600
Стандартное отклонение остатков	0,6470233
Средний отклик	0,0000006
Стандартное отклонение в % от среднего отклика	999,999
Степени свободы	23
Значение определителя	1,0000001

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	24	18,2233910		
Регрессия	1	8,5948906	8,5946906	20,5300660
Остаток	23	9,6287007	0,4186392	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
2	—0,00000044	0,2028154	0,2954271	0,0447616	20,5300700

Свободный член в предсказывающем уравнении равен 0,0000006892387.

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	0,1749640	—0,1773135	0,3522775	0,5444587
2	—0,1220760	—0,2189537	0,0968777	0,1497283
3	1,3457350	0,3976717	0,9480633	1,4652691
4	—0,5290620	—0,0025734	—0,5264886	—0,8137088
5	0,5484930	0,2195749	0,3289181	0,5083559
6	0,7987970	0,4960041	0,3027929	0,4679783
7	—0,3237340	—1,7119142	0,3881802	0,5999478
8	0,9998720	0,7389725	0,2608995	0,4032305
9	—0,1591000	0,2887272	—0,4478272	—0,6921346
10	0,1071610	—0,0122398	0,1194008	0,1845387
11	—1,6789370	—0,0947768	—1,5841603	—2,4483821
12	0,8740610	—0,0220869	0,8961479	1,3850318
13	0,5001980	—0,0280355	0,5282335	0,8164057
14	—0,9316870	—0,3518730	—0,5798140	—0,8961253
15	1,0529940	0,5166436	0,5363504	0,8289507
16	—0,1712970	—0,0791616	—0,0921354	—0,1423989
17	1,2008530	0,4067751	0,7940780	1,2272788
18	0,0750200	0,4863377	—0,4113177	—0,6357078
19	—1,2049800	—1,7446314	0,5396514	0,8340525
20	1,2042490	0,7226138	0,4816353	0,7443863
21	—0,1873400	0,0963220	—0,2836620	—0,4384107
22	—0,5149420	0,1950369	—0,7099789	—1,0973002
23	—1,2026290	—0,1081612	—1,0944679	—1,6915432
24	—0,5967100	—0,1914414	—0,4052686	—0,6263586
25	—0,2598880	0,1784977	—0,4383857	—0,6775423

Двумерная регрессия $\hat{X}_1$, или $\hat{Y} = f(X_6, X_8)$.

Вычисляются остатки и вычерчиваются в зависимости от номера наблюдения.

Информация для управления

Число наблюдений	25
Номер отклика	1
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	2, 3, 4, 5, 7, 9, 10

Включаемые переменные	8, 6
Последовательный F-критерий	19,6374510
Доля объясненной вариации R^2 в %	84,9117300
Стандартное отклонение остатков	0,6615651
Средний отклик	9,424 0000
Стандартное отклонение в % от среднего отклика	7,020
Степени свободы	22
Значение определителя	0,9559998

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	24	63,8158000		
Регрессия	2	54,1870990	27,0935490	61,9042930
Остаток	22	9,6287033	0,4376683	

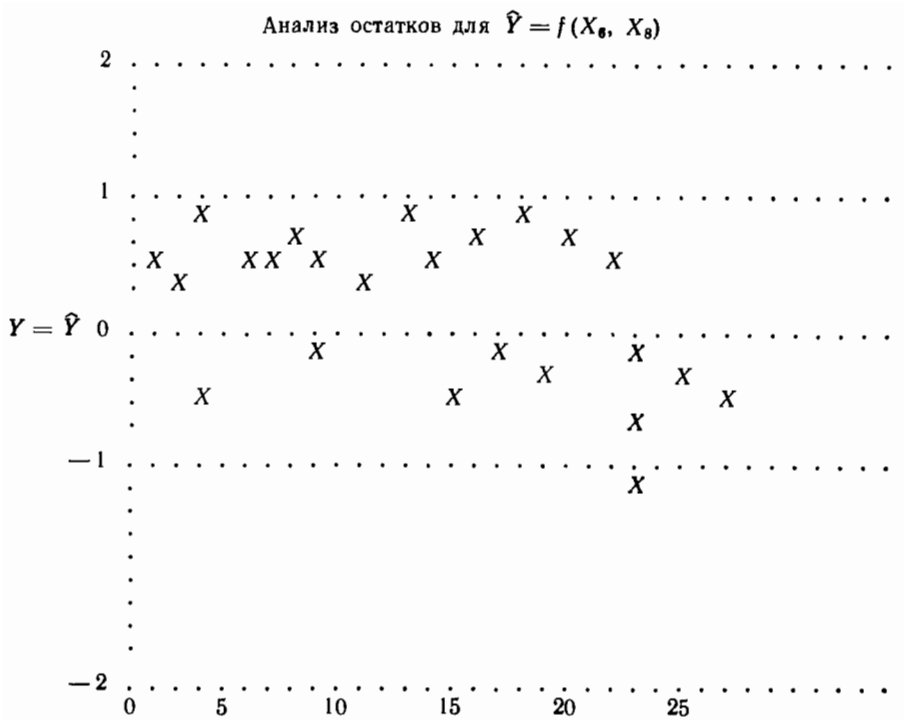
B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
8	52,6000000	—0,0723929	—0,0558022 —0,0889837	0,0079994	81,8992020
6	20,2400000	0,2028154	0,2977374 0,1078933	0,0457676	19,6374520

Свободный член в предсказывающем уравнении равен 9,1268861.

Анализ остатков для $\hat{Y} = f(X_6, X_8)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	10,9800000	10,6277220	0,3522780	0,5324918
2	11,1300000	11,0331220	0,0968780	0,1464376
3	12,5100000	11,5619360	0,9480640	1,4330623
4	8,4000000	8,9264882	—0,5264882	—0,7958223
5	9,2700000	8,9410818	0,3289182	0,4971819
6	8,7300000	8,4272071	0,3027929	0,4576918
7	6,3600000	5,9718199	0,3881801	0,5867603
8	8,5000000	8,2391005	0,2608995	0,3943671
9	7,8200000	8,2678274	—0,4478274	—0,6769212
10	9,1400000	9,0205990	0,1194010	0,1804826
11	8,2400000	9,8241607	—1,5841607	—2,3945652
12	12,1900000	11,2938520	0,8961480	1,3545878
13	11,8800000	11,3517660	0,5282340	0,7984612
14	9,5700000	10,1498130	—0,5798130	—0,8764263
15	10,9400000	10,4036490	0,5363510	0,8107305
16	9,5800000	9,6721355	—0,0921355	—0,1392690
17	10,0900000	9,2959224	0,7940780	1,2003022
18	8,1100000	8,5213179	—0,4113179	—0,6217346
19	6,8300000	6,2903489	0,5396511	0,8157189
20	8,8800000	8,3983649	0,4816351	0,7280238
21	7,6800000	7,9636620	—0,2836620	—0,4287742
22	8,4700000	9,1799785	—0,7099785	—1,0731801
23	8,8600000	9,9544680	—1,0944680	—1,6543618
24	10,3600000	10,7652690	—0,4052690	—0,6125913
25	11,0800000	11,5183850	—0,4383850	—0,6626484



ПРИЛОЖЕНИЕ Б

Эксперимент из книги: H a l d A. Statistical Theory with Engineering Applications, p. 647.

Кодирование данных

X_1 — колнчество трикальций-алюмината $3CaO \cdot Al_2O_3$;
X_2 — количество трикальций-силиката $3CaO \cdot SiO_2$;
X_3 — количество тетракальций-алюминиум-феррита $4CaO \cdot Al_2O_3 \cdot Fe_2O_3$;
X_4 — количество дикальций-силиката $2CaO \cdot SiO_2$
Отклик $Y = X_5$ — выделившееся тепло в калориях на грамм цемента; X_1, X_2, X_3 и X_4 измеряются в процентах от веса клинкера, из которого производится цемент.
Эти данные были впервые опубликованы в статье: W o o d s H., S t e i n o u r H. H., S t a r k e H. R. Effect of Composition of Portland on Heat Evolved during Hardening. Industrial and Engineering Chemistry, **24**, 1932, p. 1207—14, Table 1.

Исходные и /или преобразованные данные

	X_1	X_2	X_3	X_4	X_5
1	7,00000000	26,000000000	6,00000000	60,00000000	78,50000000
2	1,00000000	29,000000000	15,00000000	52,00000000	74,30000000
3	11,00000000	56,000000000	8,00000000	20,00000000	104,30000000
4	11,00000000	31,000000000	8,00000000	47,00000000	87,60000000
5	7,00000000	52,000000000	6,00000000	33,00000000	95,90000000
6	11,00000000	55,000000000	9,00000000	22,00000000	109,20000000
7	3,00000000	71,000000000	17,00000000	6,00000000	102,70000000
8	1,00000000	31,000000000	22,00000000	44,00000000	72,50000000
9	2,00000000	54,000000000	18,00000000	22,00000000	93,10000000
10	21,00000000	47,000000000	4,00000000	26,00000000	115,90000000
11	1,00000000	40,000000000	23,00000000	34,00000000	83,80000000
12	11,00000000	66,000000000	9,00000000	12,00000000	113,30000000
13	10,00000000	68,000000000	8,00000000	12,00000000	109,40000000

Средние преобразованных переменных

1	7,46153830	48,15384500	11,76923000	29,99999900	95,42307500

Стаидартиые отклонения преобразованных переменных

1	5,88239440	15,56087900	6,40512590	16,73817800	15,04372400

Корреляционная матрица

1	0,99999991	0,22857948	—0,82413372	—0,24544512	0,73071745
2	0,22857948	1,00000010	—0,13924238	—0,97295516	0,81625268
3	—0,82413372	—0,13924238	0,99999991	0,02953700	—0,53467065
4	—0,24544512	—0,97295516	0,02953700	1,00000010	—0,82130513
5	0,73077145	0,81625268	—0,53467065	—0,82130513	0,99999999

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	2, 3, 4
Включаемая переменная	1
Последовательный F-критерий	12,6025160
Доля объясненной вариации R^2 в %	53,3948000
Стандартное отклонение остатков	10,7267170
Средний отклик	95,4230750
Стандартное отклонение в % от среднего отклика	11,241
Степени свободы	11
Значение определителя	0,9999999

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	1	1450,0764000	1450,0764000	12,6025160
Остаток	11	1265,6870000	115,0624500	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
1	7,4615383	1,8687477	3,0273705 0,71011249	0,5264075	12,6025150

Свободный член в предсказывающем уравнении равен 81,4793430.

Квадраты частных коэффициентов корреляции для переменных, не входящих в регрессию

Переменные	Квадраты коэффициентов
2	0,95425
3	0,03051
4	0,94093
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_1)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	94,5605760	—16,0605760	—1,4972499
2	74,3000000	83,3480900	—9,0480900	—0,8435097
3	104,3000000	102,0355600	2,2644400	0,2111028
4	87,6000000	102,0355600	—14,4355600	—1,3475575
5	95,9000000	94,5605760	1,3394240	0,1248680
6	109,2000000	102,0355600	7,1644400	0,6679061
7	102,7000000	87,0855860	15,6144200	1,4556569
8	72,5000000	83,3480900	—10,8480900	—1,0113150
9	93,1000000	85,2168380	7,8831620	0,7349091
10	115,9000000	120,7230400	—4,8230400	—0,4496287
11	83,8000000	83,3480900	0,4519100	0,0421294
12	113,3000000	102,0355600	11,2644400	1,0501293
13	109,4000000	100,1668200	9,2331800	0,8607648

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	1, 3, 4
Включаемая переменная	2
Последовательный F-критерий	21,9606150
Доля объясненной вариации R^2 в %	66,6268400
Стандартное отклонение остатков	9,0771249
Средний отклик	95,4230750
Стандартное отклонение в % от среднего отклика	9,513
Степени свободы	11
Значение определителя	1,0000000

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	1	1809,4274000	1809,4274000	21,9606160
Остаток	11	906,3661700	82,3941970	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный коэффициент B	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
2	48,1538450	0,7891250	1,1597575 0,4184924	0,1683928	21,9606140

Свободный член в предсказывающем уравнении равен 57,4236730.

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
1	0,93611
3	0,54162
4	0,04133
5	1,00000

Анализ остатков для $\widehat{X}_5 = f(X_2)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	77,9409220	0,5590780	0,0615920
2	74,3000000	80,3082970	—6,0082970	—0,6619163
3	104,3000000	101,6146700	2,6853300	0,2958349
4	87,6000000	81,8865470	5,7134530	0,6294342
5	95,9000000	98,4581720	—2,5581720	—0,2818262
6	109,2000000	100,8255400	8,3744600	0,9225895
7	102,7000000	113,4515400	—10,7515400	—1,1844653
8	72,5000000	81,8865470	—9,3865470	—1,0340881
9	93,1000000	100,0364200	—6,9364200	—0,7641649
10	115,9000000	94,5125470	21,3874600	2,3561932
11	83,8000000	88,9886720	—5,1886720	—0,5716206
12	113,3000000	109,5059200	3,7940800	0,4179826
13	109,4000000	111,0841700	—1,6841700	—0,1855400

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	1, 2, 4
Включаемая переменная	3
Последовательный F-критерий	4,4034159
Доля объясненной вариации R^2 в %	28,5872800

Стандартное отклонение остатков	13,2781460
Средний отклик	95,4230750
Стандартное отклонение в % от среднего отклика	13,915
Степени свободы	11
Значение определителя	0,9999998

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	1	776,3629100	776,3629100	4,4034177
Остаток	11	1939,4008000	176,3091600	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
3	11,7692300	—1,2557812	0,0613807 —2,5729431	0,5984380	4,4034159

Свободный член в предсказывающем уравнении равен 110,2026500.

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
1	0,36729
2	0,78579
4	0,90939
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_3)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	102,6679700	—24,1679700	—1,8201313
2	74,3000000	91,3659400	—17,0659400	—1,2852652
3	104,3000000	100,1564100	4,1435900	0,3120609
4	87,6000000	100,1564100	—12,5564100	—0,9456448
5	95,9000000	102,6679700	—6,7679700	—0,5097075
6	109,2000000	98,9006200	10,2993800	0,7756640
7	102,7000000	88,8543700	13,8456300	1,0427381
8	72,5000000	82,5754700	—10,0754700	—0,7588010
9	93,1000000	87,5985900	5,5014100	0,4143206
10	115,9000000	105,1795300	10,7204700	0,8073770
11	83,8000000	81,3196900	2,4803100	0,1867964
12	113,3000000	98,9006200	14,3993800	1,0844420
13	109,4000000	100,1564100	9,2435900	0,6961507

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключениых переменных	1, 2, 3
Включаемая переменная	4
Последовательный F-критерий	22,7985280
Доля объясненной вариации R^2 в %	67,4542100
Стандартное отклонение остатков	8,9639014
Средний отклик	95,4230750
Стандартное отклонение в % от среднего отклика	9,394
Степени свободы	11
Значение определителя	0,9999999

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	1	1831,8968000	1831,8968000	22,7985300
Остаток	11	883,8668200	80,3515290	

***B*-коэффициенты и доверительные пределы**

Номер переменной	Среднее	Натуральный *B*-коэффициент	Суммы квадратов	Средние квадраты	Общий F
4	29,9999990	—0,7381619	—0,3978962 —1,0784277	0,1545960	22,7985270

Свободный член в предсказывающем уравнении равен 117,5679300.

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
1	0,91541
2	0,01696
3	0,80117
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_4)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	73,2782200	5,2217800	0,5825343
2	74,3000000	79,1823100	—4,8835100	—0,5447974
3	104,3000000	102,8047000	1,4953000	0,1668135
4	87,6000000	82,8743200	4,7256800	0,5271901
5	95,9000000	93,2085900	2,6914100	0,3002498
6	109,2000000	101,3283700	7,8716300	0,8781478
7	102,7000000	113,1389600	—10,4389600	—1,1645554
8	72,5000000	85,0888100	—12,5888100	—1,4043896
9	93,1000000	101,3283700	—8,2283700	—0,9179452
10	115,9000000	98,3757200	17,5242800	1,9549835
11	83,8000000	92,4704300	—8,6704300	—0,9672608
12	113,3000000	108,7099900	4,5900100	0,5120549
13	109,4000000	108,7099900	0,6900100	0,0769765

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	3,4
Включаемая перемеиная	1
Последовательный F-критерий	146,5229400
Доля объясненной вариации R^2 в %	97,8678500
Стаидартное отклонение остатков	2,4063327
Средний отклик	25,4230750
Стаидартное отклонение в % от среднего отклика	2,522
Степени свободы	10
Значение определителя	0,9477514

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	2	2657,8593000	1328,9296000	229,5042100
Остаток	10	57,9043680	5,7904368	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний нижний	Стандартная ошибка	Частный F-критерий
2	48,1538450	0,6622507	0,7644149 0,5600865	0,0458547	208,5823200
1	7,4615383	1,4683057	1,7385638 1,1980476	0,1213008	146,5229400

Свободный член в предсказывающем уравнении равен 52,5773400.

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
3	0,16914
4	0,17152
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_2, X_1)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	80,0739960	—1,5739960	—0,6541057
2	74,3000000	73,2509140	1,0490860	0,4359688
3	104,3000000	105,8147300	—1,5147300	—0,6294765
4	87,6000000	89,2584720	—1,6584720	—0,6892114
5	95,9000000	97,2925130	—1,3925130	—0,5786868
6	109,2000000	105,1524800	4,0475200	1,6820284
7	102,7000000	104,0020500	—1,3020500	—0,5410931
8	72,5000000	74,5754150	—2,0754150	—0,8624805
9	93,1000000	91,2754870	1,8245130	0,7582131
10	115,9000000	114,5375100	1,3624600	0,5661977
11	83,8000000	80,5356710	3,2643490	1,3565576
12	113,3000000	112,4372400	0,8627600	0,3585373
13	109,4000000	112,2934400	—2,8934400	—1,2024272

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	2,4
Включаемая переменная	3
Последовательный F-критерий	0,3146887
Доля объясненной вариации R^2 в %	54,8166700
Стандартное отклонение остатков	11,0773310
Средний отклик	95,4230750
Стандартное отклонение в % от среднего отклика	11,609
Степени свободы	10
Значение определителя	0,3208036

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	2	1488,6911000	744,3455500	6,0660270
Остаток	10	1227,0726000	122,7072600	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
1	7,4615383	2,3124675	4,4508523 0,1740827	0,9597778	5,8051026
3	11,7692300	0,4944674	2,4583356 —1,4694008	0,8814489	0,3146887

Свободный член в предсказывающем уравнении равен 72,3490110.

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
2	0,96079
4	0,95857
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_1, X_3)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	91,5030870	—13,0030870	—1,1738465
2	74,3000000	82,0784890	—7,7784890	—0,7021988
3	104,3000000	101,7418900	2,5581100	0,2309320
4	87,6000000	101,7418900	—14,1418900	—1,2766513
5	95,9000000	91,5030870	4,3969130	0,3969289
6	109,2000000	102,2363500	6,9636500	0,6286397
7	102,7000000	87,6923590	15,0076500	1,3548073
8	72,5000000	85,5397610	—13,0397610	—1,1771572
9	93,1000000	85,8743590	7,2256410	0,6522908
10	115,9000000	122,8886900	—6,9886900	—0,6309002
11	83,8000000	86,0342280	—2,2342280	—0,2016937
12	113,3000000	102,2363500	11,0636500	0,9987650
13	109,4000000	99,4294250	9,9705800	0,9000887

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	2,3
Включаемая переменная	4
Последовательный F-критерий	159,2951900
Доля объясненной вариации R^2 в %	97,2471100
Стандартное отклонение остатков	2,7342662
Средний отклик	95,4230750
Стандартное отклонение в % от среднего отклика	2,865
Степени свободы	10
Значение определителя	0,9397566

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	2	2641,0015000	1320,5007000	176,6269800
Остаток	10	74,7621170	7,4762117	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
1	7,4615383	1,4399582	1,7483504 1,1315660	0,1384166	108,2238900
4	29,9999990	—0,6139537	—0,5055737 —0,7223338	0,0486446	159,2952400

Свободный член в предсказывающем уравнении равен 103,0973800.

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
2	0,35833
3	0,32003
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_1, X_4)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	76,3398700	2,1601300	0,7900218
2	74,3000000	72,6117500	1,6882500	0,6174417
3	104,3000000	106,6578400	—2,3578400	—0,8623301
4	87,6000000	90,0811000	—2,4811000	—0,9074098
5	95,9000000	92,9166200	2,9833800	1,0911081
6	109,2000000	105,4299300	3,7700700	1,3788233
7	102,7000000	103,7335300	—1,0335300	—0,3779917
8	72,5000000	77,8233800	—5,0233800	—1,8371949
9	93,1000000	92,4703200	0,6296800	0,2302921
10	115,9000000	117,3737000	—1,4737000	—0,5389746
11	83,8000000	83,6629200	0,1370800	0,0501341
12	113,3000000	111,5694700	1,7305300	0,6329047
13	109,4000000	110,1295100	—0,7295100	—0,2668028

Информации для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	1,4
Включаемая переменная	3
Последовательный F-критерий	11,8161580
Доля объясненной вариации R^2 в %	84,7025600
Стандартное отклонение остатков	6,4454832
Средний отклик	95,4230750
Стандартное отклонение в % от среднего отклика	6,755
Степени свободы	10
Значение определителя	0,9806113

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	2	2300,3212000	1150,1606000	27,6851910
Остаток	10	415,4425300	41,5442530	

***B*-коэффициенты и доверительные пределы**

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
2	48,1538450	0,7313298	1,0003577 0,4623019	0,1207486	36,6827680
3	11,7692300	—1,0083860	—0,3547984 —1,6619736	0,2933517	11,8161580

Свободный член в предсказывающем уравнении равен 72,0746600.

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
1	0,88419
4	0,82232
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_2, X_3)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	85,0389180	—6,5389180	—1,0144961
2	74,3000000	78,1574330	—3,8574330	—0,5984707
3	104,3000000	104,9620400	—0,6620400	—0,1027138
4	87,6000000	86,6787950	0,9212050	0,1429226
5	95,9000000	104,0534900	—8,1534900	—1,2649928
6	109,2000000	103,2223200	5,9776800	0,9274215
7	102,7000000	106,8565100	—4,1565100	—0,6448717
8	72,5000000	72,5613910	—0,0613910	—0,0095247
9	93,1000000	93,4155200	—0,3155200	—0,0489521
10	115,9000000	102,4136100	13,4863900	2,0923784
11	83,8000000	78,1349730	5,6650270	0,8789142
12	113,3000000	111,2669500	2,0330500	0,3154224
13	109,4000000	113,7379000	—4,3379900	—0,6730279

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	1,3
Включенная переменная	4
Последовательный F-критерий	0,4310840
Доля объясненной вариации R^2 в %	68,0060600
Стандартное отклонение остатков	9,3213731
Средний отклик	94,4230750
Стандартное отклонение в % от среднего отклика	9,768
Степени свободы	10
Значение определителя	0,0533585

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	2	1846,8837000	923,4418500	10,6279560
Остаток	10	868,8799600	86,8879960	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный *B*-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный *F*-критерий
2	48,1538450	0,3109057	1,9787999 —1,3569887	0,7486061	0,1724847
4	29,9999990	—0,4569411	1,0936398 —2,0075220	0,6959520	0,4310840

Свободный член в предсказывающем уравнении равен 94,1600050.

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
1	0,94479
3	0,91515
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_2, X_4)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	74,8270850	3,6729150	0,3940315
2	74,3000000	79,4153310	—5,1153310	—0,5487744
3	104,3000000	102,4319000	1,8681000	0,2004104
4	87,6000000	82,3218490	5,2781510	0,5662418
5	95,9000000	95,2480440	0,6519560	0,0699421
6	109,2000000	101,2071100	7,9928900	0,8574799
7	102,7000000	113,4926600	—10,7926600	—1,1578401
8	72,5000000	83,6926720	—11,1926720	—1,2007535
9	93,1000000	100,8962000	—7,7962000	—0,8363789
10	115,9000000	96,8921020	19,0079000	2,0391738
11	83,8000000	91,0602340	—7,2602340	—0,7788803
12	113,3000000	109,1964800	4,1035200	0,4402270
13	109,4000000	109,8182900	—0,4182900	—0,0448743

Информация для управлении

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	1,2
Включаемая переменная	3
Последовательный F-критерий	40,2945330
Доля объясненной вариации R^2 в %	93,5289700
Стандартное отклонение остатков	4,1921130
Средний отклик	95,4230750
Стандартное отклонение в % от среднего отклика	4,393
Степени свободы	10
Значение определителя	0,9991272

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	2	2540,0256000	1270,0128000	72,2673510
Остаток	10	175,7381100	17,5738110	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
4	29,9999990	—0,7246003	—0,5634470 —0,8857535	0,0723309	100,3573400
3	11,7692300	—1,1998510	—0,7787179 —1,6209841	0,1890185	40,2945450

Свободный член в предсказывающем уравнении равен 131,2824000.

Квадраты частных коэффициеитов корреляции для перемеиных, не включеиных в регрессию

Переменные	Квадраты коэффициентов
1	0,94479
3	0,91515
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_4, X_3)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклоненне
1	78,5000000	80,6072800	—2,1072800	—0,5026773
2	74,3000000	75,6054300	—1,3054300	—0,3114014
3	104,3000000	107,1915900	—2,8915900	—0,6897691
4	87,6000000	87,6273800	—0,0273800	—0,0065313
5	95,9000000	100,1714900	—4,2714900	—1,0189348
6	109,2000000	104,5425400	4,6574600	1,1110053
7	102,7000000	106,5373400	—3,8373400	—0,9153713
8	72,5000000	73,0032700	—0,5032700	—0,1200516
9	93,1000000	93,7438800	—0,6438800	—0,1535932
10	115,9000000	107,6433900	8,2566100	1,9695580
11	83,8000000	79,0494200	4,7505800	1,1332184
12	113,3000000	111,7885400	1,5114600	0,3605485
13	109,4000000	112,9883900	—3,5883900	—0,8559860

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	4
Включаемая определенная	1
Последовательный F-критерий	68,7166430
Доля объясненной вариации R^2 в %	98,2284800
Стандартное отклонение остатков	2,3120568
Средний отклик	95,4230750
Стандартное отклонение в % от среднего отклика	2,423
Степени свободы	9
Значение определителя	0,3016276

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	3	2667,6532000	889,2177300	166,3455300
Остаток	9	48,1104560	5,3456062	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
3	11,7692300	0,2500169	0,6678322 —0,1677985	0,1847106	1,8321249
2	48,1538450	0,6569150	0,7569728 0,5568573	0,0442342	250,5476100
1	7,4615383	1,6958894	2,1586529 1,2331259	0,2045816	68,7166370

Свободный член в предсказывающем уравнении равен 48,1936420

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
4	0,00513
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_3, X_2, X_1)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	78,6447590	—0,1447590	—0,0626105
2	74,3000000	72,6903190	1,6096810	—0,6962117
3	104,3000000	105,6358000	—1,3358000	—0,5777540
4	87,6000000	89,2129250	—1,6129250	—0,6976148
5	95,9000000	95,7245500	0,1754500	0,0758848
6	109,2000000	105,2289000	3,9711000	1,7175616
7	102,7000000	104,1725600	—1,4725600	—0,6369048
8	72,5000000	75,7542670	—3,2542670	—1,4075203
9	93,1000000	91,5591350	1,5408650	0,6664477
10	115,9000000	115,6823900	0,2176100	0,0941197
11	83,8000000	81,9165190	1,8834810	0,8146344
12	113,3000000	112,4549600	0,8450400	0,3654928
13	109,4000000	111,8228900	—2,4228900	—1,0479370

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	3
Включаемая переменная	4
Последовательный F-критерий	1,8632545
Доля объясненной вариации R^2 в %	98,2335600
Стандартное отклонение остатков	2,3087418
Средний отклик	95,4230750
Стандартные отклонения в % от среднего отклика	2,419
Степени свободы	9
Значение определителя	0,0500394

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	3	2667,7911000	889,2637000	166,8321800
Остаток	9	47,9725980	5,3302886	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
2	48,1538450	0,4161107	0,8359611 —0,0037398	0,1856103	5,0258974
1	7,4615383	1,4519380	1,7165861 1,1872899	0,1169974	154,0080400
4	29,9999990	—0,2365395	0,1554371 —0,6285160	0,1732876	1,8632548

Свободный член в предсказывающем уравнении равен 71,6482410.

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
3	0,00227
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_2, X_1, X_4)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	78,4383160	0,0616840	0,0267176
2	74,30000000	72,8673360	1,4326640	0,6205389
3	104,3000000	106,1909600	—1,8909600	—0,8190435
4	87,6000000	89,4016340	—1,8016340	—0,7803532
5	95,9000000	95,6437590	0,2562410	0,1109873
6	109,2000000	105,3017700	3,8982300	1,6884651
7	102,7000000	104,1286700	—1,4286700	—0,6188089
8	72,5000000	75,5918720	- 3,0918720	—1,3392021
9	93,1000000	91,8182250	1,2817750	0,5551833
10	115,9000000	115,5461100	0,3538900	0,1532826
11	83,8000000	81,7022630	2,0977370	0,9086062
12	113,3000000	112,2443900	1,0556100	0,4572231
13	109,4000000	111,6246700	—2,2246700	—0,9635854

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	2
Включаемая переменная	3
Последовательный F-критерий	4,2358482
Доля объясненной варнации R^2 в %	98,1281200
Стандартное отклонение остатков	2,3766478
Средний отклик	95,4230750
Стандартное отклонение в % от среднего отклика	2,491
Степени свободы	9
Значение определителя	0,2716373

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	3	2664,9276000	888,3092000	157,2658800
Остаток	9	50,8360910	5,6484545	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
1	7,4615383	1,0518542	1,5578282 0,5458802	0,2236844	22,1126000
4	29,9999990	—0,6427963	—0,5420373 —0,7435552	0,0445442	208,2401700
3	11,7692300	—0,4100433	0,0406197 —0,8607064	0,19992321	4,2358519

Свободный член в предсказывающем уравнении равен 111,6844000.

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
2	0,05847
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_1, X_4, X_3)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	78,0193500	0,4806500	0,2022386
2	74,3000000	73,1602000	1,1398000	0,4795830
3	104,3000000	107,1185300	—2,8185300	—1,1859266
4	87,6000000	89,7630300	—2,1630300	—0,9101180
5	95,9000000	95,3748500	0,5251500	0,2209625
6	109,2000000	105,4228900	3,7771100	1,5899594
7	102,7000000	104,0124500	—1,3124500	—0,5522274
8	72,5000000	75,4322700	—2,9322700	—1,2337839
9	93,1000000	92,2658200	0,8341800	0,3509902
10	115,9000000	115,4204600	0,4795400	0,2017716
11	83,8000000	81,4501900	2,3498100	0,9887077
12	113,3000000	111,8508500	1,4491500	0,6097454
13	109,4000000	111,2090500	—1,8090500	—0,7611772

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Перечень исключенных переменных	1
Включаемая переменная	2
Последовательный F-критерий	12,4271010
Доля объясненной вариации R^2 в %	97,2819800
Стандартное отклонение остатков	2,8638569
Средний отклик	95,4230750
Стандартное отклонение в % от среднего отклика	3,001
Степени свободы	9
Значение определителя	0,0411008

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	3	2641,9485000	880,6495000	107,3743300
Остаток	9	73,8150840	8,2016760	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный B-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный F-критерий
4	29,9999990	—1,5570434	—1,0113254 —2,1027614	0,2412547	41,6533540
3	11,7692300	—1,4479704	—1,1153091 —1,7806317	0,1470651	96,9393040
2	48,1538450	—0,9234143	—0,3308925 —1,5159360	0,2619460	12,4270980

Свободный член в предсказывающем уравнении равен 203,6418100.

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
1	0,35157
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_4,\ X_3,\ X_2)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	77,5226200	0,9773800	0,3412810
2	74,3000000	74,1769900	0,1230100	0,0429526
3	104,3000000	109,2059800	—4,9059800	—1,7130674
4	87,6000000	90,2511700	—2,6511700	—0,9257341
5	95,9000000	95,5540200	0,3459800	0,1208091
6	109,2000000	105,5673400	3,6326600	1,2684502
7	102,7000000	104,1216500	—1,4216500	—0,4964110
8	72,5000000	74,6507200	—2,1507200	—0,7509872
9	93,1000000	93,4590200	—0,3590200	—0,1253624
10	115,9000000	113,9663400	1,9336600	0,6751944
11	83,8000000	80,4624500	3,3375500	1,1654038
12	113,3000000	110,9802200	2,3197800	0,8100195
13	109,4000000	110,5813600	—1,1813600	—0,4125066

Информация для управления

Число наблюдений	13
Номер отклика	5
Уровень риска для доверительного интервала β	5 %
Включаемая переменная	1
Последовательный F-критерий	4,3375998
Доля объясненной вариации R^2 в %	98,2375700
Стандартное отклонение остатков	2,4460044
Средний отклик	95,4230750
Стандартное отклонение в % от среднего отклика	2,563
Степени свободы	8
Значение определителя	0,0010677

ANOVA

Источник	Степени свободы	Суммы квадратов	Средние квадраты	Общий F
Общий	12	2715,7635000		
Регрессия	4	2667,9000000	666,975000	111,4795200
Остаток	8	47,8634980	5,9829372	

B-коэффициенты и доверительные пределы

Номер переменной	Среднее	Натуральный *B*-коэффициент	Пределы верхний/нижний	Стандартная ошибка	Частный *F*-критерий
4	29,9999990	—0,1440588	1,4909970 —1,7791144	0,7090441	0,0412794
3	11,7692300	0,1019111	1,8422494 —1,6384272	0,7547001	0,0182345
2	48,1538450	0,5101700	2,1792063 —1,1588665	0,7237799	0,4968402
1	7,4615383	1,5511043	3,2685233 —0,1663147	0,7447611	4,3375858

Свободный член в предсказывающем уравнении равен 62,4051530.

Квадраты частных коэффициентов корреляции для переменных, не включенных в регрессию

Переменные	Квадраты коэффициентов
5	1,00000

Анализ остатков для $\hat{X}_5 = f(X_4, X_3, X_2, X_1)$

Номер наблюдения	Наблюдаемый Y	Предсказываемый Y	Остаток	Нормальное отклонение
1	78,5000000	78,4952410	0,0047590	0,0019456
2	74,3000000	72,7887950	1,5112050	0,6178260
3	104,3000000	105,9709300	—1,6709300	—0,6831263
4	87,6000000	89,3270940	—1,7270940	—0,7060879
5	95,9000000	95,6492470	0,2507530	0,1025154
6	109,2000000	105,2745500	3,9254500	1,6048417
7	102,7000000	104,1486600	—1,4486600	—0,5922557
8	72,5000000	75,6749840	—3,1749840	—1,2980287
9	93,1000000	91,7216450	1,3783550	0,5635129
10	115,9000000	115,6184400	0,2815600	0,1151102
11	83,8000000	81,8090130	1,9909870	0,8139752
12	113,3000000	112,3270100	0,9729900	0,3977875
13	109,4000000	111,6943300	—2,2943300	—0,9379910

ПРИЛОЖЕНИЕ В

Шаговая регрессия для данных о цементе из приложения Б, построенная с помощью программы BMDP 2R

[В этом приложении демонстрируются возможности новой версии пакета BMDP от декабря 1977 г. при решении задачи шаговой регрессии. В СССР адаптирована версия этого пакета от 1975 г., см.: Математическое обеспечение ЕС ЭВМ.— Минск: Ин-т математики АН БССР, 1980, вып. 1.— 202 с. Характеристики пакета см.: Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Исследование зависимостей.— М.: Финансы и статистика, 1985.— 488 с. (особо с. 433); Дайитбегов Д. М., Калмыкова С. В., Черепанов А. И. Программное обеспечение статистической обработки данных.— М.: Финансы и статистика, 1984.— 192 с. (особо с. 141—152).

Возможности новой версии существенно расширены, приведенная распечатка демонстрирует это.

В распечатке использованы термины, перевод которых нуждается в пояснении. Для обозначения новой возможности употребляется термин «option», который здесь переводится не обычным своим значением «вариант», а с помощью кальки «опция». Под словом «paragraph» понимается «раздел программы или листинга». Термином INTLEV обозначен оператор, задающий уровень промежуточной печати. В новой версии программы он отсутствует, ибо потребность в нем отпала. Уровень печати определяется теперь в задании или в интерактивном режиме. При использовании данных из файла хранения применяются «код» (длиной не более 8 символов), идентифицирующий файл, и «метка» (label) — BMDP-файла (не менее 40 символов). Упоминаемые в распечатке нормальные графики остатков и методы оценивания и элиминирования временных треидов обсуждаются в гл. 3. В «ЗАГЛАВИИ ЗАДАЧИ» и в «НАЗВАНИИ РЕГРЕССИИ» приведен номер страницы по английскому оригиналу. В русском переводе это соответствует приложению Б.— *Примеч. пер.*]

* Звездочкой обозначены работы, переведенные на русский язык. Переводы указаны в списке дополнительной библиографии и тоже отмечены звездочкой.— *Примеч. пер.*

BMDP2R – ШАГОВАЯ РЕГРЕССИЯ ПРОГРАММА, ПЕРЕСМОТРЕННАЯ В ДЕКАБРЕ 1977

МЕДИЦИНСКИЙ ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР РУКОВОДСТВО 1977

УНИВЕРСИТЕТ ШТАТА КАЛИФОРНИЯ, ЛОС АНЖЕЛЕС

В ЭТОЙ ВЕРСИИ ПРОГРАММЫ BMDP2R

– НОВАЯ ОПЦИЯ – ВЫДАЧА НА ПЕЧАТЬ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ РЕГРЕССИОННЫХ КОЭФФИЦИЕНТОВ ЗАДАЕТСЯ СИМВОЛОМ RREG В СООТВЕТСТВУЮЩЕМ РАЗДЕЛЕ ЛИСТИНГА.

  ЕСЛИ УРАВНЕНИЕ СОДЕРЖИТ МЕНЕЕ ЧЕМ ДВЕ ВХОДНЫЕ ПЕРЕМЕННЫЕ, ТО КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ ДЛЯ КОЭФФИЦИЕНТОВ РЕГРЕССИИ НЕ ПЕЧАТАЮТСЯ.

– СТАРАЯ ОПЦИЯ "УРОВЕНЬ ПРОМЕЖУТОЧНОЙ ПЕЧАТИ" (INTLEV) ОТМЕНЯЕТСЯ.

– НОВАЯ ОПЦИЯ – ЗАДАТЬ "НОРМАЛЬНЫЙ" (NORMAL) В РАЗДЕЛЕ О ГРАФИКАХ И БУДЕТ НАПЕЧАТАН НОРМАЛЬНЫЙ ГРАФИК ОСТАТКОВ.

– НОВАЯ ОПЦИЯ – ЗАДАТЬ "ЭНОРМАЛЬНЫЙ" (DNORMAL) В РАЗДЕЛЕ О ГРАФИКАХ И БУДЕТ НАПЕЧАТАН НОРМАЛЬНЫЙ ГРАФИК ОСТАТКОВ С ЭЛИМИНИРОВАННЫМ ТРЕНДОМ.

– НОВАЯ ОПЦИЯ – ЗАДАТЬ НЕЗАВИСИМЫЕ ПЕРЕМЕННЫЕ В РАЗДЕЛЕ О РЕГРЕССИИ, УСТАНОВИВ ИХ СПИСОК (INDEP= =VARIABLE LIST).

  ТЕМ НЕ МЕНЕЕ УРОВНИ ОПЦИИ ДЛЯ ЗАДАНИЯ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ ОСТАЮТСЯ, ЕСЛИ ОТСУТСТВУЕТ УКАЗАННЫЙ ВЫШЕ СПИСОК НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ (INDEP=STATEMENT).

  ИНФОРМАЦИЯ ДЛЯ УПРАВЛЕНИЯ ПРОГРАММОЙ

```
./ЗАГЛАВИЕ ЗАДАЧИ = "ДРЕЙПЕР И СМИТ, СТР. 630, # DS36603, WFC"
/ВХОДНОЙ КОД = СМИТ.
/УСТРОЙСТВО ВВОДА=3.
/ЗАВИСИМАЯ ПЕРЕМЕННАЯ (REGR DEPENDENT) =X5.
/ПЕЧАТЬ (PRINT) КОВАРИАЦИИ.
                КОРРЕЛЯЦИИ.
                ЧАСТНЫЕ
                F – ОТНОШЕНИЯ.
                ДАННЫЕ.
/ПОСТРОИТЬ (PLOT) ГРАФИКИ ОСТАТКОВ.
                  НОРМАЛЬНЫЙ.
                  ЭНОРМАЛЬНЫЙ.
/КОНЕЦ
```

ЗАГЛАВИЕ ЗАДАЧИ ..... ДРЕЙПЕР И СМИТ, СТР. 630, # DS36603, WFC
ЧИСЛО УЧИТЫВАЕМЫХ ПЕРЕМЕННЫХ. 5
ЧИСЛО ПЕРЕМЕННЫХ, ДОБАВЛЯЕМЫХ С ПОМОЩЬЮ ПРЕОБРАЗОВАНИЙ 0
ОБЩЕЕ ЧИСЛО ПЕРЕМЕННЫХ 5
ЧИСЛО ВОЗМОЖНЫХ ВАРИАНТОВ УЧЕТА ПЕРЕМЕННЫХ 1000000
СЛУЧАИ ПЕРЕМЕННЫХ С МЕТКАМИ .....
ПЕРЕД ПРЕОБРАЗОВАНИЕМ ПРОБЕЛОВ ПРЕДЕЛЬНЫЕ И ПРОПУЩЕННЫЕ ЗНАЧЕНИЯ РАВНЫ. .....НУЛЯМ
НОМЕР УСТРОЙСТВА ВВОДА 3
ВОЗВРАТ К НАЧАЛУ ФАЙЛА ПЕРЕД СЧИТЫВАНИЕМ ДАННЫХ
ФОРМАТ ВВОДИМОГО ФАЙЛА
BMDP ФАЙЛ ..... КОД – СМИТ
СОДЕРЖАНИЕ – ДАННЫЕ
МЕТКА – СИСТЕМА СТАТИСТИЧЕСКОГО АНАЛИЗА

ПЕРЕМЕННЫЕ 1 X1 2 X2 3 X3 4 X4 5 X5

ИСПОЛЬЗУЕМЫЕ ПЕРЕМЕННЫЕ

1 X1 2 X2 3 X3 4 X4 5 X5

СВОБОДНЫЙ ЧЛЕН РЕГРЕССИИ	НЕ НУЛЬ
ВЕСА ПЕРЕМЕННЫХ	
ПЕЧАТАТЬ КОВАРИАЦИОННУЮ МАТРИЦУ	ДА
ПЕЧАТАТЬ КОРРЕЛЯЦИОННУЮ МАТРИЦУ	ДА
ПЕЧАТАТЬ ТАБЛИЦУ ДИСПЕРСИОННОГО АНАЛИЗА НА КАЖДОМ ШАГЕ	ДА
ПЕЧАТАТЬ РЕЗУЛЬТАТ КАЖДОГО ШАГА	ДА
ПЕЧАТАТЬ СВОДНУЮ ТАБЛИЦУ РЕГРЕССИОННЫХ КОЭФФИЦИЕНТОВ	ДА
ПЕЧАТАТЬ СВОДНУЮ ТАБЛИЦУ ЧАСТНЫХ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ	ДА
ПЕЧАТАТЬ СВОДНУЮ ТАБЛИЦУ F – ОТНОШЕНИЙ	ДА
ПЕЧАТАТЬ ОБЩУЮ СВОДНУЮ ТАБЛИЦУ	ДА
ПЕЧАТАТЬ ОСТАТКИ И ИСХОДНЫЕ ДАННЫЕ	ДА
ПЕЧАТАТЬ КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ РЕГРЕССИОННЫХ КОЭФФИЦИЕНТОВ	НЕТ
ПЕЧАТАТЬ ГРАФИК НОРМАЛЬНЫХ ОСТАТКОВ	ДА
ПЕЧАТАТЬ ГРАФИК ЭНОРМАЛЬНЫХ ОСТАТКОВ	ДА
ЧИСЛО ОПЫТОВ В МАССИВЕ	13

N	ПЕРЕМЕННАЯ ИМЯ	СРЕДНЕЕ	СТАНДАРТНОЕ ОТКЛОНЕНИЕ	КОЭФФИЦИЕНТ ВАРИАЦИИ	АССИМЕТРИЯ	ЭКСЦЕСС	НАИМЕНЬШЕЕ ЗНАЧЕНИЕ	НАИБОЛЬШЕЕ ЗНАЧЕНИЕ	НАИМЕНЬШАЯ СТ.ОЦЕНКА	НАИБОЛЬШАЯ СТ.ОЦЕНКА
1	X1	7.4615	5.8824	0.7884	0.6099	−0.3801	1.0000	21.0000	−1.0985	2.3015
2	X2	48.1538	15.5609	0,3231	−0.0419	−1.5707	26.0000	71.0000	−1.4237	1.4682
3	X3	11.7692	6.4051	0.5442	0.5416	−1.3633	4.0000	23.0000	−1.2130	1.7534
4	X4	30.0000	16.7382	0.5579	0.2923	−1.3078	6.0000	60.0000	−1.4338	1.7923
5	X5	95.4230	15.0437	0.1577	−0.1728	−1.5876	72.5000	115.9000	−1 5238	1.3612

ПРИМЕЧАНИЕ — ЕСЛИ ЭКСЦЕСС БОЛЬШЕ НУЛЯ, ТО РАСПРЕДЕЛЕНИЕ СЛУЧАЙНОЙ ВЕЛИЧИНЫ ИМЕЕТ БОЛЕЕ ТЯЖЕЛЫЕ ХВОСТЫ, ЧЕМ НОРМАЛЬНОЕ

КОВАРИАЦИОННАЯ МАТРИЦА

		X1	X2	X3	X4	X5
		1	2	3	4	5
X1	1	34.6025				
X2	2	20.9231	242.1413			
X3	3	−31.0512	−13.8781	41.0256		
X4	4	−24.1666	−253.4167	3.1667	280.1660	
X5	5	64.6634	191.0798	−51.5191	−206.8083	226.3139

КОРРЕЛЯЦИОННАЯ МАТРИЦА

		X1	X2	X3	X4	X5
		1	2	3	4	5
X1	1	1.0000				
X2	2	0.2296	1.0000			
X3	3	−0.8241	−0.1392	1.0000		
X4	4	−0.2454	−0.9730	0.0295	1.0000	
X5	5	0.7307	0.8163	−0.5347	−0.8213	1.0000

НАЗВАНИЕ РЕГРЕССИИ ..... ДРЕЙПЕР И СМИТ, СТР. 630, # DS36603, WFC
ШАГОВЫЙ АЛГОРИТМ ..... F
МАКСИМАЛЬНОЕ ЧИСЛО ШАГОВ 10
ЗАВИСИМАЯ ПЕРЕМЕННАЯ 5 X5
МИНИМАЛЬНО ДОПУСТИМОЕ F ДЛЯ ВКЛЮЧЕНИЯ 4.000, 4.000
МАКСИМАЛЬНО ДОПУСТИМОЕ F ДЛЯ ИСКЛЮЧЕНИЯ 3.900, 3.900
МИНИМАЛЬНЫЙ ПРИЕМЛЕМЫЙ ДОПУСК (ТОЛЕРАНТНОСТЬ) 0.01000
ИНДЕКСЫ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ 1 2 3 4

ШАГ N 0
СТАНДАРТНАЯ ОШИБКА ОЦЕНКИ 15,0437
ДИСПЕРСИОННЫЙ АНАЛИЗ

	СУММА КВАДРАТОВ	ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ	СРЕДНИЙ КВАДРАТ
ОСТАТОК	2715.7666	12	226.3139

ПЕРЕМЕННЫЕ, ВКЛЮЧЕННЫЕ В УРАВНЕНИЕ				ПЕРЕМЕННЫЕ, НЕ ВКЛЮЧЕННЫЕ В УРАВНЕНИЕ				
ПЕРЕМЕННАЯ	КОЭФФИЦИЕНТ	СТ. ОШИБКА КОЭФФИЦ. РЕГРЕССИИ	F ДЛЯ ДОПУСК	ВКЛЮЧЕНИЯ УРОВЕНЬ	ПЕРЕМ.	ЧАСТНАЯ КОРРЕЛЯЦИЯ	F ДЛЯ ИСКЛ.	УРОВ.
(Y — СВОБОДНЫЙ ЧЛЕН 95.423)					X1	1 0.73072	1.00000	12.601
					X2	2 0.81625	1.00000	21.96
					X3	3 –0.53467	1.00000	4.40
					X4	4 –0.82131	1.00000	22.80

ШАГ 1 1
ВКЛЮЧЕННАЯ ПЕРЕМЕННАЯ 4 X4
МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ R 0.8213
МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R•R 0.6745
ПРИВЕДЕННЫЙ R-КВАДРАТ 0.6450
СТАНДАРТНАЯ ОШИБКА ОЦЕНКИ 8.9639

ДИСПЕРСИОННЫЙ АНАЛИЗ

	СУММА КВАДРАТОВ	ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ	СРЕДНИЙ КВАДРАТ	F — ОТНОШЕНИЕ
РЕГРЕССИЯ	1831.8994	1	1831.899	22.80
ОСТАТОК	883.86694	11	80.35153	

ПЕРЕМЕННЫЕ, ВКЛЮЧЕННЫЕ В УРАВНЕНИЕ				ПЕРЕМЕННЫЕ, НЕ ВКЛЮЧЕННЫЕ В УРАВНЕНИЕ				
ПЕРЕМЕННАЯ	КОЭФФИЦИЕНТ	СТ. ОШИБКА КОЭФФИЦ. РЕГРЕССИИ	F ДЛЯ ДОПУСК	ВКЛЮЧЕНИЯ УРОВЕНЬ	ПЕРЕМ.	ЧАСТНАЯ КОРРЕЛЯЦИЯ	F ДЛЯ ИСКЛ.	УРОВ.
(Y — СВОБОДНЫЙ ЧЛЕН 117.568)								

X4	4	−0.738	0.155	−0.821	1.00000	22.80	1	X1	1	0.95677	0.93976	108.22	1
								X2	2	0.13021	0.05336	0.17	1
								X3	3	−0.89508	0.99913	40.29	1

ШАГ N 2

ВКЛЮЧЕННАЯ ПЕРЕМЕННАЯ 1 X1

МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИЙ R 0.9861

МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R*R 0.9725

ПРИВЕДЕННЫЙ R — КВАДРАТ 0.9670

СТАНДАРТНАЯ ОШИБКА ОЦЕНКИ 2.7343

ДИСПЕРСИОННЫЙ АНАЛИЗ

	СУММА КВАДРАТОВ	ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ	СРЕДНИЙ КВАДРАТ	F — ОТНОШЕНИЕ
РЕГРЕССИЯ	2641.0032	2	1320.501	176 62
ОСТАТОК	74.763428	10	7.476342	

ПЕРЕМЕННЫЕ, ВКЛЮЧЕННЫЕ В УРАВНЕНИЕ							ПЕРЕМЕННЫЕ, НЕ ВКЛЮЧЕННЫЕ В УРАВНЕНИЕ					
ПЕРЕМЕННАЯ	КОЭФФИЦИЕНТ	СТ. ОШИБКА КОЭФФИЦ.	СТ. КОЭФФИЦ. РЕГРЕССИИ	ДОПУСК	F ДЛЯ ВКЛЮЧЕНИЯ	УРОВЕНЬ	ПЕРЕМ.	КОРРЕЛЯЦИЯ	ЧАСТНАЯ КОРРЕЛЯЦИЯ	F ДЛЯ ИСКЛ	УРОВ
(Y — СВОБОДНЫЙ ЧЛЕН 103.097)											
X1 1	1.440	0.138	0,563	0.93976	108,22	1	X2	2	0.59860	0.05325	5 03	1
X4 4	−0.614	0.049	−0.683	0.93976	159.22	1	X3	3	−0.56571	0.28905	4.23	1

ШАГ N 3

ВКЛЮЧЕННАЯ ПЕРЕМЕННАЯ 2 X2

МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ R 0.9911

МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R*R 0.9823

ПРИВЕДЕННЫЙ R — КВАДРАТ 0.9764

СТАНДАРТНАЯ ОШИБКА ОЦЕНКИ 2.3088

ДИСПЕРСИОННЫЙ АНАЛИЗ

	СУММА КВАДРАТОВ	ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ	СРЕДНИЙ КВАДРАТ	F — ОТНОШЕНИЕ
РЕГРЕССИЯ	2667.7920	3	889.2639	166.83
ОСТАТОК	47.974396	9	5.330488	

ПЕРЕМЕННЫЕ, ВКЛЮЧЕННЫЕ В УРАВНЕНИЕ								ПЕРЕМЕННЫЕ, НЕ ВКЛЮЧЕННЫЕ В УРАВНЕНИЕ				
ПЕРЕМЕННАЯ	КОЭФФИЦИЕНТ	СТ. ОШИБКА КОЭФФИЦ.	СТ. КОЭФФИЦ. РЕГРЕССИИ	ДОПУСК	F ДЛЯ ВКЛЮЧЕНИЯ	УРОВЕНЬ		ПЕРЕМ.	КОРРЕЛЯЦИЯ	ЧАСТНАЯ КОРРЕЛЯЦИЯ	F ДЛЯ ИСКЛ.	УРОВ.
(Y – СВОБОДНЫЙ ЧЛЕН 71.648)												
X1	1	1.452	0.117	0.568	0.93780	154.00	1	X3	3	0.04767	0.02134	0.02	1
X2	2	0.416	0.186	0.430	0.05325	5.03	1						
X4	4	–0.237	0.173	– 0.263	0.05280	1.86	1						

ШАГ N	4
ВКЛЮЧЕННАЯ ПЕРЕМЕННАЯ	4 X4
МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ R	0.9893
МНОЖЕСТВЕННЫЙ КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ R*R	0.9787
ПРИВЕДЕННЫЙ R – КВАДРАТ	0.9744
СТАНДАРТНАЯ ОШИБКА ОЦЕНКИ	2.4064

ДИСПЕРСИОННЫЙ АНАЛИЗ

	СУММА КВАДРАТОВ	ЧИСЛО СТЕПЕНЕЙ СВОБОДЫ	СРЕДНИЙ КВАДРАТ	F – ОТНОШЕНИЕ
РЕГРЕССИЯ	2657.8606	2	1328.930	229.50
ОСТАТОК	57.905914	10	5.790591	

ПЕРЕМЕННЫЕ, ВКЛЮЧЕННЫЕ В УРАВНЕНИЕ								ПЕРЕМЕННЫЕ, НЕ ВКЛЮЧЕННЫЕ В УРАВНЕНИЕ				
ПЕРЕМЕННАЯ	КОЭФФИЦИЕНТ	СТ. ОШИБКА КОЭФФИЦ.	СТ. КОЭФФИЦ. РЕГРЕССИИ	ДОПУСК	F ДЛЯ ВКЛЮЧЕНИЯ	УРОВЕНЬ		ПЕРЕМ.	КОРРЕЛЯЦИЯ	ЧАСТНАЯ КОРРЕЛЯЦИЯ	F ДЛЯ ИСКЛ.	УРОВ.
(Y – СВОБОДНЫЙ ЧЛЕН 52.577)												
X1	1	1.468	0.121	0,574	0.94775	146.52	1	X3	3	0.41125	0.31826	1.83	1
X2	2	0.662	0.046	0.685	0.94777	208.58	1	X4	4	–0.41414	0.05280	1.86	1

КОЭФФИЦИЕНТЫ ШАГОВОЙ РЕГРЕССИИ

ПЕРЕМЕННЫЕ	0 Y – СВОБОДНЫЙ ЧЛЕН	1 X1	2 X2	3 X3	4 X4
ШАГ					
0	95.4230 *	1.8687	0.7891	–1.2558	–0.7382
1	117.5679 *	1.4400	0.3109	–1.1599	–0.7382 *
2	103.0973 *	1.4400 *	0.4161	–0.4100	–0.6140 *
3	71.6482 *	1.4519	0.4161 *	0.1019	–0.2365 *
4	52.5773 *	1.4683 *	0.6623 *	0.2500	–0.2365 *

ПРИМЕЧАНИЕ –

1) КОЭФФИЦИЕНТЫ РЕГРЕССИИ ДЛЯ ПЕРЕМЕННЫХ, ВКЛЮЧЕННЫХ В УРАВНЕНИЕ НА ДАННОМ ШАГЕ, ОТМЕЧЕНЫ ЗВЕЗДОЧКОЙ
2) ОСТАЛЬНЫЕ КОЭФФИЦИЕНТЫ БЫЛИ ПОЛУЧЕНЫ ПРИ ВКЛЮЧЕНИИ ПЕРЕМЕННОЙ В МОДЕЛЬ НА СЛЕДУЮЩЕМ ШАГЕ

F ДЛЯ ВКЛЮЧЕНИЯ ИЛИ F ДЛЯ ИСКЛЮЧЕНИЯ КАЖДОЙ ПЕРЕМЕННОЙ НА КАЖДОМ ШАГЕ

ПЕРЕМЕННЫЕ	1 X1	2 X2	3 X3	4 X4
ШАГ				
0	12.6025	21.9606	4.4034	22.7986
1	108.2221	0.1725	40.2938	22.7986 *
2	108.2218 *	5.0256	4.2357	159.2928 *
3	154.0019 *	5.0256 *	0.0182	1.8632 *
4	146.5185 *	208.5819 *	1.8320	1.8632

ЧАСТНЫЕ КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ

ПЕРЕМЕННЫЕ	1 X1	2 X2	3 X3	4 X4
ШАГ				
0	0.7307	0.8163	–0.5347	–0.8213
1	0.9568	0.1302	–0.8951	–0.8213 *
2	0.9568 *	0.5986	–0.5657	–0.9700 *
3	0.9720 *	0.5986 *	0.0477	–0.4141 *
4	0.9675 *	0.9769 *	0.4113	–0.4141 *

СВОДНАЯ ТАБЛИЦА

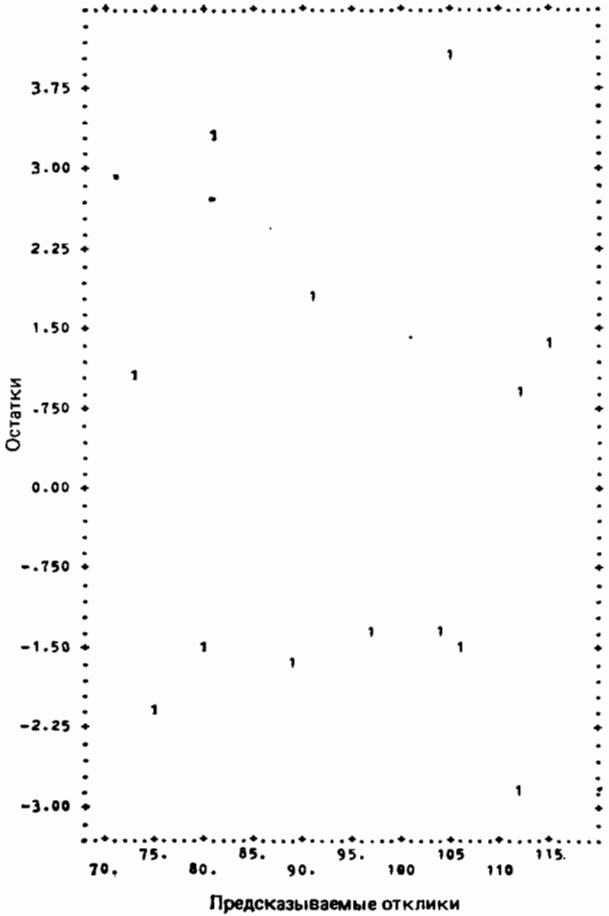
ШАГ N	ПЕРЕМЕННАЯ ВКЛЮЧЕННАЯ	ПЕРЕМЕННАЯ ИСКЛЮЧЕННАЯ	МНОЖЕСТВЕННЫЙ R	R*R	ПРИРОСТ R*R	F ДЛЯ ВКЛЮЧЕНИЯ	F ДЛЯ ИСКЛЮЧЕНИЯ	ЧИСЛО ВКЛЮЧЕННЫХ НЕЗАВИСИМЫХ ПЕРЕМЕННЫХ
1	4 X4		0.8213	0.6745	0.6745	22.7986		1
2	1 X1		0.9661	0.9725	0.2979	108.2218		2
3	2 X2		0.9911	0.9823	0.0099	5.0256		3
4		4 X4	0.9893	0.9787	–0.0037		1.8632	2

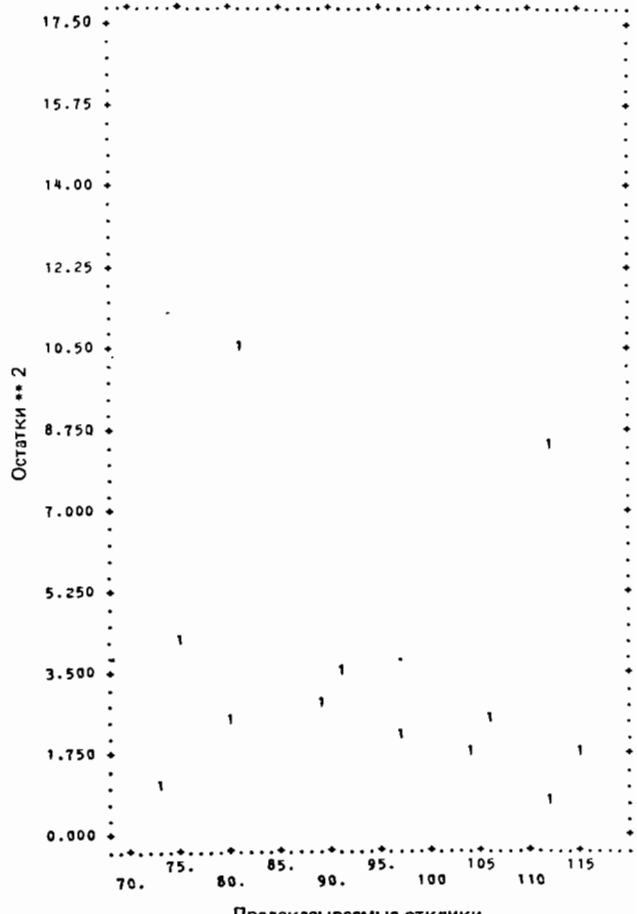
СВОДКА ПРЕДСКАЗЫВАЕМЫХ ЗНАЧЕНИЙ, ОСТАТКОВ И ПЕРЕМЕННЫХ

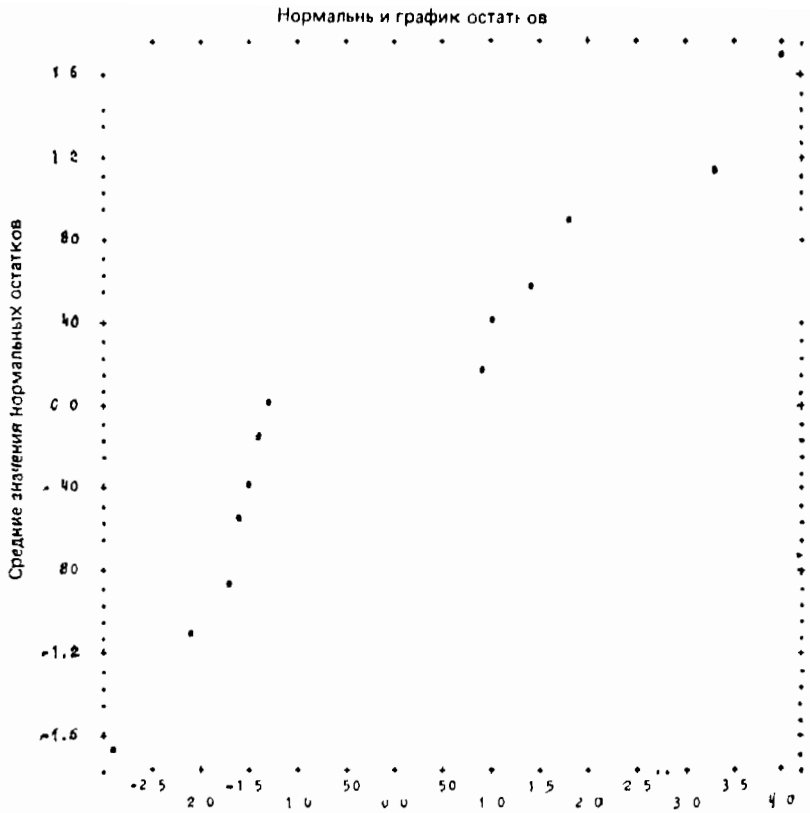
ПРИМЕЧАНИЕ – ОТРИЦАТЕЛЬНЫЕ НОМЕРА ОПЫТОВ СООТВЕТСТВУЮТ ОПЫТАМ С ПРОПУЩЕННЫМИ ЗНАЧЕНИЯМИ, ЧИСЛО СТАНДАРТНЫХ ОТКЛОНЕНИЙ ОТ СРЕДНЕГО ОБОЗНАЧАЕТСЯ ЗВЕЗДОЧКАМИ (ОТ ОДНОЙ ДО ТРЕХ) СПРАВА ОТ СООТВЕТСТВУЮЩЕГО ОСТАТКА ИЛИ ПЕРЕМЕННОЙ.
ПРОПУЩЕННЫЕ ЗНАЧЕНИЯ ЗАМЕНЯЮТСЯ БОЛЕЕ ЧЕМ ТРЕМЯ ЗВЕЗДОЧКАМИ.

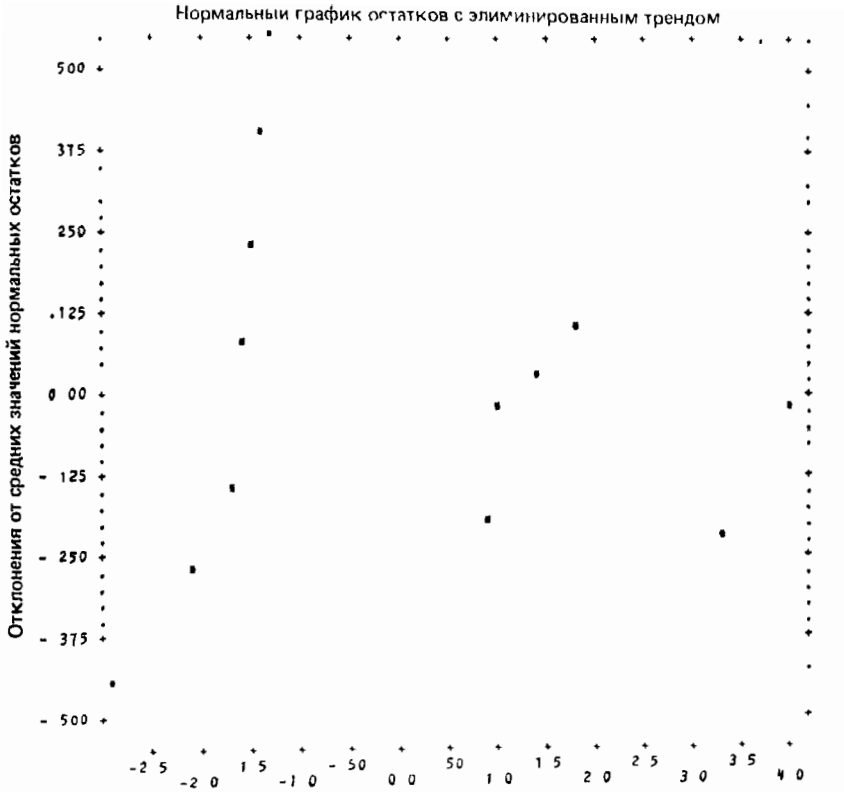
N	МЕТКА	ПРЕДСКАЗАННЫЙ	ОСТАТОК	ВЕС	5 X5	1 X1	2 X2	3 X3	4 X4
1		80.0740	−1.5740	1.000	78.5000	7.0000	26.0000	6.0000	60.0000
2		73.2509	1.0491	1.000	74.3000	1.0000	29.0000	15.0000	52.0000
3		105.8147	−1.5147	1.000	104.3000	11.0000	56.0000	8.0000	20.0000
4		89.2584	−1.6584	1.000	87.6000	11.0000	31.0000	8.0000	47.0000
5		97.2925	1.3925	1.000	95.9000	7.0000	52.0000	6.0000	33.0000
6		105.1525	4.0475 *	1.000	109.2000	11.0000/	55.0000	9.0000	22.0000
7		104.0021	1.3021	1.000	102.7000	3.0000	71.0000	17.0000	6.0000
8		74.5754	−2.0754	1.000	72.5000	1.0000	31.0000	22.0000	44.0000
9		91.2755	1.8245	1.000	93.1000	2.0000	54.0000	18.0000	22.0000
10		114.5375	1.3625	1.000	115.9000	21.0000	47.0000	4.0000	26.0000
11		80.5356	3.2643 *	1.000	83.8000	1.0000	40.0000	23.0000	34.0000
12		112.4372	0.8626	1.000	113.3000	11.0000	66.0000	9.0000	12.0000
13		112.2934	−2.8934 *	1.000	109.4000	10.0000	68.0000	8.0000	12.0000

КАЖДАЯ ЗВЕЗДОЧКА СООТВЕТСТВУЕТ ОДНОМУ СТАНДАРТНОМУ ОТКЛОНЕНИЮ


Остатки
3.75
3.00
2.25
1.50
.750
0.00
-.750
-1.50
-2.25
-3.00
70.
75.
80.
85.
90.
95.
100
105
110
115
Предсказываемые отклики


Остатки ** 2
17.50
15.75
14.00
12.25
10.50
8.750
7.000
5.250
3.500
1.750
0.000
70.
75.
80.
85.
90.
95.
100
105
110
115
Предсказываемые отклики


Нормальнь и график остать ов
Средние значения нормальных остатков
1 6
1 2
80
40
0 0
- 40
80
-1,2
-1,6
-2 5
2 0
-1 5
1 0
50
0 0
50
1 0
1 5
2 0
2 5
3 0
3 5
4 0


Нормальный график остатков с элиминированным трендом
Отклонения от средних значений нормальных остатков
500
375
250
.125
0 00
- 125
- 250
- 375
- 500
-2 5
-2 0
1 5
-1 0
- 50
0 0
50
1 0
1 5
2 0
2 5
3 0
3 5
4 0

Книги

Acton, F S (1959) *Analysis of Straight-Line Data* New York Dover

*Afifi, A A and S P Azen (1979) *Statistical Analysis A Computer Oriented Approach* New York Academic Press.

*Bard, Y (1974) *Nonlinear Parameter Estimation* New York Academic Press

Barnett, V and T Lewis (1978) *Outliers in Statistical Data* New York Wiley

Beck, J V and K J. Arnold (1977) *Parameter Estimation in Engineering and Science* New York. Wiley

Belsley, D A , E Kuh, and R E Welsch (1980) *Regression Diagnostics Identifying Influential Data and Sources of Collinearity* New York Wiley

Chatterjee, S and B Price (1977) *Regression Analysis by Example* New York Wiley

Cohen, J and P Cohen (1975) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences* New York Halsted

Daniel, C and F S Wood (1980) *Fitting Equations to Data* 2nd ed New York Wiley

Dixon, W J , Ed (1979) *BMD Biomedical Computer Programs, P Series* Berkeley University of California Press

Dunn, O J and V A Clark (1974) *Applied Statistics Analysis of Variance and Regression.* New York Wiley.

Graybill, F. A (1961) *An Introduction to Linear Statistical Models*, Vol I New York McGraw-Hill.

Gunst, R F and R L Mason (1980) *Regression Analysis and Its Application A Data-Oriented Approach* New York Marcel Dekker.

*Hannan, E J (1960) *Time Series Analysis* London Methuen

Hawkins, D M (1980) *Identification of Outliers* London Chapman and Hall

*Himmelblau, D M (1970) *Process Analysis by Statistical Methods* New York Wiley

Kleinbaum, D G and L L Kupper (1978) *Applied Regression Analysis and Other Multivariable Methods* North Scituate, Ma Duxbury Press

*Mosteller, F and J. W. Tukey (1977) *Data Analysis and Regression* Reading, Ma Addison-Wesley.

Neter, J and W. Wasserman (1974) *Applied Linear Statistical Models* Homewood, Ill Irwin

Nie, N H (1975) *SPSS Statistical Package for Social Sciences* New York McGraw-Hill.
Plackett, R L (1960) *Regression Analysis*, Oxford, England Clarendon Press.
*Rao, C R (1973) *Linear Statistical Inference and Its Applications*, 2nd ed New York. Wiley.
Ryan, T A B L Joiner, and B F Ryan (1976) *MINITAB Student Handbook* North Scituate, Ma Duxbury.
Searle, S R (1971) *Linear Models* New York Wiley
*Seber, G A F (1977) *Linear Regression Analysis* New York Wiley.
Sprent, P (1969) *Models in Regression and Related Topics* London. Methuen.
Weisberg S (1980) *Applied Linear Regression* New York Wiley.
Wesolowsky, G O (1976) *Multiple Regression and Analysis of Variance* New York Wiley.
Williams, E J (1959) *Regression Analysis* New York Wiley.
Younger, M S (1979) *Handbook for Linear Regression* North Scituate, Ma Duxbury.

Работы общего характера

Box, G E P (1966) Use and abuse of regression *Technometrics*, **8**, 625–629.
Corlett, T (1963) Ballade of multiple regression *Appl Statist* **12**, 145
Cox, D R (1968) Notes on some aspects of regression analysis *J Roy Statist Soc*, **A-131**, 265–279, discussion 315–329
Ehrenberg, A S C (1963) Bivariate regression is useless *Appl Statist*, **12**, 161–179.
Ehrenberg, A S C (1968) The elements of lawlike relationships *J Roy Statist Soc*, **A-131**, 280–302, discussion 315–329
Harter, H L. (1974–1976) The method of least squares and some alternatives—Parts I–VI. *Int Statist Rev* Part I **42**, 147 174, Part II **42**, 235 264, Part III **43**, 1–44, Part IV **43**, 125–190, Part V **43**, 269 278, Part VI, subject and author indexes **44**, 113–159
Mullet, G M (1972) A graphical illustration of simple (total) and partial regression *Am. Statist*, **26**, 25 27
Tukey, J W (1960) Where do we go from here? *J Am Statist Assoc*, **55**, 80 93
Warren, W G (1971) Correlation or regression bias or precision *Appl Statist*, **20** 148–164.
Willan, A R and D G Watts (1978) Meaningful multicollinearity measures *Technometrics*, **20**, 407 412

Глава 1

A. Простая линейная регрессия и корреляция

Barnett, V D (1970) Fitting straight lines—the linear functional relationship with replicated observations *Appl Statist* **19** 135–144
Brown, B W (1970) Simple comparisons of simultaneous regression lines *Biometrics*, **26**, 143–144
Donner, A and B Rosner (1980) On inferences concerning a common correlation coefficient. *Applied Statistics*, **29**, 69–76

Dunn, O. J. (1968). A note on confidence bands for a regression line over a finite range. *J. Am. Statist. Assoc.*, **63**, 1028–1033.

Folks, J. L. (1967). Straight line confidence regions for linear models. *J. Am. Statist Assoc*, **62**, 1365–1374.

Gillingham, R. and D. Heien. (1971). Regression through the origin. *Am. Statist.* **25**, 54–55.

Halpern, M., S. C Rastogi, I. Ho, and Y. Y. Yang. (1967). Shorter confidence bands in linear regression. *J. Am. Statist. Assoc.*, **62**, 1050–1067.

Hollander, M. (1970). A distribution-free test for parallelism [of two regression lines]. *J. Am. Statist. Assoc.*, **65**, 387–394.

Neter, J. and E. S. Maynes. (1970). On the appropriateness of the correlation coefficient with a 0, 1 dependent variable. *J. Am. Statist. Assoc.*, **65**, 501–509.

Potthoff, R. F. (1974). A non-parametric test of whether two simple regression lines are parallel. *Ann. Statist.*, **2**, 295–310.

Scheffe, H. (1958). Fitting straight lines when one variable is controlled. *J. Amer. Statist. Assoc.*, **53**, 106–118.

Sprent, P. (1971). Parallelism and concurrence in linear regression. *Biometrics*, **27**, 440–444.

Turner, M. E. (1960). Straight-line regression through the origin. *Biometrics*, **16**, 483–485.

B. Обратная регрессия, калибровка

Andrews, D. F. (1970). Calibration and statistical inference. *J. Am. Statist. Assoc.*, **65**, 1233–1242.

Berkson, J. (1950). Are there two regressions? *J. Am. Statist. Assoc.*, **45**, 164–180.

Berkson, J. (1969). Estimation of a linear function for a calibration line: consideration of a recent proposal. *Technometrics*, **11**, 649–660.

Cox, C. P. (1971). Interval estimation for X-predictions from linear Y-on-X regression lines through the origin. *J. Am. Statist. Assoc.*, **66**, 749–751, correction, 252 (1972).

Dunsmore, I. R. (1968). A Bayesian approach to calibration. *J. Roy. Statist. Soc.*, **B-30**, 396–405.

Halperin, M. (1970). On inverse estimation in linear regression. *Technometrics*, **12**, 727–736.

Hoadley, B. (1970). A Bayesian look at inverse linear regression. *J. Am. Statist. Assoc.*, **65**, 356–369.

Krutchkoff, R. G. (1967). Classical and inverse regression methods of calibration. *Technometrics*, **9**, 425–439.

Krutchkoff, R. G. (1969). Classical and inverse regression methods of calibration in extrapolation. *Technometrics*, **11**, 605–608.

Martinelle, S. (1970). On the choice of regression in linear calibration. Comments on a paper by R. G. Krutchkoff. *Technometrics*, **12**, 157–161.

Oden, A. (1973). Simultaneous confidence intervals in inverse linear regression. *Biometrika*, **60**, 339–343.

Ott, R. L. and R. H. Myers. (1968). Optimal experimental designs for estimating the independent variable in regression. *Technometrics*, **10**, 811–823.

Pepper, M. P. G. (1973). A calibration of instruments with non-random errors. *Technometrics*, **15**, 587–599.

Perng, S. K. and Y. L. Tong. (1974). A sequential solution to the inverse linear regression problem. *Ann. Statist.*, **2**, 535–539.

Scheffe, H. (1973). A statistical theory of calibration. *Ann. Statist.*, **1**, 1–37.

Shukla, G. K. (1972). On the problem of calibration. *Technometrics*, **14**, 547–553.

Williams, E. J. (1969). A note on regression methods in calibration. *Technometrics*, **11**, 189–192.

Williams, E. J. (1969). Regression methods in calibration problems *Proc. 37th Session, Bull. Int. Statist. Inst.*, **43**, Book 1, 17–28.

C. Ошибки в откликах и факторах

Bartlett, M. S. (1949). Fitting a straight line when both variables are subject to error. *Biometrics*, **5**, 207–212.

Berkson, J. (1950). Are there two regressions? *J. Am. Statist. Assoc.*, **45**, 164–180.

Carlson, F. D., E. Sobel, and G. S. Watson. (1966). Linear relationship between variables affected by errors *Biometrics*, **22**, 252–267.

Halperin, M and J. Gurian (1971). A note on estimation in straight line regression when both variables are subject to error. *J. Am. Statist. Assoc.*, **66**, 587–589.

Karni, E. and I. Weissman. (1974). A consistent estimator of the slope in a regression model with errors in the variables. *J. Am. Statist. Assoc.*, **69**, 211–213, corrections, 840.

Kerrich, J. E. (1966). Fitting the line $Y = \alpha X$ when errors of observation are present in both variables. *Am. Statist.*, **20**, 24.

Mandansky, A (1959). The fitting of straight lines when both variables are subject to error. *J. Am. Statist. Assoc.*, **54**, 173–205.

Sampson, A. R. (1974). A tale of two regressions. *J. Am. Statist. Assoc.*, **69**, 682–689.

Stroud, T. W. F. (1972). Comparing conditional means and variances in a regression model with measurement errors of known variances. *J. Am. Statist. Assoc.*, **67**, 407–412, discussion 412–414, correction (1973) **68**, 251.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *Ann. Math. Statist*, **11**, 284–300.

Ware, J. H. (1972). The fitting of straight lines when both variables are subject to error and the ranks of the means are known. *J. Am. Statist. Assoc.*, **67**, 891–897.

Глава 2

A. Обычная многомерная регрессия

Balestra, P. (1970). On the efficiency of ordinary least squares in regression models. *J. Am. Statist. Assoc.*, **65**, 1330–1337.

Cox, D. R. and D. V. Hinkley. (1968). A note on the efficiency of least squares estimates. *J. Roy. Statist. Soc.*, **B-30**, 284–289.

Cramer, E. M. (1972). Significance tests and test of models in multiple regression. *Am. Statist.*, **26**, 26–30. October issue. Also see **27**, p. 92 (1973).

Crocker, D. C. (1972). Some interpretations of the multiple correlation coefficient. *Amer. Statist.*, **26**, 31–33. April issue. Also see October, pp. 58–59.

Geary, R. C. and C. E. V. Leser. (1968). Significance tests in multiple regression. *Amer. Statist.*, **22**, 20–21.

Goldberger, A. S. (1962). Best linear unbiased prediction in the generalized linear regression model. *J. Am. Statist. Assoc.*, **57**, 369–375.

Hill, R. C., G. G. Judge, and T. B. Fomby (1978) On testing the adequacy of a regression model. *Technometrics*, **20**, 491–494.

Hill, R. C., G. G. Judge, and T. B. Fomby. (1980). Is the regression equation adequate?—a reply. *Technometrics*. **22**, 127–128.

Hinchin, J. D. (1970). Multiple regression with unbalanced data. *J. Quality Technol.*, **2**, 22–29.

Jacquez, J. A., F. J. Mather, and C. R. Crawford. (1968). Linear regression with non-constant, unknown error variances: sampling experiments with least squares, weighted least squares and maximum likelihood estimators. *Biometrics*, **24**, 607–626.

Suich, R. and G. C. Derringer. (1977). Is the regression equation adequate?—one criterion. *Technometrics*, **19**, 213–216.

Suich, R. and G. C. Derringer. (1980). Is the regression equation adequate?—a further note. *Technometrics*, **22**, 125–126.

B. Регрессия и линейное программирование

Davies, M. (1976). Linear approximation using the criterion of least total deviations. *J. Roy. Statist. Soc.*, **B-29**, 101–109. See also p. 587 (1967).

Duncan, D. B. and R. H. Jones. (1966). Multiple regression with stationary errors. *J. Am. Statist. Assoc.*, **61**, 917–928.

Fisher, W. D. (1961). A note on curve fitting with minimum deviations by linear programming. *J. Am. Statist. Assoc.*, **56**, 359–362.

Karst, O. T. (1958). Linear curve fitting using least deviations. *J. Am. Statist. Assoc.*, **53**, 118–132.

Kiountouzis, E. A. (1973). Linear programming techniques in regression analysis. *Appl. Statist.*, 69–73.

Schlossmacher, E. J. (1973). An iterative technique for absolute deviations curve fitting. *J. Am. Statist. Assoc.*, **68**, 857–859.

Wagner, H. M. (1959). Linear programming techniques for regression analysis. *J. Am. Statist. Assoc.*, **54**, 206–212.

C. Взвешенный метод наименьших квадратов; регрессия с ограничениями

Amemiya, T. (1973). Regression analysis when the variance of the dependent variable is proportional to the square of its expectation. *J. Am. Statist. Assoc.*, **68**, 928–934.

Anderson, T. W. (1962). Least squares and best unbiased estimates. *Ann. Math. Statist.*, **33**, 266–272.

Atiqullah, M. (1969). On a restricted least squares estimator. *J. Am. Statist. Assoc.*, **64**, 964–968.

Bement, T. R. and J. S. Williams. (1969). Variance of weighted regression estimators when sampling errors are independent and heteroscedastic. *J. Am. Statist. Assoc.*, **64**, 1369–1382.

Bowden, D. C. (1970). Simultaneous confidence bands for linear regression models. *J. Am. Statist. Assoc.*, **65**, 413–421.

Bradley, E. L. (1973). The equivalence of maximum likelihood and weighted least squares estimates in the exponential family. *J. Am. Statist. Assoc.*, **68**, 199–200.

Christensen, L. R. (1973). Simultaneous statistical inference in the normal multiple linear regression model. *J. Am. Statist. Assoc.*, **68**, 457–561.

Edwards, A. W. F. (1973). The likelihood treatment of linear regression. *Biometrika*, **60**, 73–77.

Hahn, G. J. (1972). Simultaneous prediction intervals for a regression model. *Technometrics*, **14**, 203–214.

Halperin, M. and J. Gurian. (1968). Confidence bands in linear regression with constraints on the independent variables. *J. Am. Statist. Assoc.*, **63**, 1020–1027.

Hartley, H. O. and K. S. E. Jayatillake. (1973). Estimation for linear models with unequal variances. *J. Am. Statist. Assoc.*, **68**, 189–192.

Judge, G. G. and T. Takayama. (1969). Inequality restrictions in regression analysis. *J. Am. Statist. Assoc.*, **61**, 166–181.

Mantel, E. H. (1973). Exact linear restrictions on parameters in the classical linear regression model. *Am. Statist.*, **27**, 86–87. See also **28**, p. 36 (1974).

McGuire, T. W., J. U. Farley, R. E. Lucas, and L. W. Ring. (1968). Estimation and inference for linear models in which subsets of the dependent variable are constrained. *J. Am. Statist. Assoc.*, **63**, 1201–1213.

Obenchain, R. L. (1975). Residual optimality: Ordinary vs. weighted vs. biased least squares. *J. Am. Statist. Assoc.*, **70**, 375–379.

Schmee, J. and G. J. Hahn. (1979). A simple method for regression analysis with censored data. *Technometrics*, **21**, 417–434.

Scott, A. J. and M. J. Symons. (1971). A note on shortest prediction intervals for log-linear regression. *Technometrics*, **13**, 889–894.

Waterman, M. S. (1974). A restricted least squares problem. *Technometrics*, **16**, 135–136.

Williams, J. S. (1967). The variance of weighted regression estimators. *J. Am. Statist. Assoc.*, **62**, 1290–1301.

Глава 3

Abrahamse, A. P. J. and J. Koerts. (1971). New estimators of disturbances in regression analysis. *J. Am. Statist. Assoc.*, **66**, 71–74.

Abrahamse, A. P. J. and A. S. Louter. (1971). On a new test for autocorrelation in least squares regression. *Biometrika*, **58**, 53–60.

Andrews, D. F. (1971). Significance tests based on residuals. *Biometrika*, **58**, 139–148.

Andrews, D. F. (1972). Plots of high-dimensional data. *Biometrics*, **28**, 125–136.

Andrews, D. F. and D. Pregibon. (1978). Finding the outliers that matter. *J. Roy. Statist. Soc.*, **B-40**, 85–93.

Andrews, D. F. and J. W. Tukey. (1973). Teletypewriter plots for data analysis can be fast: 6-line plots including probability plots. *Appl. Statist.*, **22**, 192–202.

Anscombe, F. J. (1961). Examination of residuals. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.*, **1**, 1–36.

Anscombe, F. J. (1967). Topics in the investigation of linear relations fitted by the method of least squares. *J. Roy. Statist. Soc.*, **B-29**, 1–29, discussion 29–52.

Anscombe, F. J. and J. W. Tukey. (1963). The examination and analysis of residuals. *Technometrics*, **5**, 141–160.

Beckman, R. J. and H. J. Trussell. (1974). The distribution of an arbitrary studentized residual and the effects of updating in multiple regression. *J. Amer. Statist. Assoc.*, **69**, 199–201.

Behnken, D. W. and N. R. Draper. (1972). Residuals and their variance patterns. *Technometrics*, **14**, 101–111.

Bennett, B. M. (1967). Use of Haldane-Smith test in examining randomness of residuals. *Metron*, **26**, No. 3–4, 1–3.

Berenblut, I. I. and G. I. Webb. (1973). A new test for autocorrelation errors in the linear regression model. *J. Roy. Statist. Soc.*, **B-35**, 33–50.

Collett, D. (1980). Outliers in circular data. *Applied Statistics*, **29**, 50–57. Corrections p. 229

Collett, D. and T. Lewis. (1976). The subjective nature of outliner rejection procedures. *Appl. Statist.*, **25**, 228–237.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics*, **19**, 15–18.

Cook, R. D. (1979). Influential observations in linear regression. *J. Am. Statist. Soc.*, **74**, 169–174.

Cox, D. R. and E. J. Snell. (1968). A general definition of residuals. *J. Roy. Statist. Soc.*, **B-30**, 248–265, discussion 265–275.

Cox, D. R. and E. J. Snell. (1971). On test statistics calculated from residuals. *Biometrika*, **58**, 589–594.

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two level experiments. *Technometrics*, **1**, 311–341.

Daniel, C. (1978). Patterns in residuals in the two-way layout. *Technometrics*, **20**, 385–395.

Draper, N. R. and J. A. John. (1980). Testing for three or fewer outliers in two-way tables. *Technometrics*, **22**, 9–15.

Draper, N. R. and W. E. Lawrence. (1970). A note on residuals in two dimensions. *Technometrics*, **12**, 394–398.

Durbin, J. (1969). Tests for serial correlation in regression analysis based on the periodogram of least squares residuals. *Biometrika*, **56**, 1–15.

Durbin, J. (1970). An alternative to the bounds test for testing for serial correlation in least squares regression. *Econometrica*, **38**, 422–429.

Durbin, J. and G. S. Watson. (1950). Testing for serial correlation in least squares regression, I. *Biometrika*, **37**, 409–428.

Durbin, J. and G. S. Watson. (1951). Testing for serial correlation in least squares regression. II. *Biometrika*, **38**, 159–178.

Durbin, J. and G. S. Watson. (1971). Testing for serial correlation in least squares regression. III. *Biometrika*, **58**, 1–19.

Elashoff, J. D. (1972). A model for quadratic outliers in linear regression. *J. Am. Statist. Assoc.*, **67**, 478–485.

Ellenberg, J. H. (1973). The joint distribution of the standardized least squares residuals from a general linear regression. *J. Am. Statist. Assoc.*, **68**, 941–943.

Gentleman, J. F. and M. B. Wilk. (1975). Detecting outliers in a two-way table: I. Statistical behavior of residuals. *Technometrics*, **17**, 1–14.

Gentleman, J. F. and M. B. Wilk. (1975). Detecting outliers. II. Supplementing the direct analysis of residuals. *Biometrics*, **31**, 387–410.

Gnanadesikan, R. and J. R. Kettenring. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics*, **28**, 81–124.

Green, J. R. (1971). Testing departure from a regression, without using replication. *Technometrics*, **13**, 609–615.

Grossman, S. I. and G. P. H. Styan. (1972). Optimality properties of Theil's BLUS residuals. *J. Am. Statist. Assoc.*, **67**, 672–673.

Guttman, I. (1973). Care and handling of univariate or multivariate outliers in detecting spuriosity — a Bayesian approach. *Technometrics*, **15**, 723–738.

Haberman, S. J. (1973). The analysis of residuals in cross-classified tables. *Biometrics*, **29**, 205–220.

Hannan, E. J. (1957). Testing for serial correlation in least squares regression. *Biometrika*, **44**, 57–66.

Hedayat, A. and D. S. Robson. (1970). Independent stepwise residuals for testing homoscedasticity. *J. Am. Statist. Assoc.*, **65**, 1573–1581.

Hoaglin, D. C. and R. E. Welsch. (1978). The hat matrix in regression and anova. *Am. Statist.* **32**, 17–22.

Huang, C. J. and B. W. Bolch. (1974). On the testing of regression disturbances for normality. *J. Am. Statist. Assoc.*, **69**, 330–335.

Jackson, J. E. and W. H. Lawton. (1967). Answer to query 22. *Technometrics*, **9**, 339–340.

John, J. A. (1978). Outliers in factorial experiments. *Appl. Statist.*, **27**, 111–119.

John, J. A. and N. R. Draper. (1978). On testing for two outliers or one outlier in two-way tables. *Technometrics*, **20**, 69–78.

Larsen, W. A. and S. J. McCleary. (1972). The use of partial residual plots in regression analysis. *Technometrics*, **14**, 781–790.

Loynes, R. M. (1969). On Cox and Snell's general definition of residuals. *J. Roy. Statist. Soc.*, **B-31**, 103–106.

Lund, R. E. (1975). Tables for an approximate test for outliers in linear models. *Technometrics*, **17**, 473–476.

Mickey, M. R., O. J. Dunn, and V. Clark. (1967). Note on the use of stepwise regression in detecting outliers. *Computers and Biomedical Research*, **1**, 105–111.

Machin, D. (1970). Regression with correlated residuals: an example from competition experiments. *Biometrics*, **26**, 835–840.

Nelson, W. (1973). The analysis of residuals from censored data. *Technometrics*, **15**, 697–715.

Obenchain, R. L. (1975). Residual optimality: ordinary vs. weighted vs. biased least squares. *J. Am. Statist. Assoc.*, **70**, 375–379.

Phillips, G. D. A. and A. C. Harvey. (1974). A simple test for serial correlation in regression analysis. *J. Am. Statist. Assoc.*, **69**, 935–939.

Pierce, D. A. (1971). Distribution of residual autocorrelations in the regression model with autoregressive-moving average errors. *J. Roy. Statist. Soc.*, **B-33**, 140–146.

Pierce, D. A. (1971). Least squares estimation in the regression model with autoregressive-moving average errors. *Biometrika*, **58**, 299–312.

Prescott, P. (1975). An approximate test for outliers in linear regression. *Technometrics*, **17**, 129–132.

Prescott, P. (1979). Critical values for a sequential test for many outliers. *Appl. Statist.*, **28**, 36–39.

Putter, J. (1967). Orthonormal bases of error spaces and their use for investigating the normality and variances of residuals. *J. Am. Statist. Assoc.*, **62**, 1022–1036.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least-squares regression analysis. *J. Roy. Statist. Soc.*, **B-31**, 350–371.

Snee, R. D. (1971). A note on the use of residuals for examining the assumptions of covariance analysis. *Technometrics*, **13**, 430–437.

Stefansky, W. (1972). Rejecting outliers in factorial designs. *Technometrics*, **14**, 469–479.

Theil, H. and A. L. Nagar. (1961). Testing the independence of regression disturbances. *J. Am. Statist. Assoc.*, **56**, 793–806.

Tietjen, G. L., R. H. Moore, and R. J. Beckman. (1973). Testing for a single outlier in simple linear regression. *Technometrics*, **15**, 717–721.

Von Neumann, J. (1941). Distribution of the ratio of the mean square successive difference to the variance. *Ann. Math. Statist.*, **12**, 367–395.

Welsch, R. E. and S. C. Peters. (1978). Finding influential subsets of data in regression models. *Proc. 11th Interface Symp. Comput. Sci. Statist.*, eds. R. Gallant and T. Gerry, 240–244.

Wood, F. S. (1973). The use of individual effects and residuals in fitting equations to data. *Technometrics*, **15**, 677–695.

Wooding, W. M. (1969). The computation and use of residuals in the analysis of experimental data. *J. Quality Technol.*, **1**, 175–188. Correction p. 294.

Zyskind, G. and P. A. Johnson. (1973). On a zero residual sum in regression. *Am. Statist.*, **27**, 43–44.

Глава 5

A. Преобразования

Andrews, D. F. (1971). A note on the selection of data transformations. *Biometrika*, **58**, 249–254.

Andrews, D. F., R. Gnanadesikan, and J. L. Warner. (1971). Transformations of multivariate data. *Biometrics*, **27**, 825–840.

Bartlett, M. S. (1947). The use of transformations. *Biometrics*, **3**, 39–52.

Bliss, C. I. and S. S. Whitman. (1968). A table for working angles. *Biometrics*, **24**, 413–422.

Box, G. E. P. and D. R. Cox. (1964). An analysis of transformations. *J. Roy. Statist. Soc.*, **B-26**, 211–243, discussion 244–252.

Box, G. E. P. and W. J. Hill. (1974). Correcting inhomogeneity of variance with power transformation weighting. *Technometrics*, **16**, 385–389.

Box, G. E. P. and P. W. Tidwell. (1962). Transformations of the independent variables. *Technometrics*, **4**, 531–550.

Davies, P. (1968). A sequential method for testing the linear trends of responses in dose trials. *Biometrics*, **24**, 663–677.

Derringer, G. C. (1974). An empirical model for viscosity of filled and plasticized elastomer compounds. *J. Appl. Polym. Sci.*, **18**, 1083–1101.

Dolby, J. L. (1963). A quick method for choosing a transformation. *Technometrics*, **5**, 317–325.

Draper, N. R. and D. R. Cox. (1969). On distributions and their transformations to normality. *J. Roy. Statist. Soc.*, **B-31**, 472–476.

Dutka, A. F. and F. J. Ewens. (1971). A method for improving the accuracy of polynomial regression analysis. *J. Quality Technol.*, **3**, 149–155.

Eisenpress, H. (1956). Regression techniques applied to seasonal corrections and adjustments for calendar shifts. *J. Am. Statist. Assoc.*, **51**, 615–620.

Green, R. D. and J. P. Doll. (1974). Dummy variables and seasonality—a curio. *Am. Statist.*, **28**, 60–62.

Heren, D. A. (1968). A note on log-linear regression. *J. Am. Statist. Assoc.*, **63**, 1034–1038.

Hinkley, D. V. (1975). On power transformations to symmetry. *Biometrika*, **62**, 101–111.

Hinz, P. N. and H. A. Eagles. (1976). Estimation of a transformation for the analysis of some agronomic and genetic experiments. *Crop. Sci.*, **16**, 280–283.

Hoyle, M. H. (1973). Transformations: an introduction and a bibliography. *International Statistical Review*, **41**, 203–223.

Huang, C. L., L. C. Moon, and H. S. Chang. (1978). A computer program using the Box-Cox transformation technique for the specification of functional form. *Am. Statist.*, **32**, 144.

Iman, R. L. and W. J. Conover. (1979). The use of the rank transform in regression. *Technometrics*, **21**, 499–510.

John, J. A. and N. R. Draper. (1980). An alternative family of transformations. *Appl. Statist.*, **29**, 190–197.

Kruskal, J. B. (1978). Transformations of data. Part II of entry "Statistical Analysis, Special Problems of" in *International Encyclopedia of Statistics*, Vol. 2, eds. W. H. Kruskal and J. M. Tanur. New York: The Free Press, pp. 1044–1056.

Lindsey, J. K. (1972). Fitting response surfaces with power transformations. *Appl. Statist.*, **21**, 234–247.

Llewelyn, F. W. M. (1968). The log(−log) transformation in the analysis of fruit retention records. *Biometrics*, **24**, 627–638.

Narula, S. C. (1979). Orthogonal polynomial regression. *International Statistical Review*, **47**, 31–36.

Nelder, J. A. (1966). Inverse polynomials, a useful group of multi-factor response functions. *Biometrics*, **22**, 128–141.

Quartermain, A. R. and A. E. Freeman. (1967). Some transformations of scale and the estimation of genetic parameters from daughter-dam regression. *Biometrics*, **23**, 823–834.

Robson, D. S. (1959). A simple method for constructing orthogonal polynomials when the independent variable is unequally spaced. *Biometrics*, **15**, 187–191.

Schlesselman, J. J. (1971). Power families: a note on the Box and Cox transformation. *J. Roy. Statist. Soc.*, **B-33**, 307–311.

Schlesselman, J. J. (1973). Data transformation in two-way analysis of variance. *J. Am. Statist. Assoc.*, **68**, 369–378.

Suits, D. B. (1957). Use of dummy variables in regression equations. *J. Am. Statist. Assoc.*, **52**, 548–551.

Tukey, J. W. (1957). On the comparative anatomy of transformations. *Ann. Math. Statist.*, **28**, 602–632.

Wilkie, D. (1965). Complete set of leading coefficients, $\lambda(r, n)$, for orthogonal polynomials up to $n = 26$. *Technometrics*, **7**, 644–648.

B. Отрезки прямых; сплайны

*Ahlberg, J. H., E. N. Nilson, and J. L. Walsh. (1967). *The Theory of Splines and their Application*, New York: Academic Press.

Bacon, D. W. and D. G. Watts. (1971). Estimating the transition between two intersecting straight lines. *Biometrika*, **58**, 525–534.

Beckman, R. J. and R. D. Cook. (1979). Testing for two phase regression. *Technometrics*, **21**, 65–69.

Bellman, R. and R. Roth. (1969). Curve fitting by segmented straight lines. *J. Am. Statist. Assoc.*, **64**, 1079–1084.

Boneva, L. I., D. G. Kendall, and I. Stefanov. (1971). Spline transformations: three new diagnostic aids for the statistical data-analyst. *J. Roy. Statist. Soc.*, **B-33**, 1–37, discussion 37–70.

Curnow, R. N. (1973). A smooth population response curve based on an abrupt threshold and plateau model for individuals. *Biometrics*, **29**, 1–10.

Ertel, J. E. and E. B. Fowlkes. (1976). Some algorithms for linear spline and piecewise multiple linear regression. *J. Am. Statist. Assoc.*, **71**, 640–648.

Fuller, W. A. (1969). Grafted polynomials as approximating functions. *Australian J. Agric. Econ.*, **13**, 35–46.

Gallant, A. R. and W. A. Fuller. (1973). Fitting segmented polynomial regression models whose join points have to be estimated. *J. Am. Statist. Assoc.*, **68**, 144–147.

Graybill, F. A. and D. C. Bowden. (1967). Linear segment confidence bands for simple linear models. *J. Am. Statist. Assoc.*, **62**, 403–408.

Greville, T. N. E. (1969). *Theory and Applications of Spline Functions*, New York: Academic Press.

Guthery, S. B. (1974). Partition regression. *J. Am. Statist. Assoc.*, **69**, 945–947.

Halpern, E. F. (1973). Bayesian spline regression when the number of knots is unknown. *J. Roy. Statist. Soc.*, **B-35**, 347–360.

Hinkley, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika*, **56**, 495–504.

Hinkley, D. V. (1971). Inference in two-phase regression. *J. Am. Statist. Assoc.*, **66**, 736–748.

Hudson, D. J. (1966). Fitting segmented curves whose join points have to be estimated. *J. Am. Statist. Assoc.*, **61**, 1097–1129.

Lerman, P. M. (1980). Fitting segmented regression models by grid search. *Applied Statistics*, **29**, 77–84.

McGee, F. E. and W. T. Carleton (1970). Piecewise regression. *J. Am. Statist. Assoc.*, **65**, 1109–1124.

Poirier, D. J. (1973). Piecewise regression using cubic splines. *J. Am. Statist. Assoc.*, **68**, 515–524, corrections (1974), **69**, 288.

Quandt, R. E. (1958). The estimation of the parameter of a linear regression system obeying two separate regimes. *J. Am. Statist. Assoc.*, **53**, 873–880.

Quandt, R. E. (1960). Test of the hypothesis that a linear regression system obeys two separate regimes. *J. Am. Statist. Assoc.*, **55**, 324–330.

Quandt, R. E. (1972). A new approach to estimating switching regressions. *J. Am. Statist. Assoc.*, **67**, 306–310.

Robison, D. E. (1964). Estimates for the points of intersection of two polynomial regressions. *J. Am. Statist. Assoc.*, **59**, 214–224.

Shaban, S. A. (1980). Change point problem and two-phase regression: an annotated bibliography. *International Statistical Review*, **48**, 83–93.

Sprent, P. (1961). Some hypotheses concerning two-phase regression lines. *Biometrics*, **17**, 634–645.

Watts, D. G. and D. W. Bacon. (1974). Using a hyperbola as a transition model to fit two-regime straight-line data. *Technometrics*, **16**, 369–373.

Глава 6

A. Выбор переменных и ридж-регрессия

Aitkin, M. A. (1974). Simultaneous inference and the choice of variable subsets in multiple regression. *Technometrics*, **16**, 221–227.

Alldredge, J. R. and N. S. Gilb. (1976). Ridge regression: an annotated bibliography. *International Statist. Rev.*, **44**, 355–360.

Allen, D. M. (1971). Mean square error of prediction as a criterion for selecting variables. *Technometrics*, **13**, 469–475; discussion, 477–481.

Allen, D. M. (1974). The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125–127.

Anderson, D. A. and R. G. Scott. (1974). The application of ridge regression analysis to a hydrologic target-control model. *Water Resources Bulletin*, **10**, 680–690.

Anderson, R. L., D. M. Allen, and F. B. Cady. (1972). Selection of predictor variables in linear multiple regression. *Statistical Papers in Honor of George Snedecor*, ed. T. A. Bancroft, Iowa State University Press.

Anscombe, F. J. (1967). Topics in the investigation of linear relations fitted by the method of least squares. *J. Roy. Statist. Soc.*, **B-29**, 1–29, discussion, 29–52.

Banerjee, K. S., and R. N. Carr. (1971). A comment on ridge regression. Biased estimator for non-orthogonal problems. *Technometrics*, **13**, 895–898.

Barnard, G. A. (1977). On ridge regression, and the general principles of estimation. *Utilitas Mathematica*, **11**, 299–311.

Beale, E. M. L. (1970). Note on procedures for variable selection in multiple regression. *Technometrics*, **12**, 909–914.

Beale, E. M. L. and P. C. Hutchinson. (1974). Note on constrained optimum regression. *Appl. Statist.*, **23**, 208–210.

Beale, E. M. L., M. G. Kendall, and D. W. Mann. (1967). The discarding of variables in multivariate analysis. *Biometrika*, **54**, 357–366.

Bendel, R. B. and A. A. Afifi. (1977). Comparison of stopping rules in forward "stepwise" regression. *J. Amer. Statist. Assoc.*, **72**, 46–53.

Berk, K. N. (1978). Comparing subset regression procedures. *Technometrics*, **20**, 1–6.

Blair, E. A., D. M. Brown, and A. Wilson. (1971). Zig-zag regression. *Amer. Statisticians*, **25**, 56.

Bock, M. E., T. A. Yancey, and G. G. Judge. (1973). The statistical consequences of preliminary test estimators in regression. *J. Am. Statist. Assoc.*, **68**, 109–116.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc.*, **A-143**, 383–404, discussion 404–430.

Brown, W. G. and B. R. Beattie. (1975). Improving estimates of economic parameters by use of ridge regression with production function applications. *Am. J. Agric. Econ.*, **57**, 21–32.

Bunke, O. (1975). Least squares estimators as robust and mimimax estimators. *Math. Operationsforsch. u. Statist.*, **6**, 687–688.

Bunke, O. (1975). Improved inference in linear models with additional information. *Math. Operationsforsch u. Statist.*, **6**, 817–829.

Cochran, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *J. Roy. Statist. Soc.*, **B-5**, 171–176.

Coniffe, D. and J. Stone. (1973). A critical view of ridge regression. *The Statistician*, **22**, 181–187.

Cox, D. R. (1960). Regression analysis when there is prior information about supplementary variables. *J. Roy. Statist. Soc.*, **B-22**, 172–176.

Daling, J. R. and H. Tamura. (1970). Use of orthogonal factors for selection of variables in a regression equation—an illustration. *Appl. Statist.*, **19**, 260–268.

Darlington, R. B. (1978). Reduced-variance regression. *Psychol. Bull.*, **85**, 1238–1255.

Dempster, A. P., M. Schatzoff, and N. Wermuth. (1977). A simulation study of alternatives to least squares. *J. Am. Statist. Assoc.*, **72**, 77–106.

Draper, N R , and R C Van Nostrand (1979) Ridge regression and James-Stein estimation review and comments *Technometrics*, **21** 451–466

Draper N R I Guttman and H Kanemasu (1971) The distribution of certain regression statistics *Biometrika*, **58**, 295–298

Duncan, D B (1970) Multiple comparison methods for comparing regression coefficients *Biometrics*, **26**, 141–143

Durbin, J (1953) A note on regression when there is extraneous information about one of the coefficients *J Am Statist Assoc* , **48**, 799–808

Dwivedi, T D , V K Srivastava and R L Hall (1980) Finite sample properties of ridge estimators *Technometrics* **22** 205–212

Feig, D G (1978) Ridge regression when biased estimation is better *Soc Sci Q* , **58**, 708–716

Forsythe, A B , L Engelman, R Jennrich, and P R A May (1973) A stopping rule for variable selection in multiple regression *J Am Statist Assoc* , **68**, 75–77

Furnival, G M and R W Wilson (1974) Regression by leaps and bounds *Technometrics*, **16**, 499–511

Garside, M J (1965) The best subset in multiple regression analysis *Appl Statist* , **14**, 196–200

Goldberger, A S (1961) Stepwise least-squares residual analysis and specification *J Am Statist Assoc* , **56**, 998–1000

Goldstein, M and A F M Smith (1974) Ridge-type estimators for regression analysis *J Roy Statist Soc* , **B-36**, 284–291

Gorman, J W and R J Toman (1966) Selection of variables for fitting equations to data *Technometrics*, **8**, 27–51

Guilkey, D K and J L Murphy (1975) Directed ridge regression techniques in cases of multicollinearity *J Am Statist Assoc* , **70**, 769–775

Gunst, R F and R L Mason (1979) Some considerations in the evaluation of alternate prediction equations *Technometrics*, **21**, 55–63

Hagar, H and C Antle (1968) The choice of the degree of a polynomial model *J Roy Statist Soc* , **B-30**, 469–471

Haitovsky, Y (1969) A note on the maximization of R^2 *Am Statist* , **23**, 20–21

Halpern, E F (1973) Polynomial regression from a Bayesian approach *J Am Statist Assoc* , **68**, 137–143

Hawkins, D M (1973) On the investigation of alternative regressions by principal component analysis *Appl Statist* , **22**, 275–286

Helms, R W (1974) The average estimated variance criterion for the selection-of-variables problem in general linear models *Technometrics*, **16**, 261–273

Hocking, R R (1972) Criteria for selection of a subset regression which one should be used? *Technometrics*, **14**, 967–970

Hocking, R R (1974) Misspecification in regression *Am Statist* , **28**, 39–40

Hocking, R R (1976) The analysis and selection of variables in linear regression *Biometrics*, **32**, 1–51

Hocking, R R and R N Leslie (1967) Selection of the best subset in regression analysis *Technometrics*, 9, 531–540

Hocking, R R , F M Speed, and M J Lynn (1976) A class of biased estimators in linear regression *Technometrics*, **18**, 425–438

Hoerl, A E (1962) Application of ridge analysis to regression problems *Chem Eng Prog* , **58**, 54 59

Hoerl, A E and R W Kennard (1970) Ridge regression biased estimation for non-orthogonal problems *Technometrics*, **12**, 55 67

Hoerl, A E and R W Kennard (1970) Ridge regression applications to non-orthogonal problems *Technometrics* **12** 69–82, correction **12** 723

Hoerl, A E and R W Kennard (1975) A note on a power generalization of ridge regression *Technometrics*, **17**, 269

Hoerl, A E, R W Kennard, and K F Baldwin (1975) Ridge regression some simulations *Comm Statist*, **4**, 105–123

Hotelling, H (1933) Analysis of a complex of statistical variables into principal components *J Ed Psych*, **24**, 417–441, 489–520

Jolliffe I T (1972, 1973) Discarding variables in a principal component analysis I Artificial data *Appl Statist* **21** 160–173 II Real data **22**, 21–31

Jones, T A (1972) Multiple regression with correlated independent variables *Math Geol*, **4**, 203–218

Kennard, R W (1971) A note on the C_p statistic *Technometrics*, **13**, 899 900, corrections, **15**, 657, (1973)

Kennedy, W J and T A Bancroft (1971) Model building for prediction in regression based upon repeated significance tests *Ann Math Stat*, **42**, 1273–1284

Khuri, A I and R H Meyers (1979) Modified ridge analysis *Technometrics*, **21**, 467–474

LaMotte, L R (1972) The SELECT routines a program for identifying best subset regressions *Appl Statist* **21** 92–93

Lawless J F and P Wang (1976) A simulation study of ridge and other regression estimators *Comm Statist Theory Methods*, **A5**, 307–323

Lindley, D V (1968) The choice of variables in multiple regression *J Roy Statist Soc*, **B-30**, 31–53, discussion, 54–66

Lindley, D V and A F M Smith (1972) Bayes estimates for the linear model *J Roy Statist Soc*, **B-34**, 1–18, discussion, 18–41

Lowerre, J (1974) On the mean square error of parameter estimates for some biased estimators *Technometrics*, **16**, 461–464

Lund, I A (1971) An application of stagewise and stepwise regression procedures to a problem of estimating precipitation in California *J Appl Meteorol* **10** 892–902

Mahajan, V, A K Jain and M Bergier (1977) Parameter estimation in marketing models in the presence of multicollinearity an application of ridge regression *J Market Res*, **14**, 586–591

Mallows C L (1973) Some comments on C_p *Technometrics* **15**, 661–675

Mantel N (1970) Why stepdown procedures in variable selection *Technometrics*, **12** 621–625

Mantel N (1971) More on variable selection procedures and an alternative approach, (letter to the editor) *Technometrics* **13** 455 457

Marquardt D W (1970) Generalized inverses ridge regression biased linear estimation, and nonlinear estimation *Technometrics* **12** 591 612

Marquardt, D W and R D Snee (1975) Ridge regression in practice *Am Statist*, **29**, 3 19

Mason R and W G Brown (1975) Multicollinearity problems and ridge regression in sociological models *Soc Sci Res* **4** 135–149

Mayer L S and T A Willke (1973) On biased estimation in linear models *Technometrics*, **15**, 497 508

McDonald, G. C. and R. C. Schwing. (1973). Instabilities of regression estimates relating air pollution to mortality. *Technometrics*, **15**, 463–481.

McKay, R. J. (1979). The adequacy of variable subsets in multivariate regression. *Technometrics*, **21**, 475–480.

Mikhail, W. M (1972). The bias of the two-stage least squares estimator. *J. Am. Statist. Assoc.*, **67**, 625–627.

Miller, W. L. (1972). Measures of electoral change using aggregate data. *J. Roy. Statist. Soc.*, **A-135**, 122–142.

Miller, W. L. (1978). Social class and party choice in England: a new analysis. *Brit. J. Polit. Sci.*, **8**, 257–284.

Morrison, D. F. (1976). *Multivariate Statistical Methods*, 2nd ed., New York: McGraw-Hill.

Narula, S. C. and J. F. Wellington. (1977). Prediction, linear regression, and the minimum sum of relative errors. *Technometrics*. **19**, 185–190. See also **22**, pp. 450 and 452.

Narula, S. C. and J. F. Wellington. (1979). Selection of variables in linear regression using the minimum sum of weighted absolute errors criterion. *Technometrics*, **21**, 299–306; correction, **22**, 452.

Newton, R. G. and D. J. Spurrell. (1967). A development of multiple regression for the analysis of routine data. *Appl. Statist.*, **16**, 51–64.

Newton, R. G. and D. J. Spurrell. (1967). Examples of the use of elements for clarifying regression analysis. *Appl. Statist.*, **16**, 165–172.

Newton, R. G. and D. J. Spurrell. (1968). Developments in the use of element analysis. Part I: A statement of the present position as illustrated by some blast furnace data. Part II: The selection of "best sub-sets," British Glass Industry Research Association, Sheffield, England.

Obenchain, R. L. (1977). Classical *F*-tests and confidence regions for ridge regression. *Technometrics*, **19**, 429–439.

Obenchain, R. L. (1978). Good and optimal ridge estimators. *Ann. Statist.*, **6**, 1111–1121.

Obenchain, R. L. (1980). Data analytic displays for ridge regression. University of Wisconsin Statistics Department Technical Report No. 605, April.

Obenchain, R. L. (1981). Generalized ridge regression computations: formulas and remarks. University of Wisconsin Statistics Department Technical Report, in preparation.

Pope, P. T. and J. T. Webster. (1972). The use of an *F*-statistic in stepwise regression procedures. *Technometrics*, **14**, 327–340.

Rao, P. (1971). Some notes on misspecification in multiple regressions. *Am. Statist.*, **25**, 37–39.

Rencher, A. C. and F. C. Pun. (1980). Inflation of R^2 in best subset regression. *Technometrics*, **22**, 49–53.

Ruse, A. (1973). Goodness of fit [R^2] in generalized least squares estimation. *Am. Statist.*, **27**, 106–108.

Smith, G. and F. Campbell. (1980). A critique of some ridge regression methods. *J. Am. Statist. Assoc.*, **75**, 74–81; discussion, 81–103.

Snee, R. D. (1973). Some aspects of nonorthogonal data analysis, Part I. Developing prediction equations. *J. Quality Technol.*, **5**, 67–79.

Spjotvoll, E. (1972). Multiple comparison of regression functions. *Ann. Math. Stat.*, **43**, 1076–1088.

Theobold, C. M. (1974). Generalizations of mean square error applied to ridge regression. *J. Roy. Statist. Soc.*, **B-36**, 103–106.

Thisted, R. A. and Morris, C. N. (1979). Theoretical results for adaptive ordinary ridge regression estimators. Technical Report No. 94, University of Chicago Department of Statistics.

Valiaho, H. (1969). A synthetic approach to stepwise regression analysis. *Comm. Phys.-Math.*, **34**, No. 12, 31–131.

Walls, R. C. and D. L. Weeks. (1969). A note on the variance of a predicted response in regression. *Am. Statist.*, **23**, 24–26.

Warren, W. G. (1973). On partial correlation. *Am. Statist.*, **27**, 239.

Webster, J. T., R. F. Gunst, and R. L. Mason. (1974). Latent root regression analysis. *Technometrics*, **16**, 513–522.

White, J. W. and R. F. Gunst. (1979). Latent root regression: large sample analysis. *Technometrics*, **21**, 481–488.

Yale, C. and A. B. Forsythe. (1976). Winsorized regression. *Technometrics*, **18**, 291–300.

B. Робастная регрессия

Abraham, B. and G. E. P. Box. (1978). Linear models and spurious observations. *Appl. Statist.* **27**, 131–138.

Andrews, D. F. (1974). A robust method for multiple linear regression. *Technometrics*, **16**, 523–531.

Andrews, D. F., P. J. Bickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton, NJ: Princeton University Press.

Boos, D. D. (1980). A new method for constructing approximate confidence intervals from *M* estimates. *J. Am. Statist. Assoc.*, **75**, 142–145.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. *Robustness in Statistics*, eds. R. L. Launer and G. N. Wilkinson, pp. 201–236. New York: Academic Press.

Box, G. E. P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *J. Roy. Statist. Soc.*, **A-143**, 383–404, discussion 404–430.

Box, G. E. P. and G. C. Tiao. (1964). A note on criterion robustness and inference robustness. *Biometrika*, **51**, 169–173.

Box, G. E. P. and G. C. Tiao. (1968). A Bayesian approach to some outlier problems. *Biometrika*, **55**, 119–129.

Chen, G. G. (1979). *Studies in Robust Estimation.* University of Wisconsin-Madison Ph.D. Thesis.

Crow, E. L. and M. M. Siddiqui. (1967). Robust estimates of location. *J. Am. Statist. Assoc.*, **62**, 353–389.

Dixon, W. J. (1950). Analysis of extreme values. *Ann. Math. Statist.*, **21**, 27–58.

Dixon, W. J. (1953). Processing data for outliers. *Biometrics*, **9**, 74–89.

Ferguson, T. S. (1961). On the rejection of outliers. *Proc. Fourth Berkeley Symp.*, **1**, 253–287.

Gastwirth, J. L. and M. L. Cohen. (1970). Small sample behavior of some robust linear estimators of location. *J. Am. Statist. Assoc.*, **65**, 946–973.

Geary, R. C. (1947). Testing for normality. *Biometrika*, **34**, 209–242.

Hinich, M. J. and P. P. Talwar. (1975). A simple method for robust regression. *J. Am. Statist. Assoc.*, **70**, 113–119.

Hodge, J. L., Jr. (1967). Efficiency in normal samples and tolerance of extreme values for some estimates of location. *Proc Fifth Berkeley Symp.*, **1**, 163–186.

Hogg, R. V. (1967). Some observations on robust estimation. *J. Am. Statist. Assoc.*, **62**, 1179 1186.

Hogg, R. V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *J. Am. Statist. Assoc.*, **69**, 909–927.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.*, **35**, 73–101.

Huber, P. J. (1972). Robust statistics: a review. *Ann. Math. Statist.*, **43**, 1041–1067.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.*, **1**, 799–821.

Jaeckel, L. A. (1971). Some flexible estimates of location. *Ann. Math. Statist.*, **42**, 1540–1552.

Moberg, T. F., Ramberg, J. S., and Randles, R. H. (1980). An adaptive multiple regression procedure based on *M*-estimators. *Technometrics*, **22**, 213–224.

Siddiqui, M. M. and K. Raghunandanan. (1967). Asymptotically robust estimators of location. *J. Am. Statist. Assoc.*, **62**, 950–953.

Stigler, S. M. (1977). Do robust estimators work with real data? *Ann. Statist.*, **5**, 1055–1098.

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In *Contributions to Probability and Statistics*, ed. I. Olkin, Stanford, CA: Stanford University Press.

Yohai, V. J. (1974). Robust estimation in the linear model. *Ann. Statist.*, **2**, 562–567.

Глава 7

Arrow, K. J. and M. Hoffenberg. (1959). *A Time Series Analysis of Interindustry Demands.* Amsterdam: North Holland Publishing Co.

Awerbuch, S., H. Smith, and W. A. Wallace. (1974). Regression analysis as an aid in managing a marine environmental protection program. *J. Environ. Sys.*, **4**, 143–153.

Box, G. E. P. (1954). The exploration and exploitation of response surfaces: some general considerations and examples. *Biometrics*, **10**, 16–60. (См. также работы этого автора и его же с соавторами: *J. Roy. Statist. Soc.*, **B-13**, 1951; *Biometrics*, **11**, 1955; *Ann. Math. Statist.*, **28**, 1957; *J. Am. Statist. Assoc.*, **54**, 1959 Кроме того см. Hill and Hunter (1966) and Mead and Pike (1975))

Eisenhart, C. (1939). Interpretation of certain regression methods and their use in biological and industrial research. *Ann. Math. Statist.*, **10**, 162–186.

Ericksen, E. P. (1974). A regression method for estimating population changes of local areas. *J. Am. Statist. Assoc.*, **69**, 867–875.

Feldstein, M. S. (1966). A binary variable multiple regression method of analyzing factors affecting peri-natal mortality and other outcomes of pregnancy. *J. Roy. Statist. Soc.*, **A-129**, 61–73.

Fieller, E. C. (1940). The biological standardization of insulin. *J. Roy. Statist. Soc.*, **B-7**, 1–54, discussion 54–64.

Fisher, R. A. (1924). The influence of rainfall on the yield of wheat at Rothamsted. *Phil. Trans. Roy. Soc.*, **213**, 89–142.

Frye, H. W. and J. D. Pugh. (1971). A new equation for the speed of sound in seawater. *J. Acoust. Soc. Am.*, **50**, 384–386.

Griffiths, D. O. (1968). The use of regression analysis in a depot location exercise. *Appl. Statist.*, **17**, 57–63.

Hill, W. J. and W. G. Hunter. (1966). A review of response surface methodology: a literature survey. *Technometrics*, **8**, 571–590.

Housworth, W. J. (1972). Hybrid polynomial and periodic regression with and without missing observations. *Biometrics*, **28**, 1025–1042.

Mandel, J. (1969). A method for fitting empirical surfaces to physical or chemical data. *Technometrics*, **11**, 411–429.

Mazur, D. P. (1972). Using regression models to estimate the expectation of life in the U.S.S.R. *J. Am. Statist. Assoc.*, **67**, 31–36.

Mead, R. and D. J. Pike. (1975). A review of response surface methodology from a biometric viewpoint. *Biometrics*, **31**, 803–851.

Prentice, R. L. and L. A. Gloeckler. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, **34**, 57–68.

Sorenson, F. A. Methods of performing multiple regression analysis. United States Steel Technical Report, (90.10–100G), Applied Research Laboratory, Monroeville, PA.

Stephenson, J. A. and H. T. Farr. (1972). Seasonal adjustment of economic data by application of the general linear statistical model. *J. Am. Statist. Assoc.*, **67**, 37–45.

Thomson, L. M. (1963). Weather and technology in the production of corn and soybeans. CAED Report 17, Iowa State University, Ames. IA.

Thonstad, T. and D. B. Jochems. (1961). The influence of entrepeneurial appraisals and expectations on production planning. *Int. Econ. Rev.*, **2**, 135–152.

Turnbull, P. and G. Williams. (1974). Sex differentials in teachers' pay. *J. Roy. Statist. Soc.*, **A-137**, 245–258.

Weber, D. C. (1971). Accident rate potential: an application of multiple regression analysis of a poisson process. *J. Am. Statist. Assoc.*, **66**, 285–288.

Williams, D. A. (1970). Discrimination between regression models to determine the pattern of enzyme synthesis in synchronous cell cultures. *Biometrics*, **23**, 23–32.

Глава 8

Allen, D. M. (1971). The prediction sum of squares as a criterion for selecting predictor variables. Technical Report No. 23, Department of Statistics, University of Kentucky.

Chambers, J. M. (1971). Regression updating. *J. Amer. Statist. Assoc.*, **66**, 744–748.

Finifter, B. M. (1972). The generation of confidence: evaluating research findings by random subsample replication. In *Sociological Methodology*, ed. Herbert L. Costner, San Francisco: Jossey-Bass.

Garbade, K. (1977). Two methods for examining the stability of regression coefficients. *J. Am. Statist. Assoc.*, **72**, 54–63.

Gardner, M. J. (1972). On using an estimated regression line in a second sample. *Biometrika*, **59**, 263–274.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Am. Statist. Assoc.*, **70**, 320–328.

Kendall, M. G. (1951). Regression, structure, and functional relationships, Part I. *Biometrika*, **38**, 11–25.

Kendall, M. G. (1952). Regression, structure, and functional relationships, Part II. *Biometrika*, **39**, 96–108.

Kish, L. and M. R. Frankel. (1974). Inference from complex samples. *J. Roy. Statist. Soc.*, **B-36**, 1–37.

Lindley, D. V. (1947). Regression lines and the linear functional relationship. *J. Roy. Statist. Soc.*, **B-9**, 218–244.

Lindley, D. V. (1953). Estimation of a functional relationship. *Biometrika*, **40**, 47–49.

Lovell, M. C. and E. Prescott. (1970). Multiple regression with inequality constraints: pretesting bias, hypothesis testing and efficiency. *J. Am. Statist. Assoc.*, **65**, 913–925.

McCarthy, P. J. (1976). The use of balanced half-sample replication in cross-validation studies. *J. Am. Statist. Assoc.*, **71**, 596–604.

Mosteller, F. and J. W. Tukey. (1968). Data analysis, including statistics. In *Handbook of Social Psychology*, Vol. 2, eds. G. Lindzey and E. Aronson. Reading MA: Addison-Wesley.

Nelder, J. A. (1968). Regression, model-building and invariance. *J. Roy. Statist. Soc.*, **A-131**, 303–315, discussion 315–329.

Novick, M. R., P. H. Jackson, D. T. Thayer, and N. S. Cole. (1972). Estimating multiple regression in *m* groups: a cross-validation study. *Brit. J. Math. Statist. Psychol.*, **25**, 33–50.

Park, C. N. and A. L. Dudycha. (1974). A cross-validation approach to sample size determination for regression models. *J. Am. Statist. Assoc.*, **69**, 214–218.

Snee, R. D. (1977). Validation of regression models: methods and examples. *Technometrics*, **19**, 415–428.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc.*, **B-36**, 111–147.

Swindel, B. F. (1974). Instability of regression coefficients illustrated. *Am. Statist.*, **28**, 63–65.

Глава 9

Anderson, V. L. and R. A. McLean. (1974). Restriction errors: another dimension in teaching experimental statistics. *Am. Statist.*, **28**, 145–152.

Box, G. E. P. (1954). The exploration of response surfaces: Some general considerations and examples. *Biometrics*, **10**, 16–60.

Box, G. E. P. (1963). The effects of errors in the factor levels and experimental design. *Technometrics*, **5**, 247–262.

Carter, W. H. and R. H. Myers. (1972). Orthogonal contrasts and the generalized inverse in fixed effects analysis of variance. *Am. Statistician*, **26**, 32–34.

Duncan, D. B. and M. Walser. (1966). Multiple regression combining within-and between-plot information. *Biometrics*, **22**, 26–43.

Francis, I. (1973). A comparison of several analysis of variance programs. *J. Am. Statist. Assoc.*, **68**, 860–865.

Hemmerle, W. J. (1974). Nonorthogonal analysis of variance using iterative improvement and balanced residuals. *J. Am. Statist. Assoc.*, **69**, 772–778.

Kutner, M H. (1974). Hypothesis testing in linear models (Eisenhart model I). *Am Statist.*, **28**, 98–100.

Mallios, W. S. (1967). A structural regression approach to covariance analysis when the covariable is uncontrolled. *J. Am. Statist. Assoc.*, **62**, 1037–1049.

Mantell, E. H. (1973). Exact linear restrictions on parameters in the classical linear regression model. *Am. Statist.*, **27**, 86-87. См. также **28**, p. 36 (1974).

Read, D. R. (1954). The design of chemical experiments. *Biometrics*, **10**, 1–15.

Schilling, E. G. (1974). The relationship of analysis of variance to regression, Part I. Balanced designs. *J. Quality Technol.*, **6**, 74–83.

Schilling, E. G. (1974). The relationship of analysis of variance to regression, Part II. Unbalanced designs. *J. Quality Technol.*, **6**, 146–153.

Searle, S. R. (1971). Topics in variance component estimation. *Biometrics*, **27**, 1–76. See p. 750.

Seegrist, D. W. (1973). Least squares analysis of experimental design models by augmenting the data with side conditions. *Technometrics*, **15**, 643–645.

Smith, H. (1969). The analysis of data from a designed experiment. *J. Quality Technol.*, **1**, 259–263.

Snee, R. D. (1973). Some aspects of nonorthogonal data analysis, Part I. Developing prediction equations. *J. Quality Technol.*, **5**, 67–79; Part II. Comparison of means, **5**, 109–122.

Swamy, P. A. V. B. and J. S. Mehta. (1973). Bayesian analysis of error components regression models. *J. Am. Statist. Assoc.*, **68**, 648–658.

Глава 10

A. Общие работы по нелинейному оцениванию

Agha, M. (1971). A direct method for fitting linear combinations of exponentials. *Biometrics*, **27**, 399–413.

Baily, R. C., G. S. Eadie, and F. H. Schmidt. (1974). Estimation procedures for consecutive first order irreversible reactions. *Biometrics*, **30**, 67–75.

Bard, Y. (1974). *Nonlinear Parameter Estimation*. New York and London: Academic Press.

Barham, R. H. and W. Drane. (1972). An algorithm for least squares estimation of nonlinear parameters when some of the parameters are linear. *Technometrics*, **14**, 757–766.

Bates, D. M. and D. G. Watts. (1980). Relative curvature measures of nonlinearity. *J. Roy. Statist. Soc.* **B-42**, 1–16; discussion, 16–25.

Beale, E. M. L. (1960). Confidence regions in non-linear estimation. *J. Roy. Statist. Soc.*, **B-22**, 41–76.

Beck, J. V. and K. J. Arnold. (1977). *Parameter Estimation in Engineering and Science*. New York: Wiley.

Behnken, D. W. (1964). Estimation of copolymer reactivity ratios: an example of non-linear estimation. *J. Polym. Sci.*, **A-2**, 645–668.

Blakemore, J. W. and A. W. Hoerl. (1963). Fitting non-linear reaction rate equations to data. *Chem. Eng. Prog. Symp. Ser.*, **59**, 14–27.

Box, G. E. P. (1958). Use of statistical methods in the elucidation of basic mechanism. *Bull. Int. Statist. Inst.*, **36**, 215–22.

Box, G. E. P. and W. G. Hunter. (1962). A useful method for model building. *Technometrics*, **4**, 301–318.

Box, G. E. P. and W. G. Hunter. (1965). The experimental study of physical mechanisms. *Technometrics*, **7**, 23–42.

Box, G. E. P. and H. L. Lucas. (1959). Design of experiments in non-linear situations. *Biometrika*, **46**, 77–90.

Box, M. J. (1971). Bias in nonlinear estimation. *J. Roy. Statist. Soc.*, **B-33**, 171–190, discussion, 190–201.

Carroll, C. W. (1961). The created response surface technique for optimizing non-linear restrained systems. *Operations Res.*, **9**, 169–185.

Chambers, J. M. (1973). Fitting nonlinear models: numerical techniques. *Biometrika*, **60**, 1–13.

Chen, E. H. and W. J. Dixon. (1972). Estimates of parameters of a censored regression sample. *J. Am. Statist. Assoc.*, **67**, 664–671.

Cochran, W. G. (1973). Experiments for nonlinear functions. *J. Am. Statist. Assoc.*, **68**, 771–781.

Cornell, R. G. (1962). A method for fitting linear combinations of exponentials. *Biometrics*, **18**, 104–113.

Curry, H. B. (1944). The method of steepest descent for non-linear minimization problems. *Q. Appl. Math.*, **2**, 258–261.

Della Corte, M., L. Buricchi, and S. Romano. (1974). On a fitting of linear combinations of exponentials. *Biometrics*, **30**, 367–369.

Dolby, G. R. (1972). Generalized least squares and maximum likelihood estimation on nonlinear functional relationships. *J. Roy. Statist. Soc.*, **B-34**, 393–400.

Duncan, G. T. (1978). An empirical study of jacknife-constructed confidence regions in nonlinear regression. *Technometrics*, **20**, 123–129.

Flanagan, P. D., P. A. Vitale, and J. Mendelsohn. (1969). A numerical investigation of several one-dimensional search procedures in nonlinear regression problems. *Technometrics*, **11**, 265–284.

Fletcher, R. and M. J. D. Powell. (1963). A rapidly convergent descent method for minimization. *Computer J.*, **6**, 163–168.

Fox, T., D. Hinkley, and K. Larntz. (1980). Jackknifing in nonlinear regression. *Technometrics*, **22**, 29–33.

Frome, E. L., M. H. Kutner, and J. J. Beauchamp. (1973). Regression analysis of Poisson-distributed data. *J. Am. Statist. Assoc.*, **68**, 935–940.

Gallant, A. R. (1968). A note on the measurement of cost/quantity relationships in the aircraft industry. *J. Am. Statist. Assoc.*, **63**, 1247–1252.

Gallant, A. R. (1975). Testing a subset of the parameters of a non-linear regression model. *J. Am. Statist. Assoc.*, **70**, 927–932.

Gehan, E. A. and M. M. Siddiqui. (1973). Simple regression methods for survival time studies. *J. Am. Statist. Assoc.*, **68**, 848–856.

Goldfeld, S. M. and R. E. Quandt. (1972). *Nonlinear Methods in Economics*. Amsterdam: North-Holland.

Guttman, I. and D. A. Meeter. (1964). Use of transformations on parameters in non-linear theory. I. Transformations to accelerate convergence in non-linear least squares. Technical Report No. 37, Department of Statistics, University of Wisconsin, Madison, WI.

Guttman, I. and D. A. Meeter. (1965). On Beale's measures of nonlinearity. *Technometrics*, **7**, 623–637.

Guttman, I., V. Pereyra, and H. D. Scolnik. (1973). Least squares estimation for a class of nonlinear models. *Technometrics*, **15**, 209–218.

Halperin, M. (1962). Confidence interval estimation in non-linear regression. Program 360, Applied Mathematics Department SRRC-RR-62-28, Sperry-Rand Research Center.

Hartley, H. O. (1948). The estimation of non-linear parameters by "internal least squares." *Biometrika*, **35**, 32–45.

Hartley, H. O. (1961). The modified Gauss-Newton method for the fitting of non-linear regression functions by least squares. *Technometrics*, **3**, 269–280.

Hartley, H. O. (1964). Exact confidence regions for the parameters in non-linear regression laws. *Biometrika*, **51**, 347–353.

Hartley, H. O. and A. Booker. (1965). Non-linear least squares estimation. *Ann. Math. Statist.*, **36**, 638–650.

Harville, D. A. (1973). Fitting partially linear models by weighted least squares. *Technometrics*, **15**, 509–515.

Heien, D. M. (1968). A note on log-linear regression. *J. Am. Statist. Assoc.*, **63**, 1034–1038.

Himmelbau, D. M. (1970). *Process Analysis by Statistical Methods*. New York: Wiley.

Hunter, W. G. (1963). Generation and analysis of data in non-linear situations. Ph.D. Thesis, University of Wisconsin, Madison, WI.

Hunter, W. G. and R. Mezaki. (1964). A model-building technique for chemical engineering kinetics. *Am. Inst. Chem. Eng. J.*, **10**, 315–322.

Hunter, W. G. and A. M. Reiner. (1965). Designs for discriminating between two rival models. *Technometrics*, **7**, 307–323.

Jennrich, R. I. (1969). Asymptotic properties of nonlinear least squares estimators. *Ann. Math. Statist.*, **40**, 633–649.

Jennrich, R. I. and P. F. Sampson. (1968). An application of stepwise regression to nonlinear estimation. *Technometrics*, **10**, 63–72.

Kubicek, M., M. Marek, and E. Eckert. (1971). Quasilinearized regression. *Technometrics*, **13**, 601–608.

Lawton, W. H. and E. A. Sylvestre. (1971). Elimination of linear parameters in nonlinear regression. *Technometrics*, **13**, 461–467; discussion, 477–481.

Lawton, W. H., E. A. Sylvestre, and M. S. Maggio. (1972). Self modeling non-linear regression. *Technometrics*, **14**, 513–532.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Q. Appl. Math.*, **2**, 164–168.

Marquardt, D. W. (1959). Solution of non-linear chemical engineering models. *Chem. Eng. Prog.*, **55**, 65–70.

Marquardt, D. W. (1963). An algorithm for least squares estimation of non-linear parameters. *J. Soc. Ind. Appl. Math.*, 11, 431–441.

Marquardt, D. W., R. G. Bennett, and E. J. Burrell. (1961). Least-squares analysis of electron paramagnetic resonance spectra. *J. Mol. Spectr.*, **7**, 269–279.

Moore, R. H. and R. K. Zeigler. (1967). The use of nonlinear regression methods for analyzing sensitivity and quantal response data. *Biometrics*, **23**, 563–567.

Morrison, D. D. (1960). Methods for non-linear least squares problems and convergence proofs. *Proc. Jet. Propuls. Lab. Sem., Track. Prob. Orbit Determin.*, 1–9.

Nelder, J. A. and R. Mead. (1965). A simplex method for function minimization. *Computer J.*, **7**, 308–313.

Olsson, D. M. (1974). A sequential simplex program for solving minimization problems. *J. Quality Technol.*, **6**, 53–57.

Powell, M. J. D. (1965). A method for minimizing a sum of squares of non-linear functions without calculating derivatives. *Computer J.*, **7**, 303–307.

Rosen, J B. (1960). The gradient projection method for non-linear programming. Part I. Linear constraints. *J. Soc. Ind. Appl. Math.*, **8**, 181–217.

Rosen, J. B. (1961). The gradient projection method for non-linear programming. Part II: Non-linear constraints. *J. Soc. Ind. Appl. Math.*, **9**, 514–532.

Ross, G. J. S. (1970). The efficient use of function minimization in nonlinear maximum-likelihood estimation. *Appl. Statist.*, **19**, 205–221.

Rubin, D. I. (1963). Non-linear least squares parameter estimation and its application to chemical kinetics. *Chem. Eng. Prog. Symp. Ser.*, **59**, 90–94.

Sclove, S. L. (1972). (Y vs X) or (log Y vs X)? *Technometrics*, **14**, 391–403.

Shah, B. K. and C. G. Khatri. (1965). A method of fitting the regression curve $E(y) = \alpha + \delta x + \beta\rho^x$. *Technometrics*, **7**, 59–65.

Smith, F. B. and D. F. Shanno. (1971). An improved Marquardt procedure for nonlinear regression. *Technometrics*, **13**, 63–74.

Smith, H. and S. D. Dubey. (1964). Some reliability problems in the chemical industry. *Ind. Quality Control*, **22**, 64–70.

Spang, H. A. (1962). A review of minimization techniques for non-linear functions. *Soc. Ind. Appl. Math. Rev.*, **4**, 343–365.

Sprent, P. Linear relationships in growth and size studies. *Biometrics*, **24**, 639–656.

Vitale, P. A. and G. Taylor. (1968). A note on the application of Davidon's method to nonlinear regression problems. *Technometrics*, **10**, 843–849.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biametrika*, **61**, 439–447.

Wilk, M. B. (1958). An identity of use in non-linear least squares. *Ann. Math. Statist.*, **29**, 618.

Williams, D. A. (1973). The estimation of relative potency from two parabolas in symmetric bioassays. *Biometrics*, **29**, 695–700.

Williams, E. J. (1962). Exact fiducial distributions in nonlinear estimation, *J. Roy. Statist. Soc.*, **B-24**, 125–139.

Ziegel, E. R. and J. W. Gorman. (1980). Kinetic modelling with multiresponse data. *Technometrics*, **22**, 139–151.

B. Модели роста

Amer, F. A. and W. T. Williams. (1957). Leaf-area growth in Perlargonium zonale. *Ann. Bot. N.S.*, **21**, 339–342.

Bertalanffy, L. von (1941). Stoffwecheseltypen und Wachstumstypen. *Biol. Zentralbl.*, **61**, 510–532.

Bertalanffy, L. von. (1957). Quantitative laws in metabolism and growth. *Q. Rev. Biol.*, **32**, 218–231.

Bliss, C. I. and A. T. James. (1966). Fitting the rectangular hyperbola. *Biometrics*, **22**, 573–602.

Bowden, D. C. and R. K. Steinhorst. (1973). Tolerance bands for growth curves. *Biometrics*, **29**, 361–371.

Brand, R. J., D. E. Pinnock, and K. L. Jackson. (1973). Large sample confidence bands for the logistic response curve and its inverse. *Am. Statist.*, 157–160.

Causton, D. R. (1969). A computer program for fitting the Richards function. *Biometrics*, **25**, 401–409. See also p. 779 (1969).

Colquhoun, D. (1969). A comparison of estimators for a two-parameter hyperbola. *Appl. Statist.*, **18**, 130–140.

Cornell, R. G. and J. A. Speckman (1967). Estimation for a simple exponential model. *Biometrics*, **23**, 717–737.

Day, N. E. (1966). Fitting curves to longitudinal data. *Biometrics*, **22**, 276–291.

Fletcher, R. I. (1974). The quadric law of damped exponential growth. *Biometrics*, **30**, 111–124.

Foss, S. D. (1969). A method for obtaining initial estimates of the parameters in exponential curve fitting. *Biometrics*, **25**, 580–584.

Foss, S D. (1970). A method of exponential curve fitting by numerical integration. *Biometrics*, **26**, 815–821.

Gallucci, V. F. and T. J. Quinn II (1979) Reparameterizing, fitting, and testing a simple growth model *Trans. Am. Fish. Soc.*, **108**, 14–25.

Garg, M. L., B. R. Rao, and C. K. Redmond. (1970). Maximum-likelihood estimation of the parameters of the Gompertz survival function. *Appl. Statist.*, **19**, 152–159.

Glasbey, C. A (1979). Correlated residuals in non-linear regression applied to growth data. *Applied Statistics*, **28**, 251–259.

Gomes, F. P. (1953). The use of Mitscherlich's regression law in the analysis of experiments with fertilizers. *Biometrics*, **9**, 498–516.

Gregory, F. G. (1928). Studies in the energy relations of plants, II. *Ann. Bot.*, **42**, 469–507.

Grizzle, J. E. and D. M. Allen. (1969). Analysis of growth and dose response curves, *Biometrics*, **25**, 357–381.

Hey, E. G. and M. H. Hey. (1960). The statistical estimation of a rectangular hyperbola. *Biometrics*, **16**, 606–617.

Hills, M. (1969). A note on the analysis of growth curves. *Biometrics*, **24**, 189–196.

Hiorns, R. W. (1965). The fitting of growth and allied curves of the asymptotic regression type by Steven's method. *Tracts for Computers*, **28**, University College, London: Cambridge University Press.

Jolicoeur, P. and A. A. Heusner. (1971). The allometry equation in the analysis of the standard oxygen consumption and body weight of the white rat. *Biometrics*, **27**, 841–855.

Krause, G. F., P. B. Siegel, and D. C. Hurst. (1967). A probability structure for growth curves. *Biometrics*, **23**, 217–225.

Llewelyn, F. W. M. (1968). The log (−log) transformation in the analysis of fruit retention records. *Biometrics*, **24**, 627–638.

Medawar, P. B. (1940). The growth, growth energy, and ageing of the chicken's heart. *Proc. Roy. Soc.*, **B-129**, 332–355.

Michelini, C. (1972). Estimating the exponential growth function by direct least squares: a comment. *Appl. Statist.*, **21**, 333–335.

Mitchell, A. F. S. (1968). Exponential regression with correlated observations. *Biometrika*, **55**, 149–162.

Moore, R. H. and R. K. Zeigler. (1967). The use of nonlinear regression methods for analyzing sensitivity and quantal response data. *Biometrics*, **23**, 563–566.

Nair, K. R. (1954). The fitting of growth curves. *Statistics and Mathematics in Biology* Iowa State College Press.

Nelder, J. A. (1961). The fitting of a generalization of the logistic curve. *Biometrics*, **17**, 89–110.

Nelder, J. A. (1968). Weighted regression, quantal response data, and inverse polynomials. *Biometrics*, **24**, 979–985.

Nelder, J. A., R. B. Austin, J. K. A. Bleasdale, and P. J. Salter. (1960). An approach to the study of yearly and other variation in crop yields. *J. Hort. Sci* , **35**, 73–82.

Oliver, F. R. (1964). Methods of estimating the logistic growth function. *Appl. Statist.*, **13**, 57-66.

Oliver, F. R. (1966). Aspects of maximum likelihood estimation of the logistic growth function. *J. Am. Statist. Assoc.*, **61**, 697-705.

Oliver, F. R. (1970). Estimating the exponential growth function by direct least squares. *Appl. Statist.*, **19**, 92-100.

Oliver, F. R. (1970). Some asymptotic properties of Colquhoun's estimators for a rectangular hyperbola. *Appl. Statist.*, **19**, 269-273.

Olsson, D. M. (1972). Fitting two widely useful nonlinear models. *J. Quality Technol.*, **4**, 113-117.

Patterson, H. D. (1956). The use of autoregression in fitting an exponential curve. *Biometrika*, **45**, 389-400.

Patterson, H. D. (1960). A further note on a simple method for fitting an exponential curve. *Biometrika*, **47**, 177-180.

Patterson, H. D. (1969). Baule's equation. *Biometrics*, **25**, 159-164.

Patterson, H. D. and S. Lipton. (1959). An investigation of Hartley's method for fitting an exponential curve. *Biometrika*, **46**, 281-292.

Riffenburgh, R. H. (1966). On growth parameter estimation for early life stages. *Biometrics*, **22**, 162-178.

Richards, F. J. (1959). A flexible growth function for empirical use. *J. Exp. Bot.*, **10**, 290-300.

Shah, B. K. (1961). A simple method of fitting the regression curve $y = \alpha + \delta x + \beta\rho^x$. *Biometrics*, **17**, 651-653.

Shah, B. K. and C. G. Khatri. (1965). A method of fitting the regression curve $E(y) = \alpha + \delta x + \beta\rho^x$. *Technometrics*, **7**, 59-65.

Shah, B. K. and I. R. Patel. (1960). The least squares estimates of the constants for the Makeham Second Modification of Gompertz's law. *J. M. S. Univ. Baroda (India)*, **9**, 1-10.

Spergeon, E. F. (1949). *Life Contingencies*. Cambridge, England: Cambridge University Press.

Sprent, P. (1969). *Models in Regression and Related Topics*. London: Methuen.

Stevens, W. L. (1951). Asymptotic regression. *Biometrics*, **7**, 247-267.

Turner, M. E., B. A. Blumenstein, and J. L. Sebaugh. (1969). A generalization of the logistic law of growth. *Biometrics*, **25**, 577-580.

Wilson, A. L., and A. W. Douglas. (1969). A note on nonlinear curve fitting. *Am. Statist.*, **23**, 37-38.

ДОПОЛНИТЕЛЬНАЯ БИБЛИОГРАФИЯ, СОСТАВЛЕННАЯ ПЕРЕВОДЧИКАМИ

I. ПЕРЕВОД РАБОТ, УПОМЯНУТЫХ В АВТОРСКОЙ БИБЛИОГРАФИИ

1*. Алберт Дж., Нильсон Э., Уолш Дж. Теория сплайнов и ее приложения/Пер. с англ.— М.: Мир, 1972.— 316 с.

2*. Афифи А., Эйзен С. Статистический анализ. Подход с использованием ЭВМ/Пер. с англ. Под ред. Г. П. Башарина.— М.: Мир, 1982.— 488 с.

3*. Бард Й. Нелинейное оценивание параметров/Пер. с англ. Под ред. В. Г. Горского.— М.: Статистика, 1979.— 349 с.

4*. Бокс Дж. Е. П. Устойчивость и стратегии построения научных моделей.//Устойчивые статистические методы оценки данных/Под ред. Р. Л. Лонера, Г. Н. Уилкинсона.; Пер. с англ. Под ред. Н. Г. Волкова.— М.: Машиностроение, 1984, с. 164—188.

5*. Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия/Пер. с англ. Под ред. Ю. П. Адлера.— М.: Финансы и статистика, 1982, вып. 1.— 317 с.; вып. 2.— 239 с.

6*. Рао С. Р. Линейные статистические методы и их применения/Пер. с англ. Под ред. Ю. В. Линника.— М.: Наука, 1968.— 548 с.

7*. Себер Дж. Линейный регрессионный анализ/Пер. с англ. Под ред. М. Б. Малютова.— М.: Мир, 1980.— 456 с.

8*. Хеннан Э. Анализ временных рядов/Пер. с англ. Под ред. Ю. А. Розанова.— М.: Наука, 1964.— 214 с.

9*. Химмельблау Д. Анализ процессов статистическими методами/Пер. с англ. Под ред. В. Г. Горского.— М.: Мир, 1979.— 957 с.

II. ДОПОЛНИТЕЛЬНЫЕ КНИГИ ПО РЕГРЕССИОННОМУ АНАЛИЗУ

1. Алберт А. Регрессия, псевдоинверсия и реккурентное оценивание/Пер. с англ. Под ред. Я. З. Цыпкина.— М.: Наука, 1977.— 224 с.

2. Аренс Х., Лейтер Ю. Многомерный дисперсионный анализ/Пер. с нем.— М.: Финансы и статистика, 1985.— 232 с.

3. Браверман Э. М., Мучник И. Б. Структурные методы обработки эмпирических данных.— М.: Наука, 1983.— 464 с.

4. Венсель В. В. Интегральная регрессия и корреляция: статистическое моделирование рядов динамики.— М.: Финансы и статистика, 1983.— 223 с.

5. Джоисон Дж. Эконометрические методы/Пер. с англ.— М.: Статистика, 1980.— 444 с.

6. Дринфельд Г. И. Интерполирование и способ наименьших квадратов.— Киев: Вища школа, 1984.— 102 с.

7. Елисеева И. И., Рукавишников В. О. Логика прикладного статистического анализа.— М.: Финансы, и статистика, 1982.— 192 с.

8. Елохин В. Р., Саиаев В. Г. Аппроксимация моделей энергетических систем. Планирование и анализ регрессионных экспериментов.— Новосибирск: Наука, 1985.— 144 с.

9. Катковник В. Я. Непараметрическая идентификация и сглаживание даиных. Метод локальиой аппроксимации.— М.: Наука, 1985.— 336 с.

10. Королев Ю. Г. Метод наименьших квадратов в социально-экономических исследованиях.— М.: Статистика, 1980.— 112 с.
11. Крастинь О. П. Изучение статистических зависимостей по многолетним данным.— М.: Финансы и статистика, 1981.— 136 с.
12. Кругликов В. К. Вероятностный машинный эксперимент в приборостроении.— Л.: Машиностроение, 1985.— 247 с.
13. Лимер Э. Статистический анализ неэкспериментальных данных. Выбор формы связи/Пер. с англ. Под ред. А. А. Рывкина.— М.: Финансы и статистика, 1983.— 381 с.
14. Миркин Б. Г. Группировки в социально-экономических исследованиях. Методы построения и анализа.— М.: Финансы и статистика, 1985.— 223 с
15. Мирский Г. Я. Характеристики стохастической взаимосвязи и их измерения.— М.: Энергоиздат, 1982.— 320 с.
16. Монтгомери Д. К. Планнрование эксперимента и анализ данных/Пер. с англ. Под. ред С. Б. Барона.— Л.: Судостроение, 1980.— 384 с.
17. Поллард Дж. Справочник по вычислительным методам статистики/Пер. с англ. Под ред. Е. М. Четыркина.— М.: Финансы и статистика, 1982.— 344 с.
18. Попечителев Е. П., Романов С. В. Анализ числовых таблиц в биотехнических системах обработки экспериментальных данных.— Л.: Наука, 1985.— 148 с.
19. Трофимов В. П. Логическая структура статистнческих моделей.— М.: Финансы и статистика, 1985.— 191 с.
20. Трухаев Р. И., Горшков И. С. Факторный анализ в организационных системах.— М.: Радио и связь, 1985.— 184 с.
21. Тьюки Дж. Анализ результатов наблюдений. Разведочный анализ/Пер. с англ. Под ред. В. Ф. Писареико.— М.: 1981.— 693 с.
22. Четыркин Е. М. Статистические методы прогнозирования.— 2-е изд.— М.: Статистика, 1977.— 200 с.
23. Фёстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа. Руководство для экономистов/Пер. с нем.— М.: Финансы и статистика, 1983.— 302 с.

III. РАБОТЫ ПО ПРОГРАММНОМУ ОБЕСПЕЧЕНИЮ РЕГРЕССИОННОГО АНАЛИЗА

1. Айвазян С. А., Енюков И. С., Мешалкин Л. Д. Прикладная статистика. Исследование зависимостей: Справочное издание/Под ред. С. А. Айвазяна.— М.: Финансы и статистика, 1985.— 487 с.
2. Алгоритмы и программы восстановления зависимостей/Под ред. В. Н. Вапника.— М.: Наука, 1984.— 816 с.
3. Александров В. В., Горский Н. Д. Алгоритмы и программы структурного метода обработки данных.— Л.: Наука, 1983.— 208 с.
4. Гришин В. Г. Образный анализ экспериментальных данных.— М.: Наука, 1982.— 237 с.
5. Дайитбегов Д. М., Калмыкова О. В., Черепанов А. И. Программное обеспечение статистической обработки данных.— М.: Финансы и статистика, 1984.— 192 с.
6. Демиденко Е. З. Линейная и нелинейная регрессия. Фортран-IV.— М.: Изд. ИМЭМО, 1979.— 82 с.
7. Демиденко Е. З. Гребневая регрессия. Препринт.— М.: Изд. ИМЭМО, 1982.— 126 с.
8. Демиденко Е. З. Нелинейная регрессия. Ч. 1. Алгоритмы. Препринт.— М.: Изд. ИМЭМО, 1984.— 73 с.
9. Демиденко Е. З. Нелинейная регрессия. Ч. 2. Программы. Препринт.— М.: Изд. ИМЭМО, 1984.— 72 с.
10. Денисов В. И., Попов А. А. Пакет программ оптимального планирования эксперимента.— М.: Финансы и статистика, 1986.— 159 с.
11. Джонсон К. Численные методы в химии/Пер. с англ. Под ред. А М. Евсеева.— М.: Мир, 1983.— 504 с.

12. Енюков И. С. Методы, алгоритмы, программы многомерного статистического анализа: Пакет ППСА.— М.: Финансы и статистика, 1986.— 13 п. л.
13. Загоруйко Н. Г., Елкина В. Н., Лбов Г. С. Алгоритмы обнаружения эмпирических закономерностей. — Новосибирск: Наука, 1985.— 110 с.
14. Иванова Г. Н. Типовой процесс проектирования технологических процессов в радиодеталестроении с использованием ЭВМ.— Электронная техника. Сер. Раднодетали и радиокомпоненты, 1982, вып. 4 (49), с. 25—29.
15. Лбов Г. С. Методы обработки разнотипных экспериментальных данных.— Новосибирск: Наука, 1981.— 160 с.
16. Пакет прикладных программ «ОТЭКС»/Загоруйко Н. Г., Елкина В. Н., Емельянов С. В., Лбов Г. С.— М.: Финансы и статистика, 1986.— 10 п. л.
17. Песаран М., Слейтер Л. Динамическая регрессия: теория и алгоритмы/Пер. с англ. Под ред. Э. Б. Ершова.—М.: Финансы и статистика, 1984.— 310 с.
18. Петровнч М. Л. Регрессионный анализ и его математнческое обеспечение на ЕС ЭВМ: Практическое руководство.— М.: Финансы и статистика. 1982.— 199 с.
19. Петрович М. Л. Анализ программного обеспечения по прикладной статистике/Обзор//Заводская лаборатория, 1985, 51, № 10, с. 47—56, библ. 28 назв.
20. Планирование эксперимента в задачах нелинейного оцениваиии и распознавания образов/Круг Г. К., Кабанов В. А., Фомин Г. А., Фомина Е. С.— М.: Наука, 1981.— 172 с.
21. Понск зависимости и оценка погрешности/Под ред. И. Ш. Пинскера.— М.: Наука, 1985.— 148 с.
22. Приходько Ю. Г. Оценка показателей качества проблемного программного обеспечения.— Минск: Респ. ИПК РРСОНХ, 1985.— 68 с.
23. Разработка пакетов промышленных программ по математической статистике /Дукарский О., Кошевник Ю., Френкель А., Шифрин Г.— Вестник статистики, 1985, № 5, с. 34—42.
24. Райс Дж. Матрнчные вычисления и математическое обеспечение/Пер. с англ. Под ред. В. В. Воеводина.— М.: Мир, 1984.— 264 с.
25. Статистическне методы для ЭВМ/Под ред. К. Эйнслейна, А. Ралстона, Т. Унлфа.— Пер. с англ/Под ред. М. Б. Малютова.— М.: Наука, 1986.—460 с.
26. Успенский А. Б. Вычислительные аспекты метода наименьших квадратов при анализе и планировании регрессионных экспериментов.— М.: МГУ, 1975.— 168 с.
27. Фолингер А. Ф. Статистнческие алгоритмы в социологических исследованиях.— Новосибирск: Наука, 1985.— 208 с.
28. Форсайт Дж., Мальколм М., Моулер К. Машинные методы математических вычислений/Пер. с англ.— М.: Мир, 1980,— 280 с.
29. Шнейдерман Б. Психология программирования. Человеческие факторы в вычислительных и информациониых системах/Пер. с англ. Под ред. В. В. Мартынюка.— М.: Радио и связь, 1984.— 304 с.
30. Bates D. M., Draper N. R. Applied regression analysis bibliography. Update 1981—1985. Technical Report № 765.— Dept. Statist. Univ. Wisconsin.— Julu 1985.— 22 p.

СЛОВАРЬ ТЕРМИНОВ, ОТНОСЯЩИХСЯ К РЕГРЕССИОННОМУ АНАЛИЗУ

Adequate representation — адекватное представленне

alias — смещение

— matrix — матрица смещения

all possible regressions — все возможные регрессии

analysis of variance (ANOVA) — дисперсионный анализ

arrangement — расположение; конфигурация (напр., о точках плана)

Backward elimination procedure — метод исключеиия (в регрессионном анализе)

badly conditioned — плохо обусловленная (о матрице)
balanced design — сбалансированный план
bias — смещение
— error — ошибка смещения; систематическая ошибка
— in estimate — смещенне оценки
Calibration curve — калибровочная кривая
canonical analysis — канонический анализ
canonical form of ridge regression — каноническая форма гребневой регрессии
canonical reduction — каноническое преобразование
centering (of the data) — центрирование (данных)
central limit theorem — центральная предельная теорема
characteristic values — характеристические значения
chi-squared distribution — распределение хи-квадрат (χ^2)
choice of design — выбор плана
choosing the first variable to enter regression — выбор первой переменной для включения в уравнение регрессии
classification with equal numbers of observation in the cells — классификация с равным числом наблюдений в ячейках
coefficient of multiple determination — множественный коэффициент детерминацни: квадрат коэффициента множественной корреляции
computer routine — программа для ЭВМ
confidence interval — доверительный интервал
— limit — доверительный предел
— region — доверительная область
conjectured model — предполагаемая модель; постулируемая модель; выдвигаемая модель; проверяемая модель; гипотетическая модель
continuity correction — поправка на непрерывность
continuous distribution — непрерывное распределение
corrected sum of squares (products) — скорректнрованная сумма квадратов (произведений)
correction factor — корректирующий фактор (множитель)
— for the mean — коррекция на среднее значение
correlation — корреляция
— analysis — корреляционный анализ
— coefficient — коэффициент корреляции
— matrix — корреляционная матрнца
control model — модель управления
construction of new variables — построенне новых переменных
covariance — ковариация
— matrix — ковариационная матрица
critical value — критическое значение
cross-product term (CPT) — взаимодействие (парное); перекрестное произведение
cut-off — останов (об ЭВМ)
cycling — зацикливанне
Data — данные
— matrix — матрнца данных
decoded b-coefficient — коэффициент b в исходном (натуральном) масштабе
definition — определение
default — по умолчанню
degrecs of freedom — степени свободы; чнсло степеней свободы
dependent variable — зависимая переменная; отклик
design matrix — матрица плана
— of experiment — планированне эксперимента
detecting influential observations — определение влияющих наблюдений
determinant — детерминант, определитель
deviation — отклонение
dichotomous classification — дихотомнческая класснфикация
discrepancy — расхождение (разность)
discrete distribution — дискретное распределение
dstribution — распределение
— of error — распределение ошибок
distributed parameter — распределенный параметр
dot diagram — точечная диаграмма
double precision arithmetic — вычисления с удвоенной точностью
«dummy» variable — «фиктивный» фактор (переменная)
Durbin—Watson test — критерий Дарбина—Уотсона
Eigenvalues — собственные значения
ellipsoidal confidence region — эллипсоидная доверительная область

ellipsoidal 100 (1—α) % boundary — 100 (1—α)%-ная эллипсоидальная граница
enlargering the model — расширение модели
empirical model — эмпирическая модель
environmental conditions — внешние условия
equal numbers of observation in the cells — равное число наблюдений в ячейках
error — ошибка, погрешность, отклонение
— mean squares — остаточный средний квадрат
— space — пространство ошибок
— sum of square — сумма квадратов ошибок
— variance — дисперсия ошибки
establish goal — установление цели
estimate — оценка
estimated regression equation — оценниваемое уравненне регрессии, оценка уравнения регрессии
estimated s. e. (b_i) — оценка среднеквадратичной ошибки коэффициента
estimated variance-covariance matrix of **b** — оценка матрицы дисперсий-ковариаций вектора **b**; выборочная матрица дисперсий-ковариаций вектора **b**
estimation — оценивание (подсчет, вычисление)
— space — пространство оценок
estimator — оцениватель (оценщик)
expected value — математическое ожидание; среднее значение
experimental design — экспериментальный план
exponential distribution (double) — экспоненциальное распределенне (двойное)
exponential model — экспоненциальная модель
extra conditions — дополнительные условия
extra sum of squares — дополнительная сумма квадратов
F-distribution — F-распределение
— percentage points — процентные точки F-распределения
F-test — F-критерий
factor — фактор, независимая переменная
2^k factorial design — факторный эксперимент типа 2^k
factorial two level experiments — факторные эксперименты на двух уровнях
families of straight lines — семейство прямых (линий)
family of transformations — семейство преобразований
fiducial — фидуциальный
fiducial limit — фидуциальный предел
fiducial interval — фидуциальный интервал; принятый за основу сравнения; надежный; отправной
first order model — модель первого порядка
fit the model — подбор модели; подгонка модели
fitted (or estimated) equation — подобранное (или оцениваемое) уравнение
— (or predicted) value — подобранное (или предсказанное) значение
fitting a straight line — подбор прямой (линии); подгонка
fixed effects analysis of variance (Model I) — дисперсионный анализ с постоянными эффектами (факторов) (модель I)
formal statements — описание формата
forward selection procedure — метод включения (в регрессионном анализе)
frequency — частота
frequency function — функция плотности

Gauss—Newton method — метод Гаусса—Ньютона
Gauss's theorem — теорема Гаусса
geometry of least squares — геометрия метода наименьших квадратов
graduating function — сглаживающая функция
graduation — сглаживание; нанесение кривой по точкам
grid — сетка

Half—normal plot — полунормальный график (диаграмма)
hypothesis — гипотеза
— testing — проверка гипотезы

Ill—condition — плохо обусловленная (о матрице)
incorrelated — некоррелированный
incorrect model — некорректная модель
independent variable — независимая переменная; фактор
idempotent matrix — идемпотентная матрица
input data matrix — матрица исходных данных

interaction — взаимодействие
intercept — свободный член (уравнения регресни)
internal sum of squares — внутренняя сумма квадратов
intrinsically linear — внутренне линейная (о модели)
inverse estimation — обратное оценнвание
inverse of matrix — обращение матрицы
inverse regression — обратная регрессия

Joint confidence — совместная доверительная область

Lack of fit — неадекватность (отсутствие согласия, рассогласование)
— mean squares — средний квадрат, обусловленный неадекватностью
— sum of squares — сумма квадратов, обусловленная неадекватностью
lag—1 serial correlation — сериальная корреляция с единичным сдвигом
— i — со сдвигом на i шагов
latent root regression — регрессия на собственных значениях
latent variable — латентная (скрытая) переменная
least squares method — метод наименьшнх квадратов (МНК)
— equation — МНК-уравнение
— estimate — МНК-оценка
level of factor — уровень фактора
linear regression (model) — линейная регрессия (модель)
— relationship — линейная зависнмость
— time trend — линейный временной дрейф (тренд)
linear trend — лннейный дрейф (тренд)
linearity in the parameter — линейность по параметрам
linearization — линеаризация
local direction — локальное направление
local positive serial correlation — локально положительная сериальная корреляция
logarithmic transformation — логарифмическое преобразование
logistic regression — логистическая регрессия
lower-tailer test — односторонний критерий для нижнего «хвоста» распределения
lumped parameter — сосредоточенный параметр

Major axes — главные оси
mallows C_p statistic — статистика C_p Маллоуса
manuel — руководство, инструкция
mathematical model building — построенне математической модели
matrix — матрица
— algebra — матричная алгебра; алгебра матриц
— of independent variables — матрица независнмых переменных
maximum likelihood — максимальное правдоподобие
— function — функция правдоподобия
mean — среднее (значение)
mean of all observations — общее среднее
— in row i — среднее по всем наблюдениям в i-й строке
— in column j — среднее по всем наблюдениям в j-м столбце
— the cell (i, j) — среднее по всем наблюдениям в ячейке (i, j)
mean square — средний квадрат
— — error — средний квадрат ошибки
mean square about regression — средннй квадрат отклонений относительно регрессии
— due to lack of fit — средннй квадрат, обусловленный неадекватностью
— due to regression — средний квадрат, обусловленный регрессией
— due to residual variation — остаточный средний квадрат (средний квадрат, обусловленный остаточной вариацией)
— for pure error — средний квадрат, характеризующий «чнстую» ошибку
method for discriminating — метод дискриминацнн (моделей)
— of least squares — метод нанменьшнх квадратов (МНК)
midmean — срединное (усеченное) среднее
minimum variance unbiased estimator — несмещенный оцениватель с минимальной дисперсией
missing observations — пропущенные наблюдения
model is correctly identified — модель правильно идентнфицирована
model validation technique — метод обоснования модели

moment — момент
multidimensional space — многомерное пространство
multiple regression calculation — множественные регрессионные вычисления
— correlation coefficient — множественный коэффициент корреляции
— regression — множественная регрессия
multiplicative model — мультипликативная модель
multivariate — многомерный

N-dimensional multivariate normal distribution — N-мерное нормальное распределение
negative sereal correlation between successive residuals — отрицательная сериальная корреляцня между последовательнымн (соседними) остаткамн
Newton—Raphson technique — метод Ньютона—Рафсона
no lack of fit — адекватность (неадекватность); согласие
nonadditivity — неаддитивность
nonintrinsically nonlinear — внешне нелинейная
nonlinear estimation — нелннейное оценивание
— least squares — нелинейный метод наименьших квадратов
nonlinear growth model — нелинейная модель роста
nonlinearity in the parameters — нелинейность по параметрам
nonsingular matrix — неособенная (невырожденная) матрнца
normal deviate — нормальное отклонение
— distribution random variable — нормально распределенная случайная величнна
— equations — нормальные уравнения (МНК)
— plot of residuals — графнк остатков
normal probability plot of residuals — нормальный вероятностный график остатков

Observations — наблюдения
one-sided test — односторонннй критерий
one-way classification — односторонняя классификация; классификация по одному признаку
order of the model — порядок модели
original data — исходные данные
orthogonal column — ортогональные столбцы (матрицы)
orthogonal linear functions — ортогональные линейные функции
orthogonal polinomials — ортогональные полиномы
orthogonalization of a matrix — ортогонализация матрицы
outlier — выброс; резко выделяющееся значение
overall direction — глобальное направление
overall F-test — полный F-критерий
overall mean square error — полная среднеквадратическая ошибка
overfitting — переподгонка (сверхподбор)
overparametrezed — перепараметрнзована

Parameter space — пространство параметров
partial correlation — частная (парциальная) корреляция
partial correlation of variables z_2 and Y after both have been adjusted for variable z_1 — частная (парциальная) корреляция переменных z_2 и Y после поправки их обеих на z_1
partial F-test — частный F-критерий
partioned matrix — блочная матрнца (разделенная на блоки)
percentage point of the distribution — процентная точка распределения
percentage variation explained — объясняемая доля разброса
planning — планирование
— of large regression studies — планирование больших регрессионных исследований
plot — график (диаграмма)
poorly conditioned surface — плохо обусловленная поверхность
population value — генеральное значение
probability level — уровень вероятностн
precision — точность
predictability — предсказуемость
predicted (mean) value — предсказанное (среднее) значение
predictive discrepancy sum of squares — сумма квадратов предсказанных расхожденнй
predictive equation (model) — предсказывающее уравнение (модель)
predictive error — ошибка предсказания
predictor — предиктор; независимая

переменная; [предсказатель]; фактор
PRESS — prediction sum of square— ПРЕСК — предсказанная сумма квадратов; сумма квадратов для предсказания
principal component regression — регрессия на главных компонентах
printed output — распечатка
printout — машинный бланк, распечатка
a priori estimate — априорная оценка
— information — априорная информация
probability — вероятность
— distribution — распределение вероятностей
problems with messy data — задачи с опущенными данными
proportion of total variation about the mean Y explained by the regression — доля общего разброса относительно Y, объясняемая регрессией
pure error — «чистая ошибка» (ошибка опыта)
— mean square — средний квадрат, связанный с «чистой» ошибкой
— sum of squares — сумма квадратов, связанная с «чистой» ошибкой (обусловленная «чистой» ошибкой)

Random — случайный
— arrangement of signs — случайное расположение знаков
— deviation — случайное отклонение
— error — случайная ошибка
— search — случайный поиск
— variable — случайная переменная (величина)
— variation — случайный разброс
real roots — действительные корни
reciprocal transformation — обратное преобразование
region of interest — область интереса; область действия
regression — регрессия
— analysis — регрессионный анализ
— curve — регрессионная кривая
— — of x_1 on x_2 — регрессионная кривая x_1 на x_2
— equation — уравнение регрессии
— estimate — регрессионная оценка
— mean squares — средний квадрат, обусловленный регрессией
— sum of squares — сумма квадратов, обусловленная регрессией
reparametrization — репараметризация
repeated samples— повторные выборки
residual — остаток
— mean squares — остаточный средний квадрат
— sum of squares — остаточная сумма квадратов
response — отклик
— surface — поверхность отклика
— variable — переменная- отклик; отклик
restricted least squares — метод наименьших квадратов (МНК) с ограничениями
restrictions on the parameters — ограничения на параметры
reversion — реверсия; обращение; движение вспять
ridge — гребень
ridge regression — гребневая регрессия; ридж-регрессия
ridge «squared bias» — квадрат гребневого смещения
ridge trace — след гребня, «хребет», ридж-след
right — hand— tail probability — вероятность правого «хвоста»
α-risk — α-риск
risk level — уровень риска; уровень значимости
robust regression — робастная регрессия; устойчивая регрессия
rotatability — ротатабельность
rotatable design — ротатабельный план
rounding error — ошибка округления
— of number — округления числа
row vector — вектор-строка
runs test — критерий знаков

Sample — выборка
— coefficient — выборочный коэффициент; оценка коэффициента
— correlation coefficient — выборочный коэффициент корреляции
— estimate — выборочная оценка
— size — объем (размер) выборки
satisfactory accuracy — достаточно точно
— approximating function — удовлетворительно аппроксимирующая функция
scaled — нормированный
scaling factor — масштабный коэффициент
scatter diagramm — диаграмма рассеяния

second—order model — модель второго порядка
second—order response surface analysis — анализ поверхности отклика второго порядка
select response — выбор отклика
sequential — последовательный
— F-test — последовательный F-критерий
serial correlation of residuals — сериальная корреляция остатков
set — набор, множество
— of residuals — множество остатков
— up — схема
significance — значимость
— level — уровень значимости
— of regression — значимость регрессии
slope — угловой коэффициент (наклон) (о линейной функции)
smooth curve — сглаженная кривая
software regression package — пакет регрессионных программ; пакет программ регрессионного анализа; программное обеспечение регрессионного анализа
solution locus — геометрическое место точек решения
— of normal equation — геометрическое место точек решения нормальных уравнений
source — источник (рассеяния) (в дисперсионном анализе)
spline — сплайн
split of — отделение
— up — разбиение
spread — разброс; вариация
square of multiple correlation coefficient — квадрат множественного коэффициента корреляции (множественный коэффициент детерминации)
square root transformation — преобразование квадратного корня
stagewise — ступенчатый
— regression procedure — ступенчатая регрессионная процедура
standard error — стандартная ошибка
— of the slope b_1 — стандартная ошибка углового коэффициента b_1
— of estimate — стандартная ошибка оценки
— of estimate s as a percentage of the mean — стандартная ошибка оценки s в процентах от среднего отклика
— F-test — стандартный F-критерий
standardized b-coefficient — стандартизованный b-коэффициент
standard probability paper — стандартная вероятностная бумага
stationary value — стационарное значение
statistical screening procedures — статистические методы отсеивания
statistical significance — статистическая значимость
steam—und—leaf display — представление «опора и консоль»
steepest descent — крутой спуск
stepwise — шаговый
— regression procedure — шаговый регрессионный метод
straight line regression — прямолинейная регрессия
summary data — усредненные данные
sum of squares (SS) — сумма квадратов
— — — of discrepancies — сумма квадратов расхождений
SS about regression — сумма квадратов относительно регрессии
SS about the mean — сумма квадратов относительно среднего
SS due to regression — сумма квадратов, обусловленная регрессией
suggest variables — предполагаемые (предлагаемые) переменные
sum of squares due to the hypothesis $C\beta = 0$ — сумма квадратов, обусловленная гипотезой $C\beta = 0$
sum of squares of $b_{q+1}, \ldots, b_p$ given $b_0, b_1, \ldots, b_q$ — сумма квадратов, связанная с $b_{q+1}, \ldots, b_p$ при заданных коэффициентах $b_0, b_1, \ldots, b_q$

Taylor series — ряд Тейлора
t-distribution — t-распределение
t-test — t-критерий
test — критерий; тест; проверка
— of hypothesis — проверка гипотезы
— of significance — проверка значимости
— statistic for H_0 — статистика для проверки гипотезы H_0
testing a general linear hypothesis in regression situations — проверка общей линейной гипотезы в регрессионных задачах
third-order model — модель третьего порядка
— sequential design — последовательный план третьего порядка
time sequence — временная последовательность
— trend — временной дрейф (треид)
total — общий (источник рассеяния)

(в дисперсионном анализе)
— sum of squares — общая (полная) сумма квадратов
— corrected for mean — общая (полная) сумма квадратов, скорректированная на среднее
transformation — преобразование
— on the observations — преобразование наблюдений
transpose of matrix — транспонирование матрицы
true model — «истинная» модель
two-side test — двусторонний критерий
two-tailed test — двусторонний критерий
two—way classification — двусторонняя классификация; классификация по двум признакам
two—way classification analysis of variance — дисперсионный анализ двусторонней классификации
two-way table — таблица сопряженности; таблица с двумя входами

Unbiased estimator — несмещенный оцениватель
uncontroled factor — неуправляемый (неконтролируемый) фактор
unexplained variation — необъясненная вариация
unit normal deviate — единичное (нормированное) нормальное отклонение
unit normal distribution — единичное (нормированное) нормальное распределение
unknown parameters — неизвестные параметры
upper-tailed test — односторонний критерий для верхнего «хвоста» распределения
upper tail of the distribution — верхний хвост распределения

Validation — обоснованность
validation technique — метод перепроверки (проверки) состоятельности
variable (dependent) — отклик; зависимая переменная
— (independent) — фактор; независимая переменная
variance — дисперсия
— about the regression — дисперсия относительно регрессии
— — covariance matrix — матрица дисперсий-ковариаций
— of a function — дисперсия функции
variance-covariance matrix of the vector **b** — матрица дисперсий-ковариаций вектора b
variation — вариация; разброс
vector of error — вектор ошибок (остатков)
— of observation — вектор наблюдений
— of parameters to be estimated — вектор оцениваемых параметров
verification — проверка; верификация

Weighted least squares — взвешенный метод наименьших квадратов (МНК)
well conditioned surface — хорошо обусловленная поверхность

X-space — пространство «*X*»; факторное пространство

ОГЛАВЛЕНИЕ

Дрейпер Н., Смит Г.

Д73 Прикладной регрессионный анализ: В 2-х кн. Кн. 2/Пер. с англ.—2-е изд., перераб. и доп.—М.: Финансы и статистика. 1987.—351 с.: ил.—(Математико-статистические методы за рубежом).

Работа американских ученых посвящена регрессионному анализу, применяемому во всех отраслях народного хозяйства и научных исследованиях. Второе издание книги (1-е изд. перевода — 1973 г.) значительно переработано и дополнено новыми алгоритмами и сравнением их достоинств. В кн. 2 приводится описание модели, нелинейной по параметрам регрессии, обширная библиография и приложения.

Для специалистов — статистиков, экономистов, социологов, научных работников.

Д $\frac{070200000-006}{010(01)-87}$ 109—86

ББК 22.172

Монография

Норман Дрейпер, Гарри Смит

ПРИКЛАДНОЙ РЕГРЕССИОННЫЙ АНАЛИЗ. Кн. 2

Книга одобрена на заседании редколлегии серии
«Математико-статистические методы за рубежом» 26.05.83 г.

Зав. редакцией *К. В. Коробов*
Редактор *А. А. Рывкин*
Мл. редакторы *О. Г. Виноградова, А. С. Шиманская*
Техн. редактор *И. В. Завгородняя*
Корректоры *Я. Б. Островский, М. А. Синяговская* и *Л. Г. Захарко*
Худож. редактор *Ю. И. Артюхов*
ИБ № 1571

Сдано в набор 25.06.86. Подписано в печать 24.11.86. Формат 60×90$^1/_{16}$. Бум. тип. № 1. Гарнитура «Литературная». Печать высокая. Усл. печ. л. 22,0. Усл. кр.-отт 22,0. Уч.-изд. л. 24,48. Тираж 12 000 экз. Заказ № 1809 Цена 2 р. 20 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 ул., 14.