Ф. Мостеллер
Дж. Тьюки

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ ЗА РУБЕЖОМ
12345
67890

АНАЛИЗ ДАННЫХ И РЕГРЕССИЯ

Ф. Мостеллер
Дж. Тьюки

2

DATA ANALYSIS AND REGRESSION

A second course in statistics

Frederick Mosteller
Harvard University

John W. Tukey
Princeton University
and Bell Telephone Laboratories

Addison-Wesley Publishing Company

Reading, Massachusetts·Menlo Park, California·
London·Amsterdam·Don Mills, Ontario·Sydney

Ф. Мостеллер, Дж. Тьюки

АНАЛИЗ ДАННЫХ И РЕГРЕССИЯ

Выпуск 2

Перевод с английского Б. Л. РОЗОВСКОГО
Под редакцией и с предисловием Ю. П. АДЛЕРА

Москва «Финансы и статистика» 1982

ББК 22.172
М84

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.

ГОТОВЯТСЯ К ПЕЧАТИ

1. Лимер Э. Статистический анализ неэкспериментальных данных. Выбор формы связи.
2. Бикел П., Доксам К. Математическая статистика. Вып. 1 и 2.

М $\frac{0702000000-165}{010(01)-82}$ 40—82

ПРЕДИСЛОВИЕ К РУССКОМУ ИЗДАНИЮ

НАУКА И ИСКУССТВО АНАЛИЗА ДАННЫХ (ПРОДОЛЖЕНИЕ)

Переходя ко второму выпуску, проследим теперь за причинами особого интереса авторов к регрессии, затем продолжим разговор о переводе терминов и приведем обещанный краткий перечень книг на русском языке, которые могут быть полезны читателю как предварительное чтение или как источники информации по ходу чтения этой книги.

Анализ данных принадлежит всей статистике и выходит, как мы уже видели, далеко за ее пределы. Почему же весь второй выпуск (как и некоторые места первого) отведен именно регрессии? Чтобы ответить на вопрос, мы будем вынуждены совершить краткое путешествие по истории и современному состоянию этого выдающегося метода, лежащего на «распутье» статистических дорог.

Гаусс (и независимо от него Лагранж) без малого 200 лет тому назад создали метод наименьших квадратов. Метод наименьших квадратов вызвал к жизни регрессионный анализ. Это случилось после того, как усилиями того же Гаусса, а потом Маркова на метод наименьших квадратов удалось надеть статистическую «смирительную рубашку». Безотказно служил он астрономам и геодезистам, был полезен химикам (например, Менделееву) и всем другим, кто нуждался в его помощи. Слыл добротным и надежным инструментом исследователя. Вот только немного громоздким, трудоемким. Для борьбы с трудоемкостью появилась известная вычислительная схема Дулитла, бывшая весьма актуальной еще каких-нибудь 25—30 лет назад, и разные ортогонализации, основанные большей частью на полиномах Чебышева (см., например, монографию: Немчинов В. С. Математическая статистика и полиномы Чебышева. М., Московская сельскохозяйственная академия им. К. А. Тимирязева, 1946).

Гальтон с его фейерверком новаторских идей и К. Пирсон с его систематичностью больше других способствовали доведению математической идеи до практической методики. Гальтону же принадлежит и сам термин «регрессионный анализ», вряд ли удачный, но теперь уже привычный.

С появлением вычислительной техники развитие алгоритмов регрессионного анализа воистину было путем «вверх по лестнице, ведущей вниз». Совершенствовались ЭВМ и с каждым новым поколением рождались новые, более совершенные алгоритмы. Метод всех возможных регрессий, шаговый метод, ступенчатый метод, метод Гарсайда — все

и не перечислить. Но всякий раз оказывалось, что никакие изощрения не позволяют получить единственный и однозначный ответ. Постепенно стало ясно, что в большинстве случаев регрессионная задача принадлежит к классу задач, которые математики называют *некорректно* поставленными. Либо их можно регуляризовать, за счет экзогенной информации, либо остается смириться с неоднозначными, разными, множественными ответами.

Так бесславно регрессионный анализ деградировал до *эвристического* метода, в котором решающую роль играют анализ остатков да здравый смысл интерпретатора. Автоматизация задач регрессионного анализа зашла в тупик.

Однако параллельно шел другой процесс, который поставил регрессионный анализ в центре проблематики многомерной статистики и связал его неразрывными узами с дисперсионным и ковариационным анализами (и их обобщениями), с многомерной классификацией данных: дискриминантный, кластерный, факторный анализы, метод главных компонент и т. п. Регрессионный анализ оказался одним из «столпов» планирования экспериментов (правда, в этом случае работают часто простейшие модификации, но все-таки не всегда и не везде). Его обобщили для решения широких классов задач нелинейной параметризации, простирающих свои интересы от химической кинетики до динамических моделей идентификации и управления производством. Он получил многочисленные эконометрические приложения, связь со спектральным анализом и теорией фильтрации, с решением уравнений математической физики и т. д.

Вот почему новый взгляд на проблемы регрессионного анализа неизбежно будет иметь глобальные последствия. Именно такую попытку нащупывания не известных ранее путей и возможностей и представляет собой данная книга (главным образом ее второй выпуск). Ясно, что это не окончательное решение вопросов. Но ясно и то, что это очередной шаг вперед, попытка использовать анализ данных в совершенствовании регрессионного анализа. Эта попытка потребовала создания новой терминологии. Вот некоторые примеры.

Одно из ключевых новых понятий — «гибкая» («управляемая») регрессия (guided regression). Для собирательного обозначения переменных в регрессионном анализе авторы используют термин «carrier»— «носитель» (информации), поскольку он может и не стать переменной, фактором (variable). А для самого фактора иногда применяется синоним «predictor» — «предиктор» («предсказатель») — по функциональному принципу: то, что включено в модель для предсказания отклика. Переменную, которая выдает себя за другую, естественно называть «proxy variable» — «заменитель», «подставная» переменная. Для набора переменных, порождающего интересующий исследователя класс моделей, мы после долгих колебаний выбрали термин «генератор» (stock), имея в виду некоторую аналогию с задачами факторного эксперимента. Тогда часть такого набора «costock» стала называться «подгенератором». Отметим еще, что термин «matcher» мы передали словом «балансир», продолжая механические аналоги, характерные для метода наименьших квадратов, а термин «catcher» — словом «уловитель».

Кроме необычных, приведем еще подборку вполне традиционных терминов, относящихся к представлениям о зависимостях, поскольку их интерпретация во втором выпуске играет решающую роль: «association» — «соответствие» (возникшее неизвестно как, то ли случайно, то ли обусловленно); «causation» — «причинная связь», «детерминированная зависимость» (т. е. зависимость, точно отражающая некий закон природы); «correlation» — «корреляция», «взаимосвязь» (имеющая неясный, может быть, индетерминированный характер); «dependence» — «зависимость» (одного ряда событий от другого или других, имеющая неслучайные элементы), «обусловленность»; «relation» — «отношение», «связь» (общее понятие без пояснений и акцентов); «relationship» — «соотношение» (часто как «формула»).

Литература по всем затронутым нами вопросам вряд ли обозрима. Поэтому ограничимся лишь кратким перечнем работ, наиболее простых, или тех, без которых трудно обойтись: Бейли Н. Статистические методы в биологии. М., Мир, 1964; Глясс Дж., Стэнли Дж. Статистические методы в педагогике и психологии. М., Прогресс, 1976; Закс Л. Статистическое оценивание М., Статистика, 1976. (Это — справочник.) А вот несколько книжек по регрессионному анализу: Езекиел М., Фокс К. А. Методы анализа корреляций и регрессий. М., Статистика, 1966; Линник Ю. В. Метод наименьших квадратов и основы математико-статистической обработки наблюдений. М., Физматгиз, 1962; Перегудов В. Н. Метод наименьших квадратов и его применение в исследованиях. М., Статистика, 1965; Дрейпер Н., Смит Г. Прикладной регрессионный анализ. М., Статистика, 1973; Себер Дж. Линейный регрессионный анализ. М., Мир, 1980; Вапник В. Н. Восстановление зависимостей по эмпирическим данным. М., Наука, 1979; Алберт А. Регрессия, псевдорегрессия и рекуррентное оценивание. М., Наука, 1977; Бард Й. Нелинейное оценивание параметров. М., Статистика, 1979; Демиденко Е. З. Линейная и нелинейная регрессия. М., Финансы и статистика, 1981. В области многомерного анализа ограничимся лишь: Андерсон Т. Введение в многомерный статистический анализ. М., Физматгиз, 1963; Иберла К. Факторный анализ. М., Статистика, 1980.

Единственный способ вырваться из «плена» какого-нибудь метода— это овладеть им. Поэтому стоит учиться анализу данных — за ним будущее.

Ю. Адлер

Глава 12 ● РЕГРЕССИЯ ДЛЯ ПОДГОНКИ

ВВЕДЕНИЕ

Обычно первые основные понятия бывают достаточно просты. К сожалению, этого нельзя сказать о понятиях *соответствие*, *причинная связь* и *зависимость*, неизбежно возникающих при анализе соотношений между двумя или более переменными. Поэтому, начав с кое-каких объяснений, сопоставлений и определений, мы приведем затем разнообразные примеры. Некоторые из них будут доведены до числа, а в других все будет ясно и без чисел.

Диктант по французскому. Давайте проверим способность детей грамотно писать по-французски под диктовку. Для этого устроим контрольный диктант. Чтобы не усложнять задачу, предположим, что все дети слушали текст, воспроизводившийся с одной и той же магнитофонной ленты, что оценки выставлялись объективно и по одной системе, а также что проверявшие могли «расшифровать» почерк каждого ребенка. Пусть y — полученная отметка, а x — вес ребенка. Как связаны y и x?

Этот вопрос, без дополнительной информации, звучит как шутка.

Большие различия в возрасте. Многое зависит от того, рассматриваем ли мы группу детей, возраст которых колеблется в широком интервале, скажем от 5 до 15 лет, или группу детей практически одного возраста — 15-летних. Если эта группа смешанная, то в целом более тяжелые дети — старше и, следовательно, их успехи должны быть бóльшими, во всяком случае там, где по-французски говорят или где изучают этот язык. Для такой группы, по всей вероятности, была бы отмечена сильная положительная связь отметки y и веса x.

Почти одинаковый возраст. Если всем детям в группе по 15 лет плюс — минус несколько недель, то возникают другие вопросы. Например, существует различие между мальчиками и девочками. Поэтому, если не оказывают влияния другие факторы, то, вероятно, будет замечено, что оценки более легких детей лучше (так как в этом возрасте девочки лучше усваивают языки).

Смесь. А что случится, если в нашу группу будут входить 15-летние дети из разных стран, где уровень владения французским языком к 15 годам различен, как, например, во Франции, Голландии и США? Если допустить, что французы легче голландцев, которые в свою очередь легче американцев, то мы получим сильную отрицательную связь между весом и отметкой за диктант.

Эти примеры показывают, что, обсуждая связь между x и y, надо уточнить обстоятельства, может быть, даже структуру исследуемой со-

вокупности и требования, которые следовало бы предъявить при выборе. Учитывая это, мы можем теперь обратиться к некоторым понятиям и их определениям.

НЕКОТОРЫЕ СТАТИСТИЧЕСКИЕ ПОНЯТИЯ

Соответствие. Это наислабейшее понятие. Если значения x и y кажутся способными составить пары каким-либо образом в совокупности, то соответствие наличествует, ну а если не видно никакого способа объединить их в пары, то соответствие отсутствует. Мы рассмотрели несколько ситуаций, в которых проявлялось соответствие между весом ребенка и его успехами во французском, в одних случаях оно было положительным, т. е. y возрастал с ростом x (разные возрасты), в других — отрицательным, т. е. y убывал с ростом x (мальчики и девочки, дети из разных стран).

Независимость. Если мы возьмем совокупность 15-летних девочек, одинаково обучавшихся французскому языку и одинаково питавшихся, то мы, по-видимому, найдем, что между их весом и успехами в языке нет никакого соответствия. Это иллюстрация понятия независимости. Доказать независимость строго, используя общее математическое понятие *статистической независимости*, может быть, не так просто. (Строго говоря, X и Y — независимые случайные величины тогда и только тогда, когда
$\Pr(X \leqslant a, Y \leqslant b) = \Pr(X \leqslant a) \Pr(Y \leqslant b)$ для любых a и b; подобное определение имеет место и в том случае, когда X и Y — дискретные случайные величины с неупорядоченной областью значений.)

Если бы нам, вообще говоря, удалось установить соответствие между весом и успехами во французском, то это вряд ли заставило бы нас думать, что увеличение веса служит «причиной» улучшения (или ухудшения, это не важно) отметок за французский диктант, во всяком случае если ситуация такая, как мы только что описали. Мы скорее будем склонны рассматривать в качестве причин, обусловливающих успехи в изучении французского языка, такие факторы , как время и качество обучения, пол, врожденные способности, но уже никак не вес. А почему?

Причинная связь. Обычно требуются две-три идеи, чтобы обосновать понятие «причина». Вот они:

1. *Непротиворечивость* (*состоятельность*). При прочих равных условиях, в совокупности, которую мы исследуем, связь между x и y постоянна от выборки к выборке по направлению, а может быть, даже и по величине.

2. *Чувствительность*. Если мы можем вмешаться и изменить x у какого-нибудь объекта в совокупности, то и его y должен соответственно измениться.

3. *Механизм*. Существует в принципе постоянный механизм, с помощью которого «причину» можно связать с «результатом», т. е. существует такая процедура, часто многошаговая, для которой на каждом шаге естественно считать, что «то-то служит причиной того-то».

Разумеется, ничто из вышеперечисленного неприменимо к изучению связи между весом и успехами во французском языке.

Из перечисленных нами идей лишь непротиворечивость всегда может быть подтверждена чисто экспериментально. Действительно, изучая различные совокупности, мы можем увидеть, постоянно ли соотношение между x и y по направлению и величине.

Чувствительность тоже подтверждается экспериментом, если только он возможен. Для этого надо, чтобы мы могли вмешаться, изменить x и посмотреть, изменится ли соответствующим образом y. Иногда эксперименты в естественных условиях так же, как и искусственные («рукотворные»), могут дать такую информацию. Правда, как правило, в естественных экспериментах она получается менее ценной. (Особенно опасны «естественные» эксперименты, в которых ничего не менялось.)

Наконец, механизм может быть верифицирован лишь процессом его детального построения и увязки каждой ступени такого построения с соответствующей стадией изучаемого процесса.

Причинную связь, столь часто составляющую предмет наших главных забот, обычно не удается установить статистическими методами (а в социальных проблемах и никакими вообще), хотя статистик часто располагает информацией, которая может в этом помочь. Подтверждение причинной связи можно пытаться извлечь из экспериментальных данных. По-видимому, эта возможность будет наиболее реальной, если соблюдены следующие три необходимых условия:

- наличие явного, непротиворечивого соответствия между x и y;
- отсутствие видимых общих причин для связи x и y или, во всяком случае, недостаточные для объяснения наблюдаемого явного непротиворечивого соответствия количественные соотношения между ними. (Попытки установить это часто затруднены частичными причинами, как, например, в классических проблемах: природа — воспитание, наследственность — окружающая среда. Ясно, что и то и другое влияет на формирование человека, но трудно определить, какой вклад дает каждый из факторов и имеют ли вообще смысл попытки определить это с помощью одного числа.) Приведем два примера возможных «общих причин»: (а) инфляция влияет как на цены, так и на процентные ставки; (б) техническая революция повлияла на увеличение населения, вследствие чего в Нью-Йорке увеличилось, с одной стороны, число священников, а с другой — потребление шотландского виски;
- бессмысленность рассмотрения y как причины x, что часто не так легко доказать. (В нашем примере, например, можно предположить, что не в меру ретивые родители стимулируют конфетами детей, изучающих французский язык, и даже оставляют их без ужина за плохие отметки!)

В последующих главах нас будет интересовать наличие или отсутствие соответствия, а также его количественные характеристики. Вопроса о причинной связи мы касаться не будем. Нам надо очень тщательно следить за терминологией, и мы будем постоянно предостерегать читателя и напоминать ему о том, *что не делается.*

Зависимость. Мы разобрались с понятиями «соответствие» и «причинная связь». Этого, однако, недостаточно. Следует прояснить также содержание термина «зависимость», которым столь часто злоупотребляют, и некоторых родственных ему. Когда мы говорим «y зависит от x», мы иногда имеем в виду *однозначную* (*детерминированную*) *зависимость*, т. е. такую, при которой значение x предопределяет значение y*. Обычно это бывает обусловлено законом (математическим, физическим и т. д.), связывающим x и y. Например, в математике если y — площадь круга, а r — радиус, то $y = \pi r^2$. (Такое употребление термина «зависимость» характерно для математических и некоторых физических текстов.)

В других случаях «y зависит от x» означает *отсутствие независимости*, как правило, «при прочих равных условиях». Например, «температура воды в кране зависит от расстояния до котла». Ясно, что она может зависеть также от температуры в доме, от облицовки котла, не говоря уже о том, проходит ли труба, соединяющая котел с краном, по холодной внешней стене дома или по теплой внутренней. Таким образом, существуют две концепции зависимости, и употребление одного слова для обозначения обеих может привести к неприятной, если не опасной, путанице.

Кроме того, есть еще математические термины «зависимая переменная» и «независимая переменная», весьма успешно запутывающие ситуацию при обработке данных. Мы постараемся совершенно их избегать.

12.1. РЕГРЕССИЯ: ДВА СМЫСЛА

Регрессионные методы позволяют выявлять связи между переменными, причем особенно эффективно, когда эти связи не совершенны, так что каждому x не соответствует единственный y. В качестве примеров переменных с несовершенными связями можно привести рост и вес людей или их рост, вес и объем талии. Исследования зависимостей между такого рода величинами проводились в науке задолго до появления термина «регрессия». Этот термин возник в исследованиях Гальтона (F. Galton) по биологической наследственности. Гальтон привел пример, показывающий, что у высоких отцов — высокие дети, но все же в среднем не столь высокие, как отцы. Аналогично у маленьких отцов — маленькие дети, но в среднем не такие маленькие, как отцы. Эту тенденцию избранных по некоторым показателям групп приближаться в следующем поколении к среднему популяции, а не воспроизводить средний показатель родителей Гальтон назвал *регрессией*, регрессией по направлению к среднему. Мы сделали это краткое историческое отступление, так как без него термин «регрессия» выглядел бы несколько загадочно.

Регрессия в первом смысле: средние по столбцам (локальные средние). Рассмотрим основные идеи регрессии и корреляции. Что такое

* Здесь и в дальнейшем авторы исключают из рассмотрения математическое понятие многозначной функции. Переводчик счел себя вправе сделать это и некоторые последующие примечания, памятуя обещание авторов «постоянно ... напоминать читателю о том, что не делается». — *Примеч. пер.*

регрессия? Для начала предположим, что имеются две переменные, например рост x и вес y, в большой популяции людей. Тогда для каждого маленького интервала значений x (например, сантиметровой длины) у нас есть набор (распределение) значений веса y. Можно подсчитать какую-нибудь суммарную характеристику (свертку) весов y для этого интервала x. Например, арифметическое или геометрическое среднее, медиану и т. д. Предположим, что для каждого из последовательных односантиметровых интервалов, на которые разбит интервал (скажем, от 162 до 194 см), вычислена избранная нами свертка. Тогда набор точек $(x_i, \bar{y}_i)$, где x_i — центр i-го интервала ростов, а $\bar{y}_i$ — средний вес для данного интервала, вполне возможно, хорошо совместится с некоторой гладкой кривой, например прямой линией. Тогда можно считать, что эта кривая сама есть суммарная характеристика зависимости веса от роста. Такая сглаженная кривая, приближающая линию регрессии, называется регрессией y по x. Более математическое описание регрессии будет дано ниже.

Пример. Возраст достижения выдающегося результата (y) в зависимости от продолжительности жизни (x). Леман [Lehman H. C. (1953)] исследовал распределение возрастов достижения выдающихся результатов в какой-либо из десяти различных областей деятельности. Учитывая, что смерть исключает возможность дальнейших достижений, он сгруппировал данные в соответствии с продолжительностью жизни. Это дало возможность избежать переоценки роли ранних лет жизни в человеческой деятельности. В таблице (илл. 12.1.1) для каждого интервала продолжительности жизни дано распределение возрастов, в которых были достигнуты выдающиеся результаты. Оно соответствует приведенному выше описанию отношения рост — вес. Роль x здесь играет продолжительность жизни, фиксируемая достаточно узкими интервалами, а y — возраст, в котором достигнут выдающийся результат.

Теперь мы постараемся представить информацию, содержащуюся в таблице, в более сжатом виде. Зачем? Например, для того, чтобы прояснить или подчеркнуть связь между соответствующими значениями обеих переменных. Простейший, грубый анализ таблицы показывает, например, что оптимальным для достижения выдающихся результатов получается возраст от 30 до 39, независимо от продолжительности жизни. Если мы хотим иметь обобщенные показатели, то можно вычислить среднее или медиану по каждому столбцу. Нанеся эти данные на график против середин соответствующих интервалов классов продолжительности жизни, мы получим изображение регрессии. (Правда, не совсем ясно, какую продолжительность жизни назначить попавшим в крайние интервалы до 50 или после 85 лет.) Использование медиан для характеристики возраста, в котором достигнут выдающийся результат, позволяет избежать трудностей с выбором возраста для тех, кто отличился, не достигнув 20 лет.

Медианы приводятся внизу таблицы илл. 12.1.1 под соответствующими столбцами. На илл. 12.1.2 показан график с аппроксимирующей точки линией и уравнение этой линии. Из рисунка видно, что каждые дополнительные пять лет жизни увеличивают медиану возраста дости-

жения успеха примерно на 1 год ($\approx$ 5 (0,19)) для умерших между 50 и 85 годами. Две крайние точки (45 и 87 1/2) выбраны произвольно.

Регрессия в первом смысле (формальное определение). Строго математически регрессия y по x определяется следующим образом. Предполагается, что для каждого значения x существует распределение Y с плотностью $f(y \mid x)$* (читается f от y при данном x). Для каждого x вычисляется среднее по формуле

$$\bar{y}(x) = \int_{-\infty}^{\infty} y f(y \mid x)\, dy.$$

Регрессией y по x называется функция, задаваемая множеством упорядоченных пар $(x, \bar{y}(x))$. Здесь мы следуем традиции использовать для определения регрессии свертку в виде среднего арифметического, однако можно было бы взять для этого, скажем, и медиану.

Пример. Среднее. Пусть плотность распределения задается функцией

$$f(y \mid x) = \frac{2y}{x^2}, \quad 0 \leqslant y \leqslant x \leqslant 1.$$

Тогда при заданном x, $0 \leqslant x \leqslant 1$, среднее равно:

$$\bar{y}(x) = \frac{2}{x^2} \int_0^x y \cdot y dy = \frac{2}{x^2}\left(\frac{x^2}{3}\right) = \frac{2}{3}\, x.$$

Следовательно, $\bar{y}$ — линейная функция x, проходящая через начало координат.

Рассмотрим теперь медиану при том же условном распределении y.

Пример. Медиана. Медиана, $y_{\text{мед}}(x)$ — это точка, расщепляющая плотность пополам. Следовательно, нужно требовать, чтобы выполнялось равенство

$$\frac{2}{x^2} \int_0^{y_{\text{мед}}(x)} y dy = \frac{1}{2},$$

или

$$\frac{2}{x^2} \frac{[y_{\text{мед}}(x)]^2}{2} = \frac{1}{2}.$$

Это дает нам

$$y_{\text{мед}}(x) = \frac{1}{\sqrt{2}}\, x.$$

Таким образом, медиана — также линейная функция x, проходящая через начало координат, но под несколько иным углом наклона — около 0,71 вместо 0,67.

---

* Авторы не упоминают, что рассматривают здесь Y как случайную величину; это подчеркивается употреблением прописной буквы вместо строчной, которая используется для значений этой величины. — *Примеч. пер.*

Иногда для каждого y существует распределение X с плотностью $g(x \mid y)$. Тогда мы можем определить регрессию X по y с помощью упорядоченного множества пар $(\bar{x}(y), y)$.

На практике мы редко сталкиваемся с непрерывными совокупностями, для которых известны виды функций. Зато экспериментальные данные могут быть весьма обширны. В этом случае область значений одной из переменных можно разбить на малые интервалы, для каждого из них подсчитать среднее и по полученным точкам провести достаточно простую кривую, которая и даст изображение регрессии, как это было в примере с продолжительностью жизни и возрастом выдающегося достижения.

Роль регрессионной кривой состоит в том, что она служит общей сверткой и иллюстрирует зависимость среднего по распределениям от значения x. Можно было бы пойти дальше и построить различные регрессионные кривые, иллюстрирующие зависимость разных процентилей распределения Y от x. Обычно этого не делается, следовательно, регрессия дает весьма неполную картину. Точно так же, как среднее дает грубую характеристику соответствующего распределения, регрессия будет грубой характеристикой семейства распределений. В этом и состоит смысл, который мы вкладывали в понятие регрессии в данном пункте.

Когда данные более скудны, мы можем увидеть, что вариация в выборке делает безнадежной задачу построения линии регрессии по обычным средним арифметическим.

Регрессия во втором смысле (подбор функции). Один из методов обработки данных — сглаживание. Оно может применяться и для средних по столбцам, и для самих y, упорядоченных по возрастанию x. Читатель может ознакомиться с основами этого метода по параграфам 3.6, 3. 7 и по *EDA* (гл. 7, 16). Этот метод дает сглаженную кривую, правда, не обязательно допускающую какое-нибудь простое функциональное описание. Иногда такой результат хорош сам по себе, а иногда он лишь подсказывает функциональный тип кривой, которую затем можно подогнать под экспериментальные данные.

По необходимости, когда данных совсем мало, рассматривая результаты сглаживания, либо, как в очень многих случаях, из-за отсутствия каких бы то ни было соображений, мы часто используем следующий подход. Задаемся формой кривой (линейная, квадратичная, логарифмическая или какая-нибудь еще), а затем подбираем конкретную кривую из этого класса одним из статистических методов, например методом наименьших квадратов.

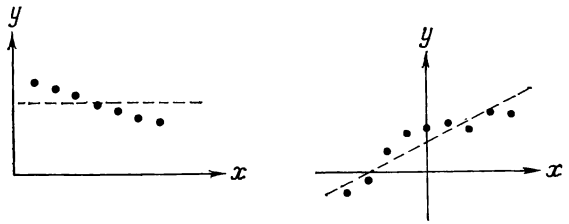
В этом случае мы *ни в коей мере* не претендуем на то, что получившаяся кривая имеет ту же форму, что и регрессионная кривая, которую можно было бы построить, имей мы неограниченные данные. Построенная таким образом кривая — не более чем аппроксимация.

Условия, вынуждающие нас использовать второй подход к регрессии — подгонку кривой определенного типа, возникают весьма часто. Поэтому мы склонны забыть о первой более фундаментальной концепции регрессии, которая, напомним, состоит в построении ломаной, соединяющей средние по распределениям столбцов. Вторая концепция

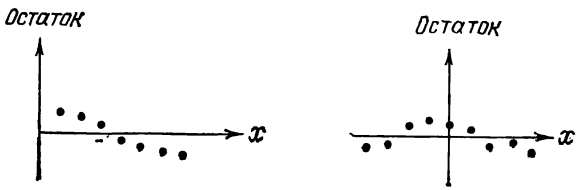
значительно расширяет область применимости регрессионных методов, так как на практике мы часто располагаем скромным набором экспериментальных данных, значительно меньшим, чем требуется для метода сглаживания, и совершенно недостаточным для эффективного применения первой концепции (здесь требуются тысячи или по крайней мере сотни пар (x, y)).

Обычно мы выбираем для построения регрессии кривую с относительно небольшим числом параметров. И мы хотим знать, как их подобрать. (Это можно сделать с помощью нескольких критериев, таких как метод наименьших квадратов, наименьших модулей, наименьших p-х степеней и т. д., в общем, любого удобного нам метода. Еще это можно сделать, рассматривая качество подгонки в терминах получающихся остатков. Кроме того, возможно и использование описания этапов построения подгонки. И наконец, есть и разные комбинации перечисленных подходов.)

В процессе подгонки простой кривой иногда выясняется, что нужна более сложная. Например, пытаясь подогнать к нашим экспериментальным данным горизонтальную прямую $y = c$, где c — константа, можно обнаружить, что требуется наклонная прямая. Или, строя наклонную прямую, убедиться в потребности криволинейной регрессии.



Проще всего выявлять это с помощью графика остатков (y — предсказанное y). Такой график полезно сгладить, проведя на глаз или с помощью формальных приемов гладкую кривую.



БОЛЕЕ ЧЕМ ОДИН НОСИТЕЛЬ

До сих пор мы рассматривали в основном (хотя были и исключения) регрессию одной переменной y (отклика) по другой переменной x (фактору), который мы будем называть носителем. Однако все, что сказано, можно обобщить и на случай более чем одного носителя. (Как правило, говоря о носителе, мы подразумеваем, что он не константа.) Важным шагом будет уже переход к двум носителям. Этот переход об-

легчается тем, что связанные с ним геометрические построения проводятся в трехмерном пространстве, а здесь можно надеяться на нашу интуицию. Привыкнув к случаю двух носителей, мы сделаем второй шаг — большое, но заданное число носителей. Третий шаг — неопределенное число носителей. Отбор носителей в этой ситуации на всех его этапах может оказаться весьма трудной задачей.

12.2. ЗАЧЕМ НУЖНА РЕГРЕССИЯ?

1. Простой, но важной целью регрессии служит обобщение накопленной информации (свертка!). Мы показали ее в параграфе 12.1 на примере анализа данных о возрасте и достижениях.

2. Тот же пример иллюстрирует другое применение регрессии — для устранения переменной, которая могла бы затруднить интерпретацию. В нашем примере такой переменной была продолжительность жизни.

3. Регрессия часто используется при попытках установить причинную связь. И это также иллюстрируется упоминавшимся примером, хотя ясно, что наличие или отсутствие выдающихся достижений не может зависеть исключительно от возраста.

4. В продолжение сказанного в предыдущем пункте можно указать еще на одно возможное использование регрессии — количественное измерение эффекта с помощью коэффициента регрессии. Это тоже делалось в нашем примере. Однако, как отмечается в гл. 13, такое измерение может наталкиваться на трудности в тех случаях, когда действует много причин наряду с «беспричинными» носителями, действующими на отклик.

5. Причинный подход дает наилучшие результаты, когда устанавливается математический или эмпирический закон.

6. Но шире всего регрессия используется для предсказаний. Например, для прогноза вероятности дождя в данном районе через определенное время на основе информации от нескольких метеостанций. Или для прогноза уровня смертности от болезней, которые, как считают, вызываются загрязнением атмосферы, по данным о составе популяции и загрязненности окружающей среды.

Использование регрессии для предсказаний широко распространено и разнообразно. Действительно, иногда можно ожидать, что такое использование выявит и причинную связь, как в пункте 4,— если уровень загрязнения окружающей среды уменьшается, то можно предполагать, что это повлияет на процент смертности. (Хотя результаты такого типа иногда бывают обманчивы.) Другой вариант использования уравнения регрессии для предсказания описан в пункте 2. В нашем примере это оценка процента смертности в другом месте («при прочих равных условиях») для сравнения со «средним» показателем.

Можем попробовать подобрать $y = f(x)$, где f определяется эмпирически среди

$$1)\ y = bx,\ 3)\ y = a + bx + cx^2,$$
$$2)\ y = a + bx,\ 4)\ y = a + bx + ct$$

и некоторых их обобщений на случай большего числа носителей (больше коэффициентов, больше переменных либо и то и другое). Задачи, которые нам придется решать, в разных случаях будут различными.

Регрессия для исключения. 7. Регрессия иногда используется, чтобы «убрать x с дороги» с помощью модели, скажем,

$$y = a + bx + cx^2.$$

Так, нас может интересовать, влияет ли на отклик переменная t, когда известно, что другая переменная, x, влияет на него существенно. (Здесь слово «влияет» используется как сокращение для «связано (возможно, но не наверняка) с помощью причинно-следственного механизма». Аналогично мы будем пользоваться и выражением «эффекты x».)

Например, мы хотим исключить эффект образовательного уровня, выясняя связь между осведомленностью в текущих событиях и чтением прессы.

Одним из подходов к решению этой задачи мог бы быть такой: вычесть из y «эффекты» x, а затем определить, как эта разность зависит от t. Мы можем попробовать, например, описать «эффекты» x соответствующим образом подогнанной квадратичной кривой

$$y = a + bx + cx^2,$$

а затем рассмотреть остаток

$$y_{\bullet x} = y - a - bx - cx^2.$$

Эта величина «свободней» от воздействия x, чем исходный y.

Иногда удается оценить условное распределение y для каждого x, а затем заменить y на некоторую его количественную характеристику относительно этого условного распределения. Примером такой характеристики может служить число стандартных отклонений от среднего. Использование этого метода, естественно, облегчается, если x может принимать малое число значений, например, в ситуации, когда x — пол (мужской или женский) или когда множество значений x разбивается на небольшое число интервалов, как это было в нашем примере с продолжительностью жизни и достижениями.

В ситуациях такого типа мы сравниваем y с тем, что можно предсказать по значению x (или нескольких x, если нам известны значения нескольких переменных). Не следует придавать большого значения коэффициентам, формулам, их интерпретации и т. д. на предварительной стадии исследования, связанной с исключением x. Естественно, надо внимательно отнестись к вопросу о выборе носителей, позаботиться об их разумном отборе, что само по себе не слишком просто. Но придавать смысл коэффициентам, по крайней мере на этом этапе, не обязательно. Если расчеты показали, что отметка за французский диктант y связана с возрастом испытуемого x формулой

$$y \approx 50 + 7\,(x - 10) - 0{,}5\,(x - 10)^2,$$

то вряд ли коэффициенты 50 (при 1), 7 (при $(x - 10)$) и — 0,5 (при $(x - 10)^2$) не изменятся в других экспериментах. Там вполне возможны

другие кривые. Исключив x, мы можем взять формулу

$$y - 50 - 7\,(x - 10) + 0{,}5\,(x-10)^2,$$

не задумываясь над тем, что будет в иных ситуациях.

Более того, часто достаточно выбрать хорошее приближение, не задумываясь над тем, будет ли оно из наилучшего класса. Так, если дети не изучают французский до семи лет, то зависимость y от x при $x < 7$ должна быть близка к $y = 0$ (если, конечно, 0 обозначает отсутствие знаний). Зависимость такого рода трудно приблизить полиномом, но если в нашем исследовании участвуют лишь дети старше 9 лет, то квадратичная функция может оказаться вполне подходящей.

Такие модели — инструмент практики. Действительно, очень часто простое выражение — уже достаточно хорошее описание. Оно может также стать верным путем к соотношению.

В следующей главе мы столкнемся с другой промежуточной проблемой: если вид зависимости выбран, то как быть с коэффициентами? Там нашей целью будет получение коэффициентов, которые допускают разумную интерпретацию, тогда как в данной главе наша цель — хорошая подгонка для исключения x или для предсказания y по x.

Может показаться, что эти две цели идентичны, но это не так. Они действительно похожи, но не тождественны.

Рассмотрим чуть внимательнее одну конкретную задачу предсказания. Предположим, что речь идет о приеме студентов в большой государственный университет. Предположим далее, что вопрос о приеме в число слушателей должен решаться на основе некоторой формулы, прогнозирующей степень успешности обучения. Как же выбрать переменные и коэффициенты в этой формуле? Предсказание, естественно, должно по крайней мере частично оправдываться.

Стоит ли, например, брать в качестве одного из x род занятий отца абитуриента? Если да, то каким должен быть коэффициент? Если это бесполезная переменная, то вопрос ясен. Если же полезная, то существуют моральные и политические соображения, по которым она может быть, а может и не быть включена в формулу. (В меньших учебных заведениях, из тех, что гордятся сбалансированным набором студентов, x может использоваться вовсе не для прогноза успешности обучения.)

Итак, выбирая переменные по их полезности, доступности, а также по моральным и политическим соображениям, мы получили некоторое их множество. Мы хотим их использовать при построении прогнозирующей формулы. Здесь нет нужды обращать особенное внимание на величину коэффициентов. Нас должен в основном заботить прогноз того, что можно сказать об y по данным x. И насколько предсказанное значение y может отличаться от истинного? Многие формулы могут работать для этой цели одинаково успешно. Нас устроит любая из них. (Если же мы попытаемся понять, что определяет студенческую успеваемость, то потребуется совершенно иной подход.)

Еще несколько слов о стратегии. Выбирая одну формулу из многих возможных, следует задуматься, не создает ли ее использование в каких-то ситуациях явной несправедливости или эффектов, которые могут послужить темой для веселых фельетонов. Следует сознавать, что

такая формула становится общественным достоянием и будет обсуждаться в лучшем случае нейтральными, а скорее всего предубежденными людьми. Например, если формула такова, что при ее применении правильное или неправильное написание одного слова оказывает такое же влияние, как и оценка по двухлетнему курсу математики, то будет трудно убедить, что эта формула столь же хороша, как и многие другие (как бы верно это ни было). Таким образом, из многих возможных формул следует выбрать ту, которую можно разумно представить и защищать.

В этой главе мы рассматриваем более легкую проблему — подбор хорошей модели, которая годится для исключения x или предсказания y, причем структуре этой модели и ее коэффициентам во всяком случае пока серьезного значения не придается.

12.3. ГРАФИЧЕСКАЯ ШАГОВАЯ ПОДГОНКА

Пусть мы хотим приблизить y с помощью линейной комбинации $\beta_1 x_1 + \beta_2 x_2$. Это можно сделать непосредственно. Но для лучшего понимания процедуры мы используем последовательный подход. Мы приблизим y, используя не x_1 и x_2, а x_1 и x_2, скорректированные по x_1, который обозначим $x_{2;1}$. А именно мы построим

$$x_{2;1} = x_2 - d_{2;1}\, x_1,$$

где $d_{2;1}$ выбрано так, чтобы освободить x_2 от влияния x_1 по крайней мере приближенно. В качестве $d_{2;1}$ можно выбрать коэффициент наклона прямой (проходящей через начало координат на плоскости x_1 — $-x_2$), приближающей точки x_2 по значениям x_1. Тогда $x_{2;1}$ — множество остатков x_2 по отношению к прогнозу $d_{2;1}\, x_1$. В этом параграфе мы не воспользуемся непосредственно методом наименьших квадратов для определения $d_{2;1}$, хотя и могли бы. Вместо этого применим к x_2 процедуру, часто применявшуюся нами к самим y и x, когда у нас были лишь y и x. Общий план состоит в следующем. Прежде всего приблизим y с помощью $\hat{\gamma}_1 x_1$

$$y = \hat{\gamma}_1 x_1 + \text{остаток}.$$

После того как это сделано, вычислим остаток

$$y_1 = y - \hat{\gamma}_1 x_1.$$

Индекс«; 1» указывает на то, что вычисляется остаток для y, а 1 означает, что в некотором смысле устранено влияние x_1. Далее, мы приближаем этот остаток

$$y_{;1} = \hat{\beta}_2\, x_{2;1} + \text{остаток}.$$

Таким образом, мы связываем y, скорректированный по x_1, с x_2, скорректированным по x_1. Объединяя вышеприведенные выкладки, видим, что

$$y = \hat{\gamma}_1\, x_1 + y_{;1}$$

приближается с помощью

$$\hat{\gamma}_1 x_1 + \hat{\beta}_2\, x_{2;1}.$$

При желании можно переписать это во вполне эквивалентной форме, где y приближается с помощью $\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$, если положить

$$\hat{\beta}_1 = \hat{\gamma}_1 - \hat{d}_{2;1}\,\hat{\beta}_2.$$

Следующий пример иллюстрирует метод последовательного исключения влияния переменных по одной.

Переменная, скорректированная по другой, будет отмечаться индексом, включающим точку с запятой. В этом разделе мы будем подбирать коэффициенты эмпирически (графическим методом), но постараемся делать это тщательно, в два этапа.

В таблице илл. 12.3.1 приведены данные за 1947—1962 гг. о ежегодных отклонениях уровня занятости (y), снижения индекса цен (x_1) и валового национального продукта (x_2) от средних значений этих показателей за те же годы. В этой таблице уровень занятости исчисляется в 1000 рабочих мест, снижение индекса цен — в процентах, валовой национальный продукт — в 1000 долларов. (Средние значения этих показателей — 65317, 101,7 и 387698 соответственно.)

Илл. 12.3.2 и 12.3.3 иллюстрируют два этапа приближения y с помощью x_1. На первом этапе на глаз проведена прямая с угловым коэффициентом + 300, приближающая зависимость между y и x_1.

Мы видим на илл. 12.3.1, что

$$y'_{;1} = -4994 - 300\,(-18{,}7) = 616.$$

Здесь и далее приближение переменной, полученное на первом этапе, обозначается штрихом над этой переменной. Следующий этап, приближение $y'_{;1} = y - 300\,x_1$ с помощью x_1, изображен на илл. 12.3.3. Здесь угловой коэффициент равен 5. Таким образом, окончательно угловой коэффициент между y и x_1 равен 305 и $y_{;1}$ есть $y - 305x_1$. Значит, в соответствии с данными таблицы илл. 12.3.1 для 1947 г.

$$y_{;1} = -4994 - 305\,(-18{,}7) \approx 710.$$

(Вся необходимая арифметика приведена в таблице илл. 12.3.4.)

Илл. 12.3.5 и 12.3.6 показывают аналогичную процедуру для x_2 и x_1. В данном случае второй член составляет 17% первого. (Вычисления приведены в таблице илл. 12.3.6.) Из данных таблицы илл. 12.3.1 следует, что

$$x'_{2;1} = -153409 - 10000\;x_1 = 33591,$$

а следовательно,

$$x_{2;1} = 33591 + 1700\;x_1 = 1861.$$

Мы можем сделать теперь заключительный шаг — провести регрессию $y_{;1}$ по $x_{2;1}$. На илл. 12.3.7 показан первый этап. Не всегда легко сразу правильно подобрать регрессионный коэффициент. Мы начнем с 0,07. Из илл. 12.3.8 видно, что это, похоже, несколько больше, чем надо. Но мы остановимся на нем, так как точки на илл. 12.3.8 расположились, в общем, вдоль горизонтали. Исключение составляют две точки — последние, 1961 и 1962 гг. (Некоторые аналитики, возможно, перешли бы к другой регрессии с угловым коэффициентом порядка —0,015 или — 0,02.)

Итак, мы получили следующие грубые приближения: 305 x_1 для y, значит $y_{;1} = y - 305x_1$; 8300 x_1 для x_2, значит $x_{2;1} = x_2 - 8300x_1$; $0{,}07x_{2;1}$ для $y_{;1}$, значит $y_{;12} = y_{;1} - 0{,}07\ x_{2;1}$, где $y_{;12}$ — остаток после корректировки y по x_1 и $x_{2;1}$ или, более кратко, по x_1 и x_2. Объединяя все вышесказанное, получим

$$y_{;12} = y_{;1} - 0{,}07\ x_{2;1} = (y - 305x_1) - 0{,}07 \times$$
$$\times (x_2 - 8300x_1) = y + 276x_1 - 0{,}07x_2.$$

Таким образом, нашим приближением y будет

$$-276x_1 + 0{,}07\ x_2.$$

Если бы мы взяли угловой коэффициент 0,01 при приближении $y'_{;12}$ с помощью $x_{2;1}$, то получили бы приближение

$$-199\ x_1 + 0{,}06\ x_2.$$

Эти цифры показывают, как сильно может повлиять на коэффициенты даже небольшое изменение в приближении. (Только одна точка, отвечающая 1962 г., лежит за границами интервала $\pm$ 200, образуя тренд в районе точки 10000.)

12.4. КОЛЛИНЕАРНОСТЬ

В регрессионных задачах кроется масса подводных камней. Один из них хорошо известен под именем «проблемы коллинеарности». В простейшем случае эта проблема возникает, когда мы исследуем под разными названиями практически одну и ту же величину, а затем используем разные ее меры для построения регрессии между почти эквивалентными носителями и откликом. По существу, эта трудность возникает тогда, когда мы пытаемся рассматривать одну порцию информации как две разные. Это неизбежно ведет к произволу в назначении весов для используемых источников информации. Такая ситуация легко может возникнуть при исследовании социальных или экономических задач, где в регрессию включается множество переменных, часто измеряющих одну и ту же вещь. Мы покажем это в конце параграфа на примере о верующих

Пример. Размеры сердца и грудной клетки. Биостатистик провел две серии измерений объема грудной клетки двенадцатилетних мальчиков с интервалом в несколько месяцев. Кроме того, были проведены прямые исследования объемов их сердец. Эти измерения использовались для построения регрессионного уравнения, предсказывающего размеры сердца. Было получено уравнение

$$H = 10 + 8C_1 + 3C_2, \qquad (1)$$

где H — объем сердца; C_1 — первый замер объема грудной клетки; C_2 — второй.

Уравнение было получено методом наименьших квадратов по данным обследования одной группы мальчиков. Однако когда в той же части страны была обследована другая группа мальчиков того же воз-

раста, то оказалось, что регрессионное уравнение выглядит примерно так:

$$H = 10 + 5C_1 + 6C_2. \qquad (2)$$

Наш биостатистик, естественно, был смущен значительной разницей в коэффициентах, так как обе группы были достаточно представительными и вполне сравнимыми. Какому же из уравнений нужно верить?

В данном случае произошло следующее. Для каждого мальчика C_1 и C_2 отличались очень мало. Причем различия были обусловлены в основном не естественным ростом, а ошибками измерений. Можно было бы использовать одно усредненное значение объема грудной клетки C, что привело бы к регрессионному уравнению вроде

$$H = 10 + 11\bar{C}. \qquad (3)$$

Отметим, что коэффициент 11 есть сумма коэффициентов C_1 и C_2 в обоих уравнениях (1) и (2) (в реальном примере суммы этих коэффициентов почти совпадали).

Геометрическое истолкование. С геометрической точки зрения экспериментальные точки лежат в трехмерном пространстве (C_1, C_2, H) практически на одной прямой. Поэтому построение плоскости по этим точкам становится невозможным. Если точки лежат близко к прямой, то задача «плохо определена» — небольшие изменения положения точек будут «раскачивать» плоскость, как качели.

Оценивание. Несмотря на значительное различие коэффициентов двух оценок H:

$$10 + 8C_1 + 3C_2 \text{ и } 10 + 5C_1 + 6C_2,$$

с точки зрения приложений они предсказывают H практически одинаково. Геометрически это можно объяснить тем, что все точки (C_1, C_2) должны быть близки к прямой $C_1 = C_2$ в плоскости (C_1, C_2). Поэтому вертикальная проекция каждой из этих точек на регрессионную плоскость (любую из большого числа возможных) будет весьма близка к первоначальной линии, проведенной по экспериментальным данным в трехмерном пространстве. А следовательно, все эти плоскости с одинаковым успехом можно использовать для приближения или предсказания.

Таким образом, если взять две надежные, одинаково распределенные меры одной и той же величины как независимые переменные в уравнении множественной регрессии, то мы видим, что

(а) сумма их регрессионных коэффициентов относительно устойчива и

(б) устойчиво предсказание отклика (ввиду устойчивости, отмеченной в пункте (а)).

Естественно, выводы, сделанные нами при изучении этого простого случая, имеют значение и для других ситуаций, иначе не стоило бы его рассматривать. Всякий раз, когда переменные, используемые в регрессии, сильно коррелированы либо некоторые из них описывают одно и то же, величины коэффициентов, по всей вероятности, будут весьма не-

определенно указывать «важность» переменных. Следует отдавать себе отчет, что тогда, мягко говоря, трудно предсказать, как повлияет на отклик изменение одной из переменных. Напомним, что в случае сильно коррелированных коэффициентов мы не можем заметно изменить значение одной переменной, не изменяя другие.

Пример. Количество прихожан. Если считать, что количество прихожан (M) зависит от активности церкви, и использовать в качестве меры проделанной церковью работы размеры выплат на заработную плату ее служащим (L) (в действительности заработная плата плюс 10% стоимости дома священника), а в качестве величины капитала принять стоимость принадлежащих ей строений минус сумма задолженности по ним (K), то экономист может попробовать связать эти величины формулой

$$M = 10^{\beta_0} L^{\beta_L} K^{\beta_K}.$$

Логарифмируя это выражение, получим

$$\log M = \beta_0 + \beta_L \log L + \beta_K \log K.$$

В таблице илл. 12.4.1 приведены значения логарифмов M, L, K в ряде штатов для методистской епископальной церкви по данным церковной переписи 1936 г. До подбора регрессионной кривой был построен график зависимости $\log L$ от $\log K$ (см. илл. 12.4.2). На рисунке мы видим, что $\log L$ и $\log K$ связаны почти линейно с угловым коэффициентом около 1. Это служит указанием на то, что мы вновь столкнулись с коллинеарностью, как и в примере с измерением объемов грудной клетки и сердца. А следовательно, коэффициенты β_L и β_K будут плохо определяться экспериментальными данными, хотя оценка для $\log M$ вполне может быть получена по $\log L$, $\log K$ или их комбинации. Рассмотрим зависимость $\log M$ от суммы $\log L$ и $\log K$. Чертеж приведен на илл. 12.4.3.Вновь зависимость имеет линейный характер с угловым коэффициентом, приблизительно равным 1/2.

12.5. ТОЧНАЯ И ПРИБЛИЖЕННАЯ ЛИНЕЙНАЯ ЗАВИСИМОСТЬ

Когда регрессия строится по нескольким носителям, их часто называют *независимыми переменными* (*факторами-предсказателями*).

Предположим, что имеет место предельный случай коллинеарности, а именно несколько факторов связаны строго линейной зависимостью. Так может быть, если факторы в сумме составляют 100% (например, компоненты сплава или процентное распределение семей по видам жилищных условий). Такая ситуация может возникнуть также в случае избытка переменных. Например, при изучении числа детей в многодетных семьях факторы «порядковый номер ребенка», B, и «число детей, старших, чем он», Y, связаны соотношением

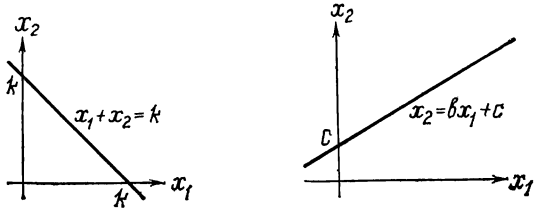
$$B + Y = N,$$

где N — число детей в семье.

Другой пример, представляющий интерес для социальных и клинических психологов, дают тесты типа описанных в книге A l l p o r t

G. W., V e r n o n P. E., L i n d z l y G. Study of Values, где сумма баллов для отдельных частей теста постоянна.

Как зависимость такого рода влияет на подбор уравнений множественной регрессии? Она создает трудности, похожие на те, что возникают в арифметике при попытках деления на нуль. Пусть переменные x_1 и x_2 в сумме равны константе K, $x_1 + x_2 = K$, или связаны линейной функцией $x_2 = b\,x_1 + c$.



Предположим, что x_1 и x_2 — факторы, а y — отклик. Тогда все значения отклика в трехмерном пространстве (x_{1i}, x_{2i}, y_i) лежат в плоскости, перпендикулярной к плоскости (x_1, x_2) и проходящей через прямую, связывающую x_1 и x_2. Это сильно напоминает пример с измерениями объема грудной клетки и сердца. Из рисунков видно, что мы имеем дело не с двумя факторами, а лишь с одним. Существует множество возможностей для описания такой ситуации. В случае когда $x_1 + x_2 = K$, мы можем обойтись только первым или, наоборот, вторым фактором либо их любой линейной комбинацией. Если мы воспользуемся линейной комбинацией $ax_1 + bx_2 + c$ этих факторов, то надо брать коэффициенты a и b неравными, так как изменение x_1 приводит к такому же по величине, но обратному по знаку изменению x_2. Поэтому, если $a = b$, то линейная комбинация будет постоянной при всех значениях факторов, и мы потеряем информацию от обоих носителей.

Во втором случае мы также можем брать любую линейную комбинацию x_1 и x_2, кроме комбинаций вида $a\,(x_2 - bx_1) + d$, так как они постоянны при всех значениях носителей.

Единственный приемлемый выход из ситуации полной коллинеарности, описанной в предыдущих абзацах, — это отказ от попыток использовать оба фактора как различные. Надо заменить их одним. Это упорядочит вычислительную сторону дела и поможет (может быть) в предсказании.

Вопрос о том, какой из носителей выбрать, никак не связан с имевшими место рассуждениями. В случае регрессии для прогноза эту проблему надо решать, исходя из соображений удобства или связей с другими носителями. В тех случаях, когда регрессия используется для целей измерения или как индикатор причинной связи, выбор носителя, как правило, весьма затруднен, поскольку здесь коэффициенты обычно должны иметь смысл.

Простейшие полиномы; почти линейная связь. Одна из форм почти коллинеарности, порожденная функциональной связью, возникает, когда мы пытаемся построить линейную функцию относительно носителей 1, x и x^2, т. е. квадратичную функцию $a + bx + cx^2$ относительно

исходных данных. Так как x и x^2 тесно связаны, мы можем столкнуться с вычислительными трудностями. Это приведет к необходимости сохранять в вычислениях много знаков, чтобы обеспечить реальную точность. Вычисляя среднее, мы обычно пользуемся следующим приемом. Выбирается какая-нибудь близкая к центру величина, подсчитываются отклонения от нее, эти отклонения усредняются и возвращаются к действительному среднему. *Пример*: усреднить 13879, 13881, 13864. Удобно в качестве центрального значения взять, например, 13870. Тогда отклонения составляют 9, 11, — 6, их сумма 14, а среднее отклонение $4\frac{2}{3}$. Значит, среднее составляет $13874\frac{2}{3}$. Люди, делающие вычисления в уме, обычно пользуются такого рода приемами. Идея здесь состоит в том, что исходные числа 5-значные, а отклонения 1-, 2-значные, что упрощает счет.

В множественной регрессии, особенно в случае сильно коррелированных переменных, для уменьшения числа десятичных знаков используется аналогичный прием, может быть, даже более нужный и полезный в данной ситуации. Подбор кривой $a + bx + cx^2$ эквивалентен подбору кривой

$$a^* + b^* x + c(x^2 - A - Bx)$$

при любых A и B, удовлетворяющих $a = a^* - Ac$ и $b = b^* - Bc$, и тем же C, что и выше. Если x_0 около центра множества данных, то $(x - x_0)^2$ будет мало по сравнению с x^2. Кроме того,

$$(x - x_0)^2 = x^2 - 2x_0x + x_0^2$$

имеет вид

$$x^2 - Bx - A$$

при $A = -x_0^2$ и $B = 2x_0$. Следовательно, должно быть выгодно подгонять

$$a^* + b^* x + c(x - x_0)^2,$$

или даже

$$a^{**} + b^*(x - x_0) + c(x - x_0)^2.$$

Помимо упрощения вычислений эта форма еще проливает дополнительный свет на неопределенность оценок коэффициентов. Если все значения x близки к x_0, то носитель $(x - x_0)^2$ очень мал по сравнению с $x - x_0$ или 1, значит, по причинам, которые мы объясним ниже, нельзя ожидать оценки c с достаточной точностью.

Как говорится в гл. 14, знаменатель в формуле для дисперсии $\hat{C}$ представляет собой сумму квадратов остатков от регрессии $(x - x_0)^2$ по 1 (что важно) и $x - x_0$ (что обычно менее важно). Иными словами, мы приближаем $(x - x_0)^2$ функцией вида $\gamma_0 + \gamma_1 (x - x_0)$, а сумма квадратов остатков образует нужный нам знаменатель. Таким образом, наши трудности связаны не только с малостью $\Sigma(x - x_0)^2$. Это не имеет, правда, большого значения для наших ближайших целей — подбора «хорошего приближения», так как величины $(x - x_0)^2$, которые мы подставляем, тоже, по-видимому, будут малы в сравнении с 1 и $x - x_0$. (Другой вопрос — экстраполяция. Здесь параболическое приближение гораздо опасней, так как вклад от квадратичного члена может погубить все дело.)

Возвращаясь к вычислительным проблемам, отметим, что использование формы

$$a^{**} + b^{*}(x - x_0) + c(x - x_0)^2$$

вместо

$$a + bx + cx^2$$

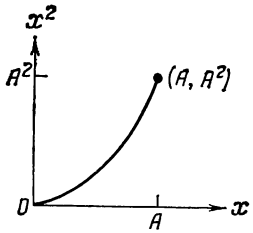
часто бывает очень существенным.

Есть несколько «стандартных» программ, которые подгоняют первую форму автоматически, а затем делают пересчет для второй и распечатывают константы a, b и c. Программы, осуществляющие подгонку непосредственно второй формы, могут дать пользователям вводящие в заблуждение результаты, если набор носителей недостаточно хорош.

Даже в тех случаях, когда интервал значений x кажется достаточно большим, корреляция между x и x^2 может быть значительной. Например, если x равномерно распределен в интервале $(0, A)$, то корреляция между x и x^2 равна 0,97, независимо от величины A. Такую степень связи вполне можно ассоциировать с идеей практически строгой линейной зависимости. Здесь дисперсия наших коэффициентов увеличивается на сомножитель, близкий к

$$\frac{1}{1-r^2} = 12,$$

по сравнению с приближением, использующим либо x, либо x^2. И снова мы сталкиваемся с почти линейной связью.



Несмотря на все эти трудности, определение и исключение влияния квадратичного члена на диаграмме рассеивания может быть полезным.

Легко предсказать, что вычислительные трудности и неопределенность в коэффициентах еще возрастут при подгонке кубического полинома или полинома еще более высокого порядка к любому набору данных.

Мы использовали x и x^2 лишь как легкий пример; почти линейные или точные функциональные связи — общее явление, когда в регрессии участвует несколько фактически эквивалентных переменных. В начале следующего параграфа покажем это на примерах.

12.6. ИСКЛЮЧЕНИЕ ПЛОХО ИЗМЕРЯЕМОГО, РЕГРЕССИЯ ДЛЯ ИСКЛЮЧЕНИЯ

(Прежде чем приступать к этому параграфу, читателю может быть полезно ознакомиться с введением и первым параграфом гл. 14.)

Чтобы достаточно чисто исключить эффект фактора (седьмое направление использования регрессии см. в 12.2), который нельзя оценить непосредственно, но который связан с чем-то, поддающимся измерению, нужны дополнительные данные и специальные приемы. Примером такой скрытой переменной может служить понятие человеческой зрелости, которая, конечно, связана с возрастом, но не определяет-

ся им. Чтобы контролировать эффект подобной переменной, помимо возраста, нужна еще хотя бы одна мера.

Когда нет отклонений. Прежде всего опишем вкратце обычный подход. Если x измеряется точно и мы хотим исключить его влияние из y, то мы должны исследовать остаток

$$y - bx$$

или

$$y - a - bx,$$

где $a + bx$ — наилучшее доступное нам приближение для y. Так как x измеряется без ошибок, процедура исключения влияния x аналогична предсказанию по x. Но поскольку оценка b не будет точной относительно истинного β, то и исключение линейной комбинации x не будет идеальным. Можно записать

$$y - bx = (y - \beta x) - (b - \beta)\, x$$

либо

$$y - a - bx = (y - \alpha - \beta x) - (a - \alpha) - (b - \beta)x,$$

где $y - \beta x$ (или $y - \alpha - \beta x$) — то, что нам нужно. Отклонение $(b - \beta)x$ будет несмещенным относительно среднего, поскольку поведение этого отклонения не зависит от набора данных так же как и $b - \beta$. Если приближение достаточно хорошее, то $(b - \beta)x$ будет то положительным, то отрицательным.

В тех случаях, когда интересен не только вклад x сам по себе, можно испробовать дополнительные члены или иные модели. Например, можно взять остаток

$$y - bx - cx^2,$$

где $bx + cx^2$ — лучшее приближение к y.

Обычная ситуация. Но как быть, если x — полезная (но не полная) мера фактора, влияние которого мы хотим исключить? Пусть мы изучаем поведение детей и хотим исключить (учесть) эффект зрелости. Обычно нам известен возраст, который, конечно, говорит многое, но не все о степени зрелости. Нам нетрудно использовать возраст в предсказании, но как быть с ним при учете зрелости?

Можно сформулировать проблему в более общих терминах, а именно: есть переменная x, приближенно измеряющая z; как использовать x для исключения влияния z на y? Оказывается, что для этого нам нужна по крайней мере еще одна переменная, мы будем называть ее инструментальной. Примером такой переменной может служить результат специального теста на зрелость.

Говоря об *инструментальной переменной* u, мы подразумеваем, что

1) u тоже связано с z;

2) остаток от вычитания из u ее «истинной» регрессии по z не коррелирован (имеет нулевую ковариацию) с остатками x и y.

Более точно, предполагается следующая статистическая модель:

$$x = \alpha_x + \beta_x z + \text{остаток}_x,$$
$$y = \alpha_y + \beta_y z + \text{остаток}_y,$$
$$u = \alpha_u + \beta_u z + \text{остаток}_u,$$

β_x, β_y, β_u — регрессионные коэффициенты при z,

$$\text{cov}\ \{\text{остаток}_x,\ \text{остаток}_u\} = 0,$$
$$\text{cov}\ \{\text{остаток}_y,\ \text{остаток}_u\} = 0.$$

Можно испробовать линейные функции от u или x для исключения z из y. Например, если мы хотим исключить влияние z из y, вычитая из последнего x с некоторым весом, то надо умножить x на β_y/β_x, в результате чего $\beta_x z$ перейдет в $\beta_y z$. Вычитая, получим

$$y - \frac{\beta_y}{\beta_x} x = \beta_y z + \text{остаток}_y - \frac{\beta_y}{\beta_x}\beta_x z - \frac{\beta_y}{\beta_x}\ \text{остаток}_x + C =$$
$$= \text{остаток}_y - \frac{\beta_y}{\beta_x}\ \text{остаток}_x + C,$$

где $C = \alpha_y - (\beta_y/\beta_x)\,\alpha_x$. (Величина C сейчас нам не важна.) Трудность состоит в том, что мы не можем непосредственно получить нужную нам оценку β_y/β_x, так как β_y и β_x — регресионные коэффициенты z, которые нам неизвестны.

Вычисляя ковариации для x и y с u, мы видим, что в нашем идеальном случае

$$\text{cov}\ \{x, u\} = \text{cov}\ \{\beta_x z, \beta_u z\} = \beta_x \beta_u\ \text{var}\ \{z\},$$
$$\text{cov}\ \{y, u\} = \text{cov}\ \{\beta_y z, \beta_u z\} = \beta_y \beta_u\ \text{var}\ \{z\},$$

в чем можно убедиться, подставляя x, y и u в первые равенства и замечая, что в обеих строчках три из четырех ковариаций исчезают. Соответственно найденные по методу наименьших квадратов регрессионные коэффициенты x и y по u выражаются формулами

$$\beta_{xu} = \frac{\text{cov}\ \{x, u\}}{\text{var}\ \{u\}} = \beta_x \frac{\beta_u\ \text{var}\ \{z\}}{\text{var}\ \{u\}}$$
$$\text{и}\ \beta_{yu} = \frac{\text{cov}\ \{y, u\}}{\text{var}\ \{u\}} = \beta_y \frac{\beta_u\ \text{var}\ \{z\}}{\text{var}\ \{u\}},$$

так что $\dfrac{\beta_{yu}}{\beta_{xu}} = \dfrac{\beta_y}{\beta_x}$.

Теперь мы можем оценить β_{yu} с помощью b_{yu} (так как y и u известны) и β_{xu} с помощью b_{xu}. Это дает нам возможность использовать b_{yu}/b_{xu} как разумную оценку β_y/β_x, а $y - (b_{yu}/b_{xu})\,x$ как разумно скорректированное (по z, а не только по x) y. Сказанное остается верным во всех случаях, когда для u выполняются условия некоррелированности

остатков и cov $(x, u) \neq 0$. (Экономисты, возможно, назвали бы проделанную нами работу переходом от прогнозирующей регрессии к структурной.)

Переменная u часто бывает довольно грубым инструментом. И все же это лучше, чем ничего. Даже если u принимает всего два-три значения, разбивающих наши данные на группы, ее вполне можно использовать. Конечно, при условии надежной некоррелированности остатков и коррелируемости с x.

Заметим, что b_{yu}/b_{xu} обычно больше, чем b_{yx}. Действительно, пусть cov {остаток$_x$, остаток$_y$} = 0. Тогда

$$\beta_{yx} = \frac{\text{cov}\,\{y,\ x\}}{\text{var}\,\{x\}} = \frac{\beta_y\,\beta_x\,\text{var}\,\{z\}}{\beta_x^2\,\text{var}\,\{z\} + \text{var}\,\{\text{остаток}_x\}},$$

а β_y/β_x можно записать в виде

$$\frac{\beta_y}{\beta_x} = \frac{\beta_y\,\beta_x\,\text{var}\,\{z\}}{\beta_x^2\,\text{var}\,\{z\}},$$

т. е. в виде дроби с тем же числителем, но большим знаменателем. Если x и y коррелированы, то оценивать β_y/β_x все же можно, хотя найти u может быть еще сложней. Но в этом случае нет уверенности, что β_y/β_x ближе к 0, чем β_{yx}.

Хотя мы и не доказывали этого, можно утверждать, что изменчивость $y - (b_{yu}/b_{xu})\,x$ обычно больше, чем изменчивость $y - b_{yx}\,x$. Это можно объяснить тем, что попытка предсказать y по скорректированному z труднее, чем по скорректированному x, ибо об x нам известно гораздо больше, чем о z.

Итак, мы разработали технику исключения эффекта скрытой переменной (здесь z) на отклик y, если найдутся два показателя, связанных с этой скрытой переменной, и выполняются некоторые другие условия. Теперь попробуем применить ее.

Наиболее удобны для демонстрации новой статистической техники примеры, структура которых «по секрету» известна. Это дает возможность увидеть, насколько эффективна наша техника, и найти источники отклонений. Приведем прежде всего набор данных, которые могли бы возникнуть в реальном примере с неизвестной структурой.

Пример. Структура известна. В таблице илл. 12.6.1 приведены исходные значения y, x и u. Мы попробуем исключить влияние z, хотя неизвестно, насколько нам удастся в этом преуспеть. Как и выше, мы предполагаем, что

$y = \beta_y z +$ остаток$_y$, $x = \beta_x z +$ остаток$_x$, $u = \beta_u z +$ остаток$_u$, а остатки некоррелированы.

Решение. В таблице илл. 12.6.1 найдено $b_{yu}/b_{xu} = 2{,}505$. Используя этот результат, можем вычислить остатки. Они приведены в первом столбце таблицы илл. 12.6.2. Истинные значения сопоставлены с их оценками остатков на илл. 12.6.4. Отметим, что, хотя на первый взгляд рассчитанные остатки соответствуют истинным лишь более или менее, на самом деле соответствие отличное.

ОБСУЖДЕНИЕ

«По секрету» мы знаем точно, как устроен этот пример, поскольку мы сами построили его так, чтобы удовлетворялись все требования модели. Мы взяли 10 значений z (= — 5, —4, —3, —2, —1, 1, 2, 3, 4, 5). (Примечание. Нуль не попал в их число.) Затем положили $\beta_y = 5$, $\beta_x = 2$, $\beta_u = 3$, $\alpha_x = \alpha_y = \alpha_u = 0$. Роль остатков играли независимые нормальные случайные величины с нулевыми средними и стандартными отклонениями 2, 1 и 3 для y, x и u соответственно. (Примечание. Речь идет о стандартных отклонениях остатков y, x и u, а не о самих переменных.) Используя значения этих случайных величин, построили, а затем округлили до ближайшего целого 10 значений каждой переменной. Все данные приведены в таблице илл. 12.6.3. В реальной ситуации известны лишь значения y, x и u.

Оценивание остатков. Величина 2,505, оценка β_y/β_x, необычно близка к истинному значению 2,5. Тем не менее точно восстановить остатки нам не удалось. Это объясняется тем, что информация об x не может полностью заменить здесь информацию о z. В связи с этим ошибка в оценке остатка y будет, как легко подсчитать, примерно равна $2{,}505 \times$ $\times$ остаток$_x$. В нашем примере остатки x равны ± 2, ± 1 или 0, а следовательно, остатки от оценки y равны ± 5, $\pm 2{,}5$ или 0. В действительности у x лишь один остаток, первый, равен 2, что приводит к ошибке на 5 единиц в первом знаке после запятой. Кроме того, по мере удаления оценки отношения β_y/β_x от истинного значения будет возрастать вклад в ошибку углового коэффициента.

Заинтересованный читатель заметил, по-видимому, что можно было также использовать x как инструментальную переменную в регрессии y по z, оценивая β_y/β_u через b_{yu}/b_{xu}. Если бы мы пошли этим путем, то оценка остатка $_y$ была бы

$$y - (b_{yx}/b_{xu})\, u.$$

Поскольку $\beta_y/\beta_u = 5/3$, новые оценки остатков$_y$ вносили бы в ошибки, грубо говоря, 5/3 от того, что дают остатки $_u$.

Можно использовать любой из этих методов или какую-либо их комбинацию с соответствующими весами. Однако здесь мы не будем обсуждать этот вопрос.

* 12.7. «ДОРОГИ, КОТОРЫЕ МЫ ВЫБИРАЕМ» (ФАКУЛЬТАТИВНО)

Предположим на минуту, что у нас всего один носитель и что y хорошо описывается прямой

$$y \equiv A + Bx.$$

Действительно ли это именно та единственная линия, которую мы ищем, которая нам нужна? Либо да, либо нет.

Пусть, например, $Y(y)$ — монотонно возрастающая * функция y. Тогда

$$Y(y) \equiv Y(A + Bx),$$

* Или убывающая, вообще — монотонная. — *Примеч. пер.*

так что

$$Y(y) \equiv A^* + B^* X(x),$$

где

$$X(x) = \frac{Y(A - Bx) - A^*}{B^*}$$

зависит только от x, если константы A, B, A^*, B^* известны. Таким образом, для любой такой функции Y существует целое семейство таких X (A^* и B^* мы выбираем по своему произволу), для которых $Y(y)$ хорошо приближается прямой от $X(x)$. Так какую же прямую линию нам взять?

Если простые преобразования y и x можно выбрать единственным способом, то можно поверить, что «вот эта...» линия. (Простой частный случай этой ситуации возникает тогда, когда первое приходящее в голову преобразование дает линейную зависимость.) Но как быть, если нет такой подсказки? Или если существуют два-три относительно простых и примерно одинаково хорошо подходящих преобразования? (Тем, кому такое предположение кажется маловероятным, напомним, что если $y = bx$, то $\log y = \log b + \log x$ и что если $\log y = \log b + k \log x$, то $y = bx^k$.)

Часто говорят, особенно часто—физики, что выбор подскажет теория. Но если теория говорит, что $y = bx$, то она говорит также и что $\log y = \log b + \log x$. Нельзя ждать помощи от теории, имея лишь один набор данных.

Иногда сделать выбор помогает наличие нескольких аналогичных ситуаций. Например, если в i-й ситуации

$$y = a_i + b_i x,$$

где a_i и b_i меняются плохо предсказуемым образом, то переход к прямым

$$\log (y - a_i) = \log b_i + \log x$$

мало кому может показаться удачным. В конце концов $\log (y - a_i)$ — различные функции для разных a_i. Описания теряют согласованность при таком «логарифмировании», когда они означают свое в каждом случае. Аналогичное описание аналогичных ситуаций — сильный довод в пользу переменных (y, x) вместо (Y, X).

Кроме того, еще один критерий кажется достаточно общим и внушающим доверие. Он состоит в том, что выбор между (Y, X) и (y, x) надо делать в зависимости от того, будет ли разброс данных относительно прямой (или иной, более сложной модели) примерно одинаковым.

В соответствии с этим критерием, если мы можем подогнать $y = bx$ и видим, что разброс y для данного x примерно пропорционален этому x, то мы вынуждены будем перейти к

$$\log y = \log b + \log x,$$

так как для этой прямой разброс будет более стабильным. Такой выбор зависит от степени совершенства зависимости, и в нем есть смысл лишь тогда, когда модель несовершенна.

Почему этот критерий внушает доверие? Прежде всего потому, что он облегчает осмысление ситуации. Можно сказать, например, что в

90 или 95% границы имеют такую-то ширину, не заботясь об x. Можно взглянуть на остатки и сразу оценить ситуацию, снова не интересуясь x.

Далее, мы часто приближаем y по x, а затем анализируем остатки, например, строя регрессию по другому носителю. Обычно наиболее полезный способ—это регрессия по остаткам нового носителя относительно регрессии по x. Здесь хорошо бы было, чтобы переменные, используемые на втором и последующих шагах, хотя бы в скромных пределах согласовывались с остатками. Например, хорошим стандартом была бы близость дисперсии к постоянной.

Если отсутствуют какие-либо более важные критерии для выбора между (y, x) и (Y, X), стоит руководствоваться следующими двумя:

- аналогичное описание аналогичных ситуаций;
- приблизительно равномерное качество модели.

Трудно надеяться на то, что можно удовлетворить обоим критериям в рамках одной совокупности данных. Когда они оба удовлетворяются, они часто гармонируют, а когда они противоречивы, следует предпочесть первый из них.

12.8. ИСПОЛЬЗОВАНИЕ ПОДВЫБОРОК

Регрессионные методы применяются при весьма разных объемах данных от 3, 4 точек до, скажем, 5 миллионов. В обычной ситуации начинать построение регрессии, имея сразу 1000 или даже 200 точек,— неважная идея. Регрессия должна быть гибкой процедурой. Должна быть возможность изучить остатки и так, и этак. Гибкость процедуры в значительной степени теряется, если мы начинаем исследование, имея слишком большой набор данных.

Если нам так повезло, что у нас есть 10000 данных, то надо начинать работу с образования подвыборок. Может быть, стоит взять подвыборки размером 1000 и 100 единиц, может быть, 2500 и 625 и подвыборку в 160 единиц или около того. В любом случае мы должны начинать анализ с наименьшей подвыборки, получить всю возможную информацию и лишь затем переходить к большей подвыборке.

Следовательно, большие наборы данных следовало бы организовать так, чтобы было легко извлекать из них подвыборки разных размеров. Когда порядок данных не играет роли, простейшим методом организации выборок будет последовательный, т. е. такой, когда каждая выборка является подвыборкой последующей.

Это можно сделать, выбирая по одной точке из каждых 2,4, 8 или вообще 2^k (k — целое положительное) последовательных значений. Причем сам выбор может осуществляться с помощью равномерно распределенных случайных чисел. Такая выборка не может быть существенно хуже простой случайной выборки. В тех же случаях, когда исходный массив данных имеет выраженную структуру (упорядоченность), она будет много лучше обычной. Программы для более изощренных выборочных методик, например точно учитывающих структуру множества расслоенных данных, приведены в статье Фана, Мюллера и Резухи (см. библиографию к данной главе).

Подвыборки используются также непосредственно или вместе с методом «складного ножа» (см. гл. 8) в целях достижения большей устойчивости.

РЕЗЮМЕ. РЕГРЕССИЯ

Мы встретились с трудностями в определении причинных связей. Было установлено, что под давлением обстоятельств мы иногда принимаем решение о наличии причинной связи, тогда как установлено лишь наличие соответствия. Обсуждалась опасность связанных с этим серьезных ошибок.

Рассматривались различные трактовки термина «зависимость» и некоторых других, а также опасности, вызванные путаницей в этих терминах.

Понятие «регрессия» имеет два различных смысла. Один из них относится к кривой (или в более общем случае — поверхности) ожидаемых значений y, отвечающих фиксированным (точно или почти точно) значениям x. Второй — к выбору семейства моделей, отличающихся значениями констант, и подбору нужных констант.

Всякая регрессия (независимо от определения) — это неполное описание. Поэтому какой бы удобной и полезной моделью она ни была, бессмысленно говорить об «истинной модели» или надеяться получить из нее всю информацию.

Регрессия может использоваться в разных целях: для свертывания информации, для измерений (коэффициентов в конкретных уравнениях регрессии), для исключения (эффектов мешающих переменных) и для предсказания, причем то, что может быть хорошо в одном случае, не обязательно хорошо в другом.

Полная или приближенная коллинеарность одного или нескольких носителей — это то:

1) к чему мы должны быть готовы;

2) что может нарушить точность оценок или простоту интерпретации коэффициентов, а может быть, и то и другое (при этом качество приближения может оставаться высоким);

3) что может привести ко многим ошибкам и недоразумениям;

4) что мы должны изучать, для того чтобы с ним бороться.

Полиномы, даже квадратичные, часто порождают почти коллинеарность (если мы недостаточно внимательны).

Хорошее соответствие прямой $y = a + bx$ данным может означать, а может и не означать, что это «правильная» прямая.

Мы можем строить регрессию шаг за шагом, удаляя графическим или алгебраическим путем вклад каждого носителя по очереди не только из отклика, но и из носителей, не использованных еще в приближении.

Наборы тесно связанных (почти коллинеарных) носителей заменяются одним комбинированным либо одним комбинированным с остатками других по нему или с другими носителями, значительно большими по величине ошибок измерения (это дает основание предполагать, что их полезно включить в рассмотрение).

В регрессии для исключения нет смысла использовать регрессионные коэффициенты, работающие в регрессии для предсказания (здесь лучше брать больше коэффициентов с особой структурой).

* Приблизительное постоянство величин остатков служит критерием при выборе между альтернативными прямыми, хорошо приближающими данные (это относится к тем случаям, когда альтернатива предполагает преобразование переменной x либо и x, и y). Но непротиворечивое описание независимых наборов данных — лучший критерий.

Для эффективной работы с достаточно большими множествами данных надо планировать подвыборки и начинать работу с наименьшей из них, достаточной для наших целей. В дальнейшем, по мере развития исследования, можно увеличивать объемы выборок в 2, 3, 5, 10 раз и т. д.

БИБЛИОГРАФИЯ

Allport G. W., Vernon P. E. and Lindzey G. (3rd ed., 1960). Study of Values. Boston, Houghton Mifflin.

EDA-Tykey J. W. (1977). Exploratory Data Analysis. Reading, Mass, Addison-Wesley.

Fan C. T., Muller M. E. and Rezucha I. (1962). Development of sampling plans by using sequential (item-by-item) selection techniques and digital computers. — J. of Amer. Statist. Assoc., 57, 387—402.

Riley M. W. and Foner A. (1968). Aging and Society, Vol. 1. New York, Russel-Sage Foundation, 437. (Shows Exhibit 1, page 263).

ИЛЛЮСТРАЦИИ

Иллюстрация 12.1.1.

Процент выдающихся достижений за каждое десятилетие жизни 980 людей, умерших в различном возрасте. Для каждой группы людей с «одинаковой» продолжительностью жизни сумма выдающихся достижений — 100%. Всего 1540 выдающихся достижений

Возраст, в котором достигнуты выдающиеся результаты	x — возраст к моменту смерти							
	до 50	50—59	60—64	65—69	70—74	75—79	80—84	85+
До 20	5		1		2			
20—29	32	23	17	8	15	10	12	8
30—39	50	39	32	38	36	28	32	29
40—49	14	28	27	28	28	27	28	26
50—59		9	20	16	13	20	15	22
60—69			4	10	6	10	9	9
70—79						4	3	3
80—89							1	2
Итоги[1]	101	99	101	100	100	99	100	99
Медианы[2]	32,7	36,8	40,2	41,4	36,4	44,3	42,1	44,8

[1] Итог может отличаться от 100 из-за округлений.

[2] Интерполирование линейное. Так, среди умерших ранее 50 лет 37% достигло замечательных результатов до 30, а 87% — до 40.

Интерполяция дает 50% достижений в возрасте 32,7.

Источник. Lehman H. C. (1953). Age and Achievement, Princeton University Press; 317 p. Copyright 1953. Воспроизведено с разрешения Princeton University Press.

Иллюстрация 12.1.2.

Регрессия медианы возраста достижения выдающегося результата по продолжительности жизни; прямая построена по 6 средним точкам (с равномерными весами); в качестве концевых x произвольно выбраны 45 и 87 1/2

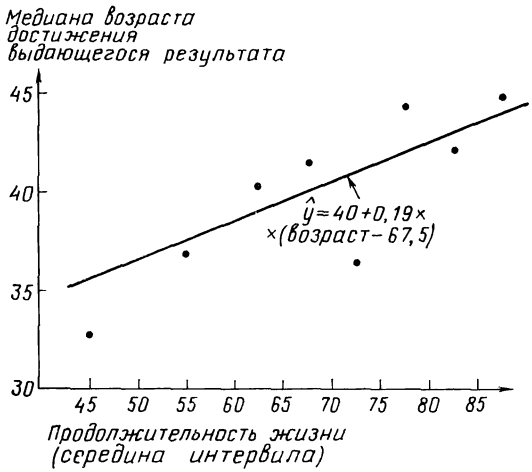


Иллюстрация 12.3.1

Занятость, снижение индекса цен, валовой национальный продукт по годам

Год	Занятость y	Снижение индекса цен x_1	Валовой национальный продукт x_2
1947	—4994	—18,7	—153400
1948	—4195	—13,2	—128272
1949	—5146	—13,5	—129644
1950	—4130	—12,2	—103099
1951	—2096	—5,5	—58723
1952	—1678	—3,6	—40699
1953	—328	—2,7	—22313
1954	—1556	—1,7	—24586
1955	—702	—0,5	9771
1956	2540	2,9	31482
1957	2852	6,7	55071
1958	1196	9,1	56848
1959	3338	10,9	95006
1960	4247	12,5	114903
1961	4014	14,0	130475
1962	5234	15,2	167196

Замечание. Точность данных та же, что и в источниках, из которых эти данные заимствованы.

Источники: Beaton A. E., Rubin D. E. and Barone J. L. (1976). The acceptability of regression solutions: another look at computational accuracy.—

J. of Amer. Statist. Assoc., 71, 158—168. Цитируется по L o n g l e y J. W. (1967). An appraisal of least-squares for the electronic computer from the point of view of the user. — J. of Amer. Statist. Assoc., 62, 819—841.

Иллюстрация 12.3.2

Регрессия y по x (первый этап)

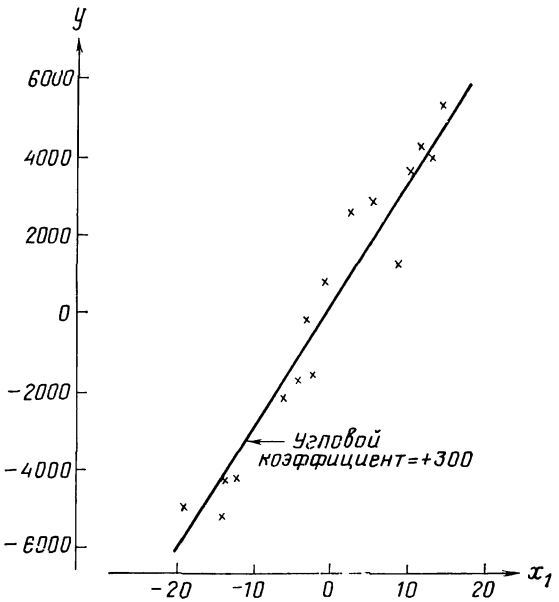


Иллюстрация 12.3.3

Регрессия $y'_{;1}$ по x_1 (второй этап)

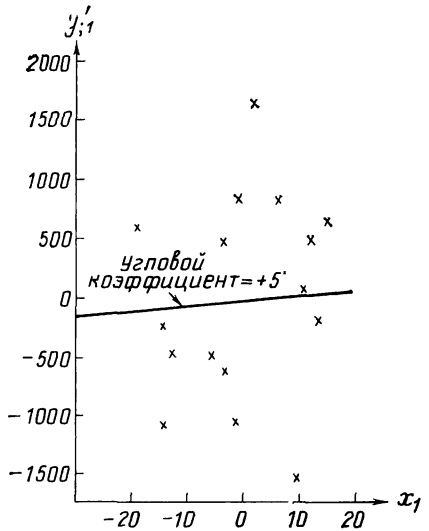


Иллюстрация 12.3.4

Первый этап приближения (расчеты на основе данных таблицы илл. 12.3.1)

Год	x_1	Первые коррекции		Предсказанные значения		Первая коррекция
		$y'_{;1}$	$x'_{2;1}$	$y_{;1}$	$x_{2;1}$	$y'_{;12}$
1947	—18,7	616	33591	710	1801	584
1948	—13,2	—235	3728	—169	—18712	1141
1949	—13,5	—1096	5356	—1028	—17594	204
1950	—12,2	—470	18901	—409	—1839	—280
1951	—5,5	—446	—3723	—418	—13073	497
1952	—3,6	—598	—4699	—580	—10819	177
1953	—2,7	482	4687	496	97	489
1954	—1,7	—1046	—7586	—1038	—10476	—305
1955	—0,5	852	14771	854	13921	—120
1956	2,9	—1670	2482	1656	1412	1137
1957	6,7	842	—11929	808	—539	846
1958	9,1	—1534	—34152	—1580	—18682	—272
1959	10,9	68	—13994	14	4536	—304
1960	12,5	497	—10097	434	11153	—1255
1961	14,0	—186	—9525	—256	14275	—2275
1962	15,2	674	—15196	598	41036	

Примечания:

$y'_{;1} = y - 300\ x_1$,

$y_{;1} = y'_{;1} - 5x_1 = y - 305x_1$,

$x'_{2;1} = x_2 - 10000\ x_1$,

$x_{2;1} = x'_{2;1} + 1700x_1 = x_2 - 8300\ x_1$,

$y'_{;12} = y_{;1} - 0{,}07x_{2;1}$.

Иллюстрация 12.3.5.

Регрессия x_2 по x_1 (первый этап)

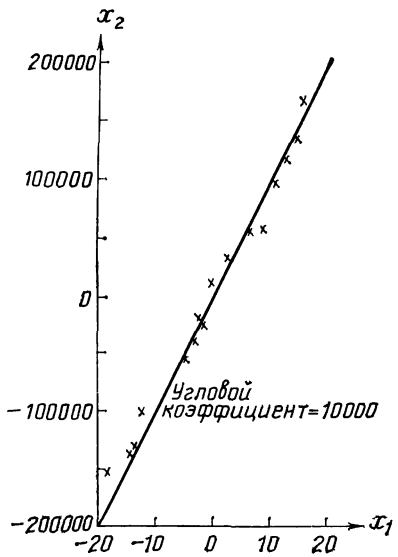


Иллюстрация 12.3.6

Регрессия $x'_{2;1}$ по x_1 (второй этап)

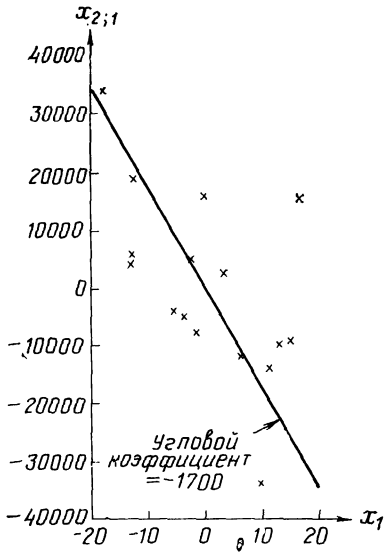


Иллюстрация 12.3.7

Регрессия $y_{;1}$ по $x_{2;1}$ (первый этап)

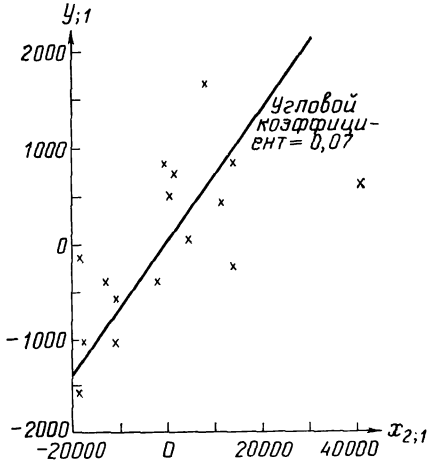


Иллюстрация 12.3.8

Основа для второго этапа регрессии $y'_{;12}$ по $x_{2;1}$

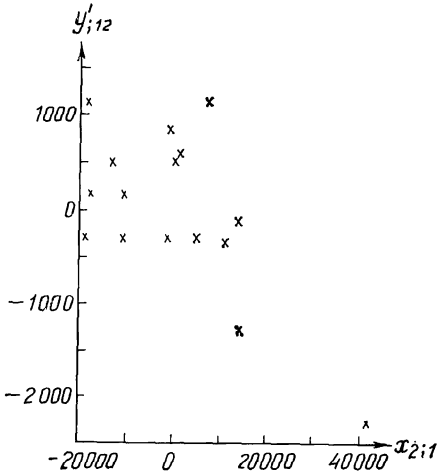


Иллюстрация 12.4.1

Логарифмы числа прихожан, выплачиваемой заработной платы и капитала для одного штата из каждого региона США по методистской епископальной церкви за 1936 г.

Штат	Число прихожан	Зарплата	Капитал
Мэн	4,30	5,33	6,20
Нью-Йорк	5,48	6,47	7,60
Огайо	5,58	6,34	7,44
Миннесота	4,87	5,75	6,77
Делавэр	4,41	5,30	6,33
Кентукки	4,38	5,19	6,13
Арканзас	3,62	4,49	5,45
Монтана	4,12	5,05	5,93
Вашингтон	4,61	5,44	6,33
Медиана	4,41	5,33	6,33
7-й минус 3-й	0,57	0,56	0,64

Источник. Mosteller G. Частное сообщение, цитируется по: Religious Bodies. 1936, Vol. II, Part 2, Bureau of the Census, United States Department of Commerce, U. S. Government Printing Office; p. 1086—1096.

Иллюстрация 12.4.2

Логарифм зарплаты в зависимости от логарифма капитала для методистской епископальной церкви за 1936 г. Представлен один штат для каждого из девяти регионов США

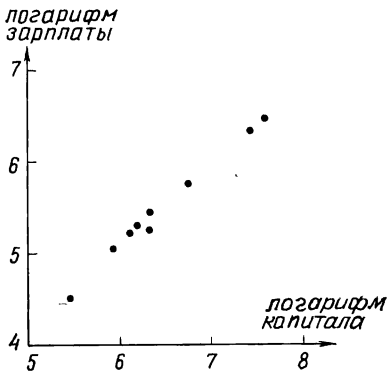


Иллюстрация 12.4.3

Логарифм числа прихожан в зависимости от суммы логарифма зарплаты плюс логарифм капитала методистской епископальной церкви за 1936 г. Представлен один штат для каждого из девяти регионов США

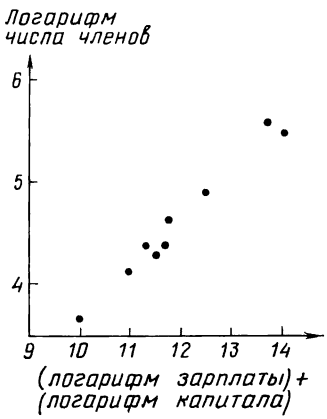


Иллюстрация 12.6.1

Данные для выделения эффекта z из y; u — инструментальная переменная

y	x	u
—26	—8	—16
—21	—7	—14
—14	—7	—8
—5	—4	—2
—8	—2	—2
7	3	3
8	5	2
13	6	11
16	8	9
25	11	14
Итоги—5	5	—3

$\Sigma yu = 1522$, $\Sigma xu = 605$,

$$\frac{b_{yu}}{b_{xu}} = \frac{152{,}2 - (-0{,}5)\ (-0{,}3)}{60{,}5 - (0{,}5)\ (-0{,}3)} = 2{,}505.$$

Иллюстрация 12.6.2

Рассчитанные и истинные значения остатков

Рассчитанный остаток $_y$ $y - 2{,}505\ x$	Истинный остаток $_y$
—6,0	—1
—3,5	—1
3,5	1
5,0	5
—3,0	—3
—0,5	2
—4,5	—2
—1,5	—2
—4,0	—4
—2,6	0
Итоги —17,1	—5

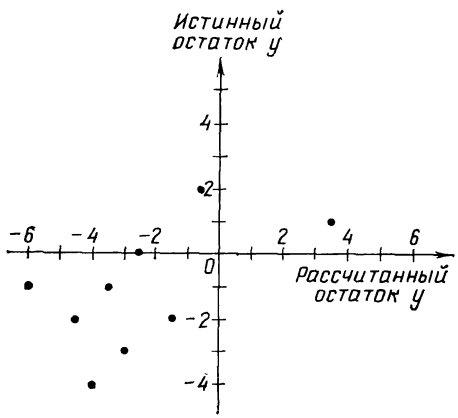
Иллюстрация 12.6.3

Расчет переменных y, x, u

z	$\beta_y z$	Остаток$_y$	y	$\beta_x z$	Остаток$_x$	x	$\beta_u z$	Остаток$_u$	u
—5	—25	—1	—26	—10	2	—8	—15	1	—16
—4	—20	—1	—21	—8	1	—7	—12	—2	—14
—3	—15	1	—14	—6	—1	—7	—9	1	—8
—2	—10	5	—5	—4	0	—4	—6	4	—2
—1	—5	—3	—8	—2	0	—2	—3	1	—2
1	5	2	7	2	1	3	3	0	3
2	10	—2	8	4	1	5	6	—4	2
3	15	—2	13	6	0	6	9	2	11
4	20	—4	16	8	0	8	12	—3	9
5	25	0	25	10	1	11	15	—1	14

Иллюстрация 12.6.4

График для данных из илл. 12.6.2.



Глава 13 ● БЕДЫ РЕГРЕССИОННЫХ КОЭФФИЦИЕНТОВ

Мы знаем, что регрессионные коэффициенты часто страдают разными «недугами». В этой главе мы подробно рассмотрим, что в связи с этим можно сделать. Мы обсудим возникающие трудности, как и почему они возникают и в какой степени мы можем подготовиться к ним, избежать их или по крайней мере узнать о них.

Основной результат, на который здесь можно надеяться, — это определение регрессионных коэффициентов, применимых и в новых ситуациях (исключение составляют ситуации с жесткой структурой).

13.1. СМЫСЛ КОЭФФИЦИЕНТОВ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Форма переменной. Рассмотрим для начала квадратичную функцию вида $\beta_0 1 + \beta_1 x_1 + \beta_2 x_2$, где $x_2 \equiv x_1^2$. Мы рассмотрим разные способы записи основной переменной x (индекс опустим) и посмотрим, какое это окажет влияние на величину коэффициентов.

Пусть в области $-2 \leqslant x \leqslant 5$ хорошим приближением служит функция $117 - 3x + 2x^2$. Какой смысл мы вкладываем при этом в коэффициент -3? Каждое из следующих выражений численно идентично трехчлену $117 - 3x + 2x^2$:

$$115 + x + 2(x - 1)^2,$$
$$109 + 5x + 2(x - 2)^2,$$
$$117 + x + 2x(x - 2),$$
$$117 + 5x + 2x(x - 4).$$

Какую бы интерпретацию мы ни дали коэффициенту -3, должна существовать параллельная интерпретация, удовлетворяющая 5 и 1, которые служат коэффициентами при x в последних четырех формулах, так как все эти формулы эквивалентны.

Если мы от выражения

$$12 - 3x_1 + 5x_2$$

перейдем к

$$12 - 8x_1 + 5x_2^*,$$

где $x_2^* = x_1 + x_2$, может возникнуть соблазн говорить и о коэффициенте «при одном и том же x_1», ибо все остальные коэффициенты в обоих выражениях равны. Вспоминая, однако, что x_2 мог бы обозначать и x^2, мы видим, что для положительных x нельзя сохранить один и тот же

x^2 при заметном варьировании x. Таким образом, соблазн никуда не ведет.

Единственное, что можно утверждать, — это то, что коэффициент при любом носителе, в данном случае при x, зависит от того, какие еще носители входят в модель. В нашем примере это были 1 и x^2, или 1 и $(x-1)^2$, или 1 и $x(x-2)$, или 1 и $x(x-4)$.

Различия между коэффициентами при x:

$$-3 \text{ для } 1 \text{ и } x^2,$$
$$+1 \text{ для } 1 \text{ и } (x-1)^2,$$
$$+5 \text{ для } 1 \text{ и } (x-2)^2.$$

указывают на то, что коэффициент в множественной регрессии, как теоретический, так и эмпирический, зависит *не только* от

- набора данных и метода подгонки,
- носителя-сомножителя, но и от того,
- что еще входит в модель.

Подмножества переменных. Часть трудностей, связанных с приданием смысла коэффициентам множественной регрессии, возникает из-за того, что коэффициенты изменяются при изменении носителей, которые *представлены* в модели. Мы уже говорили об этом выше. Поэтому мы часто оказываемся в трудной ситуации, так как не знаем заранее, какие переменные выбирать. Проиллюстрируем это утверждение простым численным примером.

Пример. Плоскость, проходящая через начало координат.

Предположим, что неизвестная нам зависимость на самом деле имеет вид

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

но мы собираемся приблизить ее с помощью метода наименьших квадратов одной из следующих функций:

$$y = \gamma_1 x_1 + \gamma_2 x_2,$$

или

$$y = \delta_1 x_1 + \delta_3 x_3,$$

или

$$y = \varepsilon_1 x_1.$$

Предположим далее, что $\beta_1 = 2$, $\beta_2 = 4$, $\beta_3 = 10$, а для облегчения счета допустим еще, что все парные произведения носителей симметричны, откуда для наших трех носителей имеем

$$\Sigma x_1^2 = \Sigma x_2^2 = \Sigma x_3^2 = 1,$$
$$\Sigma x_1 x_2 = \Sigma x_1 x_3 = \Sigma x_2 x_3 = \frac{1}{2}.$$

Какими будут γ_1, δ_1 и ε_1 и как они соотносятся с β_1?

Решение. Если мы подберем методом наименьших квадратов плоскость, проходящую через начало координат, то получим

$$\gamma_1 = \frac{(\Sigma x_1 y)\left(\sum x_2^2\right) - (\Sigma x_2 y)(\Sigma x_1 x_2)}{\left(\sum x_1^2\right)\left(\sum x_2^2\right) - (\Sigma x_1 x_2)^2}.$$

Формула для δ_1 выглядит точно так же, если только x_2 заменить на x_3. Величины $\Sigma x_i y$ для $i = 1, 2, 3$ легко найти из

$$\Sigma x_i y = \beta_1 \Sigma x_1 x_i + \beta_2 \Sigma x_2 x_i + \beta_3 \Sigma x_3 x_i, \quad i = 1, 2, 3.$$

Для наших числовых данных получим

$$\Sigma x_1 y = 9, \quad \Sigma x_2 y = 10, \quad \Sigma x_3 y = 13.$$

Таким образом, мы видим, что

$$\beta_1 = 2, \quad \gamma_1 = 5\,1/3, \quad \delta_1 = 3\,1/3.$$

Следовательно, даже в нашей весьма симметричной ситуации коэффициент при x_i заметно меняется в зависимости от того, какие переменные представлены. Если же мы ограничимся просто функцией $y = \varepsilon_1 x_1$, то, как легко видеть, $\varepsilon_1 = 9$.

Если смотреть на этот пример, что мы и делали, с чисто арифметических позиций, то, возможно, первой реакцией будет недоумение. Действительно, кто сказал, что должна иметь место инвариантность или другие простые соотношения при смене переменных. Это вполне достаточное возражение. Но в реальных ситуациях нам действительно приходится копаться в переменных, оставлять одни и отбрасывать другие. А в конце концов нам приходится давать этим коэффициентам физическую интерпретацию типа «если мы изменим x_1 определенным образом, то соответственно изменится и y_0, а следовательно, ...». Наш пример показывает, что даже в детерминированных системах (исходная зависимость $y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ не содержит ошибок) различные подмножества переменных, включаемых в регрессию, могут давать существенно различные коэффициенты для одной и той же переменной. При переходе от одного множества к другому может меняться даже знак.

Таким образом, говоря о коэффициенте при x_1, мы должны учитывать еще, включены ли в модель x_2 и x_3, только x_2, только x_3 или нет никаких других переменных. Следует понимать, что во всех вариантах коэффициенты при x_1 могут существенно отличаться.

Совокупность x. Этот вопрос столь важен, что мы рассмотрим его еще раз с более общих позиций.

Если задано несколько x, то коэффициенты c_1, c_1^*, c_1^{**}, c_1^{***} при x_1 в моделях

$$\begin{aligned} &c_1 x_1 + c_2 x_2 + c_3 x_3 + c_4 x_4, \\ &c_1^* x_1 + c_2^* x_2 + c_3^* x_3, \\ &c_1^{**} x_1 + c_2^{**} x_2 + c_6^{**} x_6, \\ &c_1^{***} x_1 + c_7^{***} x_7 \end{aligned} \tag{1}$$

обычно разные. Они совпадают крайне редко.

Мы рассматриваем четыре различные модели, т. е. мы подразумеваем, что модель, представляемая $c_1^* x_1 + c_2^* x_2 + c_3^* x_3$ (для некоторых c_1^*, c_2^*, c_3^*), существенно отличается от модели, представляемой, например, $c_1^{***} x_1 + c_7^{***} x_7$ (при всех возможных c_1^{***}, c_7^{***}). Мы хотим подчеркнуть, что без дополнительной информации невозможно перейти от модели для (c_1, c_2, c_3, c_4) к моделям для (c_1^*, c_2^*, c_3^*), или $(c_1^{**}, c_2^{**}, c_6^{**})$, или

(c_1^{***}, c_7^{***}). Нельзя также использовать одно из этих приближений, чтобы получить без дополнительной информации (c_1, c_2, c_3, c_4) или даже один из этих коэффициентов. Надо было бы знать гораздо больше о данных (или о рассматриваемой совокупности), чтобы сделать такое превращение. Даже в нашем детерминированном примере понадобились все суммы квадратов и парных произведений.

(Представления $12 - 3x_1 + 5x_2$ и $12 - 8x_1 + 5x_2^*$, где $x_2^* = x_1 + x_2$ — другое дело. Здесь мы сталкиваемся с одним и тем же выражением, записанным по-разному. Зная, что $x_2^* = x_1 + x_2$, всегда можно преобразовать одно множество коэффициентов в другое. Такая удобная с точки зрения вычислений ситуация редко встречается в конкретных исследованиях.)

Генераторы моделей. Каждая модель берется из какого-то множества возможных моделей. Так, в (1) входят четыре множества. Мы будем называть их генераторами, так как в дальнейшем они рассматриваются как собрания возможных моделей, из которых одна выделяется выбором конкретных значений c, подобно тому, как при покупке рубашки в магазине для выбора из однотипных изделий достаточно указать рост и размер воротничка. Эти генераторы и поставляют множества возможных моделей, используемых в регрессии.

Мы будем рассматривать лишь генераторы аддитивной природы вида

$$c_1x_1 + c_2x_2 + \ldots + c_kx_k.$$

При этом будет предполагаться, что x_2, x_3, ..., x_k могут зависеть от x_1. (Например, $x_2 \equiv x_1^2$, $x_3 \equiv x_3^*$, $x_4 \equiv x_1^3 + x_4^*$, где x_3^* и x_4^* не зависят от x_1, определяют генератор вида $c_1x_1 + c_2x_2 + c_3x_3 + c_4x_4$.)

Важно знать, какие еще носители, помимо того, коэффициент которого нас интересует, входят в модель. Важны ли детали представления генератора? Нет, если мы можем переходить от одного представления к другому с помощью алгебраических преобразований. Так, генераторы

$$c_1x_1 + c_2x_2 + \ldots + c_kx_k$$

и

$$c_1x_1 + c_2^*x_2^* + \ldots + c_k^*x_k^*$$

совпадают, если совпадают меньшие генераторы, составленные из всех форм вида

$$c_2x_2 + \ldots + c_kx_k$$

и

$$c_2^*x_2^* + \ldots + c_k^*x_k^*$$

(мы будем называть их подгенераторами x_1).

Значение c_1 в обоих приближениях будет одинаковым, если только критерий подгонки: а) тот же, б) зависит только от остатков, как если бы оба уравнения подбирались методом наименьших квадратов. Таким образом, при построении аддитивных моделей на фиксацию коэффициента при x_1 влияют:

- множество данных и метод подгонки (с учетом весов);
- генератор, образованный другими членами, входящими в модель.

13.2. ЛИНЕЙНАЯ КОРРЕКЦИЯ КАК МЕТОД ОПИСАНИЯ

В параграфе 12.3 мы ввели и графически продемонстрировали метод последовательного исключения переменных по одной (мы называем этот метод шаговым подбором). В данном параграфе изложим этот же метод с алгебраических позиций по следующим (и некоторым другим) причинам:

а) мы хотим использовать его в дальнейшем;

б) хотим показать, что он дает решение метода наименьших квадратов;

в) хотим подчеркнуть еще раз, что коэффициент любого носителя, полученный методом наименьших квадратов, можно рассматривать как коэффициент регрессии отклика на этот носитель, скорректированный линейно по остальным носителям из его подгенератора.

Несколько слов об обозначениях. Для y и нескольких x ($x_1, x_2, \ldots, x_k$) введем специальные обозначения остатков. Например, мы пишем $y_{;12}$ и $x_{3;12}$, чтобы указать, что влияние x_1 и x_2 исключено из y и x_3 соответственно. При использовании метода наименьших квадратов мы заменяем точку с запятой точкой; так, аналогичные остатки метода наименьших квадратов обозначаются $y_{.12}$, $x_{.12}$. Иногда x_1 бывает константой, причем чаще всего $x_1 \equiv 1$. В примере, к обсуждению которого мы приступаем, три переменные y, x, t и константа также участвуют в приближении. Мы могли бы использовать обычные обозначения, полагая $x_1 = 1$, $x_2 = x$, $x_3 = t$. Но в этом примере удобней оставить исходные. Так, $y_{.1x}$ будет означать, что влияние 1 и x удалено из y при помощи метода наименьших квадратов.

В этом примере мы описываем поэтапный метод поочередного «исключения» линейных эффектов переменных. Давайте начнем с трех переменных y, x, t и константы 1, т. е. с одного отклика и трех носителей. Допустим, что мы построили регрессию y по 1 и x вида $y = a + bx$. Теперь вычислим остатки (y линейно скорректированный по 1 и x)

$$y_{\cdot 1x} \equiv y - a - bx$$

и предположим,что их анализ свидетельствует об адекватном приближении. Тогда зададимся вопросом, зависит ли скорректированное y от t? Изучение этого вопроса разумно начать с построения графика для этих переменных.

Картина будет более ясной, если мы еще скорректируем t по 1 и x, т. е. построим

$$t_{.1x} \equiv t - c - dx.$$

Теперь можно строить график зависимости $y_{\cdot 1x}$ от $t_{.1x}$. Предположим, что мы обнаружили явную зависимость. Значит, стоит пытаться построить линейную регрессию, например

$$y_{\cdot 1x} \sim e + ft_{.1x}.$$

Что можно сказать о коэффициентах e и f, если всюду применялся метод наименьших квадратов? В конце следующей главы мы покажем, что полученный результат совпадает с тем, который дал бы метод наи-

меньших квадратов для подгонки y моделью вида $\beta_1 1 + \beta_2 x + \beta_3 t$. Более конкретно:

$$a + bx + f(t - c - dx)$$

совпадает с этим приближением (см. 14.11).

Общий случай. В параграфе 14.11 рассматривается более общий многомерный случай — с учетом константы. В частности, из доказанных там результатов вытекает, что, например, для регрессионной модели

$$y \sim a + bx + ct + du + ev + fw$$

1) регрессионный коэффициент b перед x имеет место, когда y линейно скорректировано по 1, t, u, v и w, а x линейно скорректирован по 1, t, u, v, w;

2) регрессионный коэффициент c перед t имеет место, когда y линейно скорректировано по 1, x, u, v, w, а t линейно скорректирован по 1, x, u, v, w.

. . .
. . .
. . .

6) регрессионный коэффициент a перед 1 имеет место, когда y скорректировано по x, t, u, v, w, а 1 линейно скорректирована по x, t, u, v, w.

Имея дело с конкретными ситуациями, такого сорта утверждения необходимо делать для каждого коэффициента, который мы хотели бы как-то интерпретировать. Никакой альтернативы детальному рассмотрению фактов нет.

Сделав над собой некоторое усилие, мы можем теперь резюмировать вышесказанное в следующем виде.

Коэффициент (метода наименьших квадратов) для любого носителя есть коэффициент регрессии линейно скорректированного по подгенератору этого носителя отклика по самому этому носителю, тоже скорректированному линейно по своему подгенератору. (Предполагается, что подгенератор содержит все возможные варианты из исходного генератора, куда наш носитель входит с нулевым коэффициентом.)

13.3. ПРИМЕРЫ ЛИНЕЙНОЙ КОРРЕКЦИИ

Построение осмысленных регрессионных коэффициентов по имеющемуся массиву данных — одно из требований, часто предъявляемых регрессионному анализу. Рассмотрим сейчас пример, где мы правильно выбрали вид модели, но экспериментальные данные получились округленными. Пусть $y = -1 + x + 0{,}5x^2$ — истинная модель квадратичного вида $a + bx + cx^2$. Мы будем, используя эту форму, получать значения y по значениям x. При этом значения члена $0{,}5x^2$ будем округлять до двух десятичных знаков. Округление производится так, чтобы последний десятичный знак был четным. Скажем, при $x = 0{,}9$ заменяем $0{,}5(0{,}81)$ числом 0,40 вместо точного 0,405. Мы начинаем с данных, близких к истинным значениям квадратичной формы. Давайте посмотрим, как точно они восстанавливаются регрессией.

Начнем с константы. Если для облегчения вычислений мы сочтем необходимым начать с устранения влияния константы 1, то, как и в параграфе 13.2, прежде всего вычислим

$$y_{\cdot 1} \equiv y - \bar{y},\ x_{\cdot 1} \equiv x - \bar{x},\ t_{\cdot 1} \equiv t - \bar{t}.$$

Затем устраним влияние x:

$$y_{\cdot 1x} \equiv y_{\cdot 1} - gx_{\cdot 1},\ t_{\cdot 1x} \equiv t_{\cdot 1} - hx_{\cdot 1}.$$

Наконец

$$y_{\cdot 1xt} \equiv y_{\cdot 1x} - ft_{\cdot 1x}.$$

Значит,

$$y_{\text{предсказанное}} \equiv \bar{y} + gx_{\cdot 1} + ft_{\cdot 1x}.$$

Именно эта схема используется в следующих примерах.

Пример 1. Квадратичная модель. В таблице илл. 13.3.1 приведены значения y, x и $t = x^2$. Мы хотим построить множественную регрессию y по 1, x и t.

Решение. Вычитая из y, x и x^2 их средние 0,94, 1,2 и 1,48 соответственно, получим

$$y_{\cdot 1} \equiv y - \bar{y},\ x_{\cdot 1} \equiv x - \bar{x},\ t_{\cdot 1} \equiv (x^2)_{\cdot 1} \equiv x^2 - \overline{x^2}$$

($\overline{x^2}$ — конечно, среднее квадратов).

Далее, мы корректировкой $y_{\cdot 1}$ и $t_{\cdot 1}$ по $x_{\cdot 1}$ получаем

$$y_{\cdot 1x} \equiv y_{\cdot 1} - gx_{\cdot 1} = y_{\cdot 1} - 2{,}2x_{\cdot 1},$$
$$t_{\cdot 1x} \equiv t_{\cdot 1} - hx_{\cdot 1} = t_{\cdot 1} - 2{,}4x_{\cdot 1}.$$

И наконец, приходим к регрессии $x_{\cdot 1x}$ по $t_{\cdot 1x}$

$$y_{\cdot 1xt} \equiv y_{\cdot 1x} - ft_{\cdot 1x} = y_{\cdot 1x} - 0{,}48t_{\cdot 1x}.$$

Из данных, приведенных в последнем столбце таблицы илл. 13.3.1, мы видим, что $y_{\cdot 1xt}$ равно нулю с точностью до двух десятичных знаков (напомним, что мы округляли $0{,}5x^2$ до ближайшего четного второго десятичного знака).

Окончательное приближение выглядит так:

$$y_{\text{предск}} = 0{,}94 + 2{,}2x_{\cdot 1} + 0{,}48\,(x^2)_{\cdot 1x}.$$

Его можно переписать, используя носители 1, x и x^2 в виде

$$y_{\text{предск}} = -1{,}03 + 1{,}05x + 0{,}48x^2.$$

Результат выглядит весьма обнадеживающим, если вспомнить что мы начинали с выражения $y = -1 + x + 0{,}5x^2$, но округляли $0{,}5x^2$. Начав весьма близко к истинной квадратичной модели, мы пре успевали в ее восстановлении. Стоит отметить, что возмущение одного члена слегка изменило все коэффициенты.

Вычисления проводились в указанных пределах, различные ошиб ки округления оказали соответствующее влияние на их результаты В частности, значения $y_{\cdot 1}$, приведенные в таблице, не дают в сумме 0 как должно быть в теории. Мы видим, что окончательные остатки тож

дают в сумме отрицательную величину, а не 0, как им предписывается теорией. Но следует помнить, что теория подразумевает точные вычисления с бесконечным числом знаков после запятой. Величины остатков $y_{.1xt}$ из последнего столбца таблицы илл. 13.3.1 вводят в некоторое заблуждение, так как если мы используем полученную формулу, то найдем меньшие остатки.

Рассмотрим разность между исходной неокругленной функцией и приближенной моделью. Она равна:

$$y_{\text{ист}} - y_{\text{предск}} = -1 + x + 0{,}5x^2 -$$
$$-(-1{,}03 + 1{,}05x + 0{,}48x^2) = 0{,}03 - 0{,}05x + 0{,}02x^2.$$

Дополняя до полного квадрата, получим

$$0{,}02\left[x^2 - \frac{0{,}05}{0{,}02}x + \frac{1}{4}\left(\frac{0{,}05}{0{,}02}\right)^2\right] - \frac{(0{,}05)^2}{4\,(0{,}02)} +$$
$$+0{,}03 = 0{,}02\,(x - 1{,}25)^2 + 0{,}03 - 0{,}03125 = 0{,}02\,(x - 1{,}25)^2 - 0{,}00125.$$

Между 0,9 и 1,5 наибольшее отклонение x от 1,25 имеет место при $x = 0{,}9$ и $x = 1{,}5$. При $x = 0{,}9$ мы получаем наибольшую ошибку, около 0,0012, а при $x = 1{,}5$ ошибка равна 0. Так, мы видим, что остатки в последнем столбце превышают истинные ошибки приближения, если их считать точно. В следующем примере того же типа мы проявим меньшую осторожность и посмотрим, каковы будут последствия.

Пример 2. Округленная квадратичная модель. Мы будем округлять значения квадратичной формы $y = 1 + x^2$ до одного десятичного знака (после запятой) при $x = 0{,}9;\ 1{,}0;\ 1{,}1;\ 1{,}2;\ 1{,}3;\ 1{,}4;\ 1{,}5$ и попытаемся восстановить квадратичную форму по этим данным.

Решение. Результаты приведены в таблице илл. 13.3.2, аналогичной илл. 13.3.1. График на илл. 13.3.3 показывает искажения, внесенные округлением.

Восстановленная форма

$$y_{\text{предск}} = 0{,}95 - 0{,}02x + 1{,}01x^2$$

так же, как и в предыдущем примере, весьма близка к истинной. Ошибка приближения $-0{,}05 - 0{,}02x + 0{,}01x^2 = -0{,}06 + 0{,}01\,(x - 1)^2$ при x, изменяющемся от 0,9 до 1,5, близка к $-0{,}06$. Конечно, нам существенно помогает то, что нет никаких случайных ошибок, кроме ошибок округления, и *правильно* выбран вид модели, приближающей истинную функцию. С другой стороны, остатки не столь малы, как в примере 1. Но все же они меньше, чем, например, ошибки в коэффициенте при 1. Следовательно, восстановленная функция хорошо работает во всем диапазоне, хотя коэффициенты не слишком близки к истинным. Это наводит на мысль, что даже если сама функция восстановлена неверно, приближение тем не менее может быть весьма хорошим.

Пример 3. Округление до ближайшего целого. Что изменится, если мы попробуем восстановить $y = 1 + x^2$ по тем же данным, что в примере 2, но округляя до ближайшего целого?

Решение. В таблице илл. 13.3.4 приведены результаты, аналогичные предыдущим. Окончательный результат

$$y_{\text{предск}} = 3{,}31 - 3{,}67x + 2{,}38x^2.$$

Даже авторский глаз вряд ли обнаружит сходство правой части этой формулы с $1 + x^2$.

Рассмотрим разность

$$y_{\text{ист}} - y_{\text{предск}} = 1 + x^2 - (3{,}31 - 3{,}67x + 2{,}38x^2).$$

Дополняя до полного квадрата, получим, что ошибка равна:

$$0{,}12 - 1{,}38\,(x - 1{,}33)^2.$$

Следовательно, ее максимальное абсолютное значение (при x, меняющемся от 0,9 до 1,5) равно:

$$|(0{,}12 - 1{,}38\,(-0{,}43)^2)| \approx |-0{,}14| = 0{,}14.$$

При округлении до ближайшего целого, когда мы вносим изменения такой величины, как 0,5, ошибка в приближении порядка 0,15 не кажется чрезмерно большой. Данные приближены достаточно хорошо в соответствующем интервале x, хотя различие в коэффициентах значительное. Мы вправе спросить, почему.

Возвращаясь к началу этого параграфа, мы еще раз можем убедиться в том, что в первом примере нам удалось весьма точно восстановить коэффициенты, и это еще более подчеркивает неудовлетворительные результаты третьего примера. Различие разительное и, возможно, неискушенному читателю его трудно объяснить. Но напомним ему, что возмущения, внесенные нами в третьем примере, в 100 раз больше, чем в первом. Чтобы погасить стократное увеличение разброса, обычно требуется тысячекратное увеличение объема выборки. Такое изменение должно соответствовать радикальному различию в поведении.

Общий вывод, который мы извлекли из рассмотрения этих примеров, вовсе не неожиданный, но требующий к себе постоянного внимания. Он таков: чем более размытыми данными мы пользуемся, тем меньше шансов у нас получить коэффициенты, реально отражающие вклад соответствующих им носителей *даже в том случае, когда мы точно угадали модель.* Для читателя может быть интересно установить, отвечают ли его собственные задачи ситуации примера 1, 2, 3 или даже чему-нибудь худшему.

Пример 4. Взаимоотношения рас; прогноз результатов совместной жизни двух рас и предыстория.

В таблице илл. 13.3.5 приводятся данные Герберта Хаймана [Hyman H. (1965)] о прогнозе результатов совместной жизни двух рас, сделанном белыми и черными на основе имеющегося у них жизненного опыта. Мы рассматриваем «расу» и «предысторию» как факторы, определяющие «прогноз».

Таким образом, мы вводим следующую систему баллов:

переменная x: раса $\begin{cases} \text{черный} = 1, \\ \text{белый} = 0; \end{cases}$

переменная t:	предыстория	жил раньше в районах со смешанным населением = 1, не жил раньше в районах со смешанным населением = 0;
переменная y:	прогноз	интеграция рас = 2. сосуществование рас = 1, расовый конфликт = 0.

Выбор шкалы для переменных x и t не играет роли, поскольку они содержат всего по две категории. Выбор шкалы для y играет некоторую роль. Если какой-нибудь читатель в отличие от нас предпочтет неравномерную шкалу, он без труда приспособит наш анализ к этой ситуации.

Счет упростится, если воспользоваться тем, что для упорядоченной последовательности чисел метод наименьших квадратов даст среднее. Более того, если можно провести для каждого x прямую через соответствующее среднее y, то эта прямая и есть приближение метода наименьших квадратов. Учитывая, что у нас всего два значения x, мы легко построим линию регрессии, вычисляя средние y и t при каждом x. Отрезок, отсекаемый на оси ординат, даст значение среднего при $x = 0$, а тангенс угла наклона будет равняться среднему при $x = 1$ минус среднее при $x = 0$, деленному на 1, так как разность абсцисс есть $(1 - 0)$.

В таблице илл. 13.3.5 приводятся необходимые для построения регрессии y и t по x вычисления, а также остатки $y_{\cdot 1x}$ и $t_{\cdot 1x}$. Окончательное уравнение регрессии получается на основе данных, приведенных в таблице илл. 13.3.6. Оно выглядит следующим образом:

$$y_{\text{предск}} = 0{,}514 + 0{,}029\ x + 0{,}371t.$$

Клетка, отвечающая наибольшему числу случаев, соответствует $y = 1$, $x = 1$, $t = 1$. Окончательное уравнение здесь дает

$$y_{\text{предск}} = 0{,}914 \text{ при остатке } y_{\cdot 1xt} = 1{,}09.$$

Максимальное колебание оценок (оценки не имеющего опыта белого и имеющего опыт черного индивидуумов) составляет 0,154 к 0,914, или 0,4 единицы шкалы, т. е. мы видим, что опыт и раса влияют на прогноз, но не слишком заметно. Важно помнить, что мы изучаем популяции, как они есть. События, определившие сам факт наличия или отсутствия опыта, возможно, также связаны с прогнозом. Прогнозы, делавшиеся в прошлом, сами по себе могут частично служить причиной знакомства или незнакомства с практикой совместного проживания. Поэтому надо проявлять осторожность и помнить, что мы скорее описываем ситуацию, чем определяем причины и следствия.

Пример 5. Линейные ограничения на переменные. Мы рассматривали уже в параграфе 12.5 равенство

$$B + Y = N,$$

где B — номер ребенка; Y — число детей, старших, чем он; N — число детей в семье.

Исследователь, изучающий эффект порядка рождения, должен выбирать между моделями

$$\text{отклик} = a + b_B B,$$

или

$$\text{отклик} = a + b_B^* B + b_Y Y,$$

или

$$\text{отклик} = a + b_B^{**} B + b_N N,$$

или какой-нибудь более сложной. Любую модель вида

$$\text{отклик} = a + b_B B + b_Y Y + b_N N$$

можно свести к одной из двух предыдущих (двучленных). Для интерпретации коэффициента B, который измеряет этот эффект, он должен выбирать между b_B, b_B^* и b_B^{**}. Выбор приходится производить из подгенератора, который здесь состоит либо из

$$\{\text{всех констант}\},$$

либо из

$$\{\text{всех } a + b_Y Y\},$$

либо из

$$\{\text{всех } a + b_N N\}.$$

Как же быть исследователю?

Отметим, что выбор может производиться из

$$\{\text{всех } a + b_Y Y + b_N N\},$$

так как любую модель вида $c + dB$ можно записать в форме

$$c + dY - dN.$$

Таким образом, подгенератор для B охватывает весь генератор, так что для оценки B ничего не остается.

Если исследователь полагает, что Y играет роль в исследовании лишь потому, что он связан с B и N, ему следует принять

$$a + b_B^{**} B + b_N N.$$

Аналогично, если исследователь убежден, что N влияет на отклик только через Y, ему лучше принять

$$a + b_B^* B + b_Y Y.$$

В любом случае мы должны отметить, что: а) выбор определяется здесь «взглядом на вещи», а также тем, кому этот «взгляд» принадлежит; б) данные со связью $Y = N + B$ не могут пролить свет на компетентность или разумность сделанного выбора.

Конечно, если все, чего добивается исследователь, — это аппроксимация данных, то любая из двух последних возможностей дает то же приближение (при всей неопределенности коэффициентов), что и

$$a + b_B B + b_N N + b_Y Y.$$

Никаких сложностей не возникает, пока мы не начнем задумываться о коэффициантах, чего как раз часто трудно избежать.

13.4. НЕКОТОРЫЙ ПРОИЗВОЛ В ВЫБОРЕ ХОРОШЕГО НОСИТЕЛЯ

Как мы видели выше, для описания регрессионного коэффициента нужно знать не только то, какому носителю он соответствует, но и подгенератор этого носителя. Теперь пора идти дальше.

Возьмем двух исследователей. Один из них методом наименьших квадратов строит модель

$$y \sim a + bx + ct + du.$$

Второй тем же методом —

$$y \sim A + Bx^* + Ct + Du,$$

где

$x^* = kx +$ любая линейная комбинация $(1, t, u)$. Как соотносятся b и B?

Если есть всего один набор данных (или, как мы будем говорить в дальнейшем,— одна ситуация), то, как показывает подстановка,

$$b = kB.$$

В тех случаях, когда k неизвестно, а это обычно для социальных и экономических исследований, возникают трудности в соотнесении двух описаний.

Если же есть множество ситуаций, то для каждой ситуации получаются свои коэффициенты:

ситуация $^{(1)}$,	ситуация$^{(2)}$	....,	ситуация$^{(H)}$,
$b^{(1)}$	$b^{(2)}$	....,	$b^{(H)}$,
$B^{(1)}$	$B^{(2)}$	....,	$B^{(H)}$.

Причем

$$b^{(i)} = kB^{(i)}$$

с одним и тем же k при всех i.

Это означает, что

$$\frac{b^{(1)}}{B^{(1)}} = \frac{b^{(2)}}{B^{(2)}} = \ldots = \frac{b^{(H)}}{B^{(H)}}.$$

Таким образом, с точностью до масштабного множителя $b^{(i)}$ дают то же, что и $B^{(i)}$.

Если наша цель — сравнение, а так и бывает в большинстве случаев, то нет большой разницы в использовании x и x^*. Важны:

1) подгенератор x (или x^*)

и

2) полный генератор модели.

(Они совпадают для $1, x, t, u$ и $1, x^*, t, u$).

Итак, помимо шкалирующего фактора, на конечный результат влияют подгенераторы рассматриваемого носителя (т. е. все возможные приближения, не содержащие рассматриваемого носителя) и полный генератор (собрание всех возможных моделей), а не конкретный выбор носителя.

13.5. ФЕНОМЕН «ЗАМЕСТИТЕЛЯ»

Предположим, что x_2 вносит существенный вклад в регрессию, т. е. существенно уменьшает остаточную дисперсию. Предположим далее, что x_{22} сильно коррелировано с x_2 так, что

$$x_2 \approx ex_{22}$$

при некотором e. (Присутствие или отсутствие в последней формуле постоянного члена не добавляет ничего нового к обсуждению.) Тогда мы можем ожидать близких дисперсий при включении в модель

$$\text{либо } b_2\,(ex_{22}),\ \text{либо } b_2\,(x_2).$$

Это неизбежное следствие тесной корреляции. Наличие или отсутствие установленного или предполагаемого соотношения между x_2 и x_{22} здесь роли не играет.

В таких случаях мы часто говорим, особенно если x_2 участвует в регрессии, а x_{22} — нет, что

$$x_2 \textit{ заместитель } x_{22}.$$

Иногда это удобно, но чаще — нет.

Туннельный эффект. Если x_2 и x_{22} тесно связаны и x_2 *не* связано с тем, что мы изучаем, а x_{22} *тесно* связано, причем x_2 *участвует* в регрессии, а x_{22} *не* участвует, то мы, по всей вероятности, установим, что x_2 вносит существенный вклад в регрессию.

Если это случится, мы будем склонны верить, что x_2 как раз и есть тот самый носитель, который нужен, хотя более уместна такая интерпретация: «носитель x_2 кажется относящимся к делу, так как он замещает x_{22}, который, в чем я уверен, относится к делу, потому что...».

Можно привести простой пример из геометрии. Предположим, что мы вычисляем периметры квадратов, стороны которых $x_2 = 4, 5, 6$. Делаем мы это, не зная ни структуры задачи, ни значений переменных x_2, а пользуясь лишь значениями связанной с x_2 переменной $x_{22} \equiv x_2^2 = 16, 25, 36$ соответственно. Тогда мы можем получить методом наименьших квадратов прямую, проходящую через начало координа-

$$y = 0{,}745x_{22},$$

которая дает оценки периметра 12, 19, 27 (если бы мы учли свободный член прямой, то результаты подгонки были бы гораздо лучше). Хотя $x_{22} = x_2^2$ не имеет даже нужной размерности, результаты тем не менее удовлетворительны. Несоответствие размерностей коэффициент учитывает.

Но x_{22} тесно связано здесь с x_2.

Нам часто придется встречаться с ситуациями такого типа.

Заместитель, вносящий путаницу. Если, наоборот, мы включим в регрессию и b_2x_2, и $b_{22}x_{22}$, хотя прямое отношение к делу имеет лишь x_{22}, то возникнут «разброд и шатания», как это имело место в примере с измерениями грудной клетки и сердца в параграфе 12.4. Оба члена, b_2x_2 и $b_{22}x_{22}$, будут входить в модель и оба будут использоваться для одной и той же цели, так что $b_2 + eb_{22}$ будет играть роль, которую

мы предназначали бы b_2, если бы обладали достаточной проницательностью и запасом данных, чтобы исключить x_{22}. В результате b_2 может оказаться меньше, может быть, значительно меньше, чем мы бы хотели, а может появиться даже и с неверным знаком.

Верх путаницы. Наихудший вариант неразберихи возникает тогда, когда x_{22} оказывается столь сильным соперником, что уменьшает b_2 почти до нуля. Это случается не слишком часто, когда есть только x_{22}, но если в модели несколько переменных (скажем, x_{22}, x_{23}, ..., x_{27}), сильно коррелированных с x_2, то можно быть почти уверенными, что мы близки к исключению b_2 из рассмотрения (или опять-таки он может даже перекинуться на другую сторону оси — получить неверный знак). А еще один-единственный x_{22} может, пусть и сохранив верный знак, сделать b_2 вдвое больше.

Общий случай (много переменных). Все сложности, возникающие при наличии двух сильно коррелированных переменных, могут проявиться и при наличии слабой линейной зависимости большего числа переменных. Предположим, например, что u, v, w — независимы (причинно и статистически) и

$$x_1 = u + v + 0_1,$$
$$x_2 = v + w + 0_2,$$
$$x_3 = w - u + 0_3,$$

где 0_i — представляет собой нечто достаточно малое. Тогда разность $x_2 - x_1$ очень сильно коррелирована с x_3 и нельзя избежать уже описанных трудностей.

Пример. Во время второй мировой войны исследовалась точность поражения целей при бомбовых ударах союзной авиации в Европе. В частности, для описания точности было составлено уравнение регрессии с десятью (или что-то около того) носителями. Среди носителей были высота бомбометания, тип самолета, скорость звена бомбардировщиков, размер звена, число истребителей противника. Из физических соображений можно было ожидать, что точность попадания обратно пропорциональна скорости и высоте и что для разных типов самолетов результаты бомбометания были разными. Но совершенно удивительным было то, что в соответствии с уравнением регрессии точность повышалась с ростом числа истребителей противника. Этот удивительный факт есть следствие феномена заместителя. В уравнении не было переменной, описывающей облачность. А при сильной облачности над целью истребители обычно не появлялись, зато и ошибки бомбометания были, как правило, весьма велики.

13.6. ИНОГДА x УДАЕТСЯ «СТАБИЛИЗИРОВАТЬ»

Мы были достаточно осторожны в прошлом и отметили, что, работая с x и $t = x^2$, как правило, не стоит интерпретировать коэффициенты при x_i, обсуждая, «что было бы, если бы остальные x остались неизменными». В этом параграфе мы попытаемся немного продвинуться вперед и сделать еще несколько самых необходимых предупреждений.

Полиномиальная модель. Когда приходится приближать полиномы, такие простые, как

$$b_1x + b_2x^2,$$

или такие сложные, как

$$b_0 + b_1x + b_2x^2 + b_3x^3 + b_4x^4 + b_5x^5,$$

попытки интерпретировать коэффициенты окупаются редко. График для модели или для разности двух моделей, полученных по двум наборам данных, может оказаться весьма полезным. Но коэффициенты сами по себе не стоят пристального рассмотрения.

Неподходящие x. Если x не слишком тесно связаны функциональной или статистической зависимостью, мы можем попытаться интерпретировать b_i как «эффект изменения x_i, когда все остальные x сохраняют свои значения». Если мы жаждем говорить о b_i, то интерпретация из предыдущей фразы — это максимум того, что вообще можно сделать.

В практических ситуациях, однако, надо понимать, сколь велико различие между высказываниями:

1) x_i изменяется, в то время как на остальные x *ничто больше* не воздействует;

2) x_i изменяется, а остальные x заморожены.

В социально-экономических исследованиях проявление такого рода различий не только возможно, но и весьма вероятно. Это объясняется тем, что переменные, с которыми мы здесь обычно имеем дело, не образуют полных наборов и не являются действительно определяющими. Рассмотрим в качестве примера исследование связи между тестами познавательных способностей детей и такими факторами, как образовательный уровень родителей, их материальная обеспеченность и время обучения. Отметим, что у нас нет мер для таких переменных, как врожденные способности, трудоспособность и склонность к учебе, родительское и учительское поощрение, так что нельзя не говорить о гандикапе*. Наша регрессия «работает» лишь потому, что и y, и x определяются некоторыми фундаментальными факторами, которые действуют так, как они действуют в свое время и в своем месте, безразличные к нашему сбору данных.

Учет этих факторов привел бы, по всей вероятности, к изменению регрессии, причем эффект такого изменения был бы непредсказуем в рамках тех двух регрессий, о которых говорилось выше.

Фиксация значений всех носителей, кроме одного, если это вообще возможно, может повлиять на факторы, лежащие в основе изменчивости, что в результате изменит регрессию. В большинстве ситуаций политического или экономического характера такая опасность вполне реальна.

Мультиколлинеарные носители. Если некоторая линейная комбинация носителей почти постоянна, то, как мы уже говорили, единст-

---

* Гандикапом называется соревнование заведомо неравных партнеров, в котором для уравнения шансов сильнейшие дают слабейшим фору. Здесь — синоним несправедливого сравнения. — *Примеч. пер.*

венный выход — это изменение системы координат. При этом вводятся, с одной стороны, носители, коэффициенты которых мы можем достаточно разумно оценивать, а с другой — такие носители, о которых мы можем сказать очень мало. Коэффициенты последних вряд ли нуждаются в интерпретации.

Что же касается первых, то здесь обычно есть теоретические соображения по поводу того, что «должно быть постоянно». Так, забегая вперед, обратимся к примеру из параграфа 15.5 с профессиональным футболистом. Здесь мы могли бы предсказать, насколько лучше будут обстоять дела у футболиста, размеры тела которого на 10% больше, чем у другого. Однако коэффициент носителя $x_{2\ldots90}$ не всегда может измерить этот эффект. В этом примере все три носителя — $x_{123\ldots90}$, x_{567890}, x_{7890} — увеличились бы в таком случае на 10%. Следовательно, все три коэффициента — $b_{123\ldots90}$, b_{567890}, b_{7890} — должны вносить вклад в описание эффекта увеличения размеров на 10%.

13.7. ЭКСПЕРИМЕНТЫ, ЗАМКНУТЫЕ СИСТЕМЫ, СОПОСТАВЛЕНИЕ ЕСТЕСТВЕННЫХ И ОБЩЕСТВЕННЫХ НАУК С ПРИМЕРАМИ

Джордж Бокс (G. Box) писал в 1966 г.: «Есть только один путь, чтобы узнать, что же произойдет при воздействии на сложную систему, и этот путь — воздействовать на нее, а не пассивно наблюдать». Это предостережение против «натурных пассивных экспериментов» казалось весьма сильным. Мы и теперь не видим ему разумной альтернативы, разве что оно кажется нам слишком слабым.

Регрессия — видимо, наиболее мощное оружие из тех, какие мы используем в анализе данных. Иногда даже создается впечатление, что она дает нам больше, чем мы хотим узнать, больше того, что вообще можно извлечь из наших данных. Это впечатление, конечно, обманчиво. Несколько приводимых далее примеров проиллюстрируют нам опасность таких заблуждений и прояснят позицию Бокса.

Предположим прежде всего, что мы собираемся производить какие-то измерения на совокупности (людей или какой-либо другой) и хотим использовать регрессионные коэффициенты для оценки того, как единичный скачок определяющей переменной (скажем, x_1) влияет на отклик (скажем, y). Это не так просто сделать потому, что регрессионный коэффициент зависит от остальных переменных, участвующих в регрессии. Замечание не исключает принципиальной возможности предсказания отклика y по наблюденным, *в том виде, как они существуют сейчас*, переменным x_1, x_2 и т. д. Но в нем содержится значительная доля недоверия к практике определения изменений y по изменениям x_1 для объекта (класса, города, штата, страны) без соответствующей проверки с помощью управляемого опыта, осуществляющего требуемое изменение. (Строго говоря, если мы хотим установить эффект изменения x_1, то надо произвести такой управляемый опыт, в котором изменяется только x_1 и невозможно изменение никаких других переменных под влиянием предпринятых для изменения x_1 действий. Такое пожелание, вообще говоря, нереалистично, может оказаться не-

выполнимым и не принести тех плодов, которые мы хотим получить. Мы просто хотим знать, что же происходит на самом деле при изменении x_1.)

Когда обсуждается такой спорный вопрос, энтузиасты пассивного наблюдения в сочетании с регрессионным анализом обычно ссылаются на удачные примеры применения их подхода в технике.

Представление о том, что применение регрессии для измерений в физических и технических задачах всегда приносит успех, — серьезное заблуждение. Прежде всего потому, что большинство таких применений регрессии для измерений основано на экспериментальных данных. Во-вторых, потому, что останется очень мало задач, если ограничиться системами, где:

- «переменных» немного;
- они хорошо определены;
- измеряются с небольшими ошибками.

Пример. 1. Теплоемкость. Если инженер подгоняет к экспериментальным данным модель

$$A + BT + CT^2 + DT^3$$

для изменения теплоемкости некоторого вещества в зависимости от температуры, то он уверен, что 1) температура действительно определяет теплоемкость; 2) изменяется только температура; 3) используемая шкала температур надежна (на проверку этого, возможно, ушли недели работы); 4) точность измерений очень высока по сравнению с диапазоном температур (T). В социальных, экономических или медико-хирургических исследованиях уверенность даже по одному из подобных пунктов возникает редко.

Некоторые исследователи, остро реагирующие на эти трудности, пытаются использовать регрессию скорее для качественного анализа. Признавая, что она не может измерять количественно те или иные эффекты, они пытаются определять с помощью регрессии лишь наличие или отсутствие этих эффектов. Но и на этом пути бывают неудачи.

С чем они могут быть связаны? Мы уже касались этого вопроса в примере с совместным проживанием индивидуумов, принадлежащих к разным расам. Мы отмечали там, что прогноз, делавшийся в прошлом (величина, близкая к отклику), может частично определять наличие или отсутствие опыта совместного проживания. Сформулируем эту мысль в более общем виде. Если мы опасаемся, что величина x_2, которую мы измеряем с ошибкой, может давать вклад в «причинность y» так же, как и величина x_1, которую мы измеряем с ошибкой, мало построить множественную регрессию y по x_1 и x_2. Коэффициент при x_1 может заметно отличаться от нуля, хотя x_1 совершенно не влияет на y (см. [Tukey J. W. (1973 и 1974)]).

Приведем еще примеры.

Пример 2. Развитие ребенка. Предположим, что мы изучаем какой-либо аспект уровня развития ребенка y как отклик на какой-то «недомашний» фактор x_1. Вследствие этого мы хотим отделить влияние x_1 от домашнего влияния, материального или интеллектуального. Прежде всего мы выбираем подходящую меру домашнего влияния, назовем

ее x_2 и построим регрессию y по x_1 и x_2. Если мы найдем, что x_1 явно влияет на y, то можно ли надеяться на сохранение этого обстоятельства и при «закреплении домашнего влияния на одном уровне»?

К сожалению, в этом нельзя быть уверенным. Прежде всего, как бы разумно мы ни выбрали x_2, сколь ни велика была бы проявленная при этом интуиция, мы не сможем адекватно описать влияние всех домашних факторов. Далее вполне может оказаться что x_1, влияние которого мы как раз изучаем и которое шатается под тяжестью «домашнего влияния» x_2, коррелировано в совокупности — источнике наших данных. При таких обстоятельствах даже значительная величина коэффициента при x_1 не гарантирует, что его изменение окажет влияние на отклик, «если домашнее влияние закреплено на одном уровне».

(При написании этой части параграфа 13.7 мы использовали в переработанном виде исследование Гильберта и др. [Gilbert et al (1977)].

Пример 3. Долгое обучение—или педагогическое мастерство? Рискуя переборщить с объяснениями, рассмотрим еще один пример. Предположим, что результаты тестирования в некоторых единицах приближенно определяются образовательным уровнем x_1 (измеряемым числом оконченных классов) и специальной подготовкой x_2 (измеряемой в неделях тренировки). Модель, используемая для описания этой зависимости, выглядит так:

$$y = 6 + x_1 + 2x_2.$$

Коэффициенты 1 и 2 перед носителями x_1 и x_2 могут сгодиться для оценки среднего числа дополнительных лет учебы или среднего числа недель специальных тренировок. Такое использование коэффициентов регрессии характерно для социометрических исследований.

Когда данные получены в эксперименте, такую интерпретацию можно считать обоснованной, чего нельзя утверждать, если данные получены в результате пассивных наблюдений над реальным процессом. Например, давайте рассмотрим совокупность индивидуумов с различными значениями x и соотнесем их y с их x так, как это описывалось выше. Мы можем утверждать, что люди, имеющие образовательный уровень на 2 класса выше, имеют в среднем на две единицы лучшие результаты тестирования и что те, кто получил на три недели больше специальных упражнений, в среднем имеют на 6 единиц лучший результат тестирования.

Однако нельзя утверждать с такой же определенностью, что взятый наугад индивидуум, получив еще две класса образования, обязательно увеличит на 2 единицы свою среднюю отметку за тест. Это не связано с надежностью модели. Говоря об увеличении образовательного уровня, мы имеем в виду массу людей. Трудности возникают в связи с тем, что данные не основаны на эксперименте. Может случиться, что причина, по которой человек остановился на 9 классах образования, частично объясняет и его успехи в исполнении тестового задания. Поэтому, если такие люди получают каким-то образом дополнительное образование, это может не повлиять на их результаты в выполнении теста. Если бы речь шла об эксперименте, то мы постарались бы устроить его так, чтобы иметь равные возможности воздействия на инди-

видуумов, либо ограничились бы предсказанием поведения той категории людей, которая изучалась. Классическим примером служит эксперимент с вакциной Солка (Salk), где заболеть полиомиелитом были *более* склонны добровольцы, чем люди, не выразившие желания участвовать в эксперименте. Пассивный эксперимент должен был бы показать меньшую эффективность вакцины, чем она была на самом деле, так как должны были сравниваться результаты обследования привитых добровольцев и непривитых индивидуумов, не выразивших желания сделать прививку.

Несмотря на указанную опасность, такого рода интерпретации эффекта изменения носителей часто делаются, причем без должного предостережения читателя. Предположение, что дополнительное образование улучшает результаты тестирования, скорее всего правильно. Но оно не может быть доказано данными, полученными из «эксперимента», описанного выше (по поводу серьезных попыток обосновать это см. [Human H. H., Wright C. R. and Reed J. S. (1975)]. Общие представления о мире часто могут дать гораздо больше, чем экспериментальные данные. В частности, мы обычно не можем слишком доверять величинам коэффициентов регрессионной модели.

В физике мы иногда стараемся получить лишь приближенно закон, связывающий y и x. Коэффициенты при этом могут иметь вполне определенный смысл, например ускорения силы тяжести.

Здесь мы часто можем или по крайней мере думаем, что можем, узнать переменные, которые «определяют» отклик. Так, радиус окружности определяет ее площадь. Зная радиус, нам никуда не деться от площади. Но во многих социальных задачах «причин» много и они весьма запутаны, если вообще определимы. Это приводит к большим затруднениям при выборе факторов.

Пример 4. Бедность и врожденные уродства. Какие факторы служат причиной бедности? Чтобы исследовать эту проблему, нужно как минимум иметь определенные взгляды и представления о социальных и экономических процессах. Или, взяв другое направление, можно задаться вопросом, какие переменные определяют наличие либо отсутствие врожденных уродств? Одна из возможных концепций состоит в том, что врожденные уродства связаны с биологическими процессами, приводящими к рождению ребенка. Такой подход, по-видимому, естествен для биолога или студента-медика, хотя последний может также утверждать что-нибудь вроде «улучшенный предродовой уход должен сократить число уродств в данном классе индивидуумов». Теперь в число причин включается качество ухода. Такое нащупывание причин — бесконечный процесс, причем то, что служит причиной для одного индивидуума, может быть следствием для другого.

Пример 5. Меры здоровья. Если мы решили оценить уровень здоровья групп людей, то, может быть, разумно выбрать в качестве нескольких определяющих переменных уровень смертности в этих группах, среднее число дней болезни в году, качество питания. Мы можем использовать как переменную также уровень здоровья за прошлый год. В конце концов именно эта переменная играет очень важную роль при прогнозе здоровья на следующий год. Но если мы используем все

эти переменные при составлении регрессионного уравнения, то интерпретация коэффициентов будет затруднена, ибо все переменные предназначены для оценки одного и того же. С точки зрения врача, оценивание как таковое на базе этих переменных может быть вполне разумным, но мы мало что сможем сказать о коэффициентах уравнения. Обычно, когда задача столь расплывчата, мы в поисках «причинных» переменных вводим несколько факторов, описывающих одно и то же.

Пример 6А. Размеры яиц. Полная линейная модель. Чтобы завершить ту часть нашей дискуссии, которая относилась к более точным разделам науки и была начата с рассмотрения квадратичных моделей, вернемся к примерам физических измерений. А. П. Демпстер (A. P. Dempster) измерял диаметры (продольный и поперечный) куриных яиц с целью определения зависимости их объема от этих линейных размеров. Если яйцо — эллипсоид вращения, то его объем

$$V = kLW^2,$$

где L — продольный диаметр; W — диаметр максимального кругового сечения (как курам удается *делать* яйца с круговыми поперечными сечениями?); $k=\pi/6$. Диаметры Демпстер определял с помощью штангенциркуля, а объем — так же, как вы или Архимед определяли бы объем рака.

Демпстер подгонял уравнение

$$KV = cL^{\beta_2}W^{\beta_3}$$

к данным, приведенным в таблице илл. 13.7.1. С этой целью он сначала брал десятичный логарифм, затем использовал метод наименьших квадратов. Иными словами, он подгонял к данным илл. 13.7.1 модель

$$v = \beta_1 \cdot 1 + \beta_2 x_2 + \beta_3 x_3,$$

где

$$v = \log KV,\ x_1 = 1,\ x_2 = \log L,\ x_3 = \log W,\ K = 6/\pi,\ \log c = \beta_1.$$

Если бы предположение об эллипсоидальной форме яиц было оправдано, а измерения точны, то должно было получиться $\hat{\beta}_1 = 0$, $\hat{\beta}_2 = 1$, $\hat{\beta}_3 = 3$. В действительности было получено соотношение

$$v = 0{,}320 + 0{,}728x_2 + 1{,}812x_3.$$

Это соотношение не кажется очень близким к

$$v = 0 + 1x_2 + 2x_3,$$

но обе модели достаточно хорошо приближают экспериментальные данные, что и не удивительно, если вспомнить наш опыт с округлением в квадратичной модели. Учитывая, что x_2 приблизительно равно 0,75, а x_3 — 0,62, мы видим, что вклад x_1 (=1) в v составляет примерно $1^1/_3 x_2$ или $1^1/_2 x_3$. Кроме того, сумма трех коэффициентов равна 2,860, т. е. близка к 3. Можно получить лучшее приближение к 3, если увеличить правую часть приблизительно наполовину от 0,320. Итак, ес-

ли речь идет о регрессии для прогноза, можно смело двигаться вперед. Но предсказание коэффициентов — дело дьявольски трудное. Мы убедились на этом примере, что даже в задачах физического типа при наличии замкнутой системы и почти правильном виде зависимости коэффициенты могут быть достаточно далеки от «истинных». (Полученные значения, однако, *нельзя* считать значимо отличными от 0, 1 и 2.) Мы употребили выражение «*почти правильный вид*» и взяли слово «истинных» в кавычки, так как яйца все же яйцевидны, а не эллипсоидальны, что может сыграть свою роль. (Историки утверждают, что различие между эллипсоидальной и яйцевидной формами задержало работу астронома Кеплера на несколько лет.)

Пример 6Б. Размеры яиц. Эллипсоидальная модель. Если мы положим $\hat{\beta}_1$, как ему и полагается быть в эллипсоидальной модели, нулем, то регрессионное уравнение будет иметь вид

$$v = 0{,}858x_2 + 2{,}168x_3.$$

Это уравнение значительно больше похоже на уравнение

$$v = 1x_2 + 2x_3.$$

К тому же отметим, что сумма коэффициентов 3,026 исключительно близка к 3.

Но не во всех примерах использование наших методов приводит к противоречивым результатам.

Пример 7. Расход пара, погода и число рабочих дней. Дрейпер и Смит [Draper N. and Smith H.(1966)] приводят пример расхода пара на производстве (y — количество использованных за месяц единиц пара, x_8 — среднемесячная температура воздуха по шкале Фаренгейта, x_6 — число рабочих дней в месяце).

Их приближенное уравнение

$$y = 9{,}1 - 0{,}0724x_8 + 0{,}2028x_6.$$

Отметим, что коэффициент при x_8, вполне естественно, отрицателен. Чем выше температура воздуха, тем меньше нужно пара. С другой стороны, чем больше дней работает фабрика, тем больше расходуется пара, что должно отражаться на знаке x_6, который, как ему и следует, оказался положительным. Если $x_8 = 60°\text{F}$ и $x_6 = 20$ дней, то предсказанное y равно 8,8. Знаки коэффициентов вполне разумны, но уверенности в значениях коэффициентов нет.

Пример 8. Урожай, удобрения и осадки. Т. Воннакотт и Р. Воннакотт [Wonacott T. and Wonacott R. (1969)] исследовали урожай пшеницы (y — урожай в бушелях на акр, x — удобрения в фунтах на акр, z — осадки в дюймах).

Уравнение связи

$$y = 11{,}33 + 0{,}0689x + 0{,}6038z.$$

Оба коэффициента положительны, чего и следовало ожидать, если мы не перебарщиваем с удобрениями и если почва не становится болотистой.

Величины, входящие в уравнение, изменяются в следующих пределах:

y: 40 до 80,

x: 100 до 700,

z: 32 до 37.

Выходит, что и здесь относительный порядок.

Проиллюстрируем еще раз некоторые неувязки, часто возникающие в социально-экономических задачах.

Пример 9. Влияние школы, семьи и учителей на оценки школьников. Выборочные данные по 20 школам из районов Атлантического побережья США, приведенные у Колемана [Coleman J. S and all. (1966)], были введены в вычислительную машину.

В роли переменных выступали: y — оценка за устную речь (для шестиклассника); x_1 — доход на одного ребенка в семье; x_2 — процент детей в классе из семей служащих (белые воротнички); x_3 — СЭП (социально-экономическое положение); x_4 — средняя оценка устных ответов, данная учителями; x_5 — средний уровень образования матерей (за единицу приняты 2 года обучения в школе).

Полученное регрессионное уравнение выглядело следующим образом:

$$y = 19{,}9 - 1{,}79x_1 + 0{,}0432x_2 + 0{,}556x_3 + 1{,}11x_4 - \\ - 1{,}79x_5.$$

Коэффициенты x_1 и x_5 выглядят неожиданно, во всяком случае по знаку, а может быть, и по величине.

В тех случаях, когда мы используем несколько конкурирующих переменных для описания одного и того же, как x_1 и x_2, x_3 и x_5 в данном примере, интерпретация коэффициентов очень трудна. Один из способов обойти это затруднение состоит в том, чтобы сделать из этих переменных одну. В рассматриваемом примере в качестве такой переменной можно взять какую-нибудь комбинацию социально-экономического положения и интереса к занятиям.

Использование в качестве переменной СЭП, x_3, — это уже попытка такого рода, здесь мы используем взвешенную сумму нескольких экономических переменных. Может быть, напротив, разумно использовать только переменную x_4 вместо всех остальных.

В целом ситуация может быть грубо описана следующим образом:

- если несколько конкурирующих переменных измеряют одно и то же, они, скорее всего, будут сильно коррелированы;
- если это так, то каждая из переменных может служить заместителем всех остальных;
- если они измеряются в эквивалентных единицах, то сумма их коэффициентов определяется хорошо, чего нельзя сказать о каждом из коэффициентов в отдельности;
- из имеющихся данных нельзя извлечь информацию о том, какая линейная комбинация переменных окажется подходящей и, в частности, какая из этих тесно связанных переменных по настоящему важна;

● разумно постараться построить комбинированную переменную, содержащую эти переменные, и оценивать регрессионные коэффициенты лишь для нее и других переменных, если таковые имеются; при этом не надо тщательно рассматривать отклики и обращать внимание на кажущиеся соотношения между откликами и этими переменными.

Иногда, особенно в промежуточных случаях, мы заинтересованы в том, чтобы сделать еще один шаг в этом направлении, а именно:

● выбрать составленную оценку, включающую j компонент (при этом не надо обращать внимание на отклик или их связи с откликом);

● найти остатки от регрессий составляющих компонент по составной — это даст нам j остаточных переменных;

● изучить эти остатки, в частности в их связи с ошибками измерения, известными, выведенными или предполагаемыми (если даже в небольшом проценте случаев их значения столь велики, что их никак нельзя объяснить обычными ошибками измерений, мы должны сделать все возможное, чтобы выяснить, связано ли это с необычно большими ошибками измерений или с необычными объектами; лишь в последнем случае такие остатки заслуживают использования);

● включить в регрессию а) составную переменную, б) остаточные переменные, которые оказались заслуживающими дальнейшего исследования.

Если мы пойдем по этому пути, то даже «выжившие» остатки будут, в общем, малы, а малость сумм их квадратов приведет к большим оценкам дисперсий соответствующих коэффициентов. Это печальное, но неизбежное и естественное следствие неполноты нашей информации. Развиваемый подход позволяет утилизировать информацию о том, какие из наших переменных кажутся важными. Кроме того, оказывается, что таким образом мы можем продвинуться довольно далеко.

Попытки свалить в кучу множество сильно коррелированных переменных в надежде, что полученное приближение выявит просто и быстро, какая или какие из них важны, — это обычное проявление беспочвенного оптимизма. Гораздо полезней понимать, чего «эти данные не могут нам дать», чем безосновательно верить результату, несущему больше информации, чем он в принципе может.

13.8. ОЦЕНКА ДИСПЕРСИЙ — ЭТО НЕ ВСЕ, ЧТО НУЖНО[1]

Легкий «метод борьбы» с обычными бедами регрессионных коэффициентов состоит в том, чтобы сказать: «... да ведь я всегда рассчитываю стандартные ошибки для регрессионных коэффициентов. Этого вполне достаточно, чтобы обезопасить себя». Однако мы вскоре убедимся в том, что:

● во многих ситуациях среднеквадратичные ошибки ничего не дают;

● в других ситуациях они дают слишком много.

Так что наш «легкий метод» не работает.

---

[1] В первом чтении этот параграф можно опустить.

Внешние заместители и корреляция между переменными, измеряемыми с ошибками. Затруднения возникают, как правило, если:

- носитель служит заместителем переменной, не включенной в регрессию (пример в 13.5, а также 3 и 4 в 13.7);
- два носителя представляют в модели (с ошибками) важные коррелированные переменные (примеры 4 в 13.3 и 2, 3, 5, 6 в 13.7).

В обеих ситуациях даже сколь угодно большое число данных не спасает положения. Соответственно стандартные ошибки, роль которых состоит в том, чтобы определить, насколько наши оценки по конечным данным могут отличаться от идеальных параметров для бесконечных данных, не могут служить мерой тех неопределенностей и недостоверностей, с которыми мы столкнулись. Чтобы избежать трудностей этого типа, нам нужно не увеличение числа данных, а нечто принципиально новое. И надежность своих выводов мы должны обеспечивать не такими мерами статистической неопределенности, как дисперсии, а как-то иначе.

Внутренние заместители. Есть еще одна трудная ситуация, в которой, в отличие от первых двух, увеличение числа наблюдений может оказаться полезным (пример 9 из 13.7 будет здесь удачной иллюстрацией). Простейший вариант такой ситуации — случай, когда, скажем, x_1 и x_2 сильно коррелированы. Это проявляется в том, что b_1 и b_2 определяются плохо, хотя некоторая их комбинация определена гораздо лучше. Пусть x_1 и x_2 измеряются в эквивалентных единицах, тогда, если корреляция положительна, будет хорошо определено $b_1 + b_2$, а если отрицательна — то $b_1 - b_2$. (Значительная отрицательная корреляция причиняет столько же хлопот, сколько и значительная положительная.)

Если в такой ситуации мы рассматриваем лишь оценку дисперсии b_1 и оценку дисперсии b_2, то и получим просто два больших числа. Если мы не предпримем никаких дополнительных шагов (таких, например, как расчет ковариации b_1 и b_2), то у нас не будет шансов заметить, что по крайней мере одна линейная комбинация b_1 и b_2 определена хорошо.

Если у нас всего два носителя, то рассмотреть 1 ковариацию и 2 дисперсии — вполне реальная вещь. Однако если носителей 5, то нужно рассчитать 10 ковариаций при пяти дисперсиях, а в случае 10 носителей — это уже 45 ковариаций. Хотелось бы иметь метод попроще.

Битон и Тьюки [Beaton A. E. and Tukey J. W. (1974)] предложили следующий подход. Сначала находим c, дающие настолько малые оценки дисперсий следующих ниже выражений, насколько это возможно, и сами эти оценки:

$$b_1 + c_{12}b_2 + c_{13}b_3 + \ldots + c_{1k}b_k,$$
$$c_{21}b_1 + b_2 + c_{23}b_3 + \ldots + c_{2k}b_k,$$
$$\vdots$$
$$c_{k1}b_1 + c_{k2}b_2 + c_{k3}b_3 + \ldots + b_k.$$

Если, например, вычисления показывают, что коэффициенты, остатки для минимальных дисперсий и оценки дисперсий определяются приведенной ниже таблицей

Коэффициенты	Дисперсия	Линейные комбинации	Дисперсия
b_1	(47,2)	$b_1 - 0{,}11b_2 - 0{,}07b_3$	(44,0),
b_2	(3,0)	$-0{,}04b_1 + b_2 - 0{,}01b_3$	(2,9),
b_3	(25,3)	$-0{,}08b_1 - 0{,}23b_2 + b_3$	(21,4),

то можно, по-видимому, считать, что дела обстоят почти прекрасно, во всяком случае до тех пор, пока нас волнуют лишь зависимости, мешающие правильной интерпретации дисперсий.

Если же мы получили следующую картину:

b_1	(47,2)	$b_1 - 0{,}11b_2 + 1{,}37b_3$	(1,54),
b_2	(3,0)	$-0{,}06b_1 + b_2 - 0{,}03b_3$	(2,7),
b_3	(25,3)	$0{,}67b_1 - 0{,}08b_2 + b_3$	(0,47),

то станет очевидным, что $b_1 + 1{,}4b_3$ определено относительно гораздо лучше, чем любой из коэффициентов b_1 или b_3 (так как x_1 и x_3 сильно коррелированы). Этот подход пригоден и для диагностики более сложных связей, включающих тройные взаимодействия или взаимодействия более высоких порядков.

Успешно потрудившись на поприще определения связей, мы в состоянии продвинуть проблему на шаг дальше.

После того как найдено выражение

$$b_i + \Sigma' c_{ij} b_j$$

(где Σ' означает суммирование по всем $j \neq i$), минимизирующее оценку дисперсии, мы, естественно, попытаемся исключить из этой суммы столько членов, сколько возможно (конечно, корректируя коэффициенты c после каждого исключения), не увеличивая значительно оценки дисперсии результата.

Для этих целей, видимо, подойдет метод «исключения», который мы обсуждаем в 15-й главе. Какое увеличение дисперсии приемлемо? Здесь можно предположить два правила, одно более, а другое менее осторожное. А именно: пусть V=оценка var $\{b_i\}$ и $V_{\min}$=оценка var $\{b_i + \Sigma' c_{ji} b_j\}$; примем, что

$$\text{минимум } V_{\min} \text{ достигнут, если } V_{\min} \leqslant \frac{V - V_{\min}}{K},$$

$$\text{минимум } \frac{V - V_{\min}}{K} \text{ достигнут, если } \frac{V_{\min}}{K} \leqslant \frac{V - V_{\min}}{K} \leqslant V_{\min},$$

$$\text{минимум } \frac{V_{\min}}{K} \text{ достигнут, если } \frac{V - V_{\min}}{K} \leqslant \frac{V_{\min}}{K},$$

где K равно 10 для более осторожного правила и 5 для менее осторожного.

Мы допускаем возрастание до медианы ряда

$$\left\{\frac{V_{\min}}{K}, \frac{V - V_{\min}}{K}, V_{\min}\right\}.$$

Таким образом, допускается разумное увеличение как в случае $V_{\min} \ll V$, так и в случае $V - V_{\min} \ll V$.

В наших примерах это приводит к следующим результатам. В первом случае

b_1	(47,2)	b_1	(47,2),
b_2	(3,0)	b_2	(3,0),
b_3	(25,3)	$-0{,}27b_2+b_3$	(23,0),

что указывает на возможную умеренную или слабую зависимость b_2 и b_3, которая, однако, не мешает их одновременному включению в модель. Во втором случае

b_1	(47,2)	$b_1+1{,}38b_3$	(1,97),
b_2	(3,0)	b_2	(3,0),
b_3	(25,3)	$0{,}71b_1+b_3$	(0,52).

Здесь, очевидно, выявляется зависимость между b_1 и b_3, а также хорошая определенность их соответствующей линейной комбинации.

КОММЕНТАРИЙ

В других частных случаях могут потребоваться другие подходы. Однако мы прояснили для себя следующее:

- в некоторых ситуациях стандартные ошибки не дают информации;
- в других они могут вуалировать реально полученную информацию;
- в последнем случае существует процедура, позволяющая автоматически привлечь наше внимание к возможным трудностям;
- в первом случае такая процедура нам неизвестна, поэтому не можем предложить ничего лучшего, чем внимательное и глубокое проникновение в ситуации.

РЕЗЮМЕ. БЕДЫ РЕГРЕССИОННЫХ КОЭФФИЦИЕНТОВ

Знать носитель и его регрессионный коэффициент — это очень мало. Нужно как минимум знать подгенератор, отвечающий данному носителю (один и тот же носитель может иметь разные подгенераторы даже при представлении одного генератора разными способами).

Коэффициент носителя можно найти простой линейной регрессией отклика по нему, если *они оба скорректированы* по подгенератору носителя.

Если мы представляем генератор двумя способами, так что меняется один носитель, а остальные остаются неизменными, то оценки коэффициентов будут отличаться постоянным сомножителем. Поэтому если мы сравниваем результаты по нескольким выборкам, то коэффициенты исходного носителя дадут нам ту же информацию, что и коэффициенты измененного носителя. (Все это верно также в том случае, когда меняются и остальные носители, важно лишь, чтобы не менялся весь генератор.)

Суммируя сказанное выше, нужно признать, что подгенератор носителя играет даже более важную роль, чем сам носитель.

Мы должны быть готовы к тому, что какая-то переменная (носитель) используется как заместитель другой. Мы должны, вследствие этого, задумываться, какую часть коэффициента (который мы хотели бы видеть при нужной переменной) несет в себе ее заместитель? Мы должны задумываться и о том, почему использование некоторой переменной дает хороший эффект, — только ли потому, что она служит заместителем? (В любом случае интерпретация регрессионных коэффициентов требует большого внимания.)

Стоит различать два типа ситуаций. Существует значительное различие между тем, что происходит при изменении x, если а) остальные x никак *иначе* не возмущаются, б) остальные x заморожены. В тех случаях, когда мы собираем данные в ситуации а), а хотим иметь результаты, применимые к б), может возникнуть путаница.

Внутренние стандартные ошибки ничего не дают для преодоления описанных выше трудностей.

Если рассматривать оценки дисперсий регрессионных коэффициентов без учета ковариаций, то может создаться впечатление слишком большой изменчивости. Это приводит нас к следующей процедуре:

- найти комбинации $b_i + \Sigma' c_{ij} b_j$ (где $j \neq i$) с такими оценками регрессионных коэффициентов, которые минимизируют оценку дисперсии;
- сравнить дисперсии этих комбинаций с оценками дисперсий исходных коэффициентов;
- попытаться упростить эти комбинации, опуская отдельные слагаемые и корректируя оставшиеся так, чтобы не получилось большого увеличения оценки дисперсии результата;
- рассмотреть найденную комбинацию как источник информации о важных зависимостях между оценками регрессионных коэффициентов.

БИБЛИОГРАФИЯ

Beaton A. E. and Tukey J. W. (1974). The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data. — Technometrics, 16, 147—185.

Box G. E. P. (1966). Use and abuse of regression. — Technometrics, 8, 625—629.

Coleman J. S., Campbell E. Q., Hobson C. J., McPartland J., Mood A. M., Weinfeld F. D., York R. L. (1966). Equality of Educational Opportunity. 2 volumes. Washington, D. C. Office of Education, U. S. Department of Health, Education and Welfare, U. S. Government Printing Office. [OE—38001; Superintendent of Documents Catalog № FS 5.238: — 38001.]

Draper N. and Smith H. (1966). Applied Regression Analysis. New York, John Wiley and Sons, Inc., 362. Русский перевод: Дрейпер Н., Смит Г. Прикладной регрессионный анализ, М., Статистика, 1973., с. 355.

Gilbert J. P., Mosteller F. and Tukey J. W. (1977). Steady social progress requires quantitative evaluation to be searching. Chapter 4 of The Evaluation of Social Programs (C. C. Abt, Ed.). Beverly Hills, Ca., Sage Publications, Inc., 295—312.

Hyman H. H., Wright C. R. and Reed J. S. (1975). The Enduring Effects of Education. Chicago, University of Chicago Press.

Tukey J. W. (1973). The zig-zagging climb from initial observation to successful improvement. Frontiers of Educational Measurement and Information Systems (W. E. Coffman, Ed.). Boston, Houghton Mifflin, 113—120.

Tykey J. W. (1974). Instead of Gauss-Markov least squares, what? B: Gupta R. P. (Ed.). Applied Statistics. Proceedings of a Conference at Dalhousie University, Halifax, Nova Scotia, May 2—4, North-Holland Publishing Co., 351—372.

Wonnacott T. H. and Wonnacott R. J. (1969). Introductory Statistics. New York, John Wiley and Sons, Inc., 255.

ИЛЛЮСТРАЦИИ

Иллюстрация 13.3.1

Квадратичная регрессия, шаговый метод

y	$y_{\cdot 1}=y-\bar{y}$	$y_{\cdot 1}x_{\cdot 1}$	x	$x_{\cdot 1}=x-\bar{x}$	$(x_{\cdot 1})^2$	$t=x^2$	$t_{\cdot 1}=(x^2)_{\cdot 1}=x^2-\bar{x^2}$	$t_{\cdot 1}x_{\cdot 1}$
0,30	—0,64	0,192	0,9	—0,3	0,09	0,81	—0,67	0,201
0,50	—0,44	0,088	1,0	—0,2	0,04	1,00	—0,48	0,096
0,70	—0,24	0,024	1,1	—0,1	0,01	1,21	—0,27	0,027
0,92	—0,02	0	1,2	0	0	1,44	—0,04	0
1,14	0,20	0,020	1,3	0,1	0,01	1,69	0,21	0,021
1,38	0,44	0,088	1,4	0,2	0,04	1,96	0,48	0,096
1,62	0,68	0,204	1,5	0,3	0,09	2,25	0,77	0,231
$\bar{y}=0,94$		0,616	$\bar{x}=1,2$		0,28	$\bar{x^2}=1,48$		0,672

$$g\equiv\frac{\Sigma\, y_{\cdot 1}\, x_{\cdot 1}}{\Sigma\, x^2_{\cdot 1}}=\frac{0,616}{0,28}=2,2,\ h\equiv\frac{\Sigma\, t_{\cdot 1}\, x_{\cdot 1}}{\Sigma\, x^2_{\cdot 1}}=\frac{0,672}{0,28}=2,4.$$

$y_{\cdot 1}$	$gx_{\cdot 1}$	$y_{\cdot 1x}\equiv y_{\cdot 1}-gx_{\cdot 1}$	$t_{\cdot 1}$	$hx_{\cdot 1}$	$t_{\cdot 1x}\equiv t_{\cdot 1}-hx_{\cdot 1}$	$y_{\cdot 1x}\, t_{\cdot 1x}$	$t^2_{\cdot 1x}\equiv[(x^2)_{\cdot 1x}]^2$	$y_{\cdot 1xt}$
—0,64	—0,66	0,02	—0,67	—0,72	0,05	0,0010	0,0025	—0,004
—0,44	—0,44	0	—0,48	—0,48	0	0	0	0,000
—0,24	—0,22	—0,02	—0,27	—0,24	—0,03	0,0006	0,0009	—0,006
—0,02	0	—0,02	—0,04	0	—0,04	0,0008	0,0016	—0,001
0,20	0,22	—0,02	0,21	0,24	—0,03	0,0006	0,0009	—0,006
0,44	0,44	0	0,48	0,48	0	0	0	0,000
0,68	0,66	0,02	0,77	0,72	0,05	0,0010	0,0025	—0,004
						0,0040	0,0084	

$$f\equiv\frac{\Sigma\, y_{\cdot 1x}\, t_{\cdot 1x}}{\Sigma\, t^2_{\cdot 1x}}=\frac{0,0040}{0,0084}=0,48,$$

$$y_{\text{прибл}}=\bar{y}+gx_{\cdot 1}+f\,(x^2)_{\cdot 1x}=0,94+2,2x_{\cdot 1}+0,48\,(x^2)_{\cdot 1x},$$

$$y_{\text{прибл}}=\bar{y}-(g-fh)\,\bar{x}-f\bar{x^2}+(g-fh)\,x+fx^2=$$

$$=1,03+1,05x+0,48x^2.$$

Иллюстрация 13.3.2

Восстановление $y = 1 + x^2$ при округлении одного знака после запятой

y	$y_{\cdot 1}$	$y_{\cdot 1}x_{\cdot 1}$	x	$x_{\cdot 1}$	$(x_{\cdot 1})^2$	$t = x^2$	$t_{\cdot 1}$	$t_{\cdot 1}x_{\cdot 1}$
1,8	—0,67	0,201	0,9	—0,3	0,09	0,81	—0,67	0,201
2,0	—0,47	0,094	1,0	—0,2	0,04	1,00	—0,48	0,096
2,2	—0,27	0,027	1,1	—0,1	0,01	1,21	—0,27	0,027
2,4	—0,07	0	1,2	0	0	1,44	—0,04	0
2,7	0,23	0,023	1,3	0,1	0,01	1,69	0,21	0,021
3,0	0,53	0,106	1,4	0,2	0,04	1,96	0,48	0,096
3,2	0,73	0,219	1,5	0,3	0,09	2,25	0,77	0,231
$\bar{y}=2,47$		0,670	$\bar{x}=1,2$		0,28	$\bar{t}=1,48$		0,672

$$g = \frac{0,670}{0,28} = 2,4, \quad h = 2,4.$$

$y_{\cdot 1}$	$gx_{\cdot 1}$	$y_{\cdot 1x}$	$t_{\cdot 1}$	$hx_{\cdot 1}$	$t_{\cdot 1x}$	$y_{\cdot 1x}\, t_{\cdot 1x}$	$t^2_{\cdot 1x}$	$y_{\cdot 1xt}$
—0,67	0,72	0,05	—0,67	—0,72	—0,05	0,0025	0,0025	0
—0,47	0,48	0,01	—0,48	—0,48	0	0	0	0,01
—0,27	0,24	—0,03	—0,27	—0,24	—0,03	0,0009	0,0009	0
—0,07	0	—0,07	—0,04	0	—0,04	0,0028	0,0016	—0,03
0,23	0,24	—0,01	0,21	0,24	—0,03	0,0003	0,0009	0,02
0,53	0,48	0,05	0,48	0,48	0	0	0	0,05
0,74	0,72	0,04	0,77	0,72	0,05	0,0020	0,0025	—0,01
						0,0085	0,0084	

$$f = \frac{0,0085}{0,0084} = 1,01,$$

$$y_{\text{прибл}} = \bar{y} + gx_{\cdot 1} + ft_{\cdot 1x} = 2,47 + 2,4x_{\cdot 1} + 1,01t_{\cdot 1x},$$

$$y_{\text{прибл}} = 0,95 - 0,02x + 1,01x^2.$$

Иллюстрация 13.3.3

Зависимость $y_{\text{прибл}}$ от x (при малых x) по данным илл. 12.1.2 и кривая $y_{\text{ист}} = 1 + x^2$

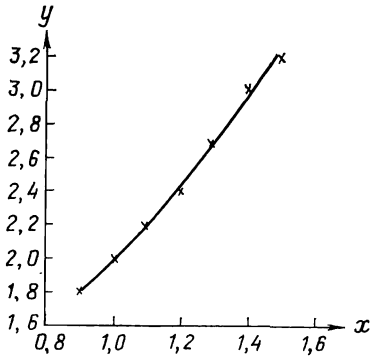


Иллюстрация 13.3.4

Восстановление $y = 1 + x^2$ по данным, округленным до целочисленных

y	$y_{\cdot 1}$	$y_{\cdot 1}x_{\cdot 1}$	x	$x_{\cdot 1}$	$(x_{\cdot 1})^2$	$t=x^2$	$t_{\cdot 1}$	$t_{\cdot 1}x_{\cdot 1}$
2	—0,43	0,129	0,9	—0,3	0,09	0,81	—0,67	0,201
2	—0,43	0,086	1,0	—0,2	0,04	1,00	—0,48	0,096
2	—0,43	0,043	1,1	—0,1	0,01	1,21	—0,24	0,027
2	—0,43	0	1,2	0	0	1,44	—0,04	0
3	0,57	0,057	1,3	0,1	0,01	1,69	0,21	0,021
3	0,57	0,114	1,4	0,2	0,04	1,96	0,48	0,096
3	0,57	0,171	1,5	0,3	0,09	2,25	0,77	0,231
$\bar{y}=2,43$		0,600	$\bar{x}=1,2$		0,28	$\bar{t}=1,48$		0,672

$$g=\frac{0,600}{0,28}=2,14,\quad h=2,4.$$

$y_{\cdot 1}$	$gx_{\cdot 1}$	$y_{\cdot 1x}$	$t_{\cdot 1}$	$hx_{\cdot 1}$	$t_{\cdot 1x}$	$y_{\cdot 1x}t_{\cdot 1x}$	$t^2_{\cdot 1x}$	$y_{\cdot 1xt}$
—0,43	—0,64	0,21	—0,67	—0,72	0,05	0,0105	0,0025	0,09
—0,43	—0,43	0	—0,48	—0,48	0	0	0	0
—0,43	—0,21	—0,22	—0,27	—0,24	—0,03	0,0066	0,0009	—0,15
—0,43	0	—0,43	—0,04	0	—0,04	0,0172	0,0016	—0,33
0,57	0,21	0,36	0,21	0,24	—0,03	—0,0108	0,0009	0,43
0,57	0,43	0,14	0,48	0,48	0	0	0	0,14
0,57	0,64	—0,07	0,77	0,72	0,05	—0,0035	0,0025	—0,19
						0,0200	0,0084	

$$f=\frac{0,0200}{0,0084}=2,38,$$

$$y_{\text{прибл}} = \bar{y} + g x_{\cdot 1} + f t_{\cdot 1x} = 2{,}43 + 2{,}14 x_{\cdot 1} + 2{,}38 t_{\cdot 1x},$$

$$y_{\text{прибл}} = 3{,}31 - 3{,}67 x + 2{,}38 x^2.$$

Иллюстрация 13.3.5

Предыстория, раса, прогноз (в скобках — число случаев)

y: прогноз	t: 1 — ранее жил в районах со смешанным населением		t: 0 — ранее не жил в районах со смешанным населением	
	x: 1 — черный	0 — белый	x: 1 — черный	0 — белый
2 — интеграция рас	11% (21)	9% (10)	5% (1)	5% (5)
1 — сосуществование рас	69% (133)	72% (83)	50% (9)	39% (35)
0 — расовый конфликт	20% (39)	19% (22)	45% (8)	56% (50)
	193	115	18	90

Источник. Hyman H. (1965). Survey Design and Analysis, N. Y., Данные воспроизводятся с разрешения автора и Macmillan Publishing Co., Inc. Copyright 1955 by the Free Press, a Division of the Macmillan Company.

Иллюстрация 13.3.6

Исключение y и t из x

А. Числа f_y

	x		
y	0	1	Итоги
2	15	22	37
1	118	142	260
0	72	47	119
Итоги	205	211	416

$\Sigma y f_y =$ 148 186

$\bar{y}_{x=0} = 0{,}722, \bar{y}_{x=1} = 0{,}882$

$\hat{y}_x = 0{,}722 + 0{,}160x$

$y_{\cdot 1x}$

y	$x=0$	$x=1$
2	1,278	1,118
1	0,278	0,118
0	—0,722	—0,882

Б. Числа f_t

	x		
t	0	1	Итоги
1	115	193	308
0	90	18	108
Итоги	205	211	416

$\Sigma t f_t = 115$ 193

$\bar{t}_{x=0} = 0{,}56, \bar{t}_{x=1} = 0{,}915$

$\hat{t}_x = 0{,}56 + 0{,}354x$

$t_{\cdot 1x}$

t	$x=0$	$x=1$
1	0,439	0,085
0	—0,561	—0,915

Иллюстрация 13.3.7

Удаление $y_{\cdot 1x}$ из $t_{\cdot 1x}$ для описания объединенного примера

y	x	t	Числа f_{yxt}	$y_{\cdot 1x}$	$t_{\cdot 1x}$	$f_{yxt} y_{\cdot 1x} t_{\cdot 1x}$	$y_{\cdot 1xt}$	$y-\bar{y}$
2	1	1	21	1,118	0,085	2,0	1,09	1,2
2	1	0	1	1,118	—0,915	—1,0	1,46	1,2
2	0	1	10	1,278	0,439	5,6	1,12	1,2
2	0	0	5	1,278	—0,561	—3,6	1,49	1,2
1	1	1	133	0,118	0,085	1,3	0,09	0,2
1	1	0	9	0,118	—0,915	—1,0	0,46	0,2
1	0	1	83	0,278	0,439	10,1	0,12	0,2
1	0	0	35	0,278	—0,561	—5,5	0,49	0,2
0	1	1	39	—0,882	0,085	—2,9	—0,91	—0,8
0	1	0	8	—0,882	—0,915	6,5	—0,54	—0,8
0	0	1	22	—0,722	0,439	—7,0	—0,88	—0,8
0	0	0	50	—0,722	—0,561	20,3	—0,51	—0,8
						24,8		

$$\sum f_{yxt}\, t^2_{\cdot 1x} = 66{,}9,$$

$$f = 24{,}8/66{,}9 = 0{,}371,$$

$$y_{\text{прибл}} = 0{,}722 + 0{,}160x + 0{,}371 t_{\cdot 1x},$$

$$y_{\text{прибл}} = 0{,}514 + 0{,}029x + 0{,}371t.$$

Иллюстрация 13.7.1

Данные о размерах яиц: $x_2 = \log L$, $x_3 = \log W$, $v = \log (6V/\pi)$; логарифмы десятичные

	x_2	x_3	v
1	0,7659	0,6360	2,031
2	0,7353	0,6198	1,982
3	0,7416	0,6280	1,995
4	0,7600	0,6280	2,019
5	0,7861	0,6239	2,031
6	0,7539	0,6156	1,956
7	0,7747	0,6156	2,007
8	0,7718	0,6239	1,995
9	0,7889	0,6114	1,995
10	0,7659	0,6072	1,995
11	0,7689	0,6156	1,995
12	0,7478	0,6239	2,007

Источник. Dempster A. P. (1969). Elements of Continuous Multivariate Analysis. Reading Mass: Addison-Wesley, p. 151. Публикуется с разрешения автора и издателя.

Глава 14 ● ОДИН КЛАСС ПРОЦЕДУР ПОДГОНКИ

При изучении этой главы читателю полезно иметь в виду, что она посвящена основным идеям регрессии, но не содержит деталей выводов. Развивая эти важные идеи, мы пользовались подходами и языком, весьма отличными от обычных. Если окажется, что читатель не всегда может сконструировать описанные ниже функции и веса, он не должен пугаться. Если же читатель поймет, как они строятся, то он сможет использовать эти идеи для того, чтобы разобраться в тонкостях подгонки, особенно в случае нескольких переменных. Представленные ниже методы и результаты могут, должны и будут применяться к анализу данных, но в этой главе упор сделан на их практические и идейные следствия, а не на технику применения.

Аддитивность служит отличительной чертой как классических методов приближения, так и процедур, входящих в качестве составных частей в более гибкие методы. Предположим, что $\hat{y}$ есть предсказание для некоторого y, $\hat{z}$ — для z, а $\widehat{y+z}$ — для $\hat{y}+\hat{z}$. Может случиться так, что

$$\widehat{y+z}=\hat{y}+\hat{z}.$$

Если это имеет место для любой выборки, то такой метод приближения или элемент процедуры приближения называется аддитивным.

Предположим, что метод приближения аддитивен в указанном смысле, тогда можно представить результат в виде суммы по множествам выборочных данных.

Если

$$y = y_{(1)} + y_{(2)} + \ldots + y_{(n)},$$

то модель должна удовлетворять соотношению

$$\widehat{y}=\widehat{y_{(1)}}+\widehat{y_{(2)}}+\ldots+\widehat{y_{(n)}}.$$

Положим, что

$$y_{(j)}(i)=\text{значение } y_{(j)} \text{ для } i\text{-го набора данных}=$$
$$=\begin{cases} y_{(i)}, & \text{если } i=j; \\ 0, & \text{если } i\neq j. \end{cases}$$

Тогда $\hat{y}_{(j)}$ зависит только от $\hat{y}$ на множестве данных i.

Таким образом, $\hat{y}$ получается *аддитивно* — суммированием оценок по множествам данных.

Методы подгонки, которые можно рассматривать как основанные на суммировании информации от различных множеств данных, важны потому, что они:

- просты;
- используют классические методы;
- гибкие (в частности, они позволяют в итеративных процедурах определять, что надо добавить на следующем шаге).

Насколько они общи? Хотя они и не универсальны, но могут почти превратиться в таковые. Все варианты линейного метода наименьших квадратов аддитивны. Аддитивны и все методы, которые к нему сводятся целиком или на каждой итерации. Нелинейный метод наименьших квадратов и современные методы устойчивого или робастного приближения можно рассматривать как итерации специальным образом изменяющихся линейных процедур типа метода наименьших квадратов.

С какой же точки зрения мы будем рассматривать такие виды приближений? В данной главе предлагается концепция «балансировки». Она связана математически и эвристически со многими идеями в теории приближений. Достаточно широко понимаемая концепция балансировки может работать во всех упомянутых случаях (для итеративного подбора, например, может быть, придется менять «балансиры» на каждой итерации).

Еще одним указанием на общность итеративно модифицированного метода наименьших квадратов служит то, что, как показано в параграфе 14.1, его частным случаем будет метод наименьших модулей.

14.1. ПРИБЛИЖЕНИЕ ПРЯМЫМИ. ПРЯМАЯ, ПРОХОДЯЩАЯ ЧЕРЕЗ НАЧАЛО КООРДИНАТ

Если мы приближаем y с помощью формы βx без свободного члена, то, как показано ниже, оценка β, получаемая обычным методом наименьших квадратов, выглядит так:

$$\widehat{\beta} = \frac{\Sigma xy}{\Sigma x^2} .$$

Переменную x мы называем *носителем*. Если приближается сумма нескольких членов, например $\beta x_1 + \gamma x_2 + \delta x_2^2$, то носителями будем называть все переменные: x_1, x_2 и x_2^2. (Мы не будем утруждать себя в дальнейшем напоминаниями о том, что на нуль делить нельзя. Отличие знаменателя от нуля всегда будет молчаливо предполагаться, кроме тех нескольких случаев, которые будут отмечены особо.) Крышечка « ˆ » над величиной в данной главе — знак оценки (полученной методом наименьших квадратов).

Если дисперсии всех y равны σ^2, а ковариации y нулевые, то

$$\operatorname{var} \widehat{\beta} = \frac{\Sigma x^2 \sigma^2}{(\Sigma x^2)^2} = \frac{\sigma^2}{\Sigma x^2} = \frac{\text{остаточная дисперсия}}{\Sigma x^2} .$$

Это соотношение, фундаментально, и мы часто будем сводить другие результаты о дисперсиях к подобному виду.

Если мы подгоняем к y модель $\alpha + \beta x$, т. е. используем в качестве носителей 1 и x ($\alpha \cdot 1 + \beta \cdot x$), то метод наименьших квадратов дает

$$\widehat{\beta} = \frac{\Sigma (x - \bar{x}) y}{\Sigma (x - \bar{x})^2}.$$

Как и раньше, если дисперсии y равны σ^2, а ковариации нулевые, то мы получаем

$$\operatorname{var} \widehat{\beta} = \frac{\text{остаточная дисперсия}}{\Sigma (x - \bar{x})^2}.$$

Последнее выражение указывает, что приближение y моделью

$$\mu + \beta (x - \bar{x}),$$

где $\mu = \alpha + \beta \bar{x}$, эквивалентно приближению формой

$$\alpha + \beta x.$$

Отметим, что при оценках методом наименьших квадратов (если y_i имеют равные дисперсии) присутствие μ или β не влияет на нашу оценку другого параметра, т. е. величина $\hat{\beta}$ при приближении вида $\beta (x - \bar{x})$ та же, что и для приближения $\mu + \beta (x - \bar{x})$. (Это верно также для β в $\alpha + \beta x$, так как оценки совпадают.) Далее, оценка $\mu - \bar{y}$ при приближении y константой такая же, как и при приближении формой

$$\mu + \beta (x - \bar{x}),$$

что неверно для α и $\alpha + \beta x$.

То, что μ и β не влияют друг на друга, вместе с предыдущими результатами означает

$$\operatorname{var} \hat{\beta} = \frac{\text{остаточная дисперсия}}{\Sigma (x - \bar{x})^2}$$

и

$$\operatorname{var} \widehat{\mu} = \operatorname{var} \{\bar{y} \mid \{x_i\}\} = \frac{\text{остаточная дисперсия}}{n},$$

где $\{x_i\}$ — заданный набор значений x.

Отметим, что в модели с $\mu \cdot 1$ все значения носителя равны 1, а сумма их квадратов — n.

Таким образом:

- если мы приближаем моделью βx, то знаменатель для $\operatorname{var} \hat{\beta}$ равен Σx^2;
- если мы приближаем моделью $\mu + \beta (x - \bar{x})$, то знаменатель для $\operatorname{var} \hat{\beta}$ равен $\Sigma (x - \bar{x})^2$;
- если мы приближаем моделью $\mu \cdot 1$, то знаменатель для $\operatorname{var} \hat{\mu}$ равен n $(= \Sigma 1^2)$.

Опираясь на наши предыдущие результаты (гл. 12), видим, что оценка модели — $\hat{\mu} + \hat{\beta}\,(x - \hat{x})$. Чтобы ее получить, надо отнять 1 от y и x, построить оценку $\hat{\mu}$, найти остаток $y_{\cdot 1}$, а затем приблизить $y_{\cdot 1}$ с помощью $x - \bar{x}$. Приближение моделью $\alpha + \beta x$ нельзя интерпретировать таким образом (за исключением случая $\bar{x} = 0$).

14.2. БАЛАНСИРОВКА — МЕТОД ПОДГОНКИ

Пусть задан набор носителей

$$x_1, x_2, \ldots, x_k,$$

каждый из которых принимает множество значений. Будем предполагать, что среди x нет идентичных, хотя некоторые из них и могут зависеть от других, а один из них может быть константой (для этих целей обычно используют x_1, полагая его единицей). Наша задача состоит в получении множественной линейной регрессии

$$\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

по экспериментальным данным.

Предположим, что в результате мы имеем

$$\hat{y} = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k.$$

Чтобы найти эти $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$, нужны какие-то уравнения. Конечно, нам будет легче, если эти уравнения будут линейными.

Приступим поэтому к описанию метода, дающего нам такие уравнения.

Пусть задан набор коэффициентов $\{h\,(i)\}$ (где индекс i обозначает i-е множество наблюдений, $i = 1, 2, \ldots, n$). Мы будем называть этот набор *балансиром* некоторого приближения в том и только в том случае, если для этого приближения и любого набора данных

$$\Sigma h\,(i) y\,(i) = \Sigma h\,(i) \hat{y}\,(i). \qquad (*)$$

Таким образом, «уравновешенная» сумма наблюдаемых значений отклика должна равняться «уравновешенной» сумме предсказанных значений отклика. Любой набор коэффициентов, удовлетворяющий равенству (*) для определенного вида модели, в нашей терминологии — *балансир* этого приближения.

Приведем простые примеры. Если мы приближаем методом наименьших квадратов $y = \beta x$, то x — балансир. Действительно,

$$\Sigma x\,(i) y\,(i) = \Sigma x\,(i) \hat{y}\,(i) = \Sigma x\,(i) \hat{\beta} x\,(i) = \hat{\beta} \Sigma x\,(i) x\,(i).$$

Значит, условие балансировки эквивалентно равенству

$$\hat{\beta} = \frac{\Sigma xy}{\Sigma x^2}.$$

Но это равенство, как уже говорилось выше, выполняется при подгонке методом наименьших квадратов.

Если мы строим модель $y = \alpha + \beta\,(x)$ или, что эквивалентно, $y = \mu + \beta\,(x - \bar{x})$, то 1, x и $x - \bar{x}$ служат балансирами.

Условие балансировки дает

для единицы: $\Sigma 1 \cdot y\,(i) = \Sigma 1 \cdot \hat{y}\,(i)$,

$$n\bar{y} = n\,(\hat{\alpha} + \hat{\beta}\bar{x}) = n\hat{\mu};$$

для x:

$$\Sigma x\,(i)y\,(i) = \Sigma x\,(i)\hat{y}\,(i) = \hat{\alpha}\Sigma x\,(i) + \hat{\beta}\Sigma x\,(i)x\,(i) =$$
$$= \hat{\mu}\Sigma x\,(i) + \hat{\beta}\Sigma x\,(i)\,(x\,(i) - \bar{x});$$

для $(x - \bar{x})$:

$$\Sigma\,(x\,(i) - \bar{x})y\,(i) = \Sigma\,(x\,(i) - \bar{x})\bar{y}\,(i) = \Sigma\,(x\,(i) - \bar{x}) \times$$
$$\times\,(\hat{\mu} + \hat{\beta}\,(x\,(i) - \bar{x})) = \hat{\beta}\Sigma\,(x\,(i) - \bar{x})^2.$$

Из этих соотношений следует, что $\hat{\mu}$, $\hat{\alpha}$, $\hat{\beta}$ — оценки метода наименьших квадратов.

Алгебра балансиров. Приходится оперировать совокупностями балансиров. Предположим, что $h = \{h(i)\}$ и $k = \{k\,(i)\}$ — балансиры. Тогда

$$\Sigma h\,(i)y\,(i) = \Sigma h\,(i)\hat{y}\,(i)$$

и

$$\Sigma k\,(i)y\,(i) = \Sigma k\,(i)\hat{y}\,(i).$$

Сложение балансиров с различными весами дает

веса 1 и 1: $\Sigma\,[h\,(i) + k\,(i)]y\,(i) = \Sigma\,[h\,(i) + k\,(i)]\hat{y}\,(i)$;

веса 2 и -3 : $\Sigma\,[2h\,(i) - 3k\,(i)]y\,(i) = \Sigma\,[2h\,(i) - 3k\,(i)]\hat{y}\,(i)$;

веса c_h и c_k: $\Sigma\,[c_h h\,(i) + c_k k\,(i)]y\,(i) = \Sigma\,[c_h h\,(i) + c_k k\,(i)]\hat{y}\,(i)$.

Таким образом, любые взвешенные суммы или линейные комбинации балансиров — балансиры.

Наша задача имеет два аспекта:

- получить балансиры для составления достаточного, хотя бы теоретически, числа независимых уравнений для определения неизвестных $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$;
- выбрать такую совокупность балансиров, которая позволяла бы легко решать эти уравнения.

Модель $y = \beta x$. В модели $y = \beta x$, подбираемой методом наименьших квадратов, роль балансиров могут играть лишь величины, пропорциональные x. Причем одного балансира достаточно, второй уже будет лишним. Арифметика обычно упрощается при использовании в качестве коэффициента пропорциональности 1, хотя если вычисления делаются, например, с точностью до 0,5, то удобней взять 2 или 4.

Модель $y = \alpha + \beta x$. При приближении $y = \alpha + \beta x$ методом наименьших квадратов все выражения вида $c + dx$ будут балансирами. Двух будет достаточно, если только они не пропорциональны. Третий — лишний.

Однако при малых x, если мы будем столь неосторожны, что выберем в качестве балансиров

$$1000000 + x \text{ и } 1000001 + x,$$

то при счете мы быстро застопоримся, так как небольшое различие балансиров при округлениях потребует внимания и огромного числа десятичных знаков.

Трудности возникнут и если мы выберем в качестве балансиров x из интервала (1950, 1975) и 1. Если же мы выберем в качестве балансира $x - \bar{x}$, то после простых преобразований получим легко решаемое уравнение

$\Sigma [x(i) - \bar{x}]y(i) = \hat{\beta}\, \Sigma[x(i) - \bar{x}]^2$ (уравнение чистого углового коэффициента).

Использование $1 + cx$, где $c = -\Sigma x(i)/\Sigma (x(i))^2$, приводит к

$\Sigma (1 + cx(i))y(i) = \hat{\alpha}\Sigma (1 + cx(i))$ (уравнение чистого свободного члена),

которое сводится к

$n\bar{y} + c\Sigma x(i)y(i) = \hat{\alpha}(n + cn\bar{x})$ (уравнение чистого свободного члена),

которое снова легко решается относительно $\hat{\alpha}$.

Если же мы хотим построить $y = \mu + \beta(x - \bar{x})$, то никаких проблем нет, поскольку, взяв 1 и $(x - \bar{x})$ в качестве балансиров, получим

$$\Sigma y(i) = n\hat{\mu}$$

и

$$\Sigma [x(i) - \bar{x}]y(i) = \hat{\beta}\Sigma [x(i) - \bar{x}]^2.$$

Когда в модели k коэффициентов, то нужны и k уравнений, причем линейно-независимых. А это значит, что семейство балансиров должно иметь размерность не менее чем k. Но невозможна и бóльшая размерность, поскольку $(k + 1)$ линейно-независимых уравнений с k неизвестными должны быть несовместны.

14.3. БАЛАНСИРЫ, НАСТРОЕННЫЕ НА ОТДЕЛЬНЫЕ КОЭФФИЦИЕНТЫ, И УЛОВИТЕЛИ

Рассмотрим теперь такие балансиры, которые позволяют легко определять отдельные $\hat{\beta}_i$ и «вылавливать» в данных всю относящуюся к ним информацию. Пусть мы строим модель

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

с помощью процесса, который эквивалентен методу балансировки (к таковым относятся все модификации линейного метода наименьших квадратов). Можно ли определить $\hat{\beta}_1$ с помощью одного балансира? Если найдется такой балансир $h = \{h(i)\}$, что

$$0 = \Sigma h(i)x_2(i) = \Sigma h(i)x_3(i) = \ldots = \Sigma h(i)x_k(i), \qquad (*)$$

то нам повезло с определением $\hat{\beta}_1$, поскольку в этом случае

$$\Sigma h\,(i)y\,(i) = \hat{\beta}_1 \Sigma h\,(i)x_1\,(i), \qquad (**)$$

следовательно,

$$\hat{\beta}_1 = \Sigma h\,(i)y\,(i)/\Sigma h\,(i)x_1\,(i),\ \Sigma h\,(i)x_1\,(i) \neq 0.$$

Можно было бы сказать, что h *настроено на* $\hat{\beta}_1$, так как $\hat{\beta}_2, \hat{\beta}_3, \ldots, \hat{\beta}_k$ не фигурируют в (**). Они как радиостанции, «отстроены», неслышны. Если мы сможем отстроить все $\hat{\beta}$, кроме одного, то и его мы тоже легко найдем.

Мы уже приводили простые примеры таких балансиров. Так, если мы приближаем $\alpha + \beta x$, то балансир $x - \bar{x}$ настроен на β и отстроен от α. Правда, в очень частном случае балансир c, $1 + cx$, настроен на α и отстроен от β.

Можно привести и более сложные примеры. Пусть $x_1 \equiv 1$, $x_2 \equiv x$, $x_3 \equiv x^2$, т. е. переменными служат 0, 1 и 2-я степени x. Предположим также, что x принимает значения 1, 2, 3, ..., 10. Тогда оказывается, что

$$c\,[22 - 11x + x^2]$$

— балансир, настроенный на $\hat{\beta}_3$ при любом c.

(Для проверки равенства

$$0 = \sum_1^{10} (22 - 11x + x^2)\,1 = \sum_1^{10} (22 - 11x + x^2)\,x,$$

может быть, полезно иметь в виду, что

$$\sum_{x=1}^{n} 1 = n,\ \sum_1^n x = \frac{n\,(n+1)}{2},\ \sum_1^n x^2 = \frac{n\,(n+1)\,(2n+1)}{6},$$

$$\sum_1^n x^3 = \left[\frac{n\,(n+1)}{2}\right]^2\Big).$$

Если число балансиров достаточно для обеспечения единственности решения, то ясно, что существует балансир, настроенный именно на то $\hat{\beta}_M$, которое нас интересует. Действительно, пусть $h_1, h_2, \ldots, h_k$ — k линейно-независимых балансиров, тогда надо найти набор d_i, удовлетворяющих $k - 1$ уравнению:

$$\sum_i (d_1 h_1\,(i) + d_2 h_2\,(i) + \ldots + d_k h_k\,(i))x_J\,(i) = 0,\ J \neq M,$$
$$1 \leqslant J \leqslant k.$$

Но эти уравнения эквивалентны (*), вследствие чего отстраивают все коэффициенты, кроме $\hat{\beta}_M$.

Определим теперь общий множитель, не учтенный последней системой, из условия

$$\sum_i (d_1 h_1\,(i) + d_2 h_2\,(i) + \ldots + d_k h_k\,(i)x_M\,(i) = 1.$$

Мы хотим, чтобы детерминант получившейся системы k (= (k — —1) + 1) уравнений не был нулевым. Тогда система имела бы единственное решение. Значит, мы хотим, чтобы не вырождался детерминант

$$\begin{vmatrix} \Sigma h_1 x_1 & \Sigma h_1 x_2 \ldots \Sigma h_1 x_k \\ \Sigma h_2 x_1 & \Sigma h_2 x_2 \ldots \Sigma h_2 x_k \\ \vdots & \vdots \\ \Sigma h_k x_1 & \Sigma h_k x_2 \ldots \Sigma h_k x_k \end{vmatrix}$$

Положим

$$c_M = d_1 h_1 + d_2 h_2 + \ldots + d_k h_k.$$

Балансир c_M — это не просто балансир, настроенный на $\hat{\beta}_M$ (каковым будет всякий балансир, пропорциональный ему), он — нечто большее, он — *уловитель* $\hat{\beta}_M$. Что это означает? Если мы уравновешиваем y и $\hat{y}$ с помощью этого балансира, то, очевидно, что

$$\Sigma c_M (i) y (i) = \hat{\beta}_M \cdot 1$$

и

$$\hat{\beta}_M = \Sigma c_M (i) y (i).$$

Отсюда, а также ввиду того, что c_M по построению зависит только от x, получаем

$$\text{var } \hat{\beta}_M = \Sigma \{c_M (i)\}^2 \text{ var } y (i) = \sigma^2 \Sigma \{c_M (i)\}^2.$$

Последнее равенство выполняется при условии, что все

$$\text{var } y (i) = \sigma^2.$$

Таким образом все, что можно извлечь из данных о $\hat{\beta}_M$ в процессе подгонки с помощью балансиров, заключено в наборах пар значений $\{y (i)\}$ и $\{c_M (i)\}$. Поэтому мы говорим, что c_M улавливает всю информацию о $\hat{\beta}_M$. Кроме всего прочего, знание уловителя позволяет свести получение $\hat{\beta}_M$ к задаче регрессии с одним параметром (в параграфе 14.5 мы покажем, что она сводится даже к еще более полезной однопараметрической регрессионной задаче.)

14.4. ОБЫЧНЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

В этом параграфе мы покажем, что если выбрать в качестве балансиров $x_1, x_2, \ldots, x_k$, то будет получено приближение метода наименьших квадратов — метода, минимизирующего $\Sigma (y - \hat{y})^2$.

Предположим, что мы приближаем

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

и что мы выбрали, пока более или менее наобум, в качестве балансиров $x_1, \ldots, x_k$. Это означает, что балансиром будет и

$$g_1 x_1 + g_2 x_2 + \ldots + g_k x_k$$

для любого набора $\{g_j\}$, не все элементы которого нули. Действительно, мы видели, что линейная комбинация балансиров есть балансир. (Мы используем теперь индекс j, так как i уже занят для обозначения i-го множества наблюдений.) Если $\hat{y}$ и $\hat{\hat{y}}$ — любые две модели такого вида, то каждая из них, а также $\hat{y} - \hat{\hat{y}}$ будут балансирами, как линейные комбинации балансиров. Напомним, что множество $\hat{\beta}_i$ задает модель.

Пусть теперь $\hat{y}$ — какая-то конкретная модель, полученная с помощью балансиров $x_1, \ldots, x_k$.

Тогда

$$y - \hat{\hat{y}} \equiv (\hat{y} - \hat{\hat{y}}) + (y - \hat{y})$$

и, следовательно,

$$\Sigma\,(y - \hat{\hat{y}})^2 \equiv \Sigma\,(\hat{y} - \hat{\hat{y}})^2 + 2\Sigma\,(\hat{y} - \hat{\hat{y}})\,(y - \hat{y}) + \Sigma\,(y - \hat{y})^2.$$

Опуская временно сомножитель 2, преобразуем среднее слагаемое в правой части последнего равенства следующим образом:

$$\Sigma\,(\hat{y} - \hat{\hat{y}})\,(y - \hat{y}) = \Sigma\,(\hat{y} - \hat{\hat{y}})y - \Sigma\,(\hat{y} - \hat{\hat{y}})\hat{y}.$$

Отсюда, поскольку $(\hat{y} - \hat{\hat{y}})$ — балансир для модели $\hat{y}$, становится ясно, что это слагаемое равно нулю. (Можно было бы взять $(\hat{y} - \hat{\hat{y}})$, чтобы найти $\hat{y}$ в терминах y.)

Следовательно,

$$\Sigma\,(y - \hat{\hat{y}})^2 = \Sigma\,(\hat{y} - \hat{\hat{y}})^2 + \Sigma\,(y - \hat{y})^2,$$

а так как первый член в правой части равенства неотрицателен, мы получаем

$$\Sigma\,(y - \hat{\hat{y}})^2 \geqslant \Sigma\,(y - \hat{y})^2.$$

Таким образом, $\hat{y}$ минимизирует сумму

$$\Sigma\,(\text{наблюденное} - \text{предсказанное})^2.$$

Отсюда термин «наименьшие квадраты» для модели

$$y = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k,$$

где в качестве балансира используются любые комбинации вида

$$g_1 x_1 + g_2 x_2 + \ldots + g_k x_k.$$

14.5. НАСТРОЙКА ДЛЯ ОБЫЧНОГО МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ

Если в множественной регрессии мы предсказываем x_i по остальным x, то дисперсия $\hat{\beta}_i$ будет велика, так как ее знаменатель — остаточная сумма квадратов $x_i - \hat{x}_i$, точнее говоря,

$$\operatorname{var}\hat{\beta}_i = \frac{\sigma^2}{\Sigma\,(x_i - \hat{x}_i)^2},$$

Мы хотим, чтобы детерминант получившейся системы k (= (k — —1) + 1) уравнений не был нулевым. Тогда система имела бы единственное решение. Значит, мы хотим, чтобы не вырождался детерминант

$$\begin{vmatrix} \Sigma h_1 x_1 & \Sigma h_1 x_2 \ldots \Sigma h_1 x_k \\ \Sigma h_2 x_1 & \Sigma h_2 x_2 \ldots \Sigma h_2 x_k \\ \vdots & \vdots \\ \Sigma h_k x_1 & \Sigma h_k x_2 \ldots \Sigma h_k x_k \end{vmatrix}$$

Положим

$$c_M = d_1 h_1 + d_2 h_2 + \ldots + d_k h_k.$$

Балансир c_M — это не просто балансир, настроенный на $\hat{\beta}_M$ (каковым будет всякий балансир, пропорциональный ему), он — нечто большее, он — *уловитель* $\hat{\beta}_M$. Что это означает? Если мы уравновешиваем y и $\hat{y}$ с помощью этого балансира, то, очевидно, что

$$\Sigma c_M (i) y (i) = \hat{\beta}_M \cdot 1$$

и

$$\hat{\beta}_M = \Sigma c_M (i) y (i).$$

Отсюда, а также ввиду того, что c_M по построению зависит только от x, получаем

$$\text{var } \hat{\beta}_M = \Sigma \{c_M (i)\}^2 \text{ var } y (i) = \sigma^2 \Sigma \{c_M (i)\}^2.$$

Последнее равенство выполняется при условии, что все

$$\text{var } y (i) = \sigma^2.$$

Таким образом все, что можно извлечь из данных о $\hat{\beta}_M$ в процессе подгонки с помощью балансиров, заключено в наборах пар значений $\{y (i)\}$ и $\{c_M (i)\}$. Поэтому мы говорим, что c_M улавливает всю информацию о $\hat{\beta}_M$. Кроме всего прочего, знание уловителя позволяет свести получение $\hat{\beta}_M$ к задаче регрессии с одним параметром (в параграфе 14.5 мы покажем, что она сводится даже к еще более полезной однопараметрической регрессионной задаче.)

14.4. ОБЫЧНЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

В этом параграфе мы покажем, что если выбрать в качестве балансиров $x_1, x_2, \ldots, x_k$, то будет получено приближение метода наименьших квадратов — метода, минимизирующего $\Sigma (y - \hat{y})^2$.

Предположим, что мы приближаем

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

и что мы выбрали, пока более или менее наобум, в качестве балансиров $x_1, \ldots, x_k$. Это означает, что балансиром будет и

$$g_1 x_1 + g_2 x_2 + \ldots + g_k x_k$$

для любого набора $\{g_j\}$, не все элементы которого нули. Действительно, мы видели, что линейная комбинация балансиров есть балансир. (Мы используем теперь индекс j, так как i уже занят для обозначения i-го множества наблюдений.) Если $\hat{y}$ и $\hat{\hat{y}}$ — любые две модели такого вида, то каждая из них, а также $\hat{y} - \hat{\hat{y}}$ будут балансирами, как линейные комбинации балансиров. Напомним, что множество $\hat{\beta}_i$ задает модель.

Пусть теперь $\hat{y}$ — какая-то конкретная модель, полученная с помощью балансиров $x_1, \ldots, x_k$.

Тогда

$$y - \hat{\hat{y}} \equiv (\hat{y} - \hat{\hat{y}}) + (y - \hat{y})$$

и, следовательно,

$$\Sigma\,(y - \hat{\hat{y}})^2 \equiv \Sigma\,(\hat{y} - \hat{\hat{y}})^2 + 2\Sigma\,(\hat{y} - \hat{\hat{y}})\,(y - \hat{y}) + \Sigma\,(y - \hat{y})^2.$$

Опуская временно сомножитель 2, преобразуем среднее слагаемое в правой части последнего равенства следующим образом:

$$\Sigma\,(\hat{y} - \hat{\hat{y}})\,(y - \hat{y}) = \Sigma\,(\hat{y} - \hat{\hat{y}})y - \Sigma\,(\hat{y} - \hat{\hat{y}})\hat{y}.$$

Отсюда, поскольку $(\hat{y} - \hat{\hat{y}})$ — балансир для модели $\hat{y}$, становится ясно, что это слагаемое равно нулю. (Можно было бы взять $(\hat{y} - \hat{\hat{y}})$, чтобы найти $\hat{y}$ в терминах y.)

Следовательно,

$$\Sigma\,(y - \hat{\hat{y}})^2 = \Sigma\,(\hat{y} - \hat{\hat{y}})^2 + \Sigma\,(y - \hat{y})^2,$$

а так как первый член в правой части равенства неотрицателен, мы получаем

$$\Sigma\,(y - \hat{\hat{y}})^2 \geqslant \Sigma\,(y - \hat{y})^2.$$

Таким образом, $\hat{y}$ минимизирует сумму

$$\Sigma\,(\text{наблюденное} - \text{предсказанное})^2.$$

Отсюда термин «наименьшие квадраты» для модели

$$y = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k,$$

где в качестве балансира используются любые комбинации вида

$$g_1 x_1 + g_2 x_2 + \ldots + g_k x_k.$$

14.5. НАСТРОЙКА ДЛЯ ОБЫЧНОГО МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ

Если в множественной регрессии мы предсказываем x_i по остальным x, то дисперсия $\hat{\beta}_i$ будет велика, так как ее знаменатель — остаточная сумма квадратов $x_i - \hat{x}_i$, точнее говоря,

$$\operatorname{var}\hat{\beta}_i = \frac{\sigma^2}{\Sigma\,(x_i - \hat{x}_i)^2},$$

где $\hat{x}_i$ — оценка метода наименьших квадратов для x_i, основанная на остальных x, а σ^2 — общая дисперсия y. Используем метод балансировки для доказательства этого факта. Мы продемонстрируем также, что информация о $\hat{\beta}_i$ содержится в комбинации а) остатков y от корректировки по подгенератору x_i, б) остатков x_i после корректировки по его подгенератору.

Оценивание x_i. Если мы получаем обычное уравнение метода наименьших квадратов вида

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k,$$

используя $x_1, \ldots, x_k$ и все их линейные комбинации как балансиры, то, полагая $z = x_1$, мы можем точно так же получить и уравнение

$$\hat{z} = \hat{\gamma}_{12} x_2 + \hat{\gamma}_{13} x_3 + \ldots + \hat{\gamma}_{1k} x_k.$$

Используя в качестве балансиров этой модели $x_2, x_3, \ldots, x_k$, получим

$$\Sigma x_J (i) z (i) = \Sigma x_J (i) \hat{z} (i), \text{ для } J = 2, 3, \ldots, k,$$

что эквивалентно

$$\Sigma x_J (i) (z (i) - \hat{z}(i)) = 0, \text{ для } J = 2, 3, \ldots, k,$$

или

$$\Sigma x_J (i) (x_1 (i) - \hat{\gamma}_{12} x_2 (i) - \hat{\gamma}_{13} x_3 (i) - \ldots - \hat{\gamma}_{1k} x_k (i)) = 0,$$

для $J = 2, 3, \ldots, k$,

Рассмотрим теперь остаток x_1 после приближения по $x_2, \ldots, x_k$, а именно

$$x_{1 \cdot 23 \ldots k} = x_1 - \hat{\gamma}_{12} x_2 - \hat{\gamma}_{13} x_3 - \ldots - \hat{\gamma}_{1k} x_k = x_1 - \hat{x}_1.$$

(В факторном анализе $x_{1 \cdot 23 \ldots k}$ иногда называют зеркальным отражением x_1 в $x_1, x_2, \ldots, x_k$, а сумму $\hat{\gamma}_{1j} x_j$ — просто отражением — весьма удобное мнемоническое правило.) Введенные нами $(k - 1)$ условия — все, чего не хватало для доказательства того, что

$$x_{1 \cdot 23 \ldots k} = x_1 - \hat{\gamma}_{12} x_2 - \hat{\gamma}_{13} x_3 - \ldots - \hat{\gamma}_{1k} x_k$$

настроено на $\hat{\beta}_1$. Действительно, умножая равенство $y = \Sigma_j \beta_j x_j$ на $x_{1 \cdot 23 \ldots k}$ и суммируя по i, мы видим, что $\Sigma x_j x_{1 \cdot 23 \ldots k}$ (для $j \neq i$) обращается в нуль. Таким образом, остается

$$\Sigma y x_{1 \cdot 23 \ldots k} = \widehat{\beta_1} \Sigma x_1 x_{1 \cdot 23 \ldots k},$$

т. е. уравнение, настроенное на $\hat{\beta}_1$.

Откуда возьмутся $\hat{\beta}_1$? Так как

$$\widehat{\beta_1} = \frac{\Sigma x_{1 \cdot 23 \ldots k} y}{\Sigma x_{1 \cdot 23 \ldots k} x_1}$$

и

$$x_1 = x_{1 \cdot 23 \ldots k} + \widehat{\gamma_{12}} x_2 + \ldots + \widehat{\gamma_{1k}} x_k,$$

то

$$\Sigma x_{1 \cdot 23 \ldots k} x_1 = \Sigma x_{1 \cdot 23 \ldots k} x_{1 \cdot 23 \ldots k}.$$

Аналогично если мы подгоняем $\delta_2 x_2 + \ldots + \delta_k x_k$ к y и находим остаток $y_{\cdot 23\ldots k}$, то

$$y = y_{\cdot 23\ldots k} + \hat{\delta}_2 x_2 + \ldots + \hat{\delta}_k x_k,$$

так что

$$\Sigma x_{1\cdot 23\ldots k}\, y = \Sigma x_{1\cdot 23,\ldots k}\, y_{\cdot 23\ldots k}.$$

Таким образом,

$$\hat{\beta}_1 = \frac{\Sigma x_{1\cdot 23\ldots k}\, y_{\cdot 23\ldots k}}{\Sigma (x_{1\cdot 23\ldots k})^2}.$$

Следовательно, $\hat{\beta}_1$ можно найти из однопараметрического регрессионного уравнения, связывающего n пар значений

$$y_{\cdot 23\ldots k} \text{ и } x_{1\cdot 23\ldots k}.$$

График зависимости

$$y_{\cdot 23\ldots k} \text{ от } x_{1\cdot 23\ldots k}$$

скорее всего окажется гораздо более полезным, чем графики зависимостей y от $x_{1\cdot 23\ldots k}$ или y от уловителя для $\hat{\beta}_1$, так как $y_{\cdot 23\ldots k}$, по-видимому, будет гораздо меньше по величине, чем y. (Графики последних двух зависимостей будут похожими, так как $\hat{\beta}_1$ должен быть кратным $x_{1\cdot 23\ldots k}$. Поэтому все равно, каким из них пользоваться.)

Может показаться, что вычисление $y_{\cdot 23\ldots k}$, потом еще $y_{\cdot 13\ldots k}$ (без x_2) и только затем остатка y от модели без x_k — слишком громоздкое дело. Но и в нем нет необходимости, так как β_1 в

$$\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

и

$$\beta_1 x_{1\cdot 23\ldots k} + \beta_2^* x_2 + \ldots + \beta_k^* x_k$$

одинаковы, что легко проверить подстановкой $x_{1\cdot 23\ldots k}$.
Это означает, что

$$y_{\cdot 123\ldots k} = y_{\cdot 23\ldots k} - \hat{\beta}_1 x_{1\cdot 23\ldots k}.$$

Таким образом, график

$$y_{\cdot 23\ldots k} \text{ по } x_{1\cdot 23\ldots k}$$

аналогичен графику

$$y_{\cdot 123\ldots k} - \hat{\beta}_1 x_{1\cdot 23\ldots k} \text{ по } x_{1\cdot 23\ldots k}.$$

Определение последней зависимости достаточно просто, когда известны остатки x и y: $y_{\cdot 123\ldots k}$ и $x_{1\cdot 23\ldots k}$ соответственно. Из этой однопараметрической регрессии мы не только получаем максимум информации о $\hat{\beta}_1$ в модели $\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$, но и информацию о вкладах, которые вносят разные множества данных в найденную оценку (см. [Larsen W. A. and McCleary S. J. (1972)]).

Уменьшение остатков. Так как обычный метод наименьших квадратов минимизирует

$$\Sigma\,(x_1 - \hat{x}_1)^2,$$

то

$$\Sigma\,(x_{1\cdot 23\ldots k})^2 \leqq \Sigma\,(x_{1\cdot \text{меньше}})^2 \leqq \Sigma x_1^2,$$

где «меньшее» относится к любому подмножеству переменных 2, 3, ..., k.

Во многих задачах подгонки возникают трудности, если

$$\Sigma\,(x_{1\cdot 23\ldots k})^2 \ll \Sigma x_1^2,$$

где «$\ll$» означает «много меньше».
Так, если

$$x_1 = 1,\ x_2 = x,\ x_3 = x^2,\ x_4 = x^3,\ x_5 = x^4$$

для $x = 1, 2, 3, \ldots, 10$,
то

$$\frac{\Sigma\,(x_{1\cdot 2345})^2}{\Sigma x_1^2} = 9{,}16 \times 10^{-3},$$

$$\frac{\Sigma\,(x_{2\cdot 1345})^2}{\Sigma x_2^2} = 1{,}87 \times 10^{-4},$$

$$\frac{\Sigma\,(x_{3\cdot 1245})^2}{\Sigma x_3^2} = 2{,}38 \times 10^{-5},$$

$$\frac{\Sigma\,(x_{4\cdot 1235})^2}{\Sigma x_4^2} = 1{,}70 \times 10^{-5},$$

$$\frac{\Sigma\,(x_{5\cdot 1234})^2}{\Sigma x_5^2} = 9{,}82 \times 10^{-6}.$$

Слишком много десятичных знаков, которые, казалось, уже ухвачены, текут сквозь пальцы.

Напомним, что ввиду

$$\Sigma x_{1\cdot 23\ldots k}(i)\, y(i) = \widehat{\beta}_1 \Sigma x_{1\cdot 23\ldots k}(i)\, x_1(i) = \widehat{\beta}_1 \Sigma\,(x_{1\cdot 23\ldots k}(i))^2$$

дисперсия $\hat{\beta}_1$ равна:

$$\operatorname{var}\{\widehat{\beta}_i\} = \frac{\Sigma x_{1\cdot 23\ldots k}^2(i)\operatorname{var}\{y(i)\}}{(\Sigma\,(x_{1\cdot 23\ldots k}(i))^2)^2} = \frac{\sigma^2}{\Sigma\,(x_{1\cdot 23\ldots k}(i))^2};$$

последнее — при условии, что все дисперсии y равны σ^2.

Шаг вперед или шаг назад? Отметим, что при преобразовании

$$\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

к виду

$$\widehat{\beta_1}\, x_{1\cdot 23\ldots k} + \widehat{\beta_2^*}\, x_2 + \ldots + \widehat{\beta_k^*}\, x_k$$

β_i^* совпадают с оценками, которые мы могли бы получить, подгоняя

$$\widehat{\beta}_2^* x_2 + \ldots + \widehat{\beta}_k^* x_k$$

без x_1. Это вытекает из (1) $x_2, x_3, \ldots, x_k$ — балансиры для обоих приближений и (2) они отстраивают $x_{1 \cdot 2 \ldots k}$.

14.6. МЕТОД ВЗВЕШЕННЫХ НАИМЕНЬШИХ КВАДРАТОВ

В топографии и астрономии, откуда и берет начало метод наименьших квадратов, давно признано,что одни наблюдения бывают «лучше», «сильнее» других и это требует принятия соответствующих мер. К таким мерам относится назначение разных весов разным наблюдениям. Это назначение может носить как объективный, так и субъективный характер. Вообще надо сказать, что история взвешенных наименьших квадратов почти столь же обширна, как и история обычного метода. Сейчас существуют новые применения метода взвешенных наименьших квадратов и новые причины для его изучения (см. параграф 14.8).

Предположим, что мы свободны выбирать практически любой набор неотрицательных весов $w = \{w(i)\}$ для каждой точки из множества данных

$$\{x_1(i), x_2(i), \ldots, x_k(i)\}.$$

Выберем балансиры вида $wx_1, wx_2, \ldots, wx_k$. Например,

$$w(1)x_2(1),\ w_2(2)\ x_2(2),\ w(3)x_2(3), \ldots, w(n)x_2(n).$$

Обозначим, как и раньше, $\hat{y}$ и $\hat{\hat{y}}$ — две модели вида $c_1x_1 + c_2x_2 + \ldots + c_kx_k$. Тогда $w\hat{y}$, $w\hat{\hat{y}}$ и $w(\hat{y} - \hat{\hat{y}})$ также будут балансирами, как линейные комбинации wx_i. Пусть теперь $\hat{y}$ — приближение, полученное методом балансировки с помощью балансиров $wx_1, wx_2, \ldots, wx_k$.

Очевидно, что

$$(y - \hat{\hat{y}}) \equiv (y - \hat{y}) + (\hat{y} - \hat{\hat{y}})$$

и

$$\Sigma w (y - \hat{\hat{y}})^2 \equiv \Sigma w (y - \hat{y})^2 + 2\Sigma w (\hat{y} - \hat{\hat{y}})(y - \hat{y}) + \Sigma w (\hat{y} - \hat{\hat{y}})^2.$$

Как и в параграфе 14.4, можно показать, что средний член — нуль. Из рассуждений, аналогичных тем, что приведены в конце параграфа 14.4, следует, что ввиду

$$\Sigma w (y - \hat{\hat{y}})^2 \geqslant \Sigma w (y - \hat{y})^2$$

$\hat{y}$ минимизирует сумму

$$\Sigma w \text{ (наблюденное} - \text{предсказанное)}^2.$$

Обобщение. Мы рассмотрели взвешенный и невзвешенный (т. е. с равными весами) методы отдельно. Это было сделано для того, чтобы более глубоко осмыслить идею и технику применения весов.

Что касается чисто математической стороны дела, то взвешенный случай всегда сводится к невзвешенному.

Пусть мы строим

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

с весом w. Тогда для любой линейной комбинации $x\,(i)x_j\,(i)$ балансировочное уравнение есть

$$\Sigma w\,(i)x\,(i)y\,(i) = \Sigma w\,(i)x\,(i)\hat{y}\,(i).$$

Его можно переписать в виде

$$\Sigma\,[\sqrt{w(i)}\,x\,(i)]\,[\sqrt{w(i)}\,y(i)] = \Sigma\,[\sqrt{w(i)}\,x\,(i)]\,[\sqrt{w(i)}\,\widehat{y}\,(i)],$$

соответствующем модели

$$(\sqrt{w}\,y) = \beta_1(\sqrt{w}\,x_1) + \beta_2(\sqrt{w}\,x_2) + \ldots + \beta_k(\sqrt{w}\,x_k) \qquad (***)$$

с балансирами

$\sqrt{w}x = \sqrt{w}$ (любая линейная комбинация $x_1, x_2, \ldots, x_k$) = (любая линейная комбинация $\sqrt{w}x_1, \sqrt{w}x_2, \ldots, \sqrt{w}x_k$).

Отсюда видно, что модель (***) имеет равные веса.

Отметим, в частности, что

$$\Sigma wyx_{1\cdot 23\ldots k} = \widehat{\beta_1}\,\Sigma w\,(x_{1\cdot 23\ldots k})^2,$$

где $x_{1\cdot 23\ldots k}$ — остаток, соответствующий различным весам w.

Так как последнее равенство можно переписать в виде

$$\Sigma\sqrt{w}\,y\sqrt{w}\,x_{1\cdot 23\ldots k} = \widehat{\beta_1}\,\Sigma\,(\sqrt{w}\,x_{1\cdot 23\ldots k})^2,$$

то и формуле для $\hat{\beta}_1$ можно придать вид

$$\widehat{\beta_1} = \frac{\Sigma wyx_{1\cdot 23\ldots k}}{\Sigma w\,(x_{1\cdot 23\ldots k})^2}$$

так же, как и вид

$$\widehat{\beta_1} = \frac{\Sigma\,(\sqrt{w}\,y)\,(\sqrt{w}\,x_{1\cdot 23\ldots k})}{\Sigma\,(\sqrt{w}\,x_{1\cdot 23\ldots k})^2}.$$

Таким образом, при наличии весов график зависимости

$$\sqrt{w}\,y \text{ от } \sqrt{w}\,x_{1\cdot 23\ldots k}$$

или еще лучше $\sqrt{w}y_{\cdot 23\ldots k}$ от $\sqrt{w}x_{1\cdot 23\ldots k}$ должен полностью описывать картину.

Настройка для взвешенных наименьших квадратов. Используя переход к равным весам, мы видим: если вес w выбран так, что

$$w \operatorname{var}\{y\} = \sigma^2,$$

а значит, $\operatorname{var}\{\sqrt{w}y\} = \sigma^2$,

то

$$\operatorname{var}\widehat{\beta_1} = \frac{\sigma^2}{\Sigma w\,(x_{1\cdot 23\ldots k})^2}$$

для взвешенных наименьших квадратов, где $x_{1\cdot 23\ldots k}$ — остаток для модели x_1 по $x_2, x_3, \ldots, x_k$ с весом w.

ЗАМЕЧАНИЕ

В тех случаях, когда среди массы наблюдений с примерно одинаковым разбросом встречается несколько имеющих значительно больший разброс, мы часто ставим в соответствие этим последним малые веса, а для остальных выбираем равные и значительно большие. Такая тактика эквивалентна «уничтожению» информации, которой на самом деле и нет.

Предположим, например, что у нас из 100 откликов три имеют очень большой разброс, а остальные — значительно меньший и примерно равный. Если мы для всех переменных выберем одинаковые веса, то это приведет к результату с весьма большой дисперсией, построенному по 100 наблюдениям. Если мы *отбросим* три наблюдения с большой дисперсией (т. е. придадим им нулевые веса), то получим, повидимому, хорошие результаты, основанные на 97 наблюдениях. Если мы используем для наблюдений с большой дисперсией маленькие, но не нулевые веса, мы, быть может, добьемся даже лучших результатов, отвечающих, скажем, 97,13 наблюдениям.

Три рассмотренные операции можно было бы описать так:

Эквивалентное число наблюдений с равными дисперсиями[1]

Веса для трех плохих значений	Предполагаемое	Фактическое
1	100	Немного (3, 4, может быть, 5)
0	97	97
0,041	97,13	97,13

[1] Предполагается, что третий набор весов оптимален.

Мы не можем получить эквивалент 100 одинаково хороших наблюдений. Всякая попытка использовать $99 = 100 - 1$ степеней свободы для определения σ^2 ведет к заблуждениям, если мы знаем, что три наблюдения из 100 , имеющихся у нас, дают значительно большую, чем остальные, дисперсию. Можно признать одинаково законным использование $96 = 97 - 1$ (если мы отбрасываем 3 наблюдения) или $96{,}13 = 97{,}13 - 1$ (если мы придаем им малые веса) степеней свободы. Но если мы не удалим каким-то образом эту троицу, то можем оказаться в весьма тяжелой ситуации. Действительно, если они имеют достаточно большую дисперсию, то ценность нашего σ^2 может оказаться соответствующей весьма малому числу степеней свободы, например 3, и остальные оценки будут очень плохими. Использование больших весов там, где должны быть маленькие, может нас очень сильно подвести, так как мы получим плохие результаты, которые будут казаться более надежными, чем они есть (а часто и могут быть в принципе).

* ЕЩЕ ОБОБЩЕНИЕ (НЕОБЯЗАТЕЛЬНОЕ)

Быть может, нам захочется еще большего обобщения. Если есть таблица с двумя входами, состоящая из «весов» $\{w_{ij}\}$, то как с ее помощью можно было бы производить подгонку? Надо минимизировать

$$\sum_i \sum_j w_{ij} (y(i) - \widehat{y}(i)) (y(j) - \widehat{y}(j)). \qquad (*)$$

Без потери общности можно предполагать, что набор весов обладает свойством симметрии. Действительно, можно заменить w_{ij} на

$$w'_{ij} = \frac{1}{2} (w_{ij} + w_{ji}).$$

Это создает нужную нам симметрию, не изменяя двойной (общей) суммы.

Условившись, что $w_{ij} = w_{ji}$, возьмем

$$h_1(i) = \sum_j w_{ij} x_1(j),$$

$$h_2(i) = \sum_j w_{ij} x_2(j).$$

$$\vdots$$

$$h_k(i) = \sum_j w_{ij} x_k(j)$$

в качестве балансиров, определяющих $\widehat{y}$. Пусть далее $\widehat{\widehat{y}}$ — другая модель того же вида

$$b_1 x_1 + b_2 x_2 + \ldots + b_k x_k,$$

т. е. модель с теми же носителями, но с другими β. Тогда

$$\widehat{h}(i) = \sum_j w_{ij} \widehat{y}(j)$$

и

$$\widehat{\widehat{h}}(i) = \sum_j w_{ij} \widehat{\widehat{y}}(j)$$

— снова балансиры так же, как и их разность

$$h'(i) = \sum_j w_{ij} [\widehat{y}(j) - \widehat{\widehat{y}}(j)].$$

Мы хотим показать, что $\widehat{y}$ минимизируют выражение (*). С этой целью рассмотрим

$$\sum_i \sum_j w_{ij} [y(j) - \widehat{\widehat{y}}(j)] [y(i) - \widehat{\widehat{y}}(i)]. \qquad (**)$$

Как это уже делалось раньше, запишем

$$y - \widehat{\widehat{y}} \equiv (\widehat{y} - \widehat{\widehat{y}}) + (y - \widehat{y})$$

и подставим это выражение в (**):

$$\sum_i \sum_j w_{ij} [\widehat{y}(j) - \widehat{\widehat{y}}(j)] [\widehat{y}(i) - \widehat{\widehat{y}}(i)] + \sum_i \sum_j w_{ij} [\widehat{y}(j) - \widehat{\widehat{y}}(j)] [y(i) - \widehat{y}(i)] +$$

$$+ \sum_i \sum_j w_{ij} [y(j) - \widehat{y}(j)] [\widehat{y}(i) - \widehat{\widehat{y}}(i)] +$$

$$+ \sum_i \sum_j w_{ij} [y(j) - \widehat{y}(j)] [y(i) - \widehat{y}(i)]. \qquad (***)$$

Во втором слагаемом в (***)

$$\Sigma w_{ij} [\widehat{y}(j) - \widehat{\widehat{y}}(j)]$$

будет балансиром для $\hat{y}$, следовательно, это слагаемое обращается в нуль. Ввиду симметрии весов то же происходит и с третьим слагаемым.

Нам потребуется специальное условие, гарантирующее неотрицательность исходного выражения (*). Когда мы имели дело с суммами квадратов, неотрицательность подавалась нам «на блюдечке с голубой каемочкой», но взвешенные суммы парных произведений могут быть и отрицательными. Мы должны потребовать, чтобы

$$\sum_i \sum_j w_{ij} u_i u_j \geqslant 0 \qquad (****)$$

для всех $\{u_i\}$. (Говоря на матричном языке, мы должны потребовать положительной полуопределенности матрицы весов.) Честно говоря, это условие не так легко проверить. Однако если w есть элементы матрицы, обратной к матрице дисперсий-ковариаций, то это условие выполняется автоматически. В общем, мы будем предполагать, что w обладают свойством неотрицательной определенности (****).

Тогда первая строка в (***) должна быть $\geqslant 0$. Таким образом, мы можем написать

(**) = первая строка (***) + последняя строка (***).

Условие неотрицательной определенности гарантирует неотрицательность обеих частей, но сейчас нам важно подчеркнуть, что первая строка в (***) $\geqslant 0$. Заменяя ее нулем, получим

$$(**) \geqslant \text{последняя строка } (***),$$

или, более подробно,

$$\sum_i \sum_j w_{ij} [y(j) - \widehat{\widehat{y}}(j)] [y(i) - \widehat{\widehat{y}}(i)] \geqslant \sum_i \sum_j w_{ij} [y(j) - \widehat{y}(j)] \times$$

$$\times [y(i) - \widehat{y}(i)].$$

Таким образом, мы получили желаемое обобщение результата параграфа 14.4: если таблица с двумя входами для весов $\{w_{ij}\}$ задает не-

отрицательно-определенную квадратичную форму (условие (****), то балансиры

$$h_m(i) = \sum w_{ij} x_m(j), \quad m = 1, 2, \ldots, k,$$

задают оценку $\hat{y}$, которая соответствует дважды обобщенной оценке метода наименьших квадратов для любых моделей вида

$$\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k.$$

14.7. КРИВЫЕ ВЛИЯНИЯ ДЛЯ МЕР ПОЛОЖЕНИЯ

Попробуем представить себе, как меняются некоторые статистики при изменении набора данных. Предположим, например, что одно значение в выборке пробегает непрерывно весь спектр значений. Для удобства рассмотрим выборку из 11 значений, десять из которых

$$10, 7, 3, 3, 3, -2, -5, -5, -6, -8,$$

в сумме дают 0, а 11-е изменяется от самого меньшего до самого большего.

Этот пример, несмотря на его чисто модельный характер, должен дать общее представление о *кривых влияния* — кривых, которые отражают влияние изменения какого-нибудь x на те или иные статистики.

1. СРЕДНЕЕ, $\bar{x}$

Ввиду того что сумма десяти данных — нуль, общая сумма должна совпадать с x.

Следовательно, выборочное среднее

$$\bar{x} \equiv \frac{\Sigma x_i}{n} = \frac{x}{n} = \frac{x}{11}.$$

Верхний график на илл. 14.7.1 — прямая линия с наклоном (в данном случае 1/11 показывает, как зависит $\bar{x}$ от x).

2. МЕДИАНА, x

Ввиду нечетности числа элементов выборки медианой служит средний ее элемент, в данном случае — шестой с любого края.

Медиана равняется

-2, если $x \leqslant -2$;
x, если $-2 \leqslant x \leqslant 3$;
3, если $3 \leqslant x$.

Таким образом, кривая влияния (изображенная на среднем графике илл. 14.7.1) состоит из трех линейных отрезков, два из которых параллельны горизонтальной оси, а третий идет под углом с коэффициентом наклона 1, соединяя два первых. Расстояние между параллельными отрезками равно разности между средними элементами базисной выборки из десяти элементов.

3. БИВЕС-ОЦЕНКА, $\hat{x}$

Бивес — это аббревиатура выражения «биквадратный вес».

Давайте взвешивать наблюдения, используя веса

$$w(u)=\begin{cases}(1-u^2)^2 & |u|\leqslant 1,\\ 0 & |u|>1,\end{cases}$$

где

$$u_i=\frac{x_i-\hat{x}}{cS},$$

причем положим c равным 6 (более или менее произвольно), а $S=\frac{1}{2}\,x\times$ $\times$ (интерквартильный размах). Для простоты примем за интерквартильный размах разность между третьими с каждого края элементами выборки. Для распределений, близких к нормальному, средний интерквартильный размах близок к $\frac{4}{3}\sigma$, а, следовательно, 6σ в среднем близко к 4σ; эти соотношения пригодятся нам для прикидок. Определим оценку соотношением

$$\hat{x}=\frac{\Sigma_{i=1}^{n} w(u_i)\,x_i}{\Sigma w(u_i)}.$$

Рассмотрим значения S. Для вычисления интерквартильного размаха найдем расстояние (I) между третьими с краев элементами выборки. Оно равно для x:

$$\begin{array}{lll}
-\infty\leqslant x\leqslant -6, & I=9, & S=4{,}5;\\
-6\leqslant x\leqslant -5, & I=3-x, & 4{,}0\leqslant S\leqslant 4{,}5,\ S=\dfrac{3-x}{2};\\
-5\leqslant x\leqslant 3, & I=8, & S=4{,}0;\\
3\leqslant x\leqslant 7, & I=x+5, & 4\leqslant S\leqslant 6,\ S=\dfrac{x+5}{2};\\
7\leqslant x\leqslant +\infty, & I=12, & S=6.
\end{array}$$

Учитывая, что $\hat{x}$ зависит от весов, а веса зависят от $\hat{x}$, придется воспользоваться итерацией. В таблице илл. 14.7.2 приведены вычисления, используемые для построения кривой, изображенной на нижнем графике илл. 14.7.1.

Заметим, что поведение кривой влияния на этом графике как раз такое, какое нам часто хотелось бы иметь. Если изменяющееся значение x становится слишком большим или слишком маленьким, его влияние приближается к нулю, и наши результаты основываются на оставшихся 10 измерениях. Значит, влияние x приближается к нулю при $x<-27$, $x\geqslant 36$, а также, конечно, при $x\approx 0$.

Таким образом, по кривой влияния мы видим, что

- среднее может уходить до $+\infty$ или $-\infty$, если есть хоть одно достаточно плохое измерение;

● медиана не изменяется при изменении одного измерения, если только оно не попадает в довольно узкий интервал;

● бивес-оценка среднего аналогично медиане не меняется при изменении данных вне некоторого довольно широкого интервала и, наоборот, чувствительна к их поведению внутри этого интервала.

Чтобы выяснить, как может отразиться небольшое изменение масштаба кривой влияния, заменим S в u на медиану абсолютного отклонения от медианы. Полученный результат изображен на илл.14.7.3. Он весьма близок к графику бивеса на илл. 14.7.1.

Некоторые подробности читатель найдет в работе [Hampel E. E. (1974)] — см. список литературы в конце главы.

14.8. ИТЕРАТИВНЫЙ ЛИНЕЙНЫЙ МЕТОД ВЗВЕШЕННЫХ НАИМЕНЬШИХ КВАДРАТОВ

Обозначения. Как и раньше, будем использовать обозначение $\hat{y}$ для любых оценок, которые получаются обычным методом наименьших квадратов, с помощью взвешенных наименьших квадратов или обобщения метода наименьших квадратов, обсуждавшегося в параграфе 14.6. Мы распространим это правило и на последовательности приближений, получающихся методом наименьших квадратов, но с разными весами.

На каждом шаге мы будем использовать обозначения $\hat{y}$, $\hat{\beta}_j$ для оценок метода наименьших квадратов, сохраняя двойные шапки $\hat{\hat{y}}$, $\hat{\hat{\beta}}_i$ — для оценок, полученных другими методами.

Обозначения y^*, β_j^* и т. д. введем для окончательных результатов, полученных в итеративном методе наименьших квадратов. Для обозначения остатка последнего приближения или промежуточного шага в процедуре приближения будем использовать e_i.

Мы уже отмечали в начале главы, говоря о взвешенных наименьших квадратах, что веса могут и должны отражать «качество» отдельных наблюдений. В этом параграфе мы рассмотрим другой аспект применения весов, а именно для обеспечения нужных нам кривых влияния, которые в свою очередь обеспечивают высокое качество оценок.

Такое использование весов отличается от их стандартного применения как технически, так и с точки зрения интерпретации, но цель остается той же — получение наилучших оценок. Мы можем в процессе анализа применять оба метода одновременно, при этом нам, может быть, придется умножать вес одного типа на вес другого, чтобы получить «вес», входящий в окончательный результат.

Предположим для начала, что все веса первого типа одинаковы. Мы сознаем как достоинства, так и недостатки приближения методом наименьших квадратов. Этот метод легко описать в терминах балансиров, он сводится к решению простой системы уравнений, хорошо изучен и обеспечен множеством программ для ЭВМ (не все они идеальны, но многие достаточно надежны). При использовании этого метода мы можем настраивать некоторые балансиры, что существенно упрощает ситуацию. Все это очень хорошо. Но, с другой стороны, следует отме-

тить, что метод наименьших квадратов не слишком гибок, хотя здесь и может помочь использование *априорных* весов. Как показывают наши примеры, метод наименьших квадратов весьма чувствителен к серьезным возмущениям наблюдений. Можно ли сохранить преимущества метода и избавиться от его недостатков?

Предположим, что мы

- выбрали $w_1 = \{w_1(i)\}$ и с этими весами построили приближение $\hat{y}^{(1)}$;
- затем выбрали веса $w_2 = \{w_2(i)\}$, опираясь, быть может, на результаты предыдущего приближения, и построили приближение $\hat{y}^{(2)}$;
- затем выбрали w_3 ... и т.д.

Можно либо делать заданное число шагов, либо продолжать до тех пор, пока приближения начнут мало отличаться друг от друга. (Можно подумать и о переходе к пределу, но на практике мы вряд ли захотим израсходовать ресурс нашего карандаша или ЭВМ, не говоря уже о заказчике.)

Реально вопрос состоит, конечно, в том, как выбирать последовательные веса. Первое множество приходится выбирать из общих соображений. Здесь мы предположим, что все веса равны.

Очень полезно выбирать следующий вес так, чтобы i-й вес зависел от отношения i-го остатка в предыдущей итерации к некоторой общей мере остатков в этой итерации, например от

$$u_i = \frac{y(i) - \hat{y}^{(k)}(i)}{cS_k},$$

где c — некоторая константа; S_k — мера разброса остатков в k-й итерации; $\hat{y}^{(k)}(i)$ — оценка $y(i)$ в k-й итерации.

Бивзвешивание. Как выбирать $w(i)$ в зависимости от u_i? Как всегда, здесь существует много хороших возможностей. Если есть компьютер, то удобно брать биквадратное взвешивание, или сокращенно — бивзвешивание, а именно:

$$w(u) = (1 - u^2)^2, \quad u^2 < 1;$$
$$= 0, \quad u^2 \geqslant 1.$$

Общая идея состоит в том, чтобы малым отклонениям придать большие веса, а по мере роста отклонений — веса последовательно уменьшать, пока, наконец, при отклонениях, бо́льших единицы, они не сойдут постепенно на нет.

Мы часто берем

$$S_k = \text{медиана } |y(i) - \hat{y}^k(i)|,$$
$$c = 6 \text{ или } 9,$$

хотя, конечно, есть сколько угодно других столь же хороших вариантов. Отметим, что S_k — медиана абсолютных значений остатков в k-й итерации. Таким образом, S_k измеряет вариацию относительно центра множества измерений, а u_i — отклонения в долях этой меры. Константа c определяет, при каких отклонениях вес должен обратиться

в 0. Если $c = 6$, то вес обращается в нуль при отклонениях, больших, чем $6S_k$, от последней оценки центра.

Пример. Резко выделяющееся наблюдение. Лишь простейшие примеры поддаются ручному счету. Поэтому начнем с подгонки к экспериментальным данным прямой, проходящей через начало координат. На илл. 14.8.1 нанесены пять экспериментальных точек. В таблице илл. 14.8.2 приведены вычисления для приближения с весом $w(u) = (1 - u^2)^2$. Так как 5-я точка далеко отклоняется от прямой, к которой близки 4 оставшиеся, задача подгонки — уточнение $\hat{\beta}^{(1)} = 1{,}13$, отвечающее малому весу для этой точки. Здесь потребуется сделать как минимум два шага, после чего последовательные оценки будут $\hat{\beta}^{(4)} = 1{,}02$, $\hat{\beta}^{(5)} = 1{,}00$ и $\hat{\beta}^{(6)} = 0{,}99$. Если читатель сделает еще один шаг, то он убедится, что $\hat{\beta}^{(7)}$ снова равно 0,99. Тем самым процесс итераций с точностью до двух десятичных знаков замкнулся.

Потребовавшееся число итераций больше, чем хотелось бы. Нельзя ли решить проблему попроще? В данном случае — да. Изучая илл. 14.8.1, можно прийти к выводу, что ввиду сильного отклонения пятой точки ей следует придать нулевой вес, а остальным — веса, равные 1.

Выбор начального нулевого веса для какой-то точки в принципе не исключает ее из рассмотрения. В процессе последующих итераций она может приобрести ненулевой вес.

Поступив так, как говорилось выше, мы сразу получим $\hat{\beta}^{(1)} = 29/30 = 0{,}987$. Эту оценку вряд ли можно изменить в ходе следующих итераций. В более общих ситуациях при более сложных зависимостях трудно предсказать, какая точка должна получить малый вес.

Шаговое взвешивание. При ручной обработке важно сократить счет. (Вообще ручная обработка мыслима лишь в простейших ситуациях.) Для этих целей мы предлагаем следующий набор весов:

$$w(u) = \begin{cases} 4 & |u| \leqslant 0{,}2, \\ 3 & 0{,}2 < |u| \leqslant 0{,}4, \\ 2 & 0{,}4 < |u| \leqslant 0{,}6, \\ 1 & 0{,}6 < |u| \leqslant 0{,}8, \\ 0 & 0{,}6 < |u|. \end{cases}$$

В таблице илл. 14.8.3 приведены вычисления с этими весами для подгонки прямой к данным илл. 14.8.1.

Теперь все вычисления свелись, по существу, к вычислению остатков, уж без этого нам не обойтись.

Если мы положим в нашей процедуре шагового взвешивания $c = 5$, то счет станет еще немного проще, ибо правило вычисления весов будет выглядеть так:

$$\begin{aligned} w_i &= 4, \text{ если } && |e_i| \leqslant S_k, \\ w_i &= 3, \text{ если } && S_k < |e_i| \leqslant 2S_k, \\ w_i &= 2, \text{ если } && 2S_k < |e_i| \leqslant 3S_k, \\ w_i &= 1, \text{ если } && 3S_k < |e_i| \leqslant 4S_k, \\ w_i &= 0, \text{ если } && 4S_k < |e_i|. \end{aligned}$$

Соответствующие вычисления приведены в таблице илл. 14.8.4. На каждом этапе, помимо расчета остатков, надо найти S_k медиану (остатка) и величины, кратные ей, с сомножителями 1, 2, 3 и 4. Далее надо еще определить, чему равны u_i — 4, 3, 2, 1 или 0. Для этого мы сравниваем $|e_i|$ с упоминавшимися кратными (см. столбец (*)). Затем вычисляются целочисленные множители для xy и x^2, используемые при вычислениях соответствующих сумм, и, наконец, $\hat{\beta}^{(k+1)}$ как их отношение. Когда мы можем вручную применить стандартный метод наименьших квадратов (а он должен включать вычисление остатков), мы, как правило, можем позволить себе и ручную процедуру последовательного взвешивания.

Иногда выгодней использовать $c = 9$ и несколько иную систему шагов, а именно:

$$
\begin{aligned}
w_i &= 4, \text{ если } && |e_i| \leqslant 2{,}4S_k,\\
w_i &= 3, \text{ если } && 2{,}4S_k < |e_i| \leqslant 4{,}2S_k,\\
w_i &= 2, \text{ если } && 4{,}2S_k < |e_i| \leqslant 5{,}6S_k,\\
w_i &= 1, \text{ если } && 5{,}6S_k < |e_i| \leqslant 7{,}4S_k,\\
w_i &= 0, \text{ если } && 7{,}4S_k < |e_i|.
\end{aligned}
$$

После вычисления этих 4 сомножителей S_k мы действуем, как и раньше.

Веса для весов. Пусть нам заранее известно, что разные измерения будут делаться с разной точностью, поэтому следует начинать с различных w_k. Что при этом изменится?

Если мы возьмем веса, обратно пропорциональные ожидаемым дисперсиям, как это обычно делается в простейших случаях, то следует ориентироваться на соответствующим образом взвешенные остатки, а именно

$$\sqrt{w(i)}\,(y(i) - \hat{y}(i)).$$

Действительно, сумму квадратов этих самых величин мы и хотим минимизировать.

На k-м шаге возьмем

$$
\begin{aligned}
e_i^{(k)} &= \sqrt{w(i)}\,[y^{(k)}(i) - \hat{y}^{(k)}(i)],\\
S_k &= \text{медиана } |e_i^{(k)}|,\\
c &, \text{ как и ранее,}\\
u_i^{(k)} &= e_i^{(k)}/cS_k,\\
w(u) &, \text{ как и ранее.}
\end{aligned}
$$

В качестве веса на следующем шаге возьмем не $w(u_i^{(k)})$, а произведение $w(u_i^k)\,w(i)$ — нового веса на старый. Вычисления здесь достаточно просты. Отметим, что в $e_i^{(k)}$ входят $w(i)$, но не $w(u_i^{(k-1)})$.

✻ 14.9. МЕТОД НАИМЕНЬШИХ АБСОЛЮТНЫХ ОТКЛОНЕНИЙ (МОДУЛЕЙ) (НЕОБЯЗАТЕЛЬНОЕ)

Иногда вместо метода наименьших квадратов желательно применить метод наименьших абсолютных отклонений (модулей), т. е. использовать метод подгонки, при котором минимизируется

$$(\text{константа})\ \Sigma\,|\,y\,(i)-\hat{\hat{y}}\,(i)\,| = \Sigma\,\frac{\text{константа}}{|\,y\,(i)-\hat{\hat{y}}\,(i)\,|}\,(y\,(i)-\hat{\hat{y}}\,(i))^2.$$

Для краткости этот метод мы будем называть методом наименьших модулей. Этот метод по сравнению с методом наименьших квадратов имеет как преимущества, так и недостатки. Так, он не слишком реагирует на большие остатки, но неоправданно чувствителен к очень малым (мы покажем далее, как бороться с этой трудностью).

Предположим, что действует итеративная подгонка из параграфа 14.7 с весами

$$w_{k+1}\,(i) = \frac{\text{константа}}{|\,y\,(i)-\hat{\hat{y}}^{(k)}\,(i)\,|}$$

и пусть в результате получено приближение $\hat{\hat{y}}$. Тогда очевидно, что это приближение минимизирует выражение

$$\Sigma w \text{ последнее } (i)\,(y\,(i)-\hat{\hat{y}}\,(i))^2 = \Sigma\,\frac{\text{константа}}{|\,y\,(i)-\hat{\hat{y}}\,(i)\,|}\,(y\,(i)-\hat{\hat{y}}\,(i))^2.$$

Следовательно, $\hat{\hat{y}}\,(i)$ — приближение метода наименьших модулей. Таким образом, как мы видим, итеративная процедура взвешенных наименьших квадратов может привести к оценкам метода наименьших модулей. Наши выкладки указывают также на одну неприятную особенность метода наименьших модулей, а именно: наибольшие веса соответствуют откликам с наименьшими остатками; действительно, малое значение $|y\,(i) - \hat{\hat{y}}\,(i)|$ в знаменателе приводит к большому весу.

Пример 1. Одно резко выделяющееся наблюдение. Используя метод наименьших модулей, подгоним прямую $y = \beta x$ к числовым данным из таблицы илл. 14.9.1.

Обсуждение. Рисунок, расположенный под таблицей илл. 14.9.1, подсказывает, что хорошим приближением к нашим данным была бы прямая $y = x$, т. е. $\hat{\beta} = 1$. Если мы начнем с этого приближения, то четыре из имеющихся у нас пяти точек лягут точно на прямую. Это создаст трудности при вычислении величин, обратных к остаткам. С точки зрения нашей интерпретации метода подгонки надо было бы придать четырем точкам бесконечные веса, а только одной — пятой — конечный вес. Все это, однако, не слишком нас беспокоит, так как ясно, что $y = x$ — хорошее приближение. Но предположим, что мы начали нашу процедуру с прямой, проходящей через пятую точку. Тогда этой

точке надо было бы приписать бесконечный вес (веса остальных не играют роли). На этом процесс итераций был бы закончен, но результат явно неприемлем как с точки зрения здравого смысла, так и с вычислительной стороны.

Если мы хотим получить приближение метода наименьших модулей, то можно было бы начать с приближения, где $\hat{\beta} = 1 + d$ с малым d (меньшим чем 0,2). Остатки для этого приближения приведены в таблице илл. 14.9.1, там же приведена их сумма — $1 + 5d$. Таким образом, мы можем минимизировать сумму абсолютных отклонений, взяв $d = 0$. Полученная при этом прямая будет как раз $y = x$.

Пример 2. Трудности. Итеративная процедура может быть проиллюстрирована также на примере из параграфа 14.7. Вычисления, относящиеся к данным этого примера, собраны в таблице илл. 14.9.2. Мы начинаем с $\hat{\beta}^{(1)} = 1{,}02$.

Все дело в том, что придание больших весов точкам с малыми остатками не только вызывает вычислительные трудности, но и бессмысленно со статистической точки зрения. Предположим, что мы хотим минимизировать сумму членов и пусть абсолютные отклонения удовлетворяют нас как мера больших остатков, но для малых остатков все же нужна более разумная мера. Мы будем приписывать средним значениям измерений достаточно заметные веса и снижать их с ростом остатков.

Выравнивающие веса. Такую программу можно осуществить, используя веса

$$w_{h+1}(i) = \begin{cases} 1, \text{ если } |y(i) - \hat{y}^{(h)}(i)| \leqslant k; \\ \dfrac{k}{|y(i) - \hat{y}^{(h)}(i)|} \text{ — в противном случае,} \end{cases}$$

где k — некоторая, не слишком большая, величина, например медиана абсолютных остатков (в параграфе 14.8 мы обозначали ее S_k).

В результате мы минимизируем

$$\Sigma\psi(y - \hat{y}),$$

где

$$\psi(u) = \begin{cases} u^2, \text{ если } -k \leqslant u \leqslant k, \\ k|u| \text{ — в противном случае.} \end{cases}$$

Пример 3. Веса. Применим предложенную методику к данным примера 2.

Обсуждение. Начнем с округленного значения последней из наших оценок β, а именно с $\hat{\beta} = 1{,}03$. Положим $k = 0{,}07$, так как из данных четвертого столбца таблицы илл. 14.9.3 следует, что медиана ошибок равна 0,07. В результате получим, что трем ближайшим к началу координат точкам надо приписать единичные веса, четвертой точке — небольшой вес, а последней — маленький. Даже с небольшим весом эта точка сыграла свою роль. Действительно, разность между отвечающим ей xy и x^2 (с учетом весов) дает значительный вклад в отклонение углового коэффициента полученной прямой от 1,000.

Впрочем, различие между 1,03 и 1,032 не столь велико, чтобы из-за него проводить дополнительные итерации.

Другие степени. Если мы хотим минимизировать $\Sigma |y - \hat{\hat{y}}|^{2-p}$ для какого-нибудь p, то в качестве весов надо брать

$$w_{h+1}(i) = \frac{\text{константа}}{|y(i) - \hat{\hat{y}}^{(h)}(i)|^p}.$$

Как и раньше, чтобы не придавать больших весов хорошо приближенным точкам, надо выравнять веса.

Веса для весов. Предположим, что $p = 0{,}5$ и мы хотим минимизировать не

$$\Sigma |y - \hat{\hat{y}}|^{1,5},$$

а

$$\Sigma w(i) |y(i) - \hat{\hat{y}}(i)|^{1,5},$$

где веса $w(i)$ заданы и не зависят от $y - \hat{\hat{y}}$.

Предоставляем читателю проверить, что для этого достаточно вместо весов, предлагаемых в параграфе 14.8,

$$w_{h+1}(i) = \frac{\text{константа}}{|y(i) - \hat{\hat{y}}^{(h)}(i)|^{0,5}},$$

взять веса

$$w_{h+1}(i) = w(i) \frac{\text{константа}}{|y(i) - \hat{\hat{y}}^{(h)}(i)|^{0,5}}.$$

14.10. ТРУДНОСТИ АНАЛИЗА

Если $\Sigma x_{1 \cdot 23 \ldots k}^2$ мало́, то дисперсия $\hat{\beta}_1$ велика и здесь ничего не поделаешь. Все, что можно узнать о β_1, используя обычный метод наименьших квадратов, заключено в картинке, изображающей зависимость $y_{\cdot 23 \ldots k}$ от $x_{1 \cdot 23 \ldots k}$. Предположим, что, рассматривая $i = 200$ значений, мы обнаружили следующую картинку:

$$-0{,}0002 \leqslant x_{1 \cdot 23 \ldots k}(i) \leqslant 0{,}0003 \quad \text{для 197 значений } i,$$
$$1{,}97 \leqslant x_{1 \cdot 23 \ldots k}(i) \leqslant 2{,}01 \quad \text{для 2 значений } i,$$
$$x_{1 \cdot 23 \ldots k}(i) = 47{,}34 \quad \text{для 1 значения } i.$$

Можно быть уверенным, что здесь имели место три большие ошибки в измерениях либо в воспроизведении результатов, а может быть, в вычислениях и т. п. Ясно также, при каких значениях x эти ошибки имели место. Более того, если мы будем приближать форму

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k,$$

а не

$$y = \beta_2 x_2 + \ldots + \beta_k x_k,$$

то полученные нами $\hat{\beta}_1$ и, к сожалению, столь же неотвратимо $\hat{\beta}_2$, $\hat{\beta}_3$, ..., $\hat{\beta}_k$ будут почти полностью определяться именно этими тремя легко

находимыми точками. Такая ситуация, конечно, как правило, неприемлема, а поэтому важно знать, имеет ли она место.

Для этого надо бы, как положено, проводить проверку распределения значений $x_{1\cdot 23\ldots k}$ и родственных величин.

На первый взгляд можно попробовать избежать трудностей, просто отбрасывая эти три точки. В этом случае наши результаты по крайней мере не будут основаны на явно ошибочных данных. Однако остаются другие серьезные трудности — прежде всего для $x_{1\cdot 23\ldots k}$, подсчитанного по оставшимся данным

$$(x_{1\cdot 23\ldots k})^2 \leqslant 0{,}00000009 < 10^{-7},$$

а следовательно,

$$\operatorname{var}\widehat{\beta}_1 > \frac{\sigma^2}{n\cdot 10^{-7}} = 1\,000\,000 \text{ раз } \frac{10\sigma^2}{n}.$$

Такая дисперсия делает анализ бессмысленным.

Если лишь несколько значений $x_{1\cdot 23\ldots k}$ достаточно велики, чтобы быть полезными, то приближение метода наименьших квадратов вида

$$\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

вряд ли будет пригодно для тех целей, ради которых оно строится.

Пусть каким-то образом нам стало известно, что эти несколько значений верны. В этом случае запись уравнения регрессии в виде

$$\widehat{\beta}_1 x_{1\cdot 23\ldots k} + \widehat{\beta}^*_2 x_2 + \ldots + \widehat{\beta}^*_k x_k,$$

где β^* — коэффициенты модели без x_1 (см. последний параграф), позволяет более эффективно оценить, что нам известно, а что нет. Хотя $\widehat{\beta}_1$, как и ранее, зависит практически только от упоминавшихся нескольких значений, мы можем, изучая $x_{2\cdot 3\ldots k}$ и родственные ему величины, определить, какие β^* зависят от большого объема данных, а какие — от меньшего.

После отбрасывания. Если мы полагаем, что наши несколько больших значений — результаты ошибок в каком-то одном или нескольких x_i, то мы должны отбросить эти данные, а подгонку осуществлять по оставшимся. Мы, конечно, будем не в восторге от малости

$$\Sigma (x_{1\cdot 23\ldots k})^2$$

и соответственно большого значения $\operatorname{var}\{\hat{\beta}_1\}$, но такова жизнь.

Полагая

$$x_{1\cdot 23\ldots k} = x_1 - \widehat{\gamma}_{12} x_2 - \ldots - \widehat{\gamma}_{1k} x_k,$$

получим

$$\widehat{\beta}_j = \widehat{\gamma}_{ij}\widehat{\beta}_1 + \widehat{\beta}^*_j, \quad j = 2, 3, \ldots, k.$$

Если, кроме того, $\hat{\gamma}_{1j}$ — не слишком маленькая величина, а $\operatorname{var}\{\hat{\beta}^*_j\}$ — не слишком большая, то

$$\operatorname{var}\{\widehat{\gamma}_{1j}\widehat{\beta}_1\} \gg \operatorname{var}\{\widehat{\beta}^*_j\},$$

а поэтому var $\{\hat{\beta}_1\}$ и все остальные var $\{\hat{\beta}_j\}$ будут велики. Как и раньше, возможно, более разумным будет рассматривать модель

$$\widehat{\beta_1}\, x_{1\cdot 23\ldots k} + \widehat{\beta_2^*}\, x_2 + \ldots + \widehat{\beta_k^*}\, x_k$$

вместо

$$\widehat{\beta_1}\, x_1 + \widehat{\beta_2}\, x_2 + \ldots + \widehat{\beta_k}\, x_k.$$

Остальные коэффициенты $\widehat{\beta_2^*}$, ..., $\widehat{\beta_k^*}$ могут быть тем не менее вполне хорошо определимыми в одной и той же модели. И если мы выясним, сколь малой стала теперь сумма квадратов $x_{1\cdot 23\ldots k}$ (после удаления нескольких выделяющихся точек), то прояснится, о каких коэффициентах мы мало что можем сказать (это все те коэффициенты, в которые входят не только $\widehat{\beta_2^*}$, ..., $\widehat{\beta_k^*}$), а также почему сумма квадратов $\Sigma\,(x_{1\cdot 23\ldots k})^2$ так мала.

ОБЩИЕ ЗАМЕЧАНИЯ

Таким образом, преобразования уравнений регрессии, при которых отдельные коэффициенты становятся хорошо определимыми, а другие, быть может, хуже, иногда желательны. Новые коэффициенты будут иметь новый смысл, хотя может случиться, что численно они не изменятся. Мы надеемся узнать больше о том, что можно и чего нельзя извлечь из наших данных.

Один из методов преобразования уравнения регрессии — отбросить один или несколько членов. Это может быть мудрым решением, а может быть формой увиливания от ответственности. В тех случаях, когда мы строим модель, не задумываясь об интерпретации коэффициентов, и не имеем возможностей даже для того, чтобы сравнить ее с аналогичными моделями, построенными по иному множеству данных, практика отбрасывания переменных с плохо определенными коэффициентами — разумная практика. Если, наоборот, коэффициенты регрессии допускают интерпретацию и есть основания считать, что x_1 стоит включить в рассмотрение, то отбрасывание этой переменной да еще без соответствующих комментариев безответственно. В такой ситуации лучше оставить x_1 в форме $x_{1\cdot 23\ldots k}$ и четко определить, что неизвестно об этом коэффициенте.

Какую переменную избрать? В тех случаях, когда мы полагаем целесообразным отбросить одну из переменных или скорректировать ее по другим, может возникнуть неопределенность — какую же переменную взять для этих целей. Например, мы можем выбирать между

$$\widehat{\beta_1}\, x_{1\cdot 23\ldots} + \widehat{\beta_2^*}\, x_2 + \widehat{\beta_3^*}\, x_3,$$

$$\widehat{\beta_1^{**}}\, x_1 + \widehat{\beta_2}\, x_{2\cdot 13} + \widehat{\beta_3^{**}}\, x_3$$

и

$$\widehat{\beta_1^{***}}\, x_1 + \widehat{\beta_2^{***}}\, x_2 + \widehat{\beta_3}\, x_{3\cdot 12}$$

или между

$$\widehat{\beta}_2^* x_2 + \widehat{\beta}_3^* x_3,$$

$$\widehat{\beta}_1^{**} x_1 + \widehat{\beta}_3^{**} x_3$$

и

$$\widehat{\beta}_1^{***} x_1 + \widehat{\beta}_2^{***} x_2.$$

Как сделать выбор? Мы предлагаем следующее: запишите все альтернативы, подумайте, что может дать каждая и выбирайте самую полезную. К сожалению, не всегда возможно свести дело к работе с одной переменной. Возможно, придется действовать с несколькими переменными одновременно.

Веса. Как мы уже отмечали выше, определение $x_{1\cdot 23\ldots k}$ можно распространить на случай весов $w\,(i)$. Эти веса будут либо отражать наши взгляды на то, как var $\{y\,(i)\}$ зависит от σ, либо работать в итеративной процедуре построения искомого приближения, либо использоваться в обоих случаях. И здесь у нас появятся трудности, если почти все значения $x_{1\cdot 23\ldots k}$ малы.

Единственный способ избежать затруднений, если мало́ $\Sigma\,(x_{1\cdot 23\ldots k})^2$, — добыть дополнительную информацию.

Неравенство. Предположим, что мы последовательно строим

$$y = \beta_1 x_1 \quad \text{(для определения } \hat{y}^{(1)}\text{)},$$

$$y = \beta_1 x_1 + \beta_2 x_2 \quad \text{(для определения } \hat{y}^{(2)}\text{)},$$

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \quad \text{(и т. д.)},$$

$$\vdots$$

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k.$$

Тогда помимо больших колебаний величин $\hat{\beta}_1$ при переходе от одной модели к другой будут еще выполняться неравенства

$$\Sigma x_1^2 \geqslant \Sigma\,(x_{1\cdot 2})^2 \geqslant \Sigma\,(x_{1\cdot 23})^2 \geqslant \ldots \geqslant \Sigma\,(x_{1\cdot 23\ldots k})^2.$$

Дисперсия var $\{y\}$ нам неизвестна. Все, что мы можем сделать, — это оценить ее, что мы делаем на основе значений $y-\hat{y}$ более точно, $y-\hat{y}^{(1)}$ или $y-\hat{y}^{(2)}$, или $y-\hat{y}^{(3)}$ ... в зависимости от того, какое из приближений оказалось подходящим. Так как мы производим подгонку методом наименьших квадратов,

$$\Sigma\,(y-\widehat{y^{(1)}})^2 \geqslant \Sigma\,(y-\widehat{y^{(2)}})^2 \geqslant \Sigma\,(y-\widehat{y^{(3)}})^2 \geqslant \ldots \geqslant \Sigma\,(y-\widehat{y^{(k)}})^2.$$

Отсюда, учитывая, что при оценке var $\{y\}$ мы обычно делим

$$(y-\hat{y}^{(i)})^2 \text{ на } (n-i),$$

можно заключить, что var $\{y\}$ может медленно расти с увеличением числа членов, включенных в регрессию. В действительности это бывает редко.

Если наши оценки var $\{y\}$ одинаковы, что и имеет место обычно, если $x_2, x_3, \ldots, x_k$ дают лишь случайный вклад в y, то

$$\hat{\text{var}}\,\{\hat{\beta}_1|\beta_1 x_1\} < \hat{\text{var}}\,\{\hat{\beta}_1|\beta_1 x_1 + \beta_2 x_2\} < \ldots < \hat{\text{var}}\,\{\hat{\beta}_1|\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k\}.$$

Поэтому в такой ситуации чем больше переменных мы включаем в регрессию, тем больше меняются коэффициенты ранее включенных переменных.

Это, конечно, противоречит привычному взгляду на мир. Ведь обычно, добавляя члены, мы убеждались в том, что $y-\hat{y}$ становится меньше. Это создало у нас счастливую уверенность, что «длинные» приближения дают меньшие оценки σ^2, чем «короткие». Предположим, что у нас есть довольно кандидатов на включение в модель и мы можем варьировать в ней число членов.

Очевидно, что в $\Sigma\,(y-\hat{y})^2$ входят, во-первых, систематическая составляющая (так как мы еще не включили в приближение достаточное число носителей) и, во-вторых, разброс, который можно объяснить случайными факторами, относительно слабо зависящий (если вообще зависящий) от того, как много носителей включено в модель. Предположим далее, что

$$\text{оценка var}\,\{y\} = (\text{систематическая составляющая})^2 + \text{изменчивость}.$$

Первое слагаемое — вклад в оценку дисперсии, объясняемый недостаточным числом включенных носителей. Оно должно уменьшаться с «удлинением» приближения. Наблюденная дисперсия $\hat{\beta}_1$ вычисляется по формуле

$$\frac{\text{оценка var}\,\{y\}}{\Sigma x_1^2 \ldots} = \frac{(\text{систематическая составляющая})^2 + \text{изменчивость}}{\Sigma x_1^2 \ldots}.$$

Здесь первое слагаемое в числителе уменьшается вместе со знаменателем. Часто в течение некоторого времени числитель уменьшается быстрее знаменателя, но затем в этом состязании неизбежно побеждает знаменатель. Это приводит к тому, что дробь опять начинает расти.

Мы можем так зайти слишком далеко. Сознавая, что добавление x всегда приводит к изменению значений коэффициентов, очевидно, легко выяснить, как далеко мы заходим, не сознавая этого, если нам нужны еще и осмысленные коэффициенты.

Мы можем также перестараться и в результате увеличить среднюю var $\{\hat{y}\}$ вместо того, чтобы сделать ее меньше. Правда, здесь увеличение происходит постепенно, поэтому результаты менее чувствительны к точности выбора (см. обсуждение в параграфе 15.1 и ссылки в конце гл. 15 по поводу соответствующих методов).

14.11. ДОКАЗАТЕЛЬСТВО ОДНОГО УТВЕРЖДЕНИЯ ИЗ ПАРАГРАФА 13.2

Мы докажем, что если

$$y_{\cdot 1x} = y - a - bx,$$

$$t_{\cdot 1x} = t - c - dx$$

и

$$y_{\cdot 1x} \sim e + ft_{\cdot 1x},$$

причем $y_{\cdot 1x}$ — приближение метода наименьших квадратов, то

$$a + bx + f(t - c - dx)$$

— тоже приближение метода наименьших квадратов для y (по 1, x и t). Начнем доказательство с самого начала. Так как $y_{\cdot 1x}$ — остаток приближения метода наименьших квадратов с учетом 1 (мы приближали $a \cdot 1 + bx$), то 1 — балансир, а, следовательно, сумма остатков y равна 0, так что

$$\Sigma 1\ (y_{\cdot 1x}) = 0.$$

По тем же причинам

$$\Sigma 1\ (t_{\cdot 1x}) = 0.$$

Из «первого» нормального уравнения для регрессии $y_{\cdot 1x}$ по $t_{\cdot 1x}$ и 1

$$\Sigma 1\ (y_{\cdot 1x} - (e + ft_{\cdot 1x})) = 0$$

следует, что

$$\Sigma y_{\cdot 1x} - e\Sigma 1 + f\Sigma t_{\cdot 1x} = 0.$$

Отсюда, учитывая, что суммы y-остатков и t-остатков равны нулю, легко заключить, что и $e = 0$.

Таким образом, последнее приближение должно иметь вид

$$y_{\cdot 1x} \sim ft_{\cdot 1x}.$$

Попробуем собрать все вместе. Прежде всего мы подгоняем

$$a + bx$$

к y, а затем $ft_{\cdot 1x}$ — к тому, что осталось. Рассмотрим теперь комбинацию

$$a + bx + ft_{\cdot 1x},$$

которую можно развернуть следующим образом:

$$a + bx + f(t - c - dx) = (a - fc) + (b - fd)\,x + ft.$$

Что можно сказать теперь?

Очевидно, что

$$\Sigma_1 (y) = \Sigma 1\ (a + bx),$$

а мы показали, что

$$\Sigma 1\ (t_{\cdot 1x}) = 0.$$

Таким образом, мы можем написать, «прибавляя нуль» в форме $f\Sigma t_{\cdot 1x}$,

$$\Sigma 1\ (y) = \Sigma 1\ (a + bx + ft_{\cdot 1x}).$$

Отсюда следует

$$\Sigma 1\ (y - a - bx - ft_{\cdot 1x}) = 0. \qquad (*)$$

Кроме того, ясно, что

$$\Sigma x\ (y) = \Sigma x\ (a + bx),$$

где «$x\ (y)$» означает «x раз y» и

$$\Sigma x\ (t_{\cdot 1x}) = 0.$$

Опять «прибавляя нуль» в форме $f\Sigma x t_{\cdot 1x}$, получим

$$\Sigma x\ (y) = \Sigma x\ (a + bx + ft_{\cdot 1x}).$$

Отсюда вытекает, что

$$\Sigma x\ (y - a - bx - ft_{\cdot 1x}) = 0. \qquad (**)$$

Поскольку

$$\Sigma 1\ (t_{\cdot 1x}) = 0$$

и

$$\Sigma x\ (t_{\cdot 1x}) = 0,$$

$$\Sigma t_{\cdot 1x}\ (a + bx) = 0.$$

Отсюда ввиду равенства $y_{\cdot 1x} = y - a - bx$

$$\Sigma t_{\cdot 1x}\ (y_{\cdot 1x}) = \Sigma t_{\cdot 1x}\ (y - a - bx) = \Sigma t_{\cdot 1x} y.$$

Таким образом, учитывая то, как мы выбирали f, получим

$$\Sigma t_{\cdot 1x}\ (y - a - bx - ft_{\cdot 1x}) = \Sigma t_{\cdot 1x}\ (y_{\cdot 1x} - ft_{\cdot 1x}) = 0.$$

Так как $t_{\cdot 1x} = t - c - dx$, получим $t = c + dx + t_{\cdot 1x}$, а в силу справедливости (*) и (**) имеем

$$\Sigma t\ (y - a - bx - ft_{\cdot 1x}) = 0. \qquad (***)$$

Таким образом доказано, что каждая из трех следующих сумм исчезает:

$$\Sigma \begin{Bmatrix} 1 \\ x \\ t \end{Bmatrix} (y - a - bx - ft_{\cdot 1x}) = 0. \qquad \begin{matrix} (*) \\ (**) \\ (***) \end{matrix}$$

Значит, мы доказали, что форма, полученная сложением последовательных приближений

$$a + bx + ft_{\cdot 1x} = (a - fc) + (b - fd)\, x + ft,$$

есть *множественное регрессионное приближение* для y по 1, x и t, полученное *методом наименьших квадратов*.

Как же нам интерпретировать коэффициент при t? Из приведенных выше конструкций следует, что это коэффициент регрессии $y_{\cdot 1x}$ (линейно скорректированного по x) относительно $t_{\cdot 1x}$ (линейно скорректированного по 1 и x).

Общий случай. Мы можем исключить 1, полагая

$$y \sim Bx,$$
$$t \sim Dx,$$
$$y_{\cdot 1x} \sim Ft_{\cdot 1x},$$

не меняя аргумент.

С помощью аналогичных рассуждений мы могли бы рассмотреть и случай нескольких x (взяв в качестве x каждую переменную x_i поочередно). Предложенный подход составляет часть вычислительного метода шаговой регрессии. Общий результат сформулирован в конце параграфа 13.2.

РЕЗЮМЕ. ПРОЦЕДУРЫ ПОДГОНКИ

Многие процедуры подгонки, в том числе и итеративные, обладают свойством аддитивности по использованным для построения множествам данных и могут определяться в терминах семейства *балансиров*. Для всякого балансира должно выполняться равенство

$$\Sigma \text{ (балансир) } y = \Sigma \text{ (балансир) } \hat{y},$$

где суммирование ведется по всем множествам данных. Последнее равенство может служить для выбора модели. В терминах балансиров описываются, в частности, различные модификации метода наименьших квадратов, в том числе и методы взвешивания с постоянными или переменными наборами весов.

Если *обычным* методом наименьших квадратов мы подгоняем данные к прямой, имея в качестве генератора {все $\alpha + \beta x$}, то множеством балансиров будет {все $c + dx$}.

Для любого коэффициента β_i в любом представлении генератора {все $\Sigma\beta_i x_i$} при подгонке обычным методом наименьших квадратов можно выделить специальный балансир, который позволительно называть *уловителем*, так как он удовлетворяет не только соотношению

$$\Sigma \text{ (балансир) } y = \Sigma \text{ (балансир) } \hat{y},$$

но и соотношению

$$\hat{\beta}_i = \Sigma \text{ (балансир) } y,$$

где суммирование производится по всем множествам данных.

Если наш генератор {все $\Sigma\beta_i x_i$}, то в качестве собрания всех балансиров можно взять все мыслимые модели из этого генератора. Если мы так и поступим, то сумма квадратов разностей $\Sigma (y - \hat{y})^2$ будет минимальной. Следовательно, полученное приближение и есть приближение метода наименьших квадратов.

Чтобы оценить возможные трудности при приближении с помощью генератора {все $\Sigma\gamma_i x_i$}, полезно внимательно изучить остатки

$$y - \hat{y} \text{ и } x_{i\cdot} = x_i - \hat{x}_i$$

для всех i.

Дисперсии приближения β_i в генераторе {все $\Sigma \beta_i x_i$} обычным методом наименьших квадратов определяются формулой

$$\operatorname{var} \hat{\beta}_i = \frac{\sigma^2}{\Sigma (x_i - \hat{x_i})^2},$$

где каждое y удовлетворяет соотношению

$$y = \text{неизвестная модель} + \text{возмущение}.$$

Здесь предполагается, что возмущения некоррелированы и имеют общую дисперсию σ^2, а $\hat{x_i}$ обозначает приближение метода наименьших квадратов для x_i по другим носителям (по подгенератору x_i).

Поскольку приближение осуществлялось по подгенератору x_i, разумно по тому же подгенератору построить приближение y и вычислить остатки $y_{\cdot \text{ все. кроме } i}$, после чего обычная подгонка с помощью метода наименьших квадратов данных $(x_i - \hat{x_i}, y_{\cdot \text{ все, кроме } i})$ к прямой даст нам прямую с угловым коэффициентом $\hat{\beta}_i$.

Работа по вычислению, $y_{\cdot \text{ все. кроме } i}$ не должна быть слишком большой, так как

$$y_{\cdot \text{ все, кроме } i} = y_{\cdot \text{ все}} + \hat{\beta}_i (x_i - \hat{x_i}).$$

Это представление может помочь нам использовать рассмотренный подход для лучшего проникновения в тонкости зависимости $\hat{\beta}_i$ от наших данных или найти такой вариант процедуры подгонки, который уменьшит несовместимость данных.

Ввиду того что

$$\Sigma x_i^2 \geqslant \Sigma x_{i \cdot 1}^2 \geqslant \Sigma x_{i \cdot 12}^2 \geqslant \cdots,$$

а эти величины могут иметь большие сомножители, точность подгонки $\hat{\beta}_i$ может только уменьшаться по мере добавления в генератор носителей.

Это уменьшение иногда покрывается убыванием s^2, которое мы вычисляем по $\Sigma (y - \hat{y})^2$. Более гибкое приближение лучше описывает поведение среднего y в функции x.

Если модель строится по генератору {все $\Sigma \beta_i x_i$} с использованием веса для каждого множества данных и набора балансиров {все (вес) $(\Sigma \gamma_i x_i)$}, то получающееся приближение минимизирует

$$\Sigma \text{ (вес) } (y - \hat{y})^2$$

и, следовательно, служит приближением взвешенного метода наименьших квадратов. Это означает, что приближение y взвешенным методом наименьших квадратов по генератору {все $\Sigma \beta_i x_i$} эквивалентно приближению $y \sqrt{\text{вес}}$ с помощью обычного метода наименьших квадратов по генератору

$$\{\text{все } \sqrt{\text{вес}}\ \Sigma \beta_i x_i\}.$$

Соответственно если мы выбираем веса так, что

$$(\text{вес}) \times \text{var}\,\{y\} = \sigma^2,$$

то

$$\text{var}\,\widehat{\beta_i} = \frac{\sigma^2}{\Sigma\,(\text{вес})\,x^2_{i\,\text{корр}}},$$

где $x_{i\,\text{корр}}$ — остаток после приближения x_i по его подгенератору с заданными весами.

Если мы введем «матричный вес» $\{w_{ij}\}$ и выберем в качестве балансиров все линейные комбинации

$$\{\text{балансир } y_i = \sum_j w_{ij}\,(\text{возможное приближение } y_j)\},$$

одну для каждой возможной модели, то мы минимизируем

$$\Sigma\Sigma w_{ij}\,(y_i - \hat{y}_i)\,(y_j - \hat{y}),$$

т. е. получим приближение обобщенного взвешенного метода наименьших квадратов.

Чтобы изучить, как выражения, зависящие от экспериментальных данных, реагируют на их изменения, полезно рассмотреть кривые влияния, построенные по результатам свободного изменения одной точки из множества данных при фиксированных остальных.

Кривая влияния бивеса ведет себя как раз так, как до́лжно в реальном мире, а именно:

а) почти прямолинейно, если изменяющаяся точка близка к центру множества данных;

б) уплощается при ее удалении от центра;

в) наконец, с удалением движущейся точки от центра убывает до такой величины, при которой перемещение движущейся точки уже не оказывает никакого влияния.

Очень гибкой процедурой подгонки служит итеративное применение метода наименьших квадратов, где на каждом шаге в качестве веса выбирается остаток приближения, полученного на предыдущем шаге.

Использование бивеса с S, равным, например, медиане абсолютного отклонения от медианы (или, другой вариант, — половине размаха между квартилями, см. гл. 10), и c, равным чему-то около 9 или, может быть, 6, дает эффективную, более того, устойчивую и робастную к эффективности процедуру подгонки.

Если мы используем веса для компенсации изменений в дисперсии при переходе от одного множества данных к другому и бивеса для того, чтобы обеспечить устойчивость и робастность к эффективности, го стоит

а) брать как вес произведение (вес) $\times$ (бивес);

б) в качестве знаменателя бивеса брать произведение $\sqrt{\text{вес}}\,((y - \hat{y})$ на последнем этапе).

Метод наименьших абсолютных отклонений или его модификация со сглаженными весами для более эффективного использования данных

с малыми $(y - \hat{y})$ может рассматриваться как результат итеративного применения взвешенного метода наименьших квадратов с весами, основанными на остатках предыдущего шага.

Процедура, использующая бивес, требует наличия компьютера.

Огрубленный вариант этой процедуры, использующий шаговую схему, поддается ручному счету.

В эти процедуры можно включать веса, компенсирующие изменение дисперсий при переходе от одного множества данных к другому.

БИБЛИОГРАФИЯ

H a m p e l F. E. (1974). The influence curve and its role in robust estimation. — J. Amer. Statist. Assoc., 69, 383—393.

L a r s e n W. A. and M c C l e a r y S. J. (1972). The use of partial residual plots in regression analysis. — Technometrics, 14, 781—790.

ИЛЛЮСТРАЦИИ

Иллюстрация 14.7.1

Кривые влияния

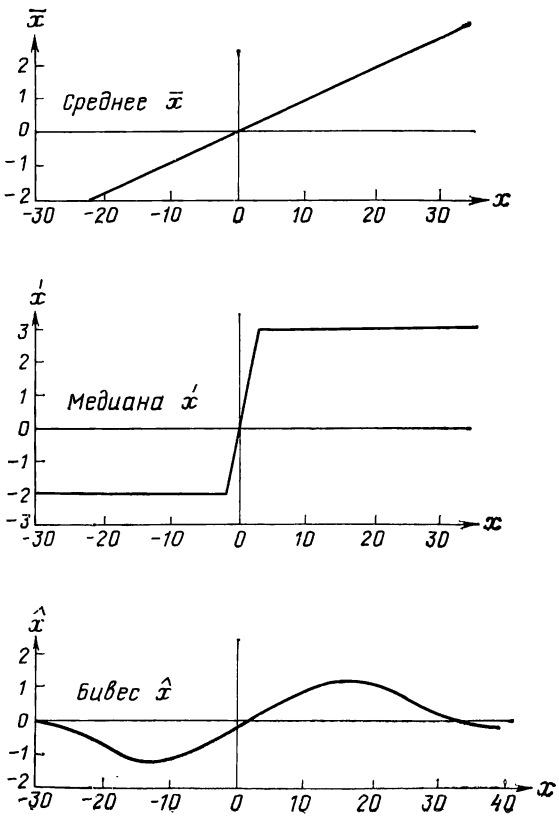


Иллюстрация 14.7.2

Вычисления для построения кривой влияния бивеса (пример: $x=-21$, $6S=27$, $\hat{x}^{(1)}=\bar{x}=-\frac{21}{11}=-1{,}9091$).

А. Первая итерация

x_i	$x_i-\hat{x}^{(1)}$	$u_i=\frac{x_i-\hat{x}^{(1)}}{6S}$	u_i^2	$1-u_i^2$	$w_i(u_i)=(1-u_i^2)^2$, если $\lvert u\rvert<1$	$w_i(u_i)x_i$
10	11,9091	0,4411	0,1945	0,8055	0,6488	6,4875
7	8,9091	0,3300	0,1089	0,8911	0,7941	5,5587
3	4,9091	0,1818	0,0331	0,9669	0,9350	2,8049
3	4,9091	0,1818	0,0331	0,9669	0,9350	2,8049
3	4,9091	0,1818	0,0331	0,9669	0,9350	2,8049
—2	—0,0909	—0,0034	0,0000	1,0000	1,0000	—2,0000
—5	—3,0909	—0,1145	0,0131	0,9869	0,9740	—4,8698
—5	—3,0909	—0,1145	0,0131	0,9869	0,9740	—4,8698
—6	—4,0909	—0,1515	0,0230	0,9770	0,9546	—5,7277
—8	—6,0909	—0,2256	0,0509	0,9491	0,9008	—7,2065
—21	—19,0909	—0,7071	0,4999	0,5001	0,2501	—5,2511
Итог					9,3014	—9,4640

$$\hat{x}^{(2)}=\frac{\Sigma w_i x_i}{\Sigma w_i}=\frac{-9{,}4640}{0{,}3014}=-1{,}0175.$$

Б. Вторая итерация: $\hat{x}^{(2)}=-1{,}0175$

x_i	$x_i-\hat{x}^{(2)}$	u_i	u_i^2	$1-u_i$	w_i	$w_i x_i$
10	11,0175	0,4081	0,1665	0,8335	0,6947	6,9471
7	8,0175	0,2969	0,0882	0,9118	0,8314	5,8200
3	4,0175	0,1488	0,0221	0,9779	0,9562	2,8686
3	4,0175	0,1488	0,0221	0,9779	0,9562	2,8686
3	4,0175	0,1488	0,0221	0,9779	0,9562	2,8686
—2	—0,9825	—0,0364	0,0013	0,9987	0,9974	—1,9947
—5	—3,9825	—0,1475	0,0218	0,9782	0,9570	—4,7848
—5	—3,9825	—0,1475	0,0218	0,9782	0,9570	—4,7848
—6	—4,9825	—0,1845	0,0341	0,9659	0,9331	—5,5983
—8	—6,9825	—0,2586	0,0669	0,9331	0,8707	—6,9657
—21	—19,9825	—0,7401	0,5477	0,4523	0,2045	—4,2954
Итог					9,3144	—7,0509

$$\hat{x}^{(3)}=\frac{-7{,}0509}{9{,}3144}=-0{,}7570.$$

В. Третья итерация $\hat{x}^{(3)} = -0{,}7570$

x_i	w_i	$w_i x_i$
10	0,7077	7,0774
7	0,8417	5,8921
3	0,9617	2,8850
3	0,9617	2,8850
3	0,9617	2,8850
—2	0,9958	—1,9915
—5	0,9512	—4,7561
—5	0,9512	—4,7561
—6	0,9260	—5,5560
—8	0,8613	—6,8900
—21	0,1917	—4,0267
Итог	9,3117	—6,3519

Г. Четвертая итерация: $\hat{x}^{(4)} = -0{,}6821$

x_i	w_i	$w_i x_i$
10	0,7114	7,1145
7	0,8446	5,9125
3	0,9632	2,8895
3	0,9632	2,8895
3	0,9632	2,8895
—2	0,9952	—1,9905
—5	0,9495	—4,7475
—5	0,9495	—4,7475
—6	0,9239	—5,5435
—8	0,8585	—6,8678
—21	0,1881	—3,9504
Итог	9,3103	—6,1517

$$\hat{x}^{(4)} = \frac{-6{,}3519}{9{,}3117} = -0{,}6821. \qquad \hat{x}^{(5)} = \frac{-6{,}1517}{9{,}3103} = -0{,}6607.$$

Д. Пятая итерация: $\hat{x}^{(5)} = -0{,}6607$

x_i	w_i	$w_i x_i$
10	0,7125	7,1251
7	0,8455	5,9183
3	0,9636	2,8907
3	0,9636	2,8907
3	0,9636	2,8907
—2	0,9951	—1,9902
—5	0,9490	—4,7450
—5	0,9490	—4,7450
—6	0,9233	—5,5399
—8	0,8577	—6,8614
—21	0,1871	—3,9287
Итог	9,3100	—6,0947

$$\hat{x}^{(6)} = \frac{-6{,}0947}{9{,}3100} = -0{,}6546.$$

Иллюстрация 14.7.3

Кривая влияния, использующая медиану абсолютного отклонения от медианы в качестве меры разброса S

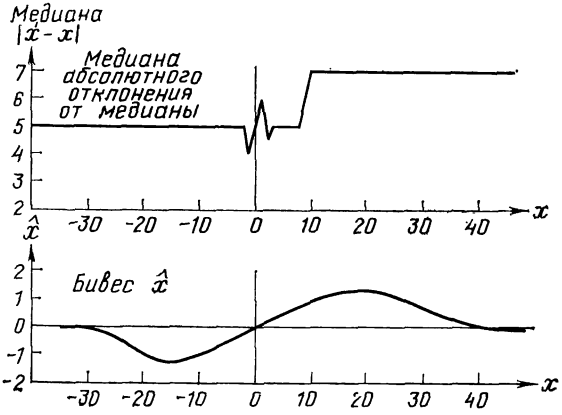


Иллюстрация 14.8.1

Прямая, проходящая через начало координат

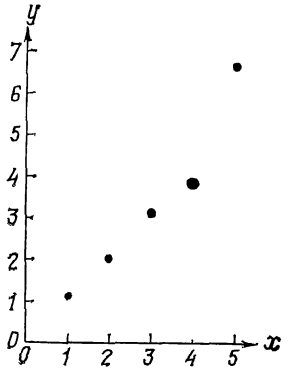


Иллюстрация 14.8.2

Расчет бивеса при подгонке прямой, проходящей через начало координат

А. Данные и простые суммы

x	y	(w)	xy	x^2
1	1,1	(1,0)	1,1	1
2	2,0	(1,0)	4,0	4
3	3,1	(1,0)	9,3	9
4	3,8	(1,0)	15,2	16
5	6,5	(1,0)	32,5	25
			62,1	55

$\hat{\beta}^{(1)} = 62,1/55 = 1,13.$

Б. Второй шаг (вычисления с точностью до двух десятичных знаков; жирный шрифт — медиана остатков)

x	y	пр.$^{(1)}$= $=1{,}13\,x$	$\lvert e\rvert=\lvert y-1{,}13\,x\rvert$	$\lvert u\rvert=\lvert e\rvert/6(0{,}29)$	$w^{(1)}=(1-u^2)^2$ или 0	$w^{(1)}xy$	$w^{(1)}x^2$
1	1,1	1,13	0,23	0,02	1,0	1,10	1,00
2	2,0	2,26	0,26	0,15	0,96	3,84	3,84
3	3,1	3,39	**0,29**	0,17	0,94	8,74	8,46
4	3,8	4,52	0,72	0,41	0,69	10,49	11,04
5	6,5	5,65	0,85	0,49	0,58	18,85	14,50
						43,02	38,84

$\beta^{(2)}=1{,}11$.

В. Третий шаг (как и ранее)

x	y	пр.$^{(2)}$= $=1{,}11\,x$	$\lvert e\rvert=\lvert y-1{,}11\,x\rvert$	$\lvert u\rvert=\lvert e\rvert/6(0{,}23)$	$w^{(2)}=(1-u^2)^2$ или 0	$w^{(2)}xy$	$w^{(2)}x^2$
1	1,1	1,11	0,01	0,01	1,00	1,10	1,00
2	2,0	2,22	0,22	0,16	0,95	3,80	3,80
3	3,1	3,33	**0,23**	0,17	0,94	8,74	8,46
4	3,8	4,44	0,64	0,46	0,62	9,42	9,92
5	6,5	5,55	0,95	0,69	0,27	8,78	6,75
						31,84	29,93

$\hat{\beta}^{(3)}=1{,}06$.

Г. Четвертый шаг (как и ранее)

x	y	пр.$^{(3)}$= $=1{,}06\,x$	$\lvert e\rvert=\lvert y-1{,}06\,x\rvert$	$\lvert u\rvert=\lvert e\rvert/6(0{,}12)$	$w^{(3)}=(1-u^2)^2$ или 0	$w^{(3)}xy$	$w^{(3)}x^2$
1	1,1	1,06	0,04	0,06	0,99	1,09	0,99
2	2,0	2,12	**0,12**	0,17	0,94	3,76	3,76
3	3,1	3,18	0,08	0,11	0,98	9,11	8,82
4	3,8	4,24	0,44	0,61	0,39	5,93	6,24
5	6,5	5,30	1,20	1,67	0	0	0
						20,28	19,81

$\hat{\beta}^{(4)}=1{,}02$.

Д. Пятый шаг (как и ранее)

x	y	пр $^{(4)}=$ $=1{,}02\ x$	$\lvert e\rvert=\lvert y-1{,}02\ x\rvert$	$\lvert u\rvert=\lvert e\rvert/6(0{,}08)$	$w^{(4)}=$ $=(1-u^2)^2$ или 0	$w^{(4)}xy$	$w^{(4)}x^2$
1	1,1	1,02	**0,08**	0,17	0,94	1,03	0,94
2	2,0	2,04	0,04	0,08	0,99	3,96	3,96
3	3,1	3,06	0,04	0,08	0,99	9,20	8,91
4	3,8	4,08	0,28	0,58	0,44	6,69	7,04
5	6,5	5,10	1,40	2,92	0	0	0
						20,88	20,85

$\hat{\beta}^{(5)}=1{,}00.$

Е. Шестой шаг (как и ранее)

x	y	пр.$^{(5)}=$ $=1{,}00\ x$	$\lvert e\rvert=\lvert y-1{,}00\ x\rvert$	$\lvert u\rvert=\lvert e\rvert/6(0{,}10)$	$w^{(5)}=$ $=(1-u^2)^2$ или 0	$w^{(5)}xy$	$w^{(5)}x^2$
1	1,1	1,0	**0,10**	0,17	0,94	1,03	0,94
2	2,0	2,0	0,00	0,00	1,00	4,00	4,00
3	3,1	3,0	0,10	0,17	0,94	8,74	8,46
4	3,8	4,0	0,20	0,33	0,79	12,01	12,64
5	6,5	5,0	1,50	2,50	0	0	0
						25,78	26,04

$\hat{\beta}^{(6)}=0{,}99.$

Иллюстрация 14.8.3

Последовательная подгонка, пример таблицы илл. 14.8.1 при $c=6$

А. Как и в части А илл. 14.8.1

Б. Второй шаг

x	y	пр.$^{(1)}=$ $=1{,}13\ x$	$\lvert e\rvert=\lvert y-1{,}13\ x\rvert$	$\lvert u\rvert=\lvert e\rvert/6(0{,}29)$	$w^{(1)}$	$w^{(1)}xy$	$w^{(1)}x^2$
1	1,1	1,13	0,03	0,02	4	4,4	4,0
2	2,0	2,26	0,26	0,15	4	16,0	16,0
3	3,1	3,39	**0,29**	0,17	4	37,2	36,0
4	3,8	4,52	0,72	0,41	2	30,4	32,0
5	6,5	5,65	0,85	0,49	2	65,0	50,0
						153,0	138,0

$\hat{\beta}^{(2)}=1{,}11.$

В. Третий шаг

x	y	пр.$^{(2)}$= $=1,11\,x$	$\lvert e\rvert=\lvert y-1,11x\rvert$	$\lvert u\rvert=$ $=\lvert e\rvert/6\ (0,23)$	$w^{(2)}$	$w_{xy}^{(2)}$	$w^{(2)}\,x^2$
1	1,1	1,11	0,01	0,01	4	4,4	4,0
2	2,0	2,22	0,22	0,16	4	16,0	16,0
3	3,1	3,33	**0,23**	0,17	4	37,2	36,0
4	3,8	4,44	0,64	0,46	2	30,4	32,0
5	6,5	5,55	0,95	0,69	1	32,5	25,0
$\beta=1,07.$						120,5	113,0

Б. Четвертый шаг

x	y	пр.$^{(3)}$= $=1,07\,x$	$\lvert e\rvert=\lvert y-1,07x\rvert$	$\lvert u\rvert=$ $=\lvert e\rvert/6\ (0,14)$	$w^{(3)}$	$w^{(3)}\,xy$	$w^{(3)}\,x^2$
1	1,1	1,07	0,03	0,04	4	4,4	4,0
2	2,0	2,14	**0,14**	0,17	4	16,0	16,0
3	3,1	3,21	0,11	0,13	4	37,2	36,0
4	3,8	4,28	0,48	0,57	2	30,4	32,0
5	6,5	5,35	1,15	1,37	0	0	0
$\beta=1,00.$						88,0	88,0

Иллюстрация 14.8.4
Расчет шаговой подгонки при $c=5$

А. Как и в части А илл. 14.8.2

Б. Второй шаг

x	y	пр.$^{(1)}$= $=1,13\,x$	$\lvert e\rvert=$ $=\lvert y-1,13\,x\rvert$	Кр.$^{(*)}$ S_1 w	$w^{(1)}$	$w^{(1)}\,xy$	$w^{(1)}\,x^2$
1	1,1	1,13	0,03	(0,29) 4	4	4,4	4,0
2	2,0	2,26	0,26	(0,58) 3	4	16,0	16,0
3	3,1	3,39	**0,29**	(0,87) 2	4	37,2	36,0
4	3,8	4,52	0,72	(1,16) 1	2	30,4	32,0
5	6,5	5,65	0,85	0	2	65,0	50,0
$\beta=1,11.$						153,0	138,0

В. Третий шаг

x	y	пр.$^{(2)}$= $=1,11\,x$	$\lvert e\rvert=$ $=\lvert y-1,11\,x\rvert$	Кр.$^{(*)}$ S_2 w	$w^{(2)}$	$w^{(2)}\,xy$	$w^{(2)}\,x^2$
1	1,1	1,11	0,01	(0,23) 4	4	4,4	4,0
2	2,0	2,22	0,22	(0,46) 3	4	16,0	16,0
3	3,1	3,33	**0,23**	(0,69) 2	4	37,2	36,0
4	3,8	4,44	0,64	(0,92) 1	2	30,4	32,0
5	6,5	5,55	0,95	0	0	0	0
$\beta=1,00.$						88,0	88,0

Г. Четвертый шаг

x	y	пр.$^{(3)}$ = $= 1,00\,x$	$\lvert e\rvert = \lvert y - 1,00\,x\rvert$	Кр. $S_3^{(*)}$	w	$w^{(3)}$	$w^{(3)}\,xy$	$w^{(3)}\,x^2$
1	1,1	1,00	0,10	(0,10)	4	4	4,4	4,0
2	2,0	2,00	0,00	(0,20)	3	4	16,0	16,0
3	3,1	3,00	**0,10**	(0,30)	2	4	37,2	36,0
4	3,8	4,00	0,20	(0,40)	1	3	45,6	48,0
5	6,5	5,00	1,50		0	0	0	0
							103,2	104,0

$\hat{\beta}^{(4)} = 0,99.$

Иллюстрация 14.9.1

Подгонка прямой с угловым коэффициентом, близким к 1

x	y	$\hat{\beta}\,x = 1 \cdot x$	$\lvert y - x\rvert$	$\hat{\beta}\,x = (1+d)\,x$	$\lvert y - (1+d)\,x\rvert$
1	1	1	0	$1 + d$	d
2	2	2	0	$2 + 2d$	$2d$
3	3	3	0	$3 + 3d$	$3d$
4	4	4	0	$4 + 4d$	$4d$
5	6	5	1	$5 + 5d$	$1 - 5d$
					Итог $1 + 5d$

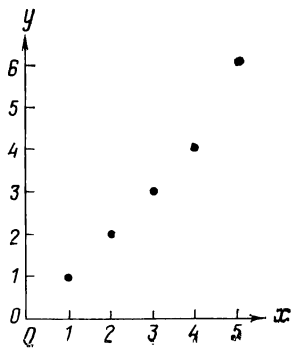


Иллюстрация 14.9.2

Подгонка прямой, проходящей через начало координат, методом наименьших абсолютных отклонений

А. Оценка метода наименьших квадратов: $62{,}1/55 \approx 1{,}13$

x	y	$\hat{\beta}^{(1)} x = 1{,}02\,x$	$\lvert y - 1{,}02\,x \rvert = e$	Вес $0{,}08/e$	wxy	wx^2
1	1,1	1,02	0,08	1,0	1,10	1,00
2	2,0	2,04	0,04	2,0	8,00	8,00
3	3,1	3,06	0,04	2,0	18,60	18,00
4	3,8	4,08	0,28	0,286	4,35	4,58
5	6,5	5,10	1,40	0,0571	1,86	1,43
			1,84		33,91	33,01

$$\hat{\beta}^{(2)} = \frac{33{,}91}{33{,}01} = 1{,}0273;$$ округлим до 1,028.

Б.

y	$\hat{\beta}^{(2)} x = 1{,}028\,x$	$\lvert y - 1{,}028\,x \rvert = e$	Вес $0{,}072/e$	wxy	wx^2
1,1	1,028	0,072	1,00	1,10	1,00
2,0	2,056	0,056	1,285	5,14	5,14
3,1	3,084	0,016	4,5	41,85	40,50
3,8	4,112	0,312	0,231	3,51	3,69
6,5	5,140	1,360	0,0529	1,72	1,32
		1,812		53,32	51,65

$$\hat{\beta}^{(3)} = \frac{53{,}32}{51{,}65} = 1{,}0323.$$

Иллюстрация 14.9.3

Выравнивающие веса, метод наименьших модулей

x	y	$\beta^{(1)} x = 1{,}03x$	$\lvert y - \hat{\beta}^{(1)} x \rvert = \lvert e \rvert$	$0{,}07/\lvert e \rvert$	$w^{(2)}$	$w_{xy}^{(2)}$	$w^{(2)} x^2$
1	1,1	1,03	0,07	1,00	1	1,1	1,0
2	2,0	2,06	0,06	1,17	1	4,0	4,0
3	3,1	3,09	0,01	7,00	1	9,3	9,0
4	3,8	4,12	0,32	0,22	0,22	3,34	3,52
5	6,5	5,15	1,35	0,05	0,05	1,63	1,25
						19,37	18,77

$$\hat{\beta}^{(2)} = \frac{19{,}37}{18{,}77} = 1{,}032.$$

Глава 15 ● ГИБКАЯ РЕГРЕССИЯ

Трудности часто возникают там, где:

● есть несколько (быть может, много) носителей;

● мы не ожидаем, что все они пригодятся;

● и готовы, после некоторых предварительных модификаций, допустить, чтобы выделенный в конце концов генератор определялся любым мыслимым подмножеством из совокупности носителей (на финальных стадиях исследования может использоваться и несколько подмножеств).

Мы не слишком заботимся о коэффициентах приближения (в противном случае следовало бы задуматься и о выборе генератора). Таким образом, наша цель — регрессия для описания или исключения.

В этой главе мы описываем лишь технику наименьших квадратов (с равными весами или с постоянным набором весов). Рассматриваются также некоторые устойчивые варианты этого метода. Конечно, сказанное не означает, что мы относимся к противникам методов, отличающихся от метода наименьших квадратов.

Нам нужны такие методы, где данные сами управляют анализом, во всяком случае на стадии выбора генератора. На практике это означает управление компьютером, так как объем счета (в таких методах) непосилен для ручного калькулятора. Эффективность комбинации «компьютер—дисплей—человек» в такого рода задачах пока мало исследовалась.

15.1. ЧЕМ МЫ МОЖЕМ РУКОВОДСТВОВАТЬСЯ ПРИ ВЫБОРЕ ПРИБЛИЖЕНИЯ?

Мы вынуждены сделать выбор из многих альтернативных генераторов. Естественно, нам бы хотелось научиться измерять качество каждого из них, что дало бы возможность выбирать целенаправленно. Мы построим такую меру, опираясь сначала на некоторые «идеальные» условия и полагая, что они на самом деле имеют место. Как обычно, i-я точка в множестве данных будет обозначаться $(x_{1i}, x_{2i}, \ldots, x_{ki}, y_i)$.

ИДЕАЛЬНЫЕ УСЛОВИЯ

1. **Модель.** Значения y представимы в виде

$$y_i = \eta_i + \text{ошибка}_i.$$

2. **Несмещенность.** Средние (математические ожидания) всех ошибок равны нулю.

3. **Однородность.** Дисперсии σ^2 всех ошибок одинаковы.

4. **Некоррелированность**. Любая пара разных ошибок, например ошибка$_i$ и ошибка$_j$ ($i \neq j$), имеет нулевую ковариацию.

5. **Структура**. «Истинное» значение есть линейная комбинация

$$\eta_i = \sum_j \beta_j x_{ji}.$$

В таком случае генератор описывает все систематические эффекты.

Эти условия вполне могут оказаться далекими от действительного положения вещей, но часто они служат разумной исходной точкой. Во всяком случае, действуя в соответствии с этими условиями, мы часто получаем полезный ответ. (Построение разумного приближения, независимо от справедливости этих условий, служит обычно первым шагом при определении их пригодности.)

Если эти идеальные условия выполнены для любого генератора, то они выполнены и для генератора, полученного расширением исходного. Если мы стремимся избежать получения неоправданно большого генератора, то разумно рассмотреть вопрос о выборе между генераторами, удовлетворяющими идеальным условиям. Подчеркнем, что выбор делается между генераторами, а не между приближениями; как уже отмечалось в этой главе, приближение всегда строится методом наименьших квадратов с фиксированным набором весов.

В любом случае удобство и простота наших идеальных условий в сочетании с идеей минимизации создают классическую основу для создания гибкой регрессии. А идеальный минимизируемой величиной мы называем математическое ожидание остаточной суммы квадратов, т. е. квадратов разностей между предсказанными значениями и тем, что наблюдалось:

$$\text{ave} \sum (\hat{y}_i - \eta_i)^2.$$

Эта величина измеряет в среднем (прямое измерение невозможно) степень неадекватности модели. Для конкретного генератора, удовлетворяющего идеальным условиям, она дает среднее по большой выборке значение суммы квадратов отклонений оценок $\hat{y}_i$ от истинных значений η_i.

Сразу возникает затруднение. Едва ли мы действительно знаем η; мы как раз и взялись за получение y, чтобы узнать хоть что-нибудь о нем. Поэтому разумно рассматривать адекватность не для произвольной идеальной минимизируемой величины, а для такой, с которой мы сможем справиться.

Оценивание неадекватности. Напомним, что $\hat{y}_i$ — несмещенная оценка η_i, т. е. ave $\hat{y}_i = \eta_i$. Следовательно, средний квадрат отклонения $\hat{y}_i$ от η — это дисперсия $\hat{y}_i$:

$$\text{ave}\,(\hat{y}_i - \eta_i)^2 = \text{var}\,\hat{y}_i.$$

Отсюда следует, что идеальная минимизируемая величина равна:

$$\sum_i \text{var}\,\hat{y}_i.$$

Мы собираемся показать, во-первых, что идеальная минимизируемая величина равна

$$\sigma^2 \times (N \text{ носителей})$$

и, во-вторых, что оценкой для нее служит

$$s^2 \times (N \text{ носителей}),$$

где s^2 — знакомая нам оценка обычной дисперсии σ^2:

$$s^2 = \widehat{\sigma^2} = \frac{\Sigma (y_i - \widehat{y_i})^2}{N \text{ измерений МИНУС } N \text{ носителей}}.$$

Например, если мы приближаем $y = \beta_0 \cdot 1 + \beta_1 x$, то используются два носителя 1 и x, и, следовательно, знаменатель правой части равен $n - 2$, где n — число измерений.

Переходим к доказательству.

Доказательство. Как и раньше, мы рассматриваем один генератор, а, следовательно, можем, не меняя его, менять x_j, если при этом остаются справедливыми равенства

$$\sum_i x_{ji} x_{Ji} = 0, \quad j \neq J, \tag{1}$$

$$\sum_i x_{Ji}^2 = 1.$$

Предположим, что изменение x произведено и (1) теперь выполняется. Поскольку $\hat{y}_i$ остаются теми же функциями $y_1, \ldots, y_n$, это не меняет Σ var $\hat{y}_i$. Так как x фиксированы (напомним), то

$$\widehat{y_i} = \sum_j \beta_j \widehat{x_{ji}}$$

и

$$\hat{\beta}_j = \sum_i x_{ji} y_i.$$

Отсюда мы получим еще несколько результатов. Прежде всего,

$$\text{cov } (\hat{\beta}_j, \hat{\beta}_J) = 0, \quad j \neq J,$$

и

$$\text{var } \{\hat{\beta}_j\} = \sigma^2.$$

Далее,

$$\text{var } \widehat{y_i} = \sum_j x_{Ji}^2 \text{ var } \{\widehat{\beta_j}\} = \sigma^2 \sum_j x_{Ji}^2,$$

следовательно,

$$\sum_i \text{var } \widehat{y_i} = \sigma^2 \sum_i \sum_j x_{Ji}^2 = \sigma^2 \sum_j \sum_i x_{Ji}^2 = \sigma^2 \sum_j 1 = \sigma^2 (N \text{ носителей}).$$

В наших идеальных условиях приближения методы наименьших квадратов удовлетворяют равенству

$$\text{cov } (y_i - \hat{y}_i, \hat{y}_i) = 0.$$

Отсюда получаем

$$\text{var } y_i = \text{var } (y_i - \hat{y_i}) + \text{var } y_i = \text{ave } (y_i - \hat{y_i})^2 + \text{var } \hat{y_i}$$

и, значит,

$$\sigma^2 = \text{ave } (y_i - \widehat{y_i})^2 + \sigma^2 \sum_j x_{ji}^2.$$

Перенося второе слагаемое правой части в левую, получим

$$\text{ave } (y_i - \widehat{y_i})^2 = \left(1 - \sum_j x_{ji}^2\right) \sigma^2,$$

поэтому

$$\sum_i \text{ave } (y_i - \widehat{y_i})^2 = \sum_i \left(1 - \sum_j x_{ji}^2\right) \sigma^2 = \left(n - \sum_j 1\right) \sigma^2,$$

так как $\sum_i x_{ji}^2 = 1$. Последнее равенство означает

$$\sum_i \text{ave } (y_i - \hat{y}_i)^2 = (N \text{ измерений МИНУС } N \text{ носителей})\, \sigma^2.$$

Так как мы предположили, что все y имеют одну и ту же дисперсию σ^2, обычной оценкой дисперсии служит

$$s^2 = \widehat{\sigma^2} = \frac{\Sigma (y_i - \widehat{y_i})^2}{N \text{ измерений МИНУС } N \text{ носителей}},$$

а оценкой фактически минимизируемой величины

$$s^2 \times (N \text{ носителей}),$$

что мы и хотели продемонстрировать.

Возможно, читателю захочется вернуться к этим результатам после приведенного ниже обсуждения меры C_P.

Если нас и устраивает любое из минимизируемых выражений, то не устраивает избыток носителей, так как это увеличивает минимизируемую величину.

В тех случаях, когда мы удовлетворены практически достигнутым минимумом как таковым или как оценкой идеального, мы предпочитаем тот генератор среди всех с заданным числом носителей, который минимизирует s^2. Такой генератор обеспечивает наиболее точное приближение. (Если идеальные условия не выполняются, то некоторые генераторы могут давать меньшие σ^2, чем другие. Тогда генератор, дающий меньшее σ^2, дает и наилучшее приближение к наблюденным данным. Однако этого нам знать не дано.) Генератор, для которого

$$\eta_i = \text{ave } \{y_i\},$$

представляется наиболее точно соответствующим идеальным условиям.

Проиллюстрируем одну важную идею на примере выборочных и генеральных средних. В этой ситуации она может показаться более знакомой, чем в отношении сумм квадратов отклонений. Пусть имеется много выборок из совокупностей с близкими генеральными средними. Тогда весьма вероятно, что максимальное выборочное среднее получится в выборке не из той совокупности, где генеральное сред-

нее максимально. Тем не менее, если мы должны по выборочным данным решить, какая из совокупностей имеет максимальное генеральное среднее, мы выберем ту, которой отвечает максимальное выборочное среднее. Здесь мы сталкиваемся с проблемой *множественности, конкуренции*. Применительно к генераторам эта идея выглядит так: тот генератор, который дает наилучшее приближение среди всех генераторов с k носителями, может в то же время не совпадать с генератором, действительно наилучшим образом приближающим наблюденные значения. Если мы используем для приближения много альтернативных генераторов, то довольно многие из них могут иметь достаточно близкие идеальные минимумы (математические ожидания реальных), и лишь случай поможет выявить, какой из них будет *выглядеть* лучшим приближением выборки. Если несколько генераторов окажется почти одинаковыми с точки зрения «идеальности» приближения, то расчет по выборочным данным может сделать почти очевидным тот факт, что генератор, оказавшийся наилучшим, *не* совпадает с наилучшим «идеальным».

Выбор среди генераторов с k носителями. Не имея дополнительной информации при выборе между нашими генераторами с k носителями и всеми возможными генераторами с k носителями, мы скорее всего выберем тот, который минимизирует s^2.

Некоторые, может быть, попробуют отобрать несколько таких генераторов с малыми s^2, а затем произвести их дополнительное изучение.

Выбор k. Как выбирать число носителей k? Можно предложить ряд критериев, во многих случаях сопровождая предложения предупреждением об опасности слепого следования им. Мы упомянем здесь три разумных критерия:

- Маллоуза (Mallows C. L.) — C_p;
- Энскамби — Тьюки (Anscombe F. J. — Tukey J. W.) — $s^2/(n-k)$;
- Аллена (Allen D. M.) — PRESS.

Энскамби [Anscombe F. J.] (1967)] предложил, а Тьюки в дискуссии по этому предложению упростил критерий для определения размерности генератора для довольно жестко упорядоченной последовательности генераторов, каждый из которых должен содержать все предыдущие. Критерии основывались на требовании малости числа

$$\frac{\text{остаточная сумма квадратов}}{(n-k)^2} = \frac{s^2}{n-k} = \frac{\text{средний квадрат}}{\text{остаточные степени свободы}}.$$

Несколько исследователей развивали подход к отбору подмножества факторов. Этот подход обычно называют PRESS по начальным буквам английских слов в выражении «сумма квадратов для предсказания» (prediction sum of squares). Работы Андерсона, Аллена и Кэди [Anderson R. L., Allen D. M. and Cady F. B. (1972)] и Аллена [Allen D. M. (1974)] подходят для ознакомления с этим направлением.

Они действовали в духе методов включения и исключения в шаговом регрессионном анализе, но с некоторыми модификациями. Так, в процедуре «включения» после добавления новой переменной происходит пересмотр всех старых с точки зрения целесообразности их

сохранения в модели. Для понимания этого критерия удобно ввести величину

$$\text{PRESS}_{\text{ген}} = \sum_{i=1} (y_i - \widehat{y_{(i)\text{ ген}}})^2.$$

Она вычисляется для каждого рассматриваемого генератора. Остаток здесь — разность между наблюденным y_i и его оценкой по данному генератору, причем i-я переменная не участвует в построении оценки. Непосредственное использование этой формы приводит к очень сложным вычислениям. Оказывается, однако, что ее можно записать в другом виде, из которого ясно, что $\text{PRESS}_{\text{ген}}$ — это взвешенная сумма квадратов обычных остатков $y_i - \hat{y}_{(i)\text{ ген}}$. Вычисление этого выражения с применением весов гораздо проще.

К. Л. Маллоуз [Mallows C. L. (1978)] собрал очень поучительную коллекцию примеров и ссылок по поводу отбора факторов. Он использовал для оценки генераторов величину

$$C_p = \frac{1}{\hat{\sigma}^2} \text{ (остаточная сумма квадратов для генератора)} - n + 2p,$$

где n — число наблюдений; p — число носителей в генераторе; $\hat{\sigma}^2$ — некоторая несмещенная оценка действительной дисперсии. Мы используем здесь p вместо нашего k, так как именно от этого p происходит название критерия Маллоуза— C_p. Однако нужно помнить, что $p \equiv k$. Маллоуз тщательно избегает использования C_p для определения момента останова* или правил выбора. Он делает ударение на желательности использования достаточно большого числа приближений, причем не менее двух с одинаковым числом носителей. В частной беседе он как-то сказал: «C_p-критерий подсказывает иногда, какое k выбрать, а иногда — что всякий выбор нелеп». Однако любящим «руководства к действию» C_p может понравиться. Маллоуз указал также, что C_p получается примерно так, как мы получили выше оценку минимума в «идеале». Можно записать

$$J = \text{ave}\,(\hat{y}_i - \eta_i)^2 = \text{смещение} + \text{дисперсия}, \qquad (*)$$

где

$$\text{смещение} = \Sigma\,(\text{ave}\,\hat{y}_i - \eta_i)^2,$$

а

$$\text{дисперсия} = (n - p)\,\sigma^2.$$

Обозначим остаточную сумму квадратов RSS. Тогда

$$\text{ave RSS} = \text{смещение} + p\sigma^2.$$

Заменяя в (*) «смещение» на RSS $- p\sigma^2$ и σ^2 на s^2, получим следующую оценку для J:

$$\hat{J} = \text{RSS} - (n - 2p)\,s^2.$$

Это выражение можно переписать так:

$$ps^2 + \text{RSS} - (n - p)\,s^2.$$

* В процессе выбора. — *Примеч. пер.*

Первое слагаемое здесь — это мера, которую мы ранее называли идеальным минимизируемым функционалом*, сумма последних двух членов образует оценку смещения (по использованному генератору). Если избранный генератор не адекватен, то это смещение может оказаться ненулевым. Возможно, читатель предпочтет воспользоваться показателем C_p вместо ps^2.

Проблематика управляемой регрессии связана с трудностями разного рода — интеллектуальными, статистическими, вычислительными и интерпретационными.

Нельзя предсказать заранее, какое приближение окажется наиболее полезным, какие в него войдут носители и с какими коэффициентами.

Разумной стратегией будет, по-видимому, следующая: с помощью некоторого формализма создать набор приближений, а затем выбирать между ними, учитывая кажущееся качество приближений и другие обстоятельства.

Можно попробовать отбросить один носитель из трех имеющихся, если, конечно, это не приведет к тяжелым последствиям. А затем к оставшейся паре можно присоединить новый. Получившаяся тройка не может быть хуже предыдущей, но может иногда оказаться много лучше. Такие колебания — включения (увеличивая число носителей) и исключения (уменьшая) — могут оказаться полезными.

15.2. ШАГОВЫЕ МЕТОДЫ

А как быть, если у нас 10, или 100, или, скажем, 1000 возможных носителей для y?

Если нужен лишь один из них, то у нас нет особого выбора. Придется провести подгонку для каждого носителя, сравнить их s^2 и взять тот носитель, который дает наименьшее s^2. Эквивалентная процедура состоит в изучении того, какое уменьшение суммы квадратов дает каждый носитель отдельно

$$\mathrm{RSS}_j = \frac{\left(\sum_i x_{ji} y_i\right)^2}{\sum_i x_{ji}^2}, \quad j = 1, 2, \ldots, k,$$

с последующим выбором того, у которого RSS — максимальна. (В любом случае выбирается носитель, имеющий максимальную корреляцию с y.)

Если мы планируем использовать два носителя, то можно попробовать перебрать все пары. Для $k = 10$ у нас 45 пар и это приемлемо. Но для $k = 1000$ уже 499500 пар, и нет надежды исследовать их все.

Один из возможных здесь подходов состоит в том, чтобы исследовать каждый из носителей поодиночке, а затем наилучший из них сопоставить с каждым из 999 оставшихся. Нет гарантий того, что одна из этих 999 пар близка к наилучшей из 499500, но, возможно, наш вы-

---

* Точнее говоря, выборочная оценка для этой меры. — *Примеч. пер.*

бор будет не так уж плох. Есть ли надежда добиться чего-то большего? Можно попробовать 998 троек, получающихся добавлением к избранной нами паре оставшихся носителей. На этом пути можно добиться улучшения и уж во всяком случае хуже не будет.

Обычный шаговый метод. Стандартные программы шаговой регрессии обычно действуют по следующему принципу.

● *Шаг включения.* Рассматривая каждый носитель, не участвующий пока в регрессии, отбираем тот, которому соответствует максимальное уменьшение остаточной суммы квадратов. Далее производим проверку, и если результаты положительны, мы включаем этот носитель в регрессию; переходим к следующему шагу (вперед или назад в зависимости от программы).

● *Шаг исключения.* Удаляя из регрессии по отдельности включенные в нее носители, отбираем тот, устранение которого приводит к минимальному увеличению суммы квадратов. Далее проверяем, и если результаты положительны, то этот носитель отбрасывается, а мы переходим к следующему шагу включения.

Проверка производится обычно с помощью F-статистики:

$$F = \frac{\text{изменение остаточной суммы квадратов}}{s^2 \text{ для большего генератора}}.$$

Каждая такая величина сравнивается с величиной (или величинами) выбранной исследователем (для исключения F-критерий обычно можно выбирать отдельно). Мы рекомендуем останавливать процедуру только после того, как проверка даст «отрицательные результаты» для нескольких последовательных шагов, например не менее трех. В любом случае исход продемонстрирует полезные для регрессии результаты, полученные на каждом шаге вперед или назад.

Используемая статистика аналогична формальному критерию для проверки значимости коэффициента регрессии для нешаговой процедуры. Естественно, в последовательном подходе шагового метода эта статистика остается полезной, но ее номинальный уровень значимости не поддается точной частотной интерпретации. В шаговом методе часто используются классические уровни значимости, например 5%, 1%. Возможно, было бы полезнее брать 10% или 20%, чтобы не потерять многих перспективных кандидатов, когда эффективность большинства переменных мала.

Ответственность исследователя. Исследователь может использовать для определения момента остановки один или несколько критериев, упомянутых в параграфе 15.1. Автоматизация этой процедуры кажется опасной. Исследователь должен каким-то образом участвовать в принимаемых решениях.

Желательные дополнения. Какую дополнительную информацию разумно предложить пользователю программой шаговой регрессии?

Стоит поинтересоваться остатками каждого из избираемых приближений. Как минимум нужно иметь: а) информацию об их распределениях; б) перечень дюжины или около того наибольших из них по абсолютной величине. Если лишь несколько остатков заметно отличаются от остальных, то, по-видимому, разумно: а) произвести пе-

перасчет по данным, не содержащим этих исключительных точек; б) рассмотреть внимательно сами эти необычные данные. Может быть, исследователь найдет ошибку или какой-нибудь существенно новый эффект.

Стоит также рассмотреть коэффициенты всех носителей, которые могут быть включены в регрессию. Это не слишком затруднительно, а может уберечь от многих неприятностей. (Необходимые вычисления все равно производятся при минимизации остаточной суммы квадратов, используемой при выборе.)

Внезапные или даже постепенные изменения коэффициентов могут послужить предупреждением о зависимости или мультиколлинеарности переменных.

Пример. Макдональд и Вард [Macdonald N. J. and Ward F. (1963)] исследовали внутреннюю структуру магнитного поля Земли. Одна из возможных мер его аномальности обозначалась C_i. Аномалии явно связаны с солнечной активностью, причем почти наверняка механизм связи состоит в возникновении частиц с высокой энергией. Можно ожидать, что данные за день обладают значительным сходством (для разных дней). То же можно утверждать и для любых 2 дней, разделенных промежутком в 27 дней (это, грубо говоря, — период возникновения вихрей в солнечной атмосфере).

В целях прогноза, а также желая иметь не слишком ярко выраженную структуру остатков, Макдональд и Вард исследовали регрессию

$$y = C_i \text{ для } (t + k)\text{-го дня}$$

по разным подмножествам носителей из

$$\begin{aligned} x_1 &= C_i \text{ для } t\text{-го дня}, \\ x_2 &= C_i \text{ для } (t - 1)\text{-го дня}, \\ x_3 &= C_i \text{ для } (i - 2)\text{-го дня}, \\ &\vdots \\ x_{98} &= C_i \text{ для } (t - 97)\text{-го дня}. \end{aligned}$$

Они использовали шаговый регрессионный анализ с некоторым правилом останова. Эта процедура применялась к нескольким пятилетним наборам данных, содержащим 1826—1827 значений t.

Для $k = 1$, т. е. предсказывая на день вперед, они получили по данным 1915—1919 гг. следующее уравнение:

$$y = 0{,}3746 + 0{,}570x_1 - 0{,}174x_2 + 0{,}113x_{27} + 0{,}071x_{54} - \\ - 0{,}063x_{86} - 0{,}057x_6.$$

Здесь x_1 и x_2 — это C_i за предыдущие два дня. Их коэффициенты определяют некоторую устойчивость (C_i) как по величине, так и по направлению изменения. Два следующих члена (x_{27} и x_{54}) определяют связь

с предшествующими периодами возмущений. Носители x_{86} и x_6 не имеют ясного смысла.

Результаты по четырем пятилетним периодам сравниваются в илл. 15.2.1. Наблюдается заметная устойчивость формы регрессионного уравнения для всех периодов. Из этих результатов видно, что в регрессию разумно включать наблюдения: а) за несколько предыдущих дней, б) 27-дневной или близкой к этому давности, и, может быть, в) 54-дневной или близкой к этому давности.

В данном случае мы можем представить довольно веские обоснования для включения в регрессию той или иной переменной, что случается не слишком часто, особенно при применении шаговой регрессии. Хотя мы и не делали попыток интерпретации величин коэффициентов регрессии, надо сказать, что при столь определенном генераторе попытка интерпретации небезнадежна.

Регрессия с весами. Большинство программ шаговой регрессии относится к случаю равных весов. Важно, однако, отметить, что вес не создает дополнительных трудностей ни с теоретической, ни с практической точки зрения. Простые суммы квадратов (исходные и уменьшенные) заменяются в этом случае взвешенными. Несколько изменяется интерпретация F, но обычно не настолько, чтобы забраковать принятое критическое значение.

Использование весов не вносит значительных изменений в методику. Правда, надо изменить пункт (3) идеальных условий так:

3. *Пропорциональность дисперсий.* Для всех i произведение веса W_i на дисперсию ошибки i постоянно и равно σ^2.

Устойчивая шаговая регрессия. Об устойчивой шаговой регрессии нам известно немного, однако и здесь можно предложить достаточно «перспективную» технику. Простейшие из известных устойчивых приближений предполагают многократное изменение весов. Это соответствующим образом модифицирует шаговый метод. Предлагаемая нами процедура должна быть довольно надежной, хотя с точки зрения экономии машинного времени она не слишком хороша.

Эта процедура состоит в следующем:

1) выбираем веса (если хотим);

2) используя их, проводим обычную шаговую процедуру регрессии (с фиксированными весами) и подбираем генератор;

3) используя этот генератор итеративно, строим устойчивые приближения. В результате получаем: а) приближение, б) остатки, в) вторые веса, — произведения исходных и возникающих в ходе устойчивой итерации. Подробно изучаем остатки и устойчивые веса. Если они кажутся разумными, то

4) повторяем шаг (2), используя вторые веса;

5) переходим к шагу (3), находим третьи веса;

6) повторяем шаг (2), используя третьи веса;

7) переходим к шагу (4), получаем четвертые веса.

Число требуемых повторений определяется опытом исследователя.

15.3. МЕТОДЫ ВСЕХ ПОДМНОЖЕСТВ

Если у нас три носителя, то мы можем построить $8 = 2^3$ уравнений регрессии (включая вырожденное), а именно уравнения с носителями:

(без носителей)
x_1
x_1 и x_2
x_2
x_2 и x_3
x_3
x_3 и x_1
x_3 и x_1 и x_2.

Таким образом, мы рассмотрели все подмножества исходного множества носителей. Обычно для обеспечения этого ненормального упорядочения при переходе от одного подмножества носителей к другому срабатывает стандартный метод включения (добавление одного носителя) или исключения (отбрасывание одного носителя). Следовательно, можно было бы переделать программу шаговой регрессии так, чтобы она построила все восемь уравнений без лишних шагов.

Такого рода программа реализуется и в более сложных ситуациях. Так Дэниел и Вуд [Daniel C., Wood F. S. (1971)] сообщили о наличии (и доступности) аналогичной программы для 12 носителей. Эта программа может дать $4096 = 2^{12}$ уравнений. Так как обычно исследователю не нужны все 4096 уравнений, программа Дэниела и Вуда может (с помощью критерия C_p Маллоуза) отобрать из них несколько, чем обеспечивается гибкость регрессии.

Когда работает этот подход, нам нечего беспокоиться о качестве управления подгонкой, поскольку никакого управления здесь просто нет. (Пропустить что-либо существенное можно лишь в том случае, если не сработает критерий C_p, но это мало вероятно.) Примеры использования этой техники можно найти в упомянутой выше книге Дэниела и Вуда.

Алгоритм, предложенный Фернивалом [Furnival G. M. (1971)], позволяет распространить этот подход на 18 и даже 20 носителей. Еще более важным продвижением оказался метод Фернивала и Уилсона [Furnival G. M. and Wilson R. W. (1974)], позволяющий упростить обсчет m важнейших подмножеств с k носителями, для разных k. По-видимому, этот метод применим для генераторов, содержащих до 35 (а то и больше) носителей.

Использование весов. Как и ранее, введение любой фиксированной системы весов, в частности использование весов, зависящих от дисперсий, не вызывает дополнительных трудностей. Обычные (равновзвешенные) суммы квадратов и парных произведений заменяются взвешенными. Если возможно, стоит изучить зависимость остатков от весов (возможно, в форме «log |остатка|» от «log веса»). Желательно было бы иметь что-нибудь вроде достаточно пологой зависимости «2 log |остатка| + log веса» от «log веса». Это будет, грубо говоря, соответствовать последовательному выбору весов, обратно пропорцио-

нальных дисперсиям. Следует также рассматривать выделяющиеся наблюдения, отдельные интересные экземпляры и другие вызывающие трудности или, напротив, полезные аспекты регрессионных уравнений. (Это может относиться и к случаю приближения, включающего все имеющие отношение к делу носители.)

Устойчивые модификации. Переход от фиксированной процедуры метода наименьших квадратов к итеративному использованию взвешенного метода всех подмножеств не столь уж прост. До тех пор пока достаточно подробно не изучены все альтернативы, относительно надежная программа (для малого числа носителей и при наличии хороших устойчивых приближений) состоит в следующем:

1) построить модель сразу для всех носителей по устойчивой методике и найти окончательные веса;

2) исследовать окончательные веса и остатки;

3) если в результате исследования не будет обнаружено серьезных отклонений от «нормы», проводить взвешенную процедуру для всех подмножеств или процедуру Фернивала и Уилсона. Брать при этом окончательные веса, полученные в результате применения устойчивой методики (из пункта 1).

Перенос этой программы на случай весов, зависящих от дисперсий, осуществляется без особых затруднений.

15.4. КОМБИНИРОВАННЫЕ МЕТОДЫ

Какой план избрать, если число возможных носителей велико?

- Одно регрессионное уравнение со всеми носителями?
- Все возможные уравнения регрессии по всем подмножествам носителей?
- Шаговую процедуру по некоторым подмножествам носителей?

В тех случаях, когда ни один из этих подходов не кажется достаточно многообещающим, иногда прибегают к комбинированным методам.

Подробное изучение носителей часто делает возможной их следующую полезную классификацию.

- *Ключевые носители*: несколько (не более шести) носителей, которые мы считаем целесообразным включать во все уравнения.
- *Перспективные носители*: второе множество (скажем, до 12, а если используется процедура Фернивала — Уилсона, то и больше) носителей, заслуживающих особого внимания.
- «*Сорняки*»: разношерстный набор носителей, заслуживающих ограниченного внимания.

Если такое разбиение возможно, то и процедуру приближения можно разбить на несколько больших этапов.

1. Освобождение от ключевых носителей («снятие»). Освобождаем y, а также перспективные носители и «сорняки» от ключевых, пользуясь устойчивой процедурой. Всю дальнейшую работу проводим с остатками от этой регрессии. Исследуем, если возможно, распределение

остатков всех типов и решаем, достаточно ли снятых переменных для решения задачи. Если попалось несколько остатков, достаточно больших, чтобы специально выяснить причины таких аномалий, то выясним это и убедимся, возможно, в необходимости отбросить некоторые части данных, после чего анализ должен повториться.

2. Выбор перспективного подмножества из перспективного множества. После определения остатков отберем несколько (от 0 до 12) носителей (иных, чем на 1-м этапе), заслуживающих особого внимания. Применим к ним какой-нибудь метод всех подмножеств, например метод Фернивала и Уилсона. Выберем перспективное подмножество из этого множества в качестве основы для дальнейшего продвижения. (Если возникнут сомнения, добавим еще один-два носителя и освободим от всех них остатки y и оставшихся носителей.) Теперь действуем с новыми остатками, как на первом этапе.

3. Последовательный поиск среди «сорняков». Строим шаговую регрессию остатков y и x (полученных на втором шаге). Используем эту процедуру для отбора нескольких дополнительных носителей. Внимательнейшим образом изучаем остатки y, оперируя с ними, если надо, как на первом этапе.

4. Проверка всех подмножеств неключевых носителей. Для того чтобы решить, не пришло ли время остановиться, отберем 12 или более носителей из числа

- тех, которые были отобраны на первом этапе;
- тех, которые были отобраны на третьем этапе;
- тех, которые рассматривались, но не были отобраны.

Используя этот «табель о рангах», применим метод всех подмножеств (к остаткам, полученным на первом этапе).

Дать стопроцентную гарантию успеха нельзя. Но можно надеяться, что такая программа будет достаточно фундаментальной. Она сочетает преимущества метода шаговой регрессии с тем, что дают здравый смысл и проникновение в суть дела.

Взвешивание. Все сказанное можно распространить и на случай работы с любыми фиксированными весами. (Выше уже обсуждались те простые изменения, которые надо сделать для этого в методе всех подмножеств и в шаговом методе.)

Устойчивая подгонка. Здесь вполне естественно начать с устойчивой подгонки и переподгонки ключевых и перспективных носителей вместе, при этом на основе устойчивого метода появляются веса для дальнейшей подгонки и переподгонки этих носителей.

Теперь приступаем к шаговым вычислениям, как это описано в конце параграфа 15.2, используя все носители — ключевые, перспективные и включавшиеся до сих пор на различных этапах устойчивых переподгонок.

Можно поступить и иначе: ограничимся устойчивыми вычислениями только на первом и четвертом этапах. В этом случае нужно будет попытаться последовательно добавлять по одному невключенные носители (в рамках устойчивой процедуры).

15.5. ПЕРЕСТРОЙКА НОСИТЕЛЕЙ, СМЫСЛОВЫЕ КОМПОНЕНТЫ

Использование большого числа носителей неудобно по двум причинам:

- оно приводит к большому объему счета;
- включение в модель большего, чем нужно для приближения η, числа переменных приводит к росту дисперсии $\hat{y}$.

Можем ли мы избежать этих трудностей?

Если мы можем отбросить часть носителей, то наше положение, конечно, облегчится.

Если мы сумеем так видоизменить носители, что при сохранении общего их числа (может быть, и всего генератора) в окончательное уравнение регрессии войдет меньше носителей, то это не только сократит объем вычислений, но и уменьшит разброс $\hat{y}$.

Рассмотрим эту ситуацию подробней.

Видоизмененные носители. Часто мы сталкиваемся с тем, что множество носителей описывают практически одно и то же. Так, изучая модель идеального футболиста*, мы, по-видимому, будем склонны использовать в качестве носителей такие показатели, как:

1) рост;
2) расстояние от пола до концов пальцев поднятой правой руки;
3) расстояние от пола до концов пальцев поднятой левой руки;
4) длина ног;
5) длина более короткой руки;
6) длина более длинной руки;
7) длина правой кисти;
8) длина левой кисти;
9) длина пальцев правой руки;
10) длина пальцев левой руки.

Кроме того, мы, по-видимому, включим носители, описывающие скорость, опыт, вес, прыгучесть, честолюбие и т. д.

Пожалуй, не стоит и пытаться определить, какие из перечисленных размеров действительно важны. Все они сильно скоррелированы, так как в профессионалы попадают обычно крупные ребята. Особенно это будет заметно, если мы будем изучать игроков университетских команд.

Однако если мы все же хотим во что бы то ни стало построить математическую модель футболиста, то выбора нет.

Следует ли включать в модель все десять перечисленных факторов в том виде, в котором они приведены? Очевидно, нет. Парные корреляции между ними будут очень высоки. Можно попробовать вычислить логарифмы этих переменных и постараться как-то их скомбинировать.

Восемь из десяти переменных парные. Поэтому мы скорее всего выиграем, образуя их суммы и разности:

$$x_{23} = \frac{1}{2}(x_2 + x_3) \text{ и } x_2 - x_3 = x_{32},$$

* Имеется в виду американский футбол. — *Примеч. пер.*

$$x_{56} = \frac{1}{2}(x_5 + x_6) \text{ и } x_5 - x_6 = x_{65},$$

$$x_{78} = \frac{1}{2}(x_7 + x_8) \text{ и } x_7 - x_8 = x_{87},$$

$$x_{90} = \frac{1}{2}(x_9 + x_{10}) \text{ и } x_9 - x_{10} = x_{09}.$$

Далее, у нас шесть общих размеров x_1, x_{23}, x_4, x_{56}, x_{78}, x_{90}. Сделаем из них три: общий рост, длина руки и длина кисти:

$$x_{1234567890} = \frac{1}{6}(x_1 + x_{23} + x_4 + x_{56} + x_{78} + x_{90}),$$

$$x_{567890} = \frac{1}{3}(x_{56} + x_{78} + x_{90}),$$

$$x_{7890} = \frac{1}{2}(x_{78} + x_{90}).$$

Теперь есть смысл посмотреть, какой вклад в регрессию по этим трем носителям могут дать оставшиеся размеры.

Возможно, разумней вместо x_{567890} использовать x_{56}, так как если речь идет о вещах, где играет роль длина руки, то мы скорее теряем от включения в рассмотрение длины пальцев и кисти, чем выигрываем.

Во всяком случае мы можем, опираясь на свой «жизненный опыт» и «понимание проблемы», небезосновательно предположить, что использование комбинированных носителей для описания качества игры футболиста не приведет к увеличению общего числа носителей, а скорее всего позволит их сократить.

Если, например, мы удовлетворимся регрессией вида

$$1{,}74x_{567890} + 0{,}12x_{1234567890},$$

то на var $\{\hat{y}_i\}$, естественно, отразится (положительно) то, что мы приближали лишь два коэффициента. Хотя эта регрессия может быть переписана в виде

$$\frac{1{,}74}{3}(x_{56} + x_{78} + x_{90}) + \frac{0{,}12}{6}(x_1 + x_{23} + x_4 + x_{56} + x_{78} + x_{90}) =$$
$$= 0{,}02x_1 + 0{,}02x_{23} + 0{,}02x_4 + 0{,}60x_{56} + 060x_{78} + 0{,}60x_{90},$$

а затем

$$0{,}02x_1 + 0{,}01x_2 + 0{,}01x_3 + 0{,}02x_4 + 0{,}30x_5 + 0{,}30x_6 + 0{,}30x_7 + \\ + 0{,}30x_8 + 0{,}30x_9 + 0{,}30x_{10},$$

на var $\{\hat{y}_{(i)}\}$ это не отразится. Приближали мы только 2, а не 10 коэффициентов.

Конечно, можно пойти по другому пути. Скрупулезный анализ данных, предварительные приближения, шаговая регрессия, направленная на уменьшение дисперсий, могут привести к выбору удачной регрессии, похожей на предложенную выше. Но тогда *анализом будут управлять данные*, а не содержательные соображения. А исследователь должен «платить» за использование всех переменных: никакого умень-

шения дисперсии коэффициентов здесь не будет. Дисперсия будет больше, что, естественно, не облегчит нам понимания сути дела.

Смысловые компоненты. Нам нужно какое-то название для комбинаций носителей, построенных исключительно на основании содержательных, неформальных соображений. По аналогии с «главными компонентами», встреча с которыми нам предстоит в следующем параграфе, мы будем называть их смысловыми компонентами.

Найденные специалистами, глубоко понимающими суть проблемы, они могут быть гораздо полезней тех, которые получаются в результате механического анализа данных. Вообще, содержательные суждения могут оказаться весьма полезными (правда, далеко не всегда). Выбор смысловых компонент, как правило, основан на детальном анализе похожих ситуаций, встречавшихся раньше. А при этом не всегда работает «голая» интуиция, может потребоваться и грубый, а то и глубокий анализ данных (числового материала).

Нелинейные комбинации. Как мы отмечали выше, часто несколько носителей очень тесно связаны друг с другом (хорошим примером такого рода будут различные характеристики экономической ситуации). Если

- мы можем шкалировать эти носители так, чтобы они стали почти эквивалентными, и
- мы опасаемся каких-то неожиданностей (вроде нефтяного эмбарго), то эффективность некоторых из наших носителей резко снижается. Тогда имеет смысл сконструировать такие смысловые компоненты, которые были бы нечувствительны к ожидаемым локальным возмущениям. Для этого можно попробовать так организовать носители (в основном путем шкалирования), чтобы они стали почти идентичными. Получившуюся последовательность носителей можно рассматривать как выборку и представить ее одной величиной, причем медиана или бивес-среднее подойдут лучше, чем среднее. Эту величину можно взять в качестве первой смысловой компоненты. (Среднее обладает свойством, относительно которого трудно сказать, хорошее оно или плохое: а именно оно не дискриминирует ни одну из входящих в него величин, и это не всегда к лучшему.)

15.6. МЕТОД ГЛАВНЫХ КОМПОНЕНТ

Не всегда удается установить связи между носителями, рассуждая неформально и привлекая интуицию (как это было в примере с футболистами). Бывают ситуации, когда такие связи нужно устанавливать лишь на основе анализа экспериментальных данных*. Можем ли мы по генератору носителей определить ту их линейную комбинацию, которая вероятно, а может быть, и по невероятной случайности, дает хорошую регрессию еще не определенного y? Можно ли найти способ сокращения числа коэффициентов? Как и раньше, гарантированного

* Методу главных компонент посвящена обширная литература. За подробностями и дальнейшими ссылками читатель может обратиться, например, к работе: Дубров А. М. Обработка статистических данных методом главных компонент. М., Статистика, 1978. — *Примеч. ред.*

ответа здесь дать нельзя. Какой-нибудь недоброжелатель, зная наши x и наш план, всегда может предложить такой y, для которого наш выбор абсолютно бессмыслен. Но природа вряд ли действует таким образом, скорее, как писал Эйнштейн*, она «коварна, но не злонамеренна». Мы предлагаем метод, который часто бывает полезен.

Перспективным кажется подход, который называют обычно *«методом главных компонент»*.

Если мы расположим наши x естественным образом (подробности ниже), то различные линейные комбинации $\Sigma c_j x_j$ будут иметь разные дисперсии. Мы можем соотнести эти дисперсии с масштабом шкалы коэффициентов более точно

$$\frac{\operatorname{var}(c_1 x_1 + c_2 x_2 + \ldots + c_k x_k)}{c_1^2 + c_2^2 + \ldots + c_k^2}$$

или, еще точнее, наши оценки этих величин

$$\frac{\widehat{\operatorname{var}}(c_1 x_1 + c_2 x_2 + \ldots + c_k x_k)}{c_1^2 + c_2^2 + \ldots + c_k^2}.$$

Естественно предположить, что линейные комбинации с большей дисперсией дадут больший вклад в регрессию, чем те, у которых дисперсия меньше. Это особенно нормально, если найдутся такие комбинации, которые почти полностью состоят из «ошибок и флуктуаций». Комбинации с большими дисперсиями могут определяться некоторым общим фактором, который оказывает влияние и на y.

Компоненты. Предлагаемая методика состоит в следующем. Надо сравнить две квадратичные формы, а именно дисперсию взвешенной суммы некоррелированных x (измеряемых в одной шкале)

$$c_1^2 + c_2^2 + \ldots + c_k^2$$

и

$$\hat{\operatorname{var}}\{c_1 x_1 + c_2 x_2 + \ldots + c_k x_k\} = c_1^2 \hat{\operatorname{var}} x_1 + c_2^2 \hat{\operatorname{var}} x_2 + \ldots + c_k^2 \hat{\operatorname{var}} x_k + 2c_1 c_2 \hat{\operatorname{cov}}\{x_1, x_2\} + 2c_1 c_3 \hat{\operatorname{cov}}\{x_1, x_3\} + \ldots + 2c_{k-1} c_k \hat{\operatorname{cov}}\{x_{k-1}, x_k\}.$$

Последнее выражение — естественная оценка дисперсии, если не предполагать некоррелированности x.

Ясно, что такое сравнение даст последовательность наборов из коэффициентов

$$(c_{11}, c_{21}, \ldots, c_{k1}), (c_{12}, c_{22}, \ldots, c_{k2}), \ldots, (c_{1k}, c_{2k}, \ldots, c_{kk}),$$

для которых отношение

$$\frac{\text{первая квадратичная форма}}{\text{вторая квадратичная форма}}$$

изменяется монотонно.

---

* Здесь речь идет об афоризме, который придумал и часто употреблял А. Эйнштейн: «Der Herrgott ist raffiniert aber boshaft ist er nicht». Он многократно обсуждался и комментировался в литературе (см., например: Винер Н. Кибернетика и общество. М., Сов. радио, 1961). — *Примеч. ред.*

Для симметричных матриц A и B, представляющих в некоторой **координатной системе** соответствующие квадратичные формы, искомые **компоненты** определяются как решения уравнения

$$|A - \lambda B| = 0.$$

В данном случае мы можем взять

$$A = \begin{pmatrix} \widehat{\operatorname{var}} x_1 & \widehat{\operatorname{cov}}(x_1, x_2) \ldots \widehat{\operatorname{cov}}(x_1, x_k) \\ & \widehat{\operatorname{var}} x_2 \quad \ldots \quad \ldots \\ & \quad \ldots \quad \ldots \\ & \qquad\qquad \widehat{\operatorname{var}} x_k \end{pmatrix};$$

$$B = \begin{pmatrix} 1 & 0 & 0 & - & - & 0 \\ & 1 & 0 & - & - & 0 \\ & & 1 & - & - & 0 \\ & & & - & - & \\ & & & - & - & 1 \end{pmatrix},$$

где **элементы** ниже диагонали зеркально-симметричны относительно элементов, стоящих выше диагонали.

Решение (вручную) такой системы — дело утомительное, но, к счастью, сейчас широкодоступны программы, позволяющие перепоручить это компьютеру.

Выбор масштаба. В тех случаях, когда об x нет никакой информации, часто бывает удобно нормировать их так, чтобы

$$\operatorname{var} x_j = 1 \text{ для всех } j.$$

Если же мы можем оценить ошибки измерений x_j, то, пожалуй, лучше всего нормировать так, чтобы предполагаемая дисперсия измерений x_j была равна 1.

Поступив таким образом, мы будем рассматривать компоненты со слишком малой дисперсией (например, такие, для которых $\hat{\operatorname{var}}\{c_1x_1 + c_2x_2 + \ldots + c_kx_k\} \leqslant k$, или, быть может, $k/2$, или, наоборот, $2k$) с некоторым подозрением. Линейная комбинация, разброс значений которой можно полностью объяснить ошибками измерений, вряд ли даст нам много (исключение здесь может составить ситуация, когда все ее экстремальные значения сосредоточены в малом числе данных).

Величина, сравнимая с ошибкой измерения, редко может оказаться полезной в анализе. Если средняя сумма квадратов отклонений состоит в основном из ошибок измерений, то порождающий ее набор значений может оказаться полезным лишь в том случае, когда несколько значений намного больше остальных. Лишь относительно этих величин можно довольно уверенно утверждать, что ошибки измерения составляют небольшую часть их значений. И только они могут дать нам кое-что небесполезное.

Ковариации измерений. До сих пор молчаливо предполагалось, что

ковариации ОШИБОК ИЗМЕРЕНИЙ между $x_j \approx 0$. (*)

(Иначе нам пришлось бы указать значения этих ковариаций на внедиагональных местах матрицы B.) В тех задачах, где такая корреляция носит умеренный характер, как это часто бывает, мы можем обычно подобрать разумную аппроксимацию. Может, однако, случиться так, что она окажется неудовлетворительной.

Предположение (*), приведенное выше, существенно отличается от следующего:

НАБЛЮДЕННЫЕ ковариации между $x_j \approx 0$. (**)

Если мы со всей тщательностью независимо измеряли естественно коррелированные величины, то скорее всего предположение (*) оказалось бы выполненным, а предположение (**) — нет. По этой причине предположение (*) имеет лучшие шансы быть ближе к истине.

Малые компоненты. Если выполнено предположение (*), то мы можем быть более или менее уверены в том, что исключение из анализа малых компонент не нанесет нам большого вреда. Даже в тех случаях, когда мы не в состоянии оценить дисперсию ошибок измерений, некоторые компоненты могут оказаться столь малыми, что лучше их не рассматривать. В любом случае мы можем использовать метод главных компонент для уменьшения числа носителей.

Применения. Если человек, к которому мы хорошо относимся, собирается применить процедуру шаговой регрессии, имея по 20 носителей для описания, скажем, влияния семьи, школы и общества в целом, то следует предостеречь его от попытки начать анализ, имея 60 (20 + 20 + 20) носителей. Если все 60 компонент выглядят уместными и хорошо измеряются, то лучшее, что мы можем сделать для него, — это посоветовать начать с 1, 2 или 3 главных компонент из каждой группы носителей.

Если наш друг извлек все, что было возможно, из этих главных компонент и собирается попробовать комбинировать отобранные компоненты со всеми первоначальными, мы постараемся отговорить его, хотя и соглашаясь под давлением, что это лучше, чем начинать с исходного набора.

Если в каждом наборе из 20 компонент хорошими выглядят лишь некоторые, мы постараемся убедить его начать с главных компонент «хороших» поднаборов. (После того как это будет сделано, можно попробовать отобрать несколько наибольших компонент из каждого «хорошего» поднабора и совсем немного (наибольших) из каждого плохого.)

Факторный анализ. Специалисты в области психометрии и экономисты разработали ряд отличных от метода главных компонент приемов, позволяющих сужать генератор. Некоторые из этих приемов приспособлены для ручного счета, другие используют преимущества устойчивых методов. Большинство из них относится к первым шагам факторного анализа, т. е. процедуры, призванной выделять линейные комбинации, которые наилучшим образом могут служить базисом в суженном генераторе.

Если следовать параграфу 15.4. Если мы следуем общим рекомендациям этого параграфа и хотим уменьшить число носителей, применяя

метод главных компонент либо ко всем сразу, либо к отдельным группам, то можно облегчить ситуацию, разбив все множество носителей на четыре группы:

- носители (их немного, может быть, и ни одного), которые «необходимо» включить;
- носители (их несколько больше), которые следует использовать в методе «всех подмножеств»;
- носители (еще несколько большее число), которые будут использованы в «шаговой» процедуре;
- оставшиеся носители, которые исключаются из регрессии.

Только после этого разбиения на четыре группы, из которых лишь одна отбрасывается, мы можем по-настоящему приступить к сортировке носителей.

Поиски новых форм. Разделение на большие и малые компоненты может быть эффективным, но вряд ли средние по величине компоненты достаточно четко отделены от тех и других. Такие названия, как *«наибольшие компоненты»*, *«компоненты второго порядка величин»*, *«компоненты третьего порядка»*, могут не иметь реального содержания. Даже в тех случаях, когда анализируются большие массивы данных, наибольшие компоненты часто значительно флуктуируют *по направлению*, тем более остальные.

Если мы

- образуем новые линейные комбинации из нескольких самых больших главных компонент,
- опустим носители с относительно малыми коэффициентами,
- скорректируем коэффициенты с помощью хорошо найденных стандартных значений (см. работу [Green B. F.]), то такие преобразования больших компонент не могут значительно уменьшить их роль, но могут значительно облегчить проблему интерпретации. Поэтому такие преобразования бывают весьма полезными.

15.7. МНОГО ЛИ МЫ СМОЖЕМ УЗНАТЬ?

Цели этой главы ограничены разумным построением регрессии для описания и исключения. Мы постоянно предупреждали раньше и предупреждаем снова, что серьезное изучение коэффициентов в этой главе в принципе не входит в наши планы. Тем не менее мы хотим отметить некоторые ситуации, в которых коэффициенты могут привлечь наше внимание.

Устойчивые коэффициенты. В предыдущих параграфах было показано, что мы не можем быть уверены ни в знаке, ни в величине коэффициента b_j для носителя x_j, приближаемого индивидуально или вместе с другими, если приближение использует много похожих носителей x. Если все же мы предприняли попытку провести шаговую регрессию и обнаружили, что b_j мало меняется от шага к шагу с самого начала и до конца, то нам стоит поразмыслить о содержании, которое *можно* было бы придать этому коэффициенту.

Если такое постоянство имеет место лишь после того, как в регрессию включено определенное множество носителей, то может быть

полезной интерпретация этого коэффициента с учетом участия в регрессии упомянутого множества. Если такая интерпретация окажется связанной с сутью проблемы, имеет смысл обратить внимание на этот коэффициент.

СМЫСЛОВЫЕ ИЛИ ГЛАВНЫЕ КОМПОНЕНТЫ

Приятная ситуация, описанная выше, наиболее вероятна при использовании главных или смысловых компонент. Весьма вероятно, что им будет соответствовать набор коэффициентов, первый (или несколько первых) из которых довольно велик, а остальные — существенно меньше. Такая структура коэффициентов часто указывает на хорошо найденный набор коэффициентов, когда по крайней мере один из них потенциально интерпретируем.

Предупреждение. Если эти условия не выполнены, интерпретация коэффициентов может быть сколь угодно ошибочной.

15.8. НЕСКОЛЬКО y ИЛИ НЕСКОЛЬКО ЗАДАЧ?

В одном исследовании может быть несколько откликов или, наоборот, в нескольких исследованиях — один и тот же отклик, а возможно и то и другое. Тогда существует выбор — порознь или вместе рассматривать y.

Обрабатывая все вместе, мы строим регрессионные коэффициенты для каждого x_j в регрессии для каждого y, но добавляем (после наилучшего возможного масштабирования) сумму квадратов остатков или ее изменение. За этим исключением все остается, как прежде. Конечно, нужно иметь параллельные картины, чтобы сохранить возможность раздельного исследования в случае если потребуются разные преобразования y. (Можно ожидать, что отклик даст в разных исследованиях одну и ту же форму приближения для одних и тех же носителей. Если данные свидетельствуют об обратном, то надо изучить ситуацию очень внимательно.)

15.9. С ЧЕГО НАЧИНАЕТСЯ РЕГРЕССИЯ?

Зачем делать все заново для каждого нового набора данных, когда это неэффективно и бессмысленно? Нужны серьезные основания, чтобы планировать проведение некоторого конкретного анализа, не ориентируясь на результаты эксперимента. Подобная ситуация может все-таки возникнуть, например, ввиду каких-то юридических ограничений, но такая «смирительная рубашка» применяется редко. В большинстве приложений регрессии мы не контролируем x, и никогда метод анализа не должен влиять на результаты эксперимента. Поэтому в поисках информации, догадок, интуиции стоит изучать и прошлые и параллельные данные.

На начальной стадии эти источники могут обеспечить нас списком x, которые следовало бы рассмотреть. Довольно часто мы можем получить список возможных кандидатов в регрессию, а также перечень x, менее перспективных в этом смысле.

Далее, прошлое может подсказать, наилучшим ли образом мы составили «пристрелочную» модель,это во всяком случае полезно на этапе, предшествующем внимательному изучению данных.

Наконец, основываясь на прошлых данных, мы можем получить некоторые оценки коэффициентов в нашей регрессии, еще не приступив к анализу текущих данных. Если мы используем гибкий подход, где данные «сами» определяют, какие носители нужно включать в регрессию и какие стоит «оставить на обочине», то предварительная информация или предварительная теория (надежная, а может быть, и шаткая) могут дать нам прикидочные значения коэффициентов. Так, если мы ожидаем, что

$$y \sim \text{constant} + 17x_1 + ?x_2 + 5x_3 - ?x_4 + 3x_5,$$

то будет более разумно приближать

$$y - (17x_1 + 5x_3 + 3x_5),$$

чем y.

Предположим, что мы приблизили эту составную переменную моделью

$$(1{,}2 \pm 1{,}5)\, x_1 + (4 \pm 1)\, x_2 + (-3 \pm 2)\, x_3 + (7 \pm 3)\, x_4 + \\ + (1 \pm 0{,}1)\, x_5,$$

где числа после знаков $\pm$ есть оценки стандартных отклонений предшествующих им оценок коэффициентов. Это приближение можно упростить и ограничиться формой

$$4x_2 + 7x_4 + x_5.$$

или более точно $(4 \pm 1)\, x_2 + (7 \pm 3)\, x_4 + (1 \pm 0{,}1)\, x_5$.

Таким образом, естественным приближением y будет

$$y \sim 17x_1 + 4x_2 + 5x_3 + 7x_4 + 4x_5.$$

Здесь два члена возникли из априорной модели, два — в процессе приближения и один, пятый, который для большей ясности можно переписать в виде

$$4x_5 = x_5 + 3x_5,$$

составлен из априорной и подгоночной компонент.

При необходимости (а она возникает неизбежно) оценивать неопределенность этих коэффициентов разумной представляется следующая процедура: если мы заменяем $b \pm s_b$ на нуль, то стоит рассмотреть среднеквадратичную ошибку

$$b^2, \text{ если } |b| \geqslant |s_b|;$$

$$s_b^2, \text{ если } |b| \leqslant |s_b|.$$

Поэтому в нашем примере получаем

$$y \sim (17 \pm 1{,}5)\, x_1 + (4 \pm 1)\, x_2 + (5 \pm 3)\, x_3 + (7 \pm 3)\, x_4 + \\ + (4 \pm 0{,}1)\, x_5,$$

где ± 3 для x_3 — это $\pm b$, в то время как $\pm 1{,}5$, ± 1, ± 3 и $\pm 0{,}1$ — это $\pm s_b$.

Отметим, что на этой стадии мы оперируем только b и s_b компонентами, полученными в результате приближения, а не суммами «априорных» и приближенных коэффициентов. Если не проводить корректировку по x_1 и x_3, то приближение будет иметь вид

$$y \sim \text{constant} + (18{,}2 \pm 1{,}5)\, x_1 + (4 \pm 1)\, x_2 + (2 \pm 2)\, x_3 + \\ + (7 \pm 3)\, x_4 + (4 \pm 0{,}1)\, x_5.$$

Всюду выше мы неявно предполагали, что можно пренебречь корреляцией между оцениваемыми коэффициентами. Конечно, так бывает не всегда и даже, как полагают некоторые, не слишком часто. (Этот подход в его раннем варианте был сформулирован впервые Е. К. Харрингтоном (E. C. Harrington) в ноябре 1956 г. на симпозиуме в университете штата Северная Каролина.)

Резюмируя вышесказанное, можно отметить, что при оценивании коэффициентов часто небесполезно иметь в виду общефилософский принцип: новые достижения в науке опираются на старые.

15.10. ПРОИЗВОЛЬНАЯ КОРРЕКТИРОВКА

Иногда, занимаясь регрессией как средством отсеивания некоторых эффектов, мы получаем не удовлетворяющую нас точность оценок важных коэффициентов по имеющимся числовым данным. В такой ситуации стоит спросить, а нельзя ли предсказать эти коэффициенты, основываясь на интуиции, порожденной имеющимся опытом? Если да, давайте серьезно обсудим вопрос о корректировке наших y в соответствии с этими «произвольно» назначенными коэффициентами. Иногда действительно из теоретических или полукачественных соображений можно получить лучшие оценки коэффициентов, чем это можно сделать, объединяя все имеющиеся данные. В такой ситуации, естественно, не стоит использовать оценки коэффициентов, полученные по имеющемуся множеству числовых данных. Исключение составляют случаи, когда такие оценки явно отличаются от оценок, построенных на основе полуэмпирических рассуждений, теории или старых данных (см. пример в [Cox D. R. (1957)].

Отметим еще раз, что нет нужды в каждом новом исследовании пересматривать все доказанное ранее*.

РЕЗЮМЕ. УПРАВЛЯЕМАЯ РЕГРЕССИЯ

Один из возможных подходов к определению числа носителей, которые имеет смысл включать в регрессию, основан на минимизации

* Расставаясь с этой главой, заметим, что многие из поставленных здесь острых вопросов нашли отражение в отечественных работах по планированию эксперимента (см., например: Налимов В. В., Чернова Н. А. Статистические методы планирования экстремальных экспериментов. М., Наука, 1965). Возможность привлечения экспертного подхода к выбору генератора, не названная явно авторами, исследовалась, например в Адлер Ю. П. Введение в планирование эксперимента. М., Металлургия, 1969, а более общие аспекты выбора генераторов — в Адлер Ю. П. Предпланирование экспериментов. М., Знание, 1980. — *Примеч. ред.*

отношения

$$\frac{\Sigma (y-\hat{y})^2}{N \text{ данных МИНУС } N \text{ носителей}}.$$

Есть и другие, лучше исследованные, подходы, основанные на минимизации таких выражений, как C_p Маллоуза, PRESS Аллена или отношения

$$\frac{\text{средний квадрат}}{\text{число степеней свободы}}.$$

Один из методов уменьшения «платы» за подгонку большого числа коэффициентов состоит в образовании линейных комбинаций из этих носителей.

Эти комбинации могут составляться на основе чисто интуитивных соображений или на основе анализа детерминантных уравнений метода главных компонент.

Существует много компьютерных программ, которые позволяют применять шаговый метод подбора,состоящий в последовательном добавлении к регрессии или, наоборот, удалении из нее отдельных носителей.

Здесь известны следующие возможности:

- процедура расчета для построения устойчивого приближения;
- процедура Дэниела и Вуда — использования всех подмножеств (до 12 носителей);
- процедура Фернивала и Уилсона — использования лучших подмножеств из k (до 35 носителей);
- модификации этих процедур, использующие веса, компенсирующие дисперсии;
- процедура расчета для устойчивых модификаций процедур всех или лучших m из k подмножеств;
- комбинированный метод анализа данных при наличии многих возможных носителей; первый шаг этого метода состоит в разделении всех возможных носителей на а) ключевые, б) перспективные,в) «сорняки», а последующие шаги основаны на этом разделении;
- замена множества заметно коррелированных носителей их линейными комбинациями, причем число этих комбинаций не должно превышать числа исходных носителей, а отбор комбинаций основан на неформализованных соображениях;
- замена множества заметно коррелированных носителей линейными комбинациями с помощью метода главных компонент;
- комбинации последних методов с методом разделения на ключевые, перспективные и «сорняки»;
- исследование эволюции «данного» коэффициента в процессе шаговой регрессии;
- регрессионный анализ, начинающийся не с нулевых коэффициентов, а с наилучшим образом предсказанных *a priori* по отношению к исследованию экспериментальных данных.

БИБЛИОГРАФИЯ

Allen D. M. (1971). Mean square error of prediction as a criterion for selecting variables. — Technometrics, 13, 469-475.

Allen D. M. (1974). The relationship between variable selection and data augmentation and a method of prediction. — Technometrics, 16, 125—127.

Anderson R. L., Allen D. M., and Cady F. B. (1972). Selection of predictor variables in linear multiple regression. B: Bankroft T. A. (Ed.) Statistical Papers in Honor of George W. Snedecor. Ames, Iowa, Iowa University Press.

Anscombe F. J. (1967). Topics in the investigation of linear relations fitted by the method of least squares. — J. Roy. Statist. Soc., Series B, 29, 1—29.

Cox D. R. (1975). The use of a concomitant variable in selecting an experimental design. — Biometrika, 44, 150—158.

Daniel C. and Wood F. S. (1971). Fitting Equations to Data. New York, Wiley and Sons.

Furnival G. M. (1971). All possible regressions with less computation.— Technometrics, 13, 403—408.

Furnival G. M. and Wilson R. W., Jr. (1974). Regression by leaps and bounds. — Technometrics, 16, 499—511.

Green B. F., Jr. Parameter sensitivity in multivariate methods. Submitted for publication in Psychometrika.

Macdonald N. J. and Ward F. (1963). The prediction of geometric disturbances indices: 1. The elimination of internally predictable variations. — J. Geophys. Res., 68, 3351—3373.

Mallows C. L. (1973). Some comments on C_p. — Technometrics, 15, 661—675.

Tukey J. W. (1967). Discussion of Anscombe's paper. — J. Roy Statist. Soc., Series B, 29, 47—48.

ИЛЛЮСТРАЦИЯ

Иллюстрация 15.2.1

Шаговая авторегрессия для C_i (прогноз на 1 день вперед)

А. Члены, отвечающие ближайшим дням

Период	Регрессия
1915—1919	$0{,}570\ x_1 - 0{,}174\ x_2 - 0{,}057\ x_6$
1920—1924	$0{,}513\ x_1 - 0{,}0114\ x_2$
1925—1929	$0{,}586\ x_1 - 0{,}137\ x_2$
1930—1934	$0{,}527\ x_1 - 0{,}123\ x_2$

Б. Члены, отвечающие 27-дневному запаздыванию (первый солнечный вихрь)

1915—1919	$+0{,}113\ x_{27}$
1920—1924	$+0{,}110\ x_{26} + 0{,}140\ x_{27}$
1925—1929	$+0{,}102\ x_{26}$
1930—1934	$+0{,}181\ x_{26} + 0{,}109\ x_{27} + 0{,}070\ x_{28}$

В. Члены, отвечающие запаздыванию на 54 дня (второй солнечный вихрь)

1915—1919	$+0{,}071\ x_{54}$
1920—1924	0
1925—1929	0
1930—1934	$+0{,}056\ x_{54}$

Г. Другие члены

1915—1919	$-0{,}063\ x_{86}$
1920—1924	0
1925—1929	0
1930—1934	0

Глава 16 ● ИССЛЕДОВАНИЕ РЕГРЕССИОННЫХ ОСТАТКОВ

В параграфе 3.10 мы строили экономные графики зависимости остатков от x. Эти графики предназначены для оценки ситуации, окончательный результат с их помощью получить нельзя. А мы нуждаемся прежде всего именно в общей оценке после того, как подбор принятой модели по данным завершен. Нужно понять, сделано ли дело или следует дополнительно включить что-либо в регрессию, и если да, то что? К сожалению, на эти важные вопросы не может быть точных ответов.

Прежде всего полезно рассмотреть y, $\hat{y}$ и $y - \hat{y}$, а затем уже остальные переменные. Мы начнем с наиболее важных — y и $\hat{y}$.

16.1. ИССЛЕДОВАНИЕ $\hat{y}$

Пусть, рассматривая простейшую регрессию y по x и найдя bx, мы получили

$$\text{ave}\,(y - bx)^2 = s^2.$$

Чтобы выяснить, что осталось после этого приближения, построим график зависимости $y - bx$ от x или от bx. Строить график от y нет смысла, так как он может оказаться вводящим в заблуждение, что мы попытаемся продемонстрировать ниже.

Предположим, что y имеет большой разброс, как показано на левом чертеже илл. 16.1.1. В то же время bx имеет небольшой разброс вокруг константы c. Тогда

$$Y \equiv (y - bx) + bx \approx (y - bx) + c,$$

и точки $(y, y - bx)$ будут близки к прямой, определяемой точками $(t + c, t)$. Это изображено на правом графике илл. 16.1.1. Явный наклон и вытянутость точек вдоль прямой как будто бы указывают на возможность улучшить линейное приближение. Но это не так. Приближение, сколь бы плохим оно ни было, все же наилучшее из возможных. График зависимости $y - bx$ от x (центральный рисунок) демонстрирует заметный вертикальный разброс, но не обнаруживает никаких следов наклона, что и отражает адекватность выбора b и подходящие величины остатков.

Предположим, что мы допускаем замену bx на x. В этом случае мы будем строить графики зависимости $y - x$ от всех прочих. Если график зависимости $y - x$ от y будет иметь наклон, близкий к 1, то это свидетельствует о плохом качестве приближения, но не говорит нам,

правильно ли мы выбрали коэффициент $b = 1$. Наклон графика $y - x$ от x (равный $b - 1$) покажет, не следует ли нам вместо x использовать bx. Поэтому имеет смысл строить графики зависимости $y - \hat{y}$ от x, а не от y.

Следующий вопрос, которым важно задаться при исследовании модели $\hat{y}$, это вопрос о том, дает ли $\hat{y}$ максимум информации? Более точно, не может ли какая-нибудь функция от $\hat{y}$ приближать y лучше?

Если ответ положительный, то мы обычно оказываемся перед выбором:

- приближать y функцией от $\hat{y}$ либо
- преобразовывать y так, чтобы была возможность более или менее легко приблизить ее с помощью какого-то подмножества от генератора, по которому было построено $\hat{y}$.

Проиллюстрируем вышесказанное на примере.

Пример. Положение банковских депозитов. В таблице илл. 16.1.2 приведены данные о размерах просроченных платежей по ссудам для каждого из 16 кварталов за 1923—1927 гг. Размер невыплаченных ссуд подсчитывается по всем банкам Федеральной резервной системы США для одного из четырех округов: Бостон, Нью-Йорк, Филадельфия, Ричмонд. Округ в каждом случае выбирался наудачу. В нижней части таблицы приведено приближение, которое имеет вид

$$y = \text{линейная функция времени ПЛЮС эффект округа.}$$

Остатки приведены в правом крайнем столбце илл. 16.1.2.

В таблице илл. 16.1.3 приведены упорядоченные (по возрастанию) значения $\hat{y}$, а также результаты сглаживания рядов $\hat{y}$ и $y - \hat{y}$. После сглаживания $\hat{y}$ и $y - \hat{y}$ складываются и результат (сглаживаемый еще раз) можно испытать как возможную функцию от $\hat{y}$.

На илл. 16.1.4 мы видим, что график сглаженной разности $y - \hat{y}$ от $\hat{y}$ явно имеет вогнутую вверх форму. (Мы будем называть вогнутыми вверх кривые типа $\cup$, или $\supset$, или $\subset$, а выпуклыми вверх $\cap$, или $\subset$, или $\supset$.)

Такая форма зависимости подсказывает нам один из следующих вариантов приближения:

- использовать вогнутую вверх функцию $\hat{y}$ для приближения нашего y или
- приближать выпуклую вверх функцию y с помощью нового $\hat{y}$.

Одной из естественных вогнутых вверх функций y служит логарифм, а подходящей выпуклой вверх функцией $\hat{y}$ будет $ae^{b\hat{y}}$.

На илл. 16.1.5 изображены $h(\hat{y})$ и

$$1500e^{0,00019\hat{y}} = \hat{\hat{y}}.$$

Нас интересуют здесь прежде всего величины остатков. Процедура подбора, дающая конкретные коэффициенты a и b, интересует нас в малой степени, а поэтому мы приводим лишь окончательный результат. Он, несмотря на не слишком хорошее качество приближения, весьма обнадеживающий. В таблице илл. 16.1.6 приведены числовые значения функции $1500e^{0,00019\hat{y}}$ и соответствующие остатки. Там же приведены значения логарифмов y, результаты приближения и остатки от этого приближения. Само приближение приводится в нижней части таблицы. Результаты сглаживания обеих последовательностей остатков даны в таблице илл. 16.1.7. На илл. 16.1.8 изображен график остатков вида

$$y - 1500e^{0,00019\hat{y}} = y - \hat{\hat{y}}.$$

Он не обнаруживает никакой определенной структуры. График остатков вида

$$\log_{10} y - \widehat{\log_{10} y}$$

изображен на илл. 16.1.9.

Медиана этих остатков равна 0,003. Эту величину надо добавить к $\widehat{\log_{10} y}$, но, кроме того, график дает нам очень мало, так же, как и график остатков экспоненциального приближения. В графике зависимости остатков от $\widehat{\log_{10} y}$ на илл. 16.1.9 отсутствует заметная структура. Следует признать, наоборот, что график зависимости сглаженных $h(\widehat{\log_{10} y})$ от $\widehat{\log_{10} y}$ (илл. 16.1.10) близок к прямой линии. В левом нижнем углу виден намек на изгиб.

Вывод, который мы можем сделать по результатам анализа данного примера, состоит в следующем:

исходные значения просроченных платежей аппроксимируются кривой вида $ae^{b\hat{y}}$, где $\hat{y}$ имеет вид

линейная функция данных ПЛЮС эффект округа,

можно также приближать непосредственно эту форму, используя log (просроченных платежей).

А если точек больше? Пример, который мы только что рассмотрели, включал 16 точек, т. е. достаточно мало, что позволило анализировать каждую точку в отдельности. Но уже для 40 точек попытка решить вопрос о качестве приближения по графику $(y, y - \hat{y})$ практически безнадежна.

В параграфе 13.10 мы, чтобы лучше изучать большие массивы данных, группировали значения остатков $y - \hat{y}$ в соответствии со значениями $\hat{y}$, суммировали данные по группам, проводили сглаживание, а затем строили графики. В таблице илл. 16.1.11 приведены результаты такой процедуры применительно к последнему примеру. Мы пытались при этом, во-первых, получать группы остатков примерно одного размера и, во-вторых, сделать разбиение $\hat{y}$ на группы равномерным. Как видно из таблицы, нам это не всегда удавалось. Полученный в результате график приведен на илл. 16.1.12. Из этого рисунка мы можем извлечь почти всю ту информацию, которую можно извлечь из данных таблицы илл. 16.1.7 с помощью довольно громоздкого анализа, приведенного в таблице илл. 16.1.8. Нужно, однако, отметить, что наш рисунок не отражает всплеска на конце. (Чтобы его сохранить, пришлось бы разбивать на слишком мелкие группы.)

Какие же методы анализа остатков $y - \hat{y}$ по $\hat{y}$ можно предложить? Грубо упорядочивая эти методы по трудоемкости, приведем следующий перечень.

1. Построение графика зависимости для экстремальных (10 максимальных и 10 минимальных точек).

2. Сглаживание данных с последующим построением графика (естественно, при большом массиве данных можно воспользоваться методом параграфа 3.10).

3. Комбинация (1) и (2).

4. Построение графика зависимости по всем точкам.

5. Комбинация (2) и (4).

6. Применение (2) с последующим построением графика отклонений от сглаженного.

Если точек достаточно много, применение (4) (самостоятельно или в (5)) нецелесообразно. Мы вообще не предполагаем часто использовать (4) и (5) ввиду их трудоемкости. Однако, если есть возможность (и желание) поработать, стоит обратиться к (6).

16.2. ПЕРЕМЕННЫЕ И ДРУГИЕ НОСИТЕЛИ

Мы часто преобразуем переменные, например заменяем t ($t > 0$) на $\log t$, или s на s^3 либо на $-1/s$. Какие свойства необходимы этим преобразованиям? Для нас это:

а) возможность вычислять значения одного выражения по значениям другого;

б) (обычно) монотонная зависимость между выражениями: если одно увеличивается, то увеличивается и другое, если одно уменьшается, то уменьшается и другое.

А что будет, если функция от переменной не есть преобразование*?

---

* В заданном выше смысле. — *Примеч. пер.*

Как правило, это приводит к какой-то свертке. Так, t^2 свертывает t, «перегибая» ось t в точке 0. Если

$$\theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right],$$

то $\cos\theta$ свертывает θ, перегибая его в точке 0, а $\sin\theta$ в точках $\frac{\pi}{2}$ и $-\frac{\pi}{2}$.

Если мы рассматриваем всю действительную ось t, то cos свертывает ее многократно, перегибая в точках, кратных π, а sin — в точках вида «$\frac{\pi}{2}$ + целое число раз π». В более общих ситуациях свертывание может включать растяжение или сжатие одной из частей свертки.

Помимо преобразований отдельных переменных, нас могут интересовать преобразования пар. Здесь дело обстоит сложнее. В частности, здесь, по-видимому, отсутствует такой удобный набор преобразований, какой существует для одномерных (степени, логарифмы, экспоненциальные функции). Пока мы не можем предложить простых и общих преобразований пар переменных.

Носители и переменные. Приближая квадратичный полином, мы вправе выбирать форму его записи. Так, мы можем рассмотреть

$$13 + 3t - 2t^2,$$

или, что эквивалентно,

$$18 + 0\,(t - 12) - (2t^2 - 3t + 5).$$

В первом выражении мы считаем носителями 1, t и t^2, а во втором — 1, $(t - 12)$ и $(2t^2 - 3t + 5)$. В более общей форме мы говорим о всех квадратичных полиномах от t с носителями 1, t и t^2, т. е. о множестве

$$\{\text{всех } a + bt + ct^2\},$$

или, что эквивалентно, о множестве

$$\{\text{всех } d + e\,(t - 12) + f\,(2t^2 - 3t + 5)\}$$

с носителями 1, $(t - 12)$ и $(2t^2 - 3t + 5)$.

Мы хотим связать генераторы или модели с переменными. Начнем со следующего примера:

$$\{\text{все } a + bt_1 + ct_1^2 + dt_2 + et_2^3\}.$$

Здесь носители 1, t_1, t_1^2, $t_2 + t_2^3$. Каждый из них определяется либо значением t_1, либо t_2 (1 определяется любым из них). В качестве переменных мы можем выбрать (t_1, t_2), или (t_1^3, t_2^3), или (t_1, t_2^3) (каждое получается простым преобразованием другого). Отметим, однако, что, например, (t_1, t_2^2) взять нельзя, если для t_2 допустимы отрицательные значения, так как t_2 не определяется тогда по t_2^2.

Из соображений простоты мы, по-видимому, выберем в качестве переменных t_1 и t_2. Этот выбор не обязателен, но скорее всего он окажется полезным.

При изменении генератора (или модели) естественно изменить и то, что мы рассматривали как переменные. В множестве

$$\{\text{все } a + bt_1 + ct_2\},$$

или полиноме

$$19 - 5t_1 + 25t_2,$$

в качестве переменных стоит рассматривать t_1 и t_2. В эквивалентном вышеприведенном множестве

$$\{\text{все } a + d(t_1 + t_2) + e(t_2 - t_1)\},$$

или соответственно полиноме

$$19 + 10(t_1 + t_2) + 15(t_2 - t_1),$$

разумно, хотя и не обязательно, считать переменными $t_1 + t_2$ и $t_2 - t_1$.

Есть ли здесь разница? В каком-то смысле это одно и то же.

В конце концов $(t_1 + t_2,\ t_2 - t_1)$ — очень простое преобразование (t_1, t_2). С другой стороны, между этими парами переменных может быть большая разница. Так, перефразируя известную мысль Линкольна* (Lincoln) и Барнума (Barnum), заметим , что мало кто из нас может думать о двух переменных одновременно. Поэтому с понятийной точки зрения преобразование, смешивающее переменные, нельзя считать тривиальным.

Зачем же мы преобразуем переменные? В основном для упрощения дела. Например, выражение

$$1000(t + s) - 0{,}004(t - s)$$

выглядит проще, естественней и понятней, чем

$$999{,}996t + 1000{,}004s.$$

Точно так же

$$\{\text{все } a + b(t + s) + c(t + s)^2 + d(t + s)^3 + e(t - s)\}$$

проще, чем

$$\{\text{все } a + b(t + s) + c(t^2 + 2st + s^2) + d(t^3 + 3t^2s + 3ts^2 + s^3) + e(t - s)\},$$

особенно, если первое записано в виде

$$\{\text{все } a + bu + cu^2 + du^3 + ev\},$$

где $u = t + s$, $v = t - s$.

Стереотипы мышления. Во всякой области, где проводятся сбор и анализ данных, складываются свои стереотипы мышления насчет того, что есть «переменная» или «переменные». Эти стереотипы меняются со временем и сами данные влияют на их изменение. В последние несколько десятилетий валовой национальный продукт, несомненно, служит переменной в глазах экономистов, бизнесменов и читателей газет.

---

* Это высказывание фигурировало в известной речи Авраама Линкольна, произнесенной им в Бостоне 8 сентября 1858 г. — *Примеч. ред.*

Хотя совсем не так давно требовались подробные объяснения по поводу того, как можно складывать совершенно различные вещи и что может означать результат. Другим примером полуторавековой давности может служить идея немецкого барона из Новой Англии, Бенджамина Румфорда (Benjamin Rumford), о разделении понятия «тепло» на два — «температура» и «теплотворная способность» (тем самым он ввел переменные, упрощавшие анализ данных и заложил основы современной термодинамики).

Разумное правило выбора переменных состоит в следующем: используйте популярные ныне переменные, пока не удается заметно упростить анализ с помощью новых, если же, напротив, последнее возможно, сделайте это.

В конце концов, у каждого свой взгляд на задачу, но какой из них лучше — иногда становится ясно лишь с течением времени.

Естественность переменных. Вопрос о том, какая переменная «естественна», зависит от наших взглядов на «мир». Например, комбинация

возраст матери ПЛЮС N живущих детей

вряд ли будет принята как естественная. С другой стороны, пара

возраст мужа $\pm$ возраст жены

выглядит наиболее естественной заменой раздельных переменных:

возраст мужа,
возраст жены.

Довод в пользу этого утверждения — популярность вопроса о разнице в продолжительности жизни мужчин и женщин.

Наиболее удобны в употреблении пары

возраст мужа,
возраст мужа — возраст жены,

или, может быть,

возраст жены,
возраст жены — возраст мужа.

Причем это не связано с алгебраической структурой. Мы, по-видимому, согласимся считать естественной переменную

N мальчиков ПЛЮС N девочек,

но вряд ли примем в этом качестве

N мальчиков МИНУС N девочек.

Нелинейный генератор. Заметим, что в обсуждавшихся до сих пор ситуациях носители можно получать как частные производные модели по соответствующим коэффициентам. Так, если

$$f(t) = a + bt,$$

то

$$\frac{\partial f}{\partial a} = 1, \quad \frac{\partial f}{\partial b} = t.$$

А что получится, если распространить эту идею на нелинейные генераторы или модели? Если, например, наш генератор

$$\{\text{все } ae^{-bt}\},$$

то, принимая в качестве носителей производные по параметрам, получим переменные

$$e^{-bt} \text{ (производная по } a),$$

$$-ate^{-bt} \text{ (производная по } b)$$

(или, что эквивалентно, e^{-bt} и te^{-bt}).

В общем случае для нелинейных генераторов этот метод дает зависимость переменных от значений генератора (в нашем примере от b).

Если, например, приближение имеет вид

$$13e^{-7t},$$

то носителями служат

$$e^{-7t} \text{ и } 13te^{-7t},$$

или, что эквивалентно,

$$e^{-7t} \text{ и } te^{-7t}.$$

16.3. СЛЕДУЮЩИЙ ШАГ: ВОЗВРАТ К СТАРОЙ ПЕРЕМЕННОЙ $t_{ст}$

Прогресс в анализе остатков облегчится, если действовать систематически. Поэтому предположим, что мы построили

$$\hat{y} = b_0 + b_1x_1 + \ldots + b_hx_h,$$

где каждый x — функция только одной переменной[1] из

$$t_1,\ t_2,\ \ldots,\ t_h \quad (h \leqslant k),$$

называемых $t_{\text{старое}}$.

Вполне возможно, что возникнут и другие переменные, каждую из которых мы будем называть $t_{\text{новое}}$.

Несколько примеров (не включающих, впрочем, переменные $t_{\text{нов}}$) могут оказаться полезными.

Пусть $k = 3$, $h = 1$, а

$$\hat{y} = 13 - 2t + 3t^2 - 4t^3,$$

где, очевидно, роль x играют 1, t, t^2 и t^3 и достаточно одной «старой» переменной t.

Другой пример с $k = 5$ и $h = 3$ дает модель

$$\hat{y} = 8 - 3t + 4t^2 - 3s + 7s^2 - st.$$

Здесь роль x играют 1, t, t^2, s, s^2, st, и требуются три переменные t, s и st. (Мы рассматриваем st как переменную лишь потому, что в принятой нами структуре каждый носитель зависит только от одной переменной.) Еще один пример с $k = 5$ и $h = 5$ дает приближение

$$\hat{y} = 4 + 3x_1 + 2x_2 + x_3 - 7x_4 - 6x_5,$$

[1] Есть или нет b_0 — не очень важно. Мы сохранили его потому, что на практике этот коэффициент обычно присутствует.

где носителями будут 1, x, x_2, x_3, x_4, x_5, а естественными t окажутся x_1, x_2, x_3, x_4, x_5.

С точки зрения $t_{ст}$. Как выяснить, получаем ли мы полную отдачу от какой-либо переменной, включенной в приближение? О какой бы из переменных t_1, t_2, ..., t_h ни шла речь, мы обозначаем ее $t_{ст}$. Вопрос о том, какова «отдача», зависит частично от того, как часто $t_{ст}$ представлено среди носителей нашей модели.

Если $t_{ст}$ представлено лишь однажды или, может быть, дважды, мы можем действовать аналогично тому, как мы поступали с $\hat{y}$. Упорядочивая $y - \hat{y}$ в зависимости от $t_{ст}$, а затем, сглаживая значения $y - \hat{y}$, мы надеемся получить ясную картину. При наличии большого числа точек может оказаться полезным до сглаживания перейти к медианам групп.

Рассмотренная процедура не зависит от того, как первоначально было представлено $t_{ст}$.

Если мы меняем выражение, то порядок и результаты сглаживания $y - \hat{y}$ не изменяются. Единственное, что может случиться с нашим графиком, — это излишняя скученность или, наоборот, разреженность точек. Если наш график действительно окажется слишком скученным, мы преобразуем $t_{ст}$ и построим новый чертеж, не делая перерасчета сглаженных $y - \hat{y}$. (Новый график нужен для того, чтобы избежать скученности.)

В каких ситуациях этот простой и прямой подход может оказаться эффективным? В основном если есть достаточно хорошо определенных точек для построения графика (непосредственно или по медианам групп). Если же число точек в три-четыре раза больше, чем число изменений направления систематического «движения», а возмущения этих точек малы по сравнению с систематическими составляющими, мы скорее всего сможем разобраться в ситуации. А когда на каждый интервал постоянного направления систематического движения приходится всего одна точка, мы не получим достаточной информации. Аналогично если на каждый такой интервал у нас приходится несколько точек, но их разброс велик по сравнению с систематической составляющей, мы также мало что сможем выяснить. В тех случаях, когда данные не богаты информацией, нужно нечто большее, чем просто сглаживание и построение графика.

Очередной носитель. Пусть нам надо извлечь из данных информацию о необъясненной зависимости от $t_{ст}$, что довольно часто случается, если $t_{ст}$ входит в модель несколько раз. В этом случае потребуется некий специальный носитель или даже несколько специальных носителей.

Пусть, например, $t_{ст} = t$ и

$$\hat{y} = a + bt_1 + ct_1^2 + dt_2 + et_3.$$

Может быть естественным добавление t_1^3 как следующего носителя, включающего $t_{ст}$. Сам по себе этот шаг может оказаться вполне разумным, но вряд ли стоит осуществлять его буквально. Так мы вряд ли получим что-нибудь существенное, строя зависимость $y - \hat{y}$ от t_1^3,

если даже t_1^3 действительно влияет на y. Дело в том, что t_1^3 — это преобразование t, а у нас мало надежд извлечь что-либо из графика зависимости $y - \hat{y}$ от t_1.

Опустим для простоты члены, содержащие t_2 и t_3. Тогда модель примет вид

$$\hat{y} = 1 + 23t_1 - 5t_1^2.$$

Если мы возьмем теперь в качестве нового носителя t_1^3, то новая модель должна стать такой:

$$\hat{y} = 1 + 23t_1 - 5t_1^2 + h\,(t_1^3 - j_2t_1^2 - j_1t_1 - j_0),$$

где выражение в скобках

$$x_{\text{корр}} = t_1^3 - j_2t_1^2 - j_1t_1 - j_0$$

есть результат «исключения» из t_1^3 ранее использованных носителей.

«Исключение» означает здесь приближение (тем же методом, например методом наименьших квадратов) t_1^3 формой вида $j_2t_1^2 + j_1t_1 + j_0$ и образование соответствующей разности. Эту процедуру иногда называют также *ортогонализацией*. Если корректируем приближение слагаемым, пропорциональным $x_{\text{корр}}$, то при исследовании остатков надо брать $x_{\text{корр}}$, а не t_1^3.

Вернемся теперь к исходному, чуть более сложному примеру, включающему t_2 и t_3. Мы должны, конечно, включить эти переменные и в $x_{\text{корр}}$, а именно задать его соотношением

$$x_{\text{корр}} = t_1^3 - \hat{t}_1^3,$$

где

$$\hat{t}_1^3 = j_2t_1^2 + j_1t_1 + j_0 + k_2t_2 + k_3t_3.$$

Таким образом, стандартная процедура, которая осуществляется, если $t_{\text{ст}}$ появляется более чем один-два раза, начинается следующими тремя шагами:

- выбор следующего носителя (или двух; может, например, случиться так, что появление t_1^3 не даст эффекта, а t_1^4 — даст);
- построение соответствующего $x_{\text{корр}}$ (или двух соответствующих $x_{\text{корр}}$) в процессе ортогонализации следующего носителя(ей) по отношению к уже включенным;
- упорядочение значений $x_{\text{корр}}$ и сглаживание упорядоченных $y - \hat{y}$ (либо эта же процедура, повторенная дважды: отдельно для каждого $x_{\text{корр}}$).

Возможно, что, найдя $x_{\text{корр}}$ и построив график зависимости 10 максимальных и 10 минимальных значений от $x_{\text{корр}}$, мы проясним ситуацию. Если же нет, то надо как-то сгладить данные и затем применить следующую схему, которая кажется необходимой и эффективной:

2) построить график по сглаженным данным;

3) построить графики зависимости экстремальных данных и сглаженных;

6) осуществить (2) и затем построить графики отклонений от этих сглаженных данных.

Дополнительные усилия, связанные с (6), часто окупаются.

Отметим, что мы не предлагаем строить график отклонений остатков от сглаженных данных по отношению к выбранному носителю. График таких отклонений даст нам мало информации о распределении явных флуктуаций в зависимости от t. Для ответа на этот вопрос нужен график зависимости от t, а не от x.

Пример. Подгонка к «чистой» экспоненте. В левой части таблицы илл. 16.3.1 приводятся 10 значений экспоненты для $t_{\text{ст}}$ (для равномерно распределенных значений последнего), значения квадратичного приближения, полученного методом наименьших квадратов, и соответствующие остатки. Графики зависимости этих остатков от $t_{\text{ст}}$ и x_t приведены на илл. 16.3.2 и 16.3.3 соответственно. Так как мы имеем здесь дело с чистыми данными, отвечающими очень гладкой кривой, то не возникает трудностей в расшифровке илл. 16.3.2 и 16.3.5. (На обоих рисунках пунктиром изображены результаты итеративного медианного сглаживания по трем соседним точкам; такое сглаживание называют также $3R$-сглаживанием. На илл. 16.3.3 изображены сглаженные остатки от квадратичного приближения, упорядоченные по $x_{\text{корр}}$.)

Что можно извлечь из этих таблиц и рисунков для улучшения нашего приближения? Из илл. 16.3.2 мы видим, что при $t_{\text{ст}}$, равном 0,8; 4,6 и 8,3, остатки нулевые. Это наводит на мысль использовать для улучшения приближения полином

$$P_3 = (t_{\text{ст}} - 0{,}8)\,(t_{\text{ст}} - 4{,}6)\,(t_{\text{ст}} - 8{,}3),$$

взятый с некоторым коэффициентом.

В экстремальных точках, а также в точках локального максимума и минимума $t_{\text{ст}}$, $y - \hat{y}$, P_3 и отношение $y - \hat{y}$ к P_3 принимает следующие значения:

$t_{\text{ст}}$:	0	2,5	7,0	9,0
$y-\hat{y}$:	—0,0064	0,0057	—0,0059	0,0070
P_3 :	—30,544	20,706	—19,344	25,256
Отношение:	0,00021	0,00028	0,00030	0,00028

Эти простые рассуждения должны подсказать нам, что: а) полином P_3, умноженный на некоторую константу, может быть хорошей добавкой к $\hat{y}$; б) в качестве первого приближения к этой константе разумно взять 0,00027. Последнее основано на величинах отношений $y - \hat{y}$ и P_3.

Обращаясь теперь к илл. 16.3.4, мы видим явную зависимость $y - \hat{y}$ от $x_{\text{корр}}$.

Взяв крайние точки, получим

$$\frac{(0{,}0070) - (-0{,}0064)}{(42) - (-42)} = 0{,}00016.$$

Поэтому разумно в качестве дополнения к $\hat{y}$ принять величину $0{,}00016x_{\text{корр}}$.

Очевидно, что в этом примере эффективно используется как зависимость от $t_{\text{ст}}$ (илл. 16.3.2), так и от $x_{\text{корр}}$ (илл. 16.3.3),причем последнее дает даже более сильную регрессию.

Пример. Подбор экспоненты с шумом. В средней части таблицы илл. 16.3.1 приведены величины возмущений для 10 значений, взятые из таблицы случайных чисел, квадратичное приближение по $t_{ст}$ и остатки. Так как в обоих примерах подгонка производится методом наименьших квадратов, то

приближение к (экспонента ПЛЮС шум)
ЕСТЬ
(приближение к экспоненте) ПЛЮС (приближение к шуму),

следовательно,

остатки для (экспонента ПЛЮС шум)
ЕСТЬ
(остатки для экспоненты) ПЛЮС (остатки для шума).

Остатки для экспоненты с шумом приведены в правой части таблицы илл. 16.3.1. График зависимости этих остатков от $t_{ст}$ изображен на илл. 16.3.4, а от $x_{корр}$ — на илл. 16.3.5.

Точки на илл. 16.3.4 не наводят сами по себе на мысль о каком-то систематическом поведении. Однако кривая, получающаяся после их 3R-сглаживания, показывает изменение знака вблизи точек 0,4, 5,2 и 7,8, достаточно близких к тем, которые были в случае чистой экспоненты.

Достаточно одного взгляда на илл. 16.3.5, чтобы убедиться в наличии явной зависимости $y - \hat{y}$ от $x_{корр}$ (сглаженная кривая лишь усиливает это впечатление).

Пример. Экспонента с удвоенным шумом. Продолжим наш процесс, увеличив шум. На илл. 16.3.6 и 16.3.7 приведены графики зависимости для величины

(экспонента) + 2 (шум)

от $t_{ст}$ и $x_{корр}$ соответственно. Из графика зависимости от $t_{ст}$ нельзя усмотреть ничего явственного, график же зависимости от $x_{корр}$ (в первоначальном и сглаженном вариантах) указывает на наличие зависимости, хотя и не так отчетливо, как на илл. 16.3.7.

Выводы по трем примерам. Приведенные примеры показывают, что кое-какие вопросы можно решить, используя графики зависимости $y - \hat{y}$ от $t_{ст}$, но в сложных задачах надо строить зависимости от соответствующего $x_{корр}$. Сглаживание делает графики более эффективными. В частности, в задачах средней сложности могут принести пользу сглаженные графики зависимости от $t_{ст}$; если же сглаживание не делать, то без зависимостей от $x_{корр}$ не обойтись.

Помимо того, что график по $x_{корр}$ лучше определяет наличие зависимости, он имеет и еще одно преимущество, а именно: коэффициент $x_{корр}$ (он такой же, как у $t_{ст}$) определяется непосредственно как наклон $y - \hat{y}$ относительно $x_{корр}$. Отметим, что даже при приближении полинома, как это делалось при анализе илл. 16.3.2, коэффициент для $t_{ст}$ не получается непосредственно из графика зависимости $y - \hat{y}$ от $t_{ст}$.

Если коэффициенты носителей, входящих в $x_{\text{корр}}$, зафиксированы и найден коэффициент c при $x_{\text{корр}}$, то новое приближение строится в виде

$$\hat{y} + cx_{\text{корр}},$$

причем $x_{\text{корр}}$ можно при нужде разложить по старым носителям.

16.4. ВВЕДЕНИЕ НОВОЙ ПЕРЕМЕННОЙ $t_{\text{нов}}$

Как нам быть, если переменная $t_{\text{нов}}$ играет важную роль?

Для переменной $t_{\text{ст}}$ у нас были две возможности:

- построить график зависимости непосредственно от $t_{\text{ст}}$;
- отобрать «следующий носитель», перейти к $x_{\text{корр}}$, а затем по нему построить график зависимости.

Было установлено, что второй подход может оказаться полезным там, где не «работает» первый.

Можно ожидать, что те же две возможности будут и в новой ситуации. Что же касается ответа на вопрос: «Работает ли график зависимости непосредственно по $t_{\text{нов}}$?», то с ответом на него не стоит торопиться. Рассмотрим прежде всего один пример, а затем вернемся к общим соображениям.

Пример. Доходы от железных дорог. В таблице илл. 16.4.1 приведены данные за 20 лет о государственных доходах США (y) за тонну груза, перевезенного по железным дорогам страны, и данные о средней протяженности грузоперевозок ($t_{\text{нов}}$ в милях). Там же приведены значения кубического приближения y по t (год) и соответствующие остатки. На илл. 16.4.2 изображена зависимость $y - \hat{y}$ от $t_{\text{нов}}$. Никакой явной тенденции здесь не усматривается. На илл. 16.4.3 построен график $y - \hat{y}$ от $x_{\text{корр}}$, полученного как остаток от $t_{\text{пов}}$ после корректировки на $t_{\text{нов}}$. Здесь видна явная регулярность, а следовательно, есть, что добавлять в регрессию.

Таким образом и здесь мы узнали намного больше из графика по $x_{\text{корр}}$.

ОБСУЖДЕНИЕ И КОММЕНТАРИИ

Почему $x_{\text{корр}}$ оказалось более полезной переменной? На илл. 16.4.4 дана картина изменения во времени $y - \hat{y}$, $t_{\text{нов}}$, $x_{\text{корр}}$, из которой мы можем извлечь много полезного. Приближение кубическим полиномом исключило «медленную» составляющую изменения y, т. е. $y - \hat{y}$ — осцилляция («быстрая» составляющая). Изучая $t_{\text{нов}}$, мы видим заметный всплеск, приходящийся на период второй мировой войны (1942—1945 гг.) и общий положительный тренд; $x_{\text{корр}}$, получающееся корректировкой на кубический полином, осциллирует примерно так же, как и $y - \hat{y}$. Наличие тренда в $t_{\text{нов}}$ делает его малопригодным в приближении $y - \hat{y}$, так как там тренд уже исключен $\hat{y}$. Мы получаем переменную, по которой имеет смысл строить регрессию $y - \hat{y}$ лишь после корректировки $t_{\text{нов}}$ на кубический полином по t, т. е. после построения $x_{\text{корр}}$.

Говоря иными словами, нет смысла добавлять величину, пропорциональную $t_{\text{нов}}$, к $\hat{y}$ без соответствующего изменения коэффициентов последнего.

В этом примере достаточно легко подметить связь между $t_{\text{нов}}$ и генератором $\hat{y}$ ввиду «медленного» характера изменений. Наоборот, связь между $t_{\text{нов}}$ и $y - \hat{y}$ такая же (с точностью до постоянного сомножителя), как и между $x_{\text{корр}}$ и $y - \hat{y}$ — функция «быстро» изменяется. Это отличительная особенность ситуации примера позволяет лучше понять суть дела. Не нужно, однако, думать, что «медленные» или «быстрые» флуктуации различают аналогичные зависимости всегда. В иных примерах могут быть другие отличия. Вот главное:

1) нет смысла добавлять к $\hat{y}$ величину $f \cdot t_{\text{нов}}$, не изменив соответствующим образом коэффициенты первого;

2) простейший способ произвести добавление «правильно» состоит в том, чтобы добавить $f \cdot x_{\text{корр}}$. Мы всегда можем это сделать, так как зависимость между $\hat{y}$ и $t_{\text{нов}}$ исключена в процессе построения $t_{\text{нов}}$. Таким образом, можно констатировать полезность графика $y - \hat{y}$ от $x_{\text{корр}}$.

Выше мы обсуждали возможности улучшения модели вида

$$\hat{y} = a + bt_1 + ct_1^2 + dt_2 + et_3$$

путем включения добавочных членов, зависящих от t_1.

При этом мы проявили достаточную осторожность, рекомендуя исключать из них:

- не только 1, t, t_1^2 (по соображениям, обсуждавшимся в параграфе 16.3),
- но также t_2 и t_3.

Теперь мы видим, что наша предосторожность была весьма разумной.

16.5. В ПОИСКАХ ДОПОЛНИТЕЛЬНЫХ МУЛЬТИПЛИКАТИВНЫХ ЧЛЕНОВ

Мультипликативная корректировка. До сих пор мы пытались лишь прибавлять что-либо к нашей модели. Теперь попытаемся умножить ее на что-нибудь. Для начала рассмотрим метод, который вряд ли сможет принести пользу. Если приближение уже достаточно хорошее, то переход от

$$\hat{y} \text{ к } \hat{y}(1 + u),$$

где u имеет, например, форму

$$c_0 + c_1x_1 + c_2x_2 + \ldots + c_kx_k$$

и относительно мало, в логарифмической шкале соответствует переходу от

$$\log \hat{y} \text{ к } \log(\hat{y}(1 + u)) = \log \hat{y} + \log(1 + u) = \log \hat{y} + (\log e) \times$$
$$\times \log_e(1 + u) = \log \hat{y} + (\log e)\, u,$$

где $e = 2{,}71828\ldots$ (Формула в таком виде остается верной при любом основании логарифма. В частности, если взять в качестве основания e, то второе слагаемое в правой части будет равно u.) Теперь посмотрим, что можно извлечь из зависимостей

$$\log y - \log \hat{y}$$

от $t_{\text{ст}}$ или $t_{\text{нов}}$.

К сожалению, сходство между $y - \hat{y}$ и $\log y - \log \hat{y}$ обычно достаточно велико, ввиду чего это преобразование не дает ничего нового.

Другой вариант. Совершенно другой характер имеет модификация

$$\hat{y}_{\text{med}} + (y - \hat{y}_{\text{med}})(1 + u),$$

где $\hat{y}_{\text{med}}$ — медиана значений $\hat{y}$. Эта модификация приводит к выражению

$$y - (\hat{y}_{\text{med}} + (\hat{y} - \hat{y}_{\text{med}})(1 + u)) = y - \hat{y} + (\hat{y}_{\text{med}} - \hat{y})\,u$$

с последующим изучением регрессии между $y - \hat{y}$ и $(\hat{y} - \hat{y}_{\text{med}})\,u$. Это новая для нас проблема, так как $\hat{y} - \hat{y}_{\text{med}}$ принимает как положительные, так и отрицательные значения с примерно равной частотой.

Значения $\hat{y}$, для которых мало $\hat{y} - \hat{y}_{\text{med}}$, можно отбросить, так как $(\hat{y} - \hat{y}_{\text{med}})\,u$ будет еще меньше.

Определим квартили $\hat{y}$ и отбросим все точки, попадающие между ними. Мы будем использовать здесь два подхода, а именно: будем заменять $y - \hat{y}$ на $Q_{\hat{y}}$ и $q_{\hat{y}}$, где

$$Q_{\widehat{y}} = \begin{cases} (y - \widehat{y})/(\widehat{y} - \widehat{y}_{\text{med}}) & \text{(в верхней и нижней четвертях } \widehat{y}), \\ \text{нуль} & \text{(в средней части } \widehat{y}); \end{cases}$$

$$q_{\widehat{y}} = \begin{cases} +(y - \widehat{y}) & \text{(в верхней четверти } \widehat{y}), \\ \text{нуль} & \text{(в средней части } \widehat{y}), \\ -(y - \widehat{y}) & \text{(в нижней четверти } \widehat{y}). \end{cases}$$

Последнее использует лишь знак $\hat{y} - \hat{y}_{\text{med}}$, а не величину этой разницы и поэтому немножко легче вычисляется, что часто существенно.

Исследование зависимости $Q_{\hat{y}}$ и $q_{\hat{y}}$ от старых и новых t потребует, конечно, здравого смысла.

Пример. Большие приливы в Гонолулу. В этом примере мы будем изучать предсказанные на 1969 г. уровни и периоды (начала) больших приливов в Гонолулу. Для того чтобы сделать данные более обозримыми, ограничимся большими приливами, выпадавшими на 2-, 4-, 6-, 8-, 10- 12- и 14-е дни января, февраля, марта, апреля, мая и июня. Это составит 77 приливов. Будем обозначать суточное время (начала) прилива — x, предсказанный уровень — y и «абсолютное» время — t.

Конечно, тщательный анализ приливов требует больше места и времени, чем заслуживает этот пример. В частности, было бы желатель-

но рассмотреть все 700 с лишним больших приливов за год (еще лучше рассмотреть несколько лет). Желательно было бы также обратить внимание на физику явления, что, вообще говоря, необходимо делать при изучении всякого физического явления. Однако наша цель — всего лишь изучение двух частных методов. Поэтому мы не вдаемся в изучение проблемы приливов глубже, чем это требуется читателю, чтобы понять, как использование величин q и Q помогает разобраться в массе вроде бы беспорядочных данных и выделить в ней важные эффекты, заслуживающие дальнейшего изучения.

С этой целью мы приводим в таблице илл. 16.5.1:

- предсказанные уровни прилива (y) в десятых фута (выше среднего минимального уровня низкой воды);
- суточное время (x) в десятых часа и в градусах (360° = 24 часа);
- абсолютное время (t), которое получается умножением на 1000 суммы номера месяца с частью дня и часа до (начала) прилива, выраженные в 1000-х долях 31-дневного месяца.

Таким образом, прилив начавшийся в 3 ч 48 мин утра (что записывается как 0,38, так как 48 мин составляют 0,8 ч) 2 января, произошел через

$$(2-1)+\frac{38}{240}=1{,}158 \text{ дня}$$

после начала января, так как январь наступает в полночь, предшествующую первому дню месяца. В долях 31-дневного месяца это даст

$$\frac{1{,}158}{31}=0{,}037,$$

а значит, абсолютное время начала прилива —

$$1000\,(1+0{,}037)=1037.$$

В этой системе счисления новогодняя полночь соответствует 1000, а не 0000.

Мы используем только доли 31-дневного месяца из-за ленности. Это означает, например, что 3,2 дня дают одинаковый вклад в абсолютное время независимо от месяца. (Это означает также, что в феврале не может, например, получиться 0,96. Мы полагаем, однако, что сделанные упрощения искупают эти сложности. При более тщательном анализе мы должны учесть разную продолжительность месяцев, что, по-видимому, принудит нас к счету в днях или долях года.)

В качестве первого шага мы разбили уровни в соответствии с градусом дня. Результаты разбиения на двадцатиградусные интервалы приведены в таблице илл. 16.5.2. На илл. 16.5.3 маленькими кружками изображены сглаженные медианы соответствующих групп. Основные особенности этого рисунка таковы:

- пики около 60° и 270° (= 180° + 90°);
- впадины в районе 170° (= 90° + 80°) и 360° (= 270° + 90°).

Попробуем грубо подогнать к этим данным форму из косинусов и синусов 2θ, так как они повторяются через каждые 180°. Любую линейную комбинацию cos 2θ и sin 2θ можно представить в виде константы, умноженной на cos 2 (θ + фазовый сдвиг). Эта функция имеет пик там, где 2 (θ + фазовый сдвиг) равно 0°. Данные не имеют пика при 0°, следовательно, фазовая константа не равна 0°. Примем 75° за фазовый сдвиг, тогда наш носитель — это cos 2 (θ — 75°). Пики сглаженных медиан достигают уровней 17 и 18, средние — 17,5, т. е. около 18, а минимальные значения — 12 и 7, в среднем 9,5, т. е. около 10. Таким образом, общий уровень 14, близкий к медиане 13,0 сглаженных медиан, служит хорошей стартовой точкой для приближения, которое, следовательно, должно иметь вид $14 + a \cos 2 (\theta - 75°)$. Расстояние между пиками и впадинами 18 — 10 = 8, а потому разумно взять $a = 4$. Таким образом, окончательный вид приближения $14 + 4 \cos 2 (\theta - 75°)$.

Значения этого приближения, так же как и результаты вычитания их из соответствующих сглаженных медиан, представлены в таблице илл. 16.5.2. График этих остатков изображен на илл. 16.5.4. Основная закономерность, которую мы здесь видим, — один пик и одна впадина. Используя ту же простую, но грубую технику, что и раньше, получим второе приближение в виде 6 cos (θ — 310°). Несколько завышенное значение в точке θ = 310° происходит от использования сомножителя 6, получающегося округлением 5,5. Значения этого второго приближения и полученные по нему и медианам 20°-ных интервалов приведены в той же таблице илл. 16.5.2. Мы могли бы попытаться еще продвинуться на этом пути улучшения нашего приближения, которое имеет теперь вид

$$14 + 6 \cos (\theta - 310°) + 4 \cos 2 (\theta - 75°),$$

но предпочитаем перейти к анализу на основе Q и q.

Анализ на основе $q_{\hat{y}}$. В таблице илл. 16.5.5, в отличие от таблицы илл. 16.5.2, где уровень прилива классифицируется по средним значениям угла поворота Земли, приведены значения $\hat{y}$ для некоторых конкретных углов поворота.

В таблице илл. 16.5.6 значения $\hat{y}$ для этих заданных углов поворота представлены в хронологическом порядке (по абсолютному времени t). Классификация по углам здесь не производится. В этой же таблице приведены значения $y - \hat{y}$, $q_{\hat{y}}$ и $Q_{\hat{y}}$, последние определяются лишь для внешних четвертей области значений $\hat{y}$, т. е. для $\hat{y} < 9{,}8$ и $\hat{y} > 17{,}6$. Для облегчения вычисления $Q_{\hat{y}}$ приводятся соответствующие значения $\hat{y} - \hat{y}_{\text{med}}$.

Медиана $\hat{y}_{\text{med}}$ для $\hat{y}$ равна 15,2.

На илл. 16.5.6 изображен график $q_{\hat{y}}$ в зависимости от t. Шести месяцам соответствуют, как и должно, шесть групп. Даже беглого взгляда достаточно, чтобы заметить:

● общий тренд вверх;

● больший разброс слева (за исключением февраля).

Более того, ранним месяцам (январю, февралю, марту и апрелю) соответствуют весьма отчетливые месячные тренды.

Стандартные приемы работы с $q_{\hat{y}}$, $Q_{\hat{y}}$, q_t или Q_t используются не для того, чтобы обнаружить какую-то добавку, которая может улучшить приближение (для этой цели можно просто продолжать процесс приближения). Основная цель состоит в том, чтобы проверить перспективность направления исследования. Здесь (в примере) общая идея, до проведения $q_{\hat{y}}$-анализа, состояла в том, чтобы строить

● не слишком сложные зависимости $\hat{y}$ от t или Q от t.

Из графика $q_{\hat{y}}$ стало вполне очевидно, что это общее направление вполне разумно. Дальнейшее продвижение и уточнение направлений исследования теперь зависит от нас.

Из графика зависимости $q_{\hat{y}}$ от t следует, что мы можем улучшить наше приближение, комбинируя: θ — время дня и t — абсолютное время (за 1969 г.), что весьма удивительно, так как в соответствии с этим y зависит только от θ. Вряд ли окажется, что полезная комбинация имеет простую форму типа

$$d\,(\hat{y} - e)\,(t - t_{\text{med}}),$$

соответствующую прямой на илл. 16.5.6. Поэтому дополнительный член такого типа не будет лишним (для того чтобы проверить, насколько это полезно, можно построить график зависимости $Q_{\hat{y}}$ от y).

Прощальное замечание. Этот пример использовался для демонстрации некоторых приемов исследования сложных структур. Мы не делали попыток проанализировать данные этого примера так глубоко, как они того требуют.

Другие типы произведений. Для того чтобы оценить произведения носителей, один из которых уже использован в приближении, выберем соответствующее x_j, обозначим его x, найдем его квартили, отбросим точки, попавшие между ними, и образуем по оставшимся одну или обе величины:

$$Q_x = (y - \hat{y})/(x - x_{\text{med}});$$

$$q_x = (y - \hat{y})\ \text{sign}\ (x - x_{\text{med}}),$$

где

$$\text{sign}\, z = \begin{cases} +1, & \text{когда } z > 0; \\ -1, & \text{“} \quad z < 0; \\ 0, & \text{“} \quad z = 0. \end{cases}$$

Теперь можно исследовать видимое поведение Q_x или q_x так же, как мы бы это делали с $y - \hat{y}$, $Q_{\hat{y}}$ или $q_{\hat{y}}$.

Если мы захотим рассмотреть произведения, включающие носители, представляющие два $t_{\text{нов}}$, скажем $x_{\text{нов}}$ и $x_{\text{сн}}$, достаточно положить $x = x_{\text{нов}}$ и соотнести $Q_{x_{\text{нов}}}$ или $q_{x_{\text{нов}}}$ с $x_{\text{сн}}$.

16.6. В КАКОМ ПОРЯДКЕ?

Мы рассмотрели несколько возможных методов уменьшения остатков за счет усложнения приближающей модели. Нет смысла использовать все на каждом приближении, которое мы строим. Было бы хорошо иметь (1) общее правило, какой из методов применять первым, какой — вторым и т. д.; (2) общее правило, указывающее, когда же пора останавливаться.

На первый вопрос можно ответить довольно уверенно, что же касается второго, то здесь имеющийся опыт пока недостаточен.

Предлагается следующий порядок исследования:

1) $y - \hat{y}$ относительно $\hat{y}$;

2) $y - \hat{y}$ относительно $t_{\text{ст}}$, иногда относительно $x_{\text{корр}}$ (это, по-видимому, важно);

3) $y - \hat{y}$ относительно ортогонализованных $t_{\text{нов}}$, т. е. относительно $x_{\text{корр}}$ (мы настаиваем на важности этого);

4А[1]) $y - \hat{y}$ относительно $t_{\text{ст}}$ (часто в виде $x_{\text{корр}}$; не представляется важным);

4Б[1]) $Q_{\hat{y}}$ или $q_{\hat{y}}$ относительно $t_{\text{ст}}$ (часто в виде $x_{\text{корр}}$; представляется важным);

4В[1]) Q_t или q_t, где t важнее $t_{\text{ст}}$ и других важных $t_{\text{ст}}$ (часто в форме $x_{\text{корр}}$);

5А) $y - \hat{\hat{y}}$ относительно $t_{\text{нов}}$ (в форме $x_{\text{корр}}$);

5Б) $Q_{\hat{\hat{y}}}$ или $q_{\hat{\hat{y}}}$ относительно $t_{\text{нов}}$ (в форме $x_{\text{корр}}$; не представляется важным).

Хотя нет возможности снабдить читателя надежным «правилом останова», мы можем предложить некоторые «правила поведения» на разных этапах, а именно:

- если на стадии (1) обнаружена заметная зависимость, надо либо преобразовать y, либо включить эту зависимость в модель (затем можно начать все сначала);
- если при сравнении на стадиях (2), (3), (4А), (4Б) или (5А) обнаружено сильное сходство, надо провести переподгонку и все повторить;
- если на этих стадиях обнаружена при одном из сравнений умеренная или небольшая зависимость, следует закончить все индивидуальные сравнения на этой стадии прежде, чем решать, какой или какие несколько носителей добавить при переподгонке и перепроверке;
- если на стадиях (4В) или (5Б) обнаружена сильная или умеренная зависимость, следует, не задумываясь, произвести переподгонку, а затем все повторить;

[1] Порядок выбора между 4А, 4Б и 4В (или между 5А и 5Б) зависит от конкретных обстоятельств. Общих указаний по этому поводу мы дать не можем.

● если на какой-то из этих стадий обнаружена небольшая зависимость, следует серьезно подумать, что делать дальше.

Отметим, что

● $x_{\text{корр}}$ для какого-либо t ($t_{\text{ст}}$ или $t_{\text{нов}}$) то же, что и для $y - \hat{y}$, Q или q (то же верно и для решения о необходимости перехода к $x_{\text{корр}}$);

● $q_{\hat{y}}$ и q_t легко вычисляются (половина множества данных отбрасывается, а у половины оставшихся $y - \hat{y}$ меняется знак). Поэтому, выполняя (2), легко и полезно заодно сделать (4Б) для $q_{\hat{y}}$ и (4В) по крайней мере для некоторых q_t.

РЕЗЮМЕ. ИССЛЕДОВАНИЕ РЕГРЕССИОННЫХ ОСТАТКОВ

Изображение зависимости $y - \hat{y}$ от y может вводить в заблуждение.

Изображение зависимости $y - \hat{y}$ от $\hat{y}$ часто приносит пользу.

Существует широкий спектр возможностей выбора «переменных» при фиксированном генераторе и даже носителях.

При нелинейном подборе носители определяются как частные производные приближения по коэффициентам.

Полезны следующие меры:

● изобразить зависимость $y - \hat{y}$ от $\hat{y}$;

● сгладить эту зависимость и объединить ее со сглаженным $\hat{y}$ для того, чтобы получить $h(\hat{y})$-модель, которая, по предположению, лучше, чем $\hat{y}$;

● отразить наши знания о $h(\hat{y})$, преобразуя $\hat{y}$ или x;

● оперировать группами точек из узких интервалов значения y, если для этого хватает точек.

При анализе целесообразности дальнейшего использования переменной, уже представленной, скажем, двумя или более носителями, может потребоваться введение одного-двух «следующих носителей» с построением по ним регрессионных остатков.

Чтобы определить целесообразность введения нового носителя, имеет смысл построить график для $t_{\text{нов}}$ или $x_{\text{корр}}$. Здесь $x_{\text{корр}}$ — это $t_{\text{нов}}$, скорректированное по носителям, участвующим в модели. Можно ожидать, что график зависимости от $x_{\text{корр}}$ будет более полезным, особенно если $t_{\text{нов}}$ было соответствующим образом преобразовано перед построением $x_{\text{корр}}$.

В поисках дополнительных мультипликативных членов стоит построить графики зависимости $y - \hat{y}$ от $Q_{\hat{y}}$ или $q_{\hat{y}}$ (см. формулы в начале параграфа 16.4).

Для подбора других дополнительных мультипликативных членов имеет смысл строить графики зависимости $y - \hat{y}$ от Q_x и q_x, где x — хорошо подобранный носитель из числа уже участвующих в модели (см. формулы в конце параграфа 16.4).

ИЛЛЮСТРАЦИИ

Иллюстрация 16.1.1

Правый рисунок иллюстрирует вводящий в заблуждение характер зависимости $y - \hat{y}$ от y при слабой зависимости y от x

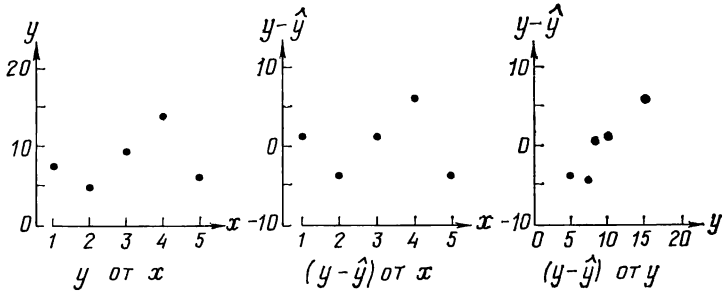


Иллюстрация 16.1.2

16 значений просроченных платежей по ссудам (по всем банкам Федеральной резервной системы США данного округа; в млн. дол.)

Округ	Обозначение и код квартала	Ссуда y	$*\hat{y}$	$y-\hat{y}$
Бостон	4К23 (—8)	3146	2740	406
Нью-Йорк	1К24 (—7)	8229	9251	—1022
Филадельфия	2К24 (—6)	1940	1624	316
Ричмонд	3К24 (—5)	1751	1269	482
Нью-Йорк	4К24 (—4)	9119	9578	—459
Бостон	1К25 (—3)	3487	3285	202
Ричмонд	2К25 (—2)	1804	1596	208
Филадельфия	3К25 (—1)	2294	2169	125
Филадельфия	4К25 (0)	2368	2278	90
Ричмонд	1К26 (1)	1873	1923	—50
Бостон	2К26 (2)	3796	3830	—34
	3К26 (3)			

Продолжение

Округ	Обозначение и код квартала	Ссуда y	*$\hat{y}$	$y-\hat{y}$
Нью-Йорк	4К26 (4)	10976	10450	526
Ричмонд	1К27 (5)	1829	2359	—530
Филадельфия	2К27 (6)	2509	2932	—423
Нью-Йорк	3К27 (7)	11731	10777	954
Бостон	4К27 (8)	4031	4484	—453

* Использованное приближение

$\hat{y}=109\times(N$ кварталов$)+$ { 3612 Бостон, 10014 Нью-Йорк, 2278 Филадельфия, 1814 Ричмонд.

Источник. Несколько ежегодных отчетов Федеральной резервной системы США.

Иллюстрация 16.1.3

Сглаженная зависимость $y-\hat{y}$ от $\hat{y}$

Упорядоченные $\hat{y}$	Сглаженные* $\hat{y}$	$y-\hat{y}$	Сглаженные $y-\hat{y}$	$h(\hat{y})$**	Сглаженные $h(\hat{y})$
1269		482		1751	
1596		208	316	1912	1832
1624		316	208	1832	1912
1923		—50	125	2048	
2169		125	90	2259	
2278		90	90	2368	
2359		—530	90	2449	
2740	То же	406	—423 90	2830	
2932		—423	202 —34	2898	
3285		202	—34	3251	
3830		—34		3796	
4484		—453		4031	
9251		—1022	—459	8792	
9578		—459		9119	
10450		526		10976	
10777		954		11731	

* Сглаживание скользящей медианой по трем текущим данным до стабилизации.

** $h(\hat{y})$=сглаженное $\hat{y}$ + сглаженное $(y-\hat{y})$.

Иллюстрация 16.1.4

Зависимость сглаженных $(y - \hat{y})$ от $\hat{y}$ из илл. 16.1.3

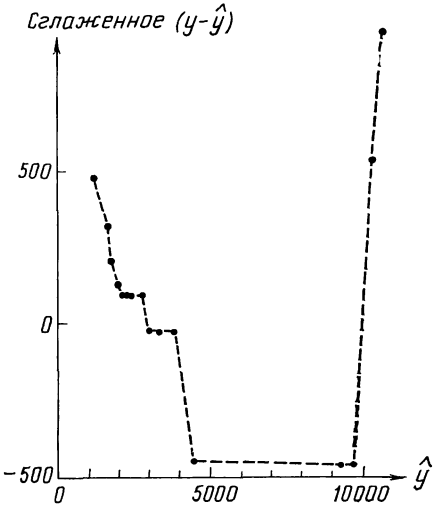


Иллюстрация 16.1.5

Зависимость $h(\hat{y})$ и 1500 $e^{0,00019\,\hat{y}}$ от y. Точки задаются парами $(\hat{y},\ h(\hat{y}))$

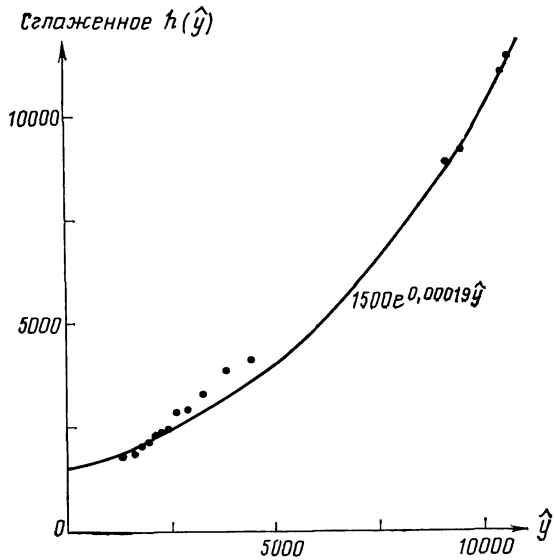


Иллюстрация 16.1.6

Экспоненциальное приближение $\hat{y}$ для y и линейное $\widehat{\log_{10} y}$ для $\log_{10} y$.

Округ	Квартал (N)	На основе исходных y				На основе $\log_{10} y$		
		y	$\hat{y}$	$1500\, e^{0,00019\hat{y}}$	Ост. $= y - ae^{b\hat{y}}$	$\log_{10} y$	$\widehat{\log_{10} y}$ *	$\log_{10} y - \widehat{\log_{10} y}$
Бостон	—8	3146	2740	2525	621	3,498	3,495	0,003
Нью-Йорк	—7	8229	9251	8699	—470	3,915	3,946	—0,031
Филадельфия	—6	1940	1624	2040	—100	3,288	3,312	—0,024
Ричмонд	—5	1751	1269	1909	—158	3,243	3,220	0,023
Нью-Йорк	—4	9119	9578	9256	—137	3,690	3,970	—0,010
Бостон	—3	3487	3285	2800	687	3,542	3,534	0,008
Ричмонд	—2	1804	1596	2025	—221	3,256	3,243	0,013
Филадельфия	—1	2294	2169	2265	29	3,361	3,350	0,011
Филадельфия	0	2368	2278	2312	56	3,374	3,358	0,016
Ричмонд	1	1873	1923	2162	—289	3,273	3,267	0,006
Бостон	2	3796	3830	3105	691	3,579	3,573	0,006
Нью-Йорк	4	10976	10450	10924	52	4,040	4,032	0,008
Ричмонд	5	1829	2359	2348	—519	3,262	3,298	—0,036
Филадельфия	6	2509	2932	2618	—109	3,400	3,405	—0,005
Нью-Йорк	7	11731	10777	11624	107	4,069	4,056	—0,013
Бостон	8	4031	4484	3516	515	3,605	3,621	—0,016

* Использованное приближение $\widehat{\log_{10} y} = 0,00786 \times (N \text{ квартала}) \text{ ПЛЮС } \begin{cases} 3,558 & \text{Бостон} \\ 4,001 & \text{Нью-Йорк} \\ 3,358 & \text{Филадельфия} \\ 3,259 & \text{Ричмонд} \end{cases}$

Иллюстрация 16.1.7

Сглаженные значения двух множеств остатков из илл. 16.1.6

По исходным y					По $\log_{10} y$					
Квартал	$\hat{y}$	Остаток	Сглаженный остаток		Квартал	$\widehat{\log_{10} y}$	Остаток	Сглаженный остаток	$h(\widehat{\log_{10} y})$	Сглаженные остатки $h(\widehat{\log_{10} y})$
—5	1269	—158			—5	3,220	0,023		3,243	
—2	1596	—221	—158		—2	3,243	0,013		3,256	
—6	1624	—100	—221	—158	1	3,267	0,006		3,273	
1	1923	—289	—100		5	3,298	—0,036	—0,024	3,274	
—1	2169	29			—6	3,312	—0,024		3,287	
0	2278	56	29		—1	3,350	0,011		3,361	
5	2359	—519	56	29	0	3,358	0,016	0,011	3,369	
—8	2740	621	—109	56	6	3,405	—0,005	0,003	3,408	
6	2932	—109	621		—8	3,495	0,003		3,498	То же
—3	3285	687			—3	3,534	0,008	0,006	3,540	
2	3830	691	687		2	3,573	0,006		3,579	
8	4484	515			8	3,621	—0,016		3,605	
—7	9251	—470	—137		—7	3,946	—0,031	—0,016	3,930	
—4	9578	—137			—4	3,970	—0,010		3,961	
4	10450	52			4	4,032	0,008		4,040	
7	10777	107			7	4,056	0,013		4,069	

$h(\widehat{\log_{10} y})$ = сглаженное $\widehat{\log_{10} y}$ + сглаженное $(\log_{10} y - \widehat{\log_{10} y})$

Иллюстрация 16.1.8

Зависимость сглаженных остатков модели $y - 1500\, e^{0,00019\hat{y}}$ от $\hat{y}$

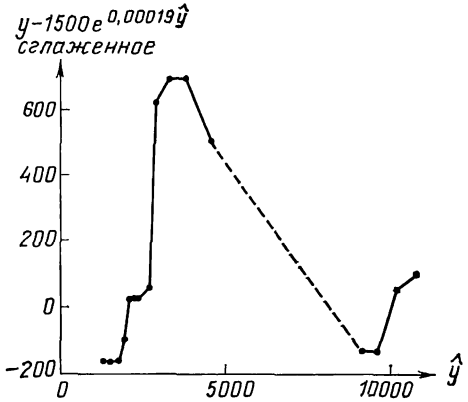


Иллюстрация 16.1.9

Зависимость сглаженных остатков модели $\log_{10} y - \widehat{\log_{10} y}$ от $\widehat{\log_{10} y}$

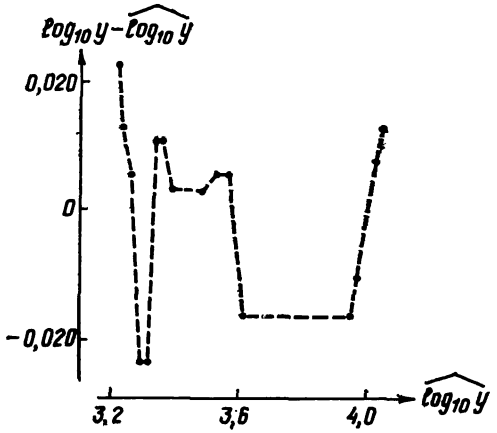


Иллюстрация 16.1.10

Сглаженная функция $h(\widehat{\log_{10} y})$ **от** $\widehat{\log_{10} y}$

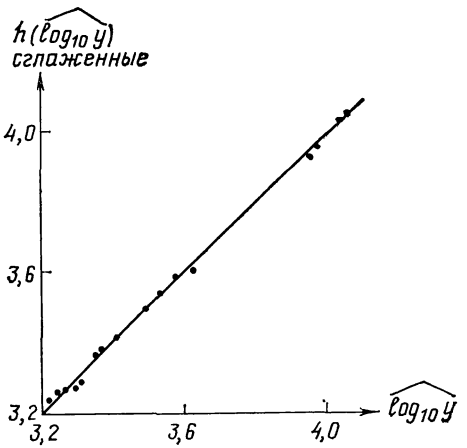


Иллюстрация 16.1.11

Сгруппированные остатки с медианами групп для исходного y **(из илл. 16.1.7)**

$\hat{y}$ (в сотнях)	Остаток $z = y - 1500\, e^{0,00019\,\hat{y}}$	Медиана остатков z'	Сглаженная медиана z'	Медиана $\hat{y}$ (в сотнях)
12—19	—158, —221, —100, —289	—190		**16**
21—29	29, 56, —519, 621, —109	29		25
32—38	687, 691	689	515	35
44	515	515		44
92—95	—470, —137	—304	80	94
104—107	52, 107	80		106

Иллюстрация 16.1.12

Сглаженные медианы z' остатков, сгруппированных для зависимости $\hat{y}$ от медиан $\hat{y}$ (в сотнях) из илл. 16.1.11

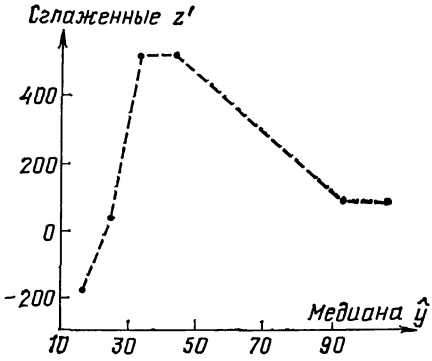


Иллюстрация 16.3.1

Остатки от приближения «чистой» экспоненты и экспоненты с наложенным шумом квадратичным полиномом и от $x_{\text{корр}}$, построенное по $(t_{\text{ст}})^3$

Экспонента[1]				Шум[3]				
$t_{\text{ст}}$	Исходн. значения	Квадратичное приближ.[2]	Остатки	Исходные значения	Квадратичное приближение[4]	Остатки	Эксп. плюс шум; остатки	$5\,x_{\text{корр}}$
0	1,000000	1,006394	—0,006394	—0,0035	—0,0006	—0,0029	—0,0093	—42
1	1,105171	1,103277	0,001894	0,0012	—0,0007	0,0019	0,0038	14
2	1,221403	1,216113	0,005290	0,0050	—0,0007	0,0057	0,0110	35
3	1,349859	1,344901	0,004958	—0,0069	—0,0007	—0,0062	—0,0012	31
4	1,491825	1,489642	0,002183	0,0051	—0,0007	0,0058	0,0079	12
5	1,648721	1,650336	—0,001615	—0,0056	—0,0006	—0,0050	—0,0067	—12
6	1,822119	1,826982	—0,004863	—0,0040	—0,0004	—0,0036	—0,0085	—31
7	2,013753	2,019581	—0,005828	0,0057	—0,0002	0,0059	0,0000	—35
8	2,225541	2,228132	—0,002591	—0,0020	0,0000	—0,0021	—0,0046	—14
9	2,459603	2,452636	0,006967	0,0008	0,0003	0,0005	0,0074	42

[1] Экспонентой мы называем $e^{\frac{t_{\text{ст}}}{10}}$.

[2] $\hat{y} = 0,0079763\,x^2 + 0,0889069\,x + 1,0063940$, причем в нормальных уравнениях $\sum x = 45$, $\sum x^2 = 285$, $\sum x^3 = 2025$, $\sum x^4 = 15333$, $\sum xy = 86,778199$, $\sum x^2 y = 589,15937$, $\sum y = 16,337995$.

[3] Шум взят из: «A Million Random Digits with 100 000 Normal Deviates» by the RAND Corporation, the Free Press, N. Y., 1955, p. 47 (последний столбец, два верхних блока).

[4] $\hat{y} = -0,00002\,x^2 - 0,0001\,x - 0,0006$.

[5] $x_{\text{корр}}$ здесь для удобства умножено на 5/3 (см.: Fisher R. A. and Yates F. (1953). Statistical Tables for Biological, Agricultural and Medical Research, 4th edition, London, Oliver and Boyd, Table XXIII, p. 80). Результаты ортогонализации $t_{\text{ст}}^3$ по 1, $t_{\text{ст}}$ и $t_{\text{ст}}^2$.

Иллюстрация 16.3.2

График зависимости $y - \hat{y}$ от $t_{ст}$ для чистой экспоненты по данным таблицы илл. 16.3.1. По сглаженным данным проведена пунктирная линия (там, где сглаженная точка отличается от исходной, стоят крестики)

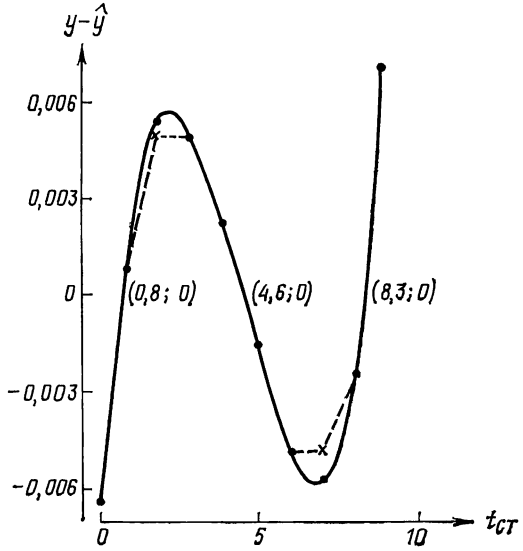


Иллюстрация 16.3.3

График зависимости $y - \hat{y}$ от $x_{корр}$ для чистой экспоненты. По сглаженным данным проведена пунктирная линия (там, где точки не совпадают, стоят крестики)

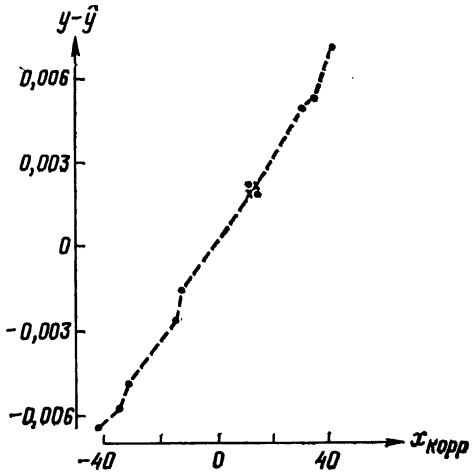


Иллюстрация 16.3.4

График зависимости $y - \hat{y}$ от $t_{ст}$ для экспоненты с шумом по данным таблицы илл. 16.3.1. Сглаженные данные соединены пунктирной линией

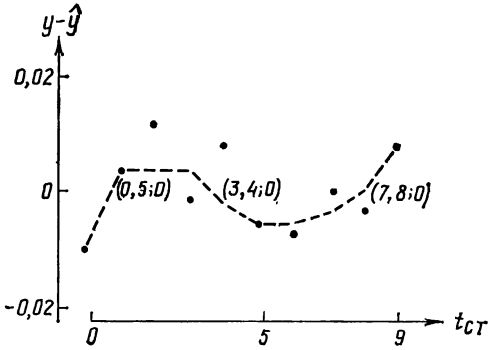


Иллюстрация 16.3.5.

График зависимости $y - \hat{y}$ от $x_{корр}$ для экспоненты с шумом по данным таблицы илл. 16.3.1. Сглаженные данные соединены пунктирной линией

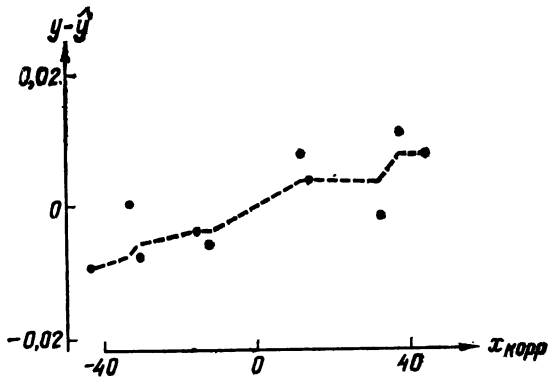


Иллюстрация 16.3.6

График зависимости $y - \hat{y}$ от $t_{ст}$ для экспоненты ПЛЮС удвоенный шум (по данным таблицы илл. 16.3.1). Сглаженные данные соединены пунктирной линией

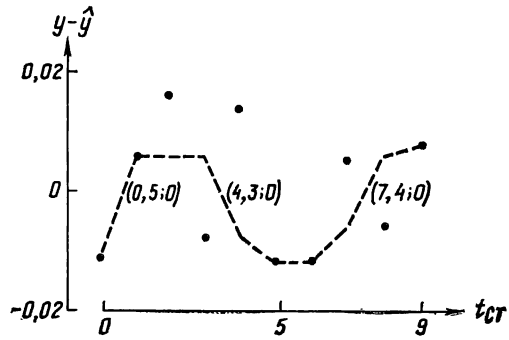


Иллюстрация 16.3.7

График зависимости $y - \hat{y}$ от $x_{корр}$ для экспоненты ПЛЮС удвоенный шум (по данным таблицы илл. 16.3.1). Сглаженные данные соединены пунктирной линией

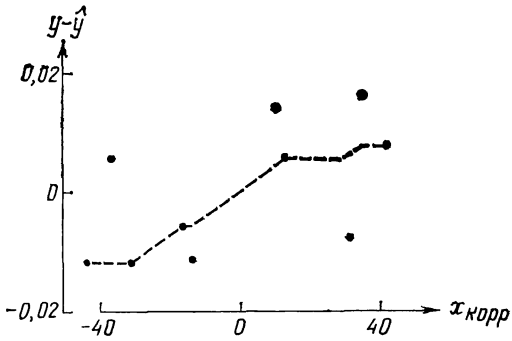


Иллюстрация 16.4.1

Доход за тонну груза, перевезенного по железным дорогам США (I, II и III класса), и средняя протяженность грузоперевозок

год t	Долларов за тонну			Миль грузоперевозки		
	фактическое y	кубическое приближение	остатки $y - \hat{y}$	фактическое $t_{нов.}$	кубическое приближение	остатки $x_{корр}$
1957	6,26	6,11	0,15	429,20	441,60	—12,40
1956	5,97	6,18	—0,21	428,08	429,60	—1,52
1955	5,95	6,18	—0,23	430,67	421,65	9,02
1954	6,19	6,14	0,05	431,65	417,18	14,47
1953	6,27	6,05	0,22	420,66	415,57	5,09
1952	6,16	5,92	0,24	426,93	416,25	10,68
1951	5,66	5,75	—0,09	419,99	418,60	1,39
1950	5,58	5,56	0,02	416,32	422,04	—5,72
1949	5,57	5,34	0,23	412,02	425,98	—13,96
1948	5,12	5,12	0,00	405,64	429,81	—24,17
1947	4,43	4,88	—0,45	407,82	432,95	—25,13
1946	4,10	4,64	—0,54	415,48	434,79	—19,31
1945	4,43	4,41	0,02	458,14	434,75	23,39
1944	4,53	4,19	0,34	473,28	432,23	41,05
1943	4,41	3,99	0,42	469,07	426,64	42,43
1942	4,02	3,81	0,21	427,76	417,37	10,39
1941	3,48	3,67	—0,19	368,54	403,84	—35,30
1940	3,35	3,56	—0,21	351,13	385,45	—34,32
1939	3,45	3,49	—0,04	351,21	361,61	—10,40
1938	3,54	3,48	0,06	356,05	331,72	24,33

Источник. y и $t_{нов}$ приводятся по «Historical Statistics of the United States, Colonial Times to 1957», U. S. Bureau of the Census, p. 431, ряды Q85 (y) и Q83 ($t_{нов}$).

Иллюстрация 16.4.2

График зависимости $y - \hat{y}$ от $t_{нов}$ для дохода от грузоперевозок (3 R — сглаженные данные соединены пунктирной линией)

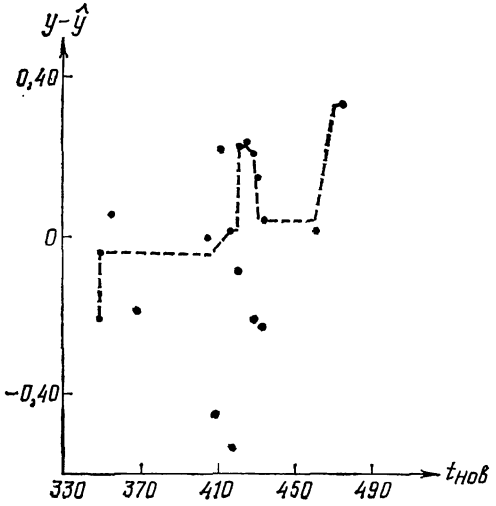


Иллюстрация 16.4.3

График зависимости $y - \hat{y}$ от $x_{корр}$ для дохода от грузоперевозок (3 R — сглаженные данные соединены пунктирной линией)

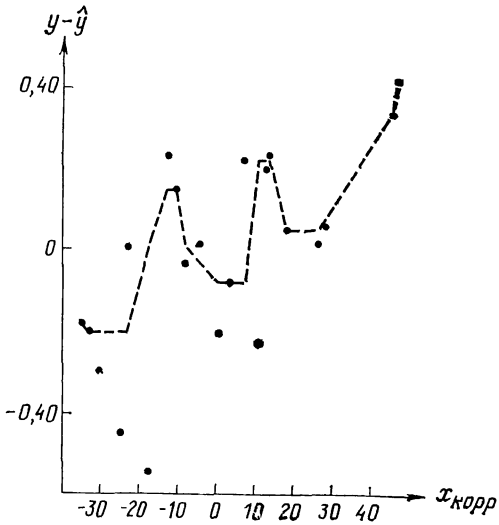


Иллюстрация 16.4.4

График зависимости $y - \hat{y}$ и $x_{корр}$ от года

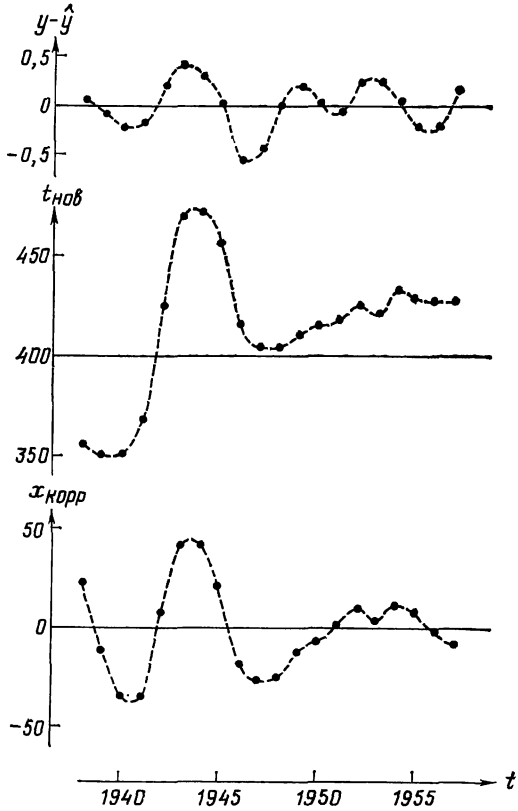


Иллюстрация 16.5.1

Предсказанные на 1969 г. большие приливы в Гонолулу (2, 4, 6, 8, 10, 12, 14 января, февраля, марта, апреля, мая, июня)

y уровень (*)	x время (**)	x время (***)	t абсол. время (****)	y уровень (*)	x время (**)	x время (***)	t абсол. время (****)	y уровень (*)	x время (**)	x время (***)	t абсол. время (****)
22	038	57°	1037	20	034	51°	3037	8	037	56°	5037
5	153	230°	1053	9	156	234°	3053	22	168	252°	5055
23	048	72°	1103	18	043	64°	3103	5	052	78°	5104

Продолжение

u уровень (*)	x время (**)	x время (***)	t абсол. время (****)	y уровень (*)	x время (**)	x время (***)	t абсол. время (****)	y уровень (*)	x время (**)	x время (***)	t абсол. время (****)
5	166	249°	1119	12	167	250°	3119	23	184	276°	5122
21	059	88°	1169	14	052	78°	3168	3	075	112°	5171
6	181	271°	1186	15	181	272°	3186	21	204	306°	5189
18	069	104°	1235	9	062	93°	3234	6	111	166°	5241
9	201	302°	1253	17	200	300°	3253	18	225	338°	5256
13	080	120°	1301	18	226	339°	3321	11	132	198°	5308
13	226	339°	1321	4	122	183°	3371	12	011	16°	5356
8	098	147°	1368	21	017	26°	3422	16	144	216°	5374
22	016	24°	1422	7	142	213°	3438	9	024	36°	5423
5	128	192°	1436	14	037	56°	4037	20	156	234°	5440
22	044	66°	2038	17	164	246°	4054	4	052	78°	6039
7	164	246°	2054	10	047	70°	4103	24	182	273°	6057
20	153	79°	2104	20	178	267°	4121	4	077	116°	6107
9	177	266°	2121	6	060	90°	4169	21	199	298°	6124
15	062	93°	2170	20	197	296°	4187	8	107	160°	6176
12	194	291°	2187	3	085	128°	4237	16	216	324°	6190
10	072	108°	2235	19	221	332°	4256	14	127	190°	6243
15	218	327°	2255	6	125	188°	4307	11	234	351°	6257
22	013	020°	2357	18	011	16°	4356	9	002	3°	6291
4	131	196°	2372	11	141	212°	4374	19	141	212°	6309
24	029	43°	2423	15	025	38°	4423	6	018	27°	6357
6	150	225°	2440	15	153	230°	4440	21	153	230°	6375
								5	032	48°	6424
								22	164	246°	6441

* Уровень выше минимального среднего уровня низкой воды в десятых фута.

** Время (24-часовое исчисление) в десятых часа.

*** Время (360° в день).

**** Абсолютное время в тысячных месяца (см. текст).

Иллюстрация 16.5.2

Уровень прилива (на таблицы илл. 16.5.1) относительно углового (градусного) времени

Код угла (*)	Средний угол	Уровень	Медиана уровня	Сглаженная медиана	(**)	Остаток$_1$ = сглаженная медиана — (**)	Сглаженный остаток$_1$	(***)	Остаток$_2$ = сглаженный остаток$_1$ — (***)
0	10°	18, 12, 9	12		11,4	0,6	1,9	3,0	—1,1
2	30°	22, 22, 21, 15, 9, 6	18	17	14,0	3,0	0,6	1,0	—0,4
4	50°	22, 24, 20, 14, 8, 5	17		16,6	0,4		—1,0	1,4
6	70°	23, 22, 20, 18, 14, 10, 5, 4	16		17,9	—1,9		—3,0	1,1
8	90°	21, 15, 9, 6	12		17,5	—5,5		—4,6	—0,9
10	110°	18, 10, 3, 4	6	8	15,4	—7,4	—5,5	—5,6	0,1
12	130°	13, 3	8		12,6	—4,6		—6,0	1,4
14	150°	8	8		10,5	—2,5	—3,1	—5,6	2,5
16	170°	6, 8	7		10,1	—3,1		—4,6	1,5
18	190°	5, 4, 4, 6, 11, 14	6	7	11,4	—4,4	—3,1	—3,0	0,1
20	210°	7, 11, 16, 19	14	12	14,0	—2,0	—2,6	—1,0	—1,6
22	230°	5, 6, 9, 15, 20, 21	12	14	16,6	—2,6		1,0	—3,6
24	250°	5, 7, 12, 17, 22, 22	14		17,9	—3,9	—2,6	3,0	—5,6
26	270°	6, 9, 15, 20, 23, 24	18		17,5	0,5		4,6	—4,1
28	290°	12, 20, 21	20	18	15,4	3,4		5,6	—2,2
30	310°	9, 17, 21	17		12,6	4,4		6,0	—1,6
32	330°	13, 15, 18, 19, 18, 16	17		10,5	6,5	4,4	5,6	—1,2
34	350°	11	11	12	10,1	1,9		4,6	—2,7
			(12)			(0,6)			
			(18)			(3,0)			
			(17)			(0,4)			

Примечание. Средний угол уменьшается на 0,5° при перенесении в следующий десятичный разряд.

---

* Код: 0 соответствует интервалу от 0° до 19°, 2 — интервалу от 20° до 39° и т. д.

** $= 14 + 4 \cos 2 (\theta - 75°)$.

*** $= 6 \cos (\theta - 310°)$.

Иллюстрация 16.5.3

Сглаженные медианы из таблицы илл. 16.5.2 и первое приближение θ первой модели относительно θ (график $14 + 4 \cos 2 (\theta - 75°)$ изображен пунктиром)

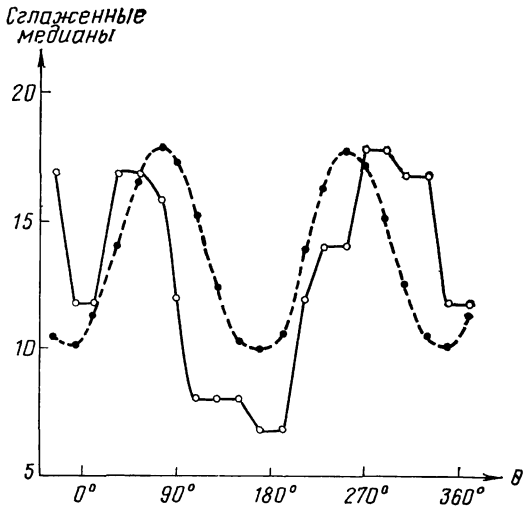


Иллюстрация 16.5.4

Первые остатки и вторая модель относительно θ (график θ (6 cos (θ—310°)) изображен пунктиром)

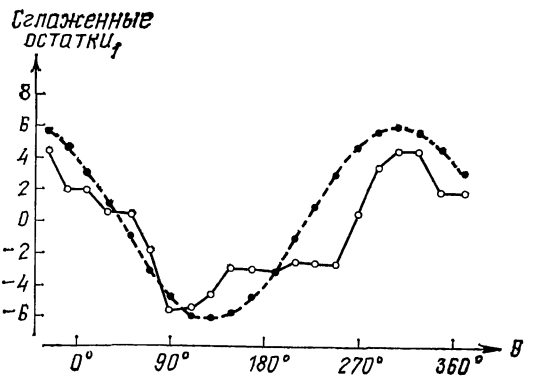


Иллюстрация 16.5.5

Расширение таблицы илл. 16.5.2 на отдельные углы из таблицы илл. 11.5.1

$\hat{y}$ для среднего угла*	$\hat{y}$ для кода**	Отдельные углы	t—абсолютное время	$\hat{y}$ для отдельных углов
14,4	14,6	16°, 16°, 3°	4356, 5356, 6291	14,7; 14,7; 14,6
15,0	14,7	24°, 20°, 26°, 38°, 36°, 27°	1422, 2357, 3422, 4423, 5423, 6357	14,8; 14,7; 14,9; 15,2; 15,2; 14,9
15,6	15,3	57°, 43°, 51°, 56°, 56°, 48°	1037, 2423, 3037, 4037, 5037, 6424	15,2; 15,3; 15,2; 15,2; 15,2; 15,3
14,9	15,2	72°, 66°, 79°, 64°, 78°, 70°, 78°, 78°	1103, 2038, 2104, 3103, 3168, 4103, 5104, 6039	14,7; 14,3; 15,1; 14,2; 15,1; 14,6; 15,1; 15,1
12,9	13,9	88°, 93°, 93°, 90°	1169, 2170, 3234, 4169	12,9; 12,3; 12,3; 12,6
9,8	11,4	104°, 108°, 112°, 116°	1235, 2235, 5171, 6107	10,8; 10,1; 9,5; 8,8
6,6	8,2	120°, 128°	1301, 4237	8,2; 7,2
4,9	5,8	147°	1368	5,6
5,5	5,2	166°, 160°	5241, 6176	5,7; 5,2
8,4	7,0	192°, 196°, 183°, 188°, 198°, 190°	1436, 2372, 3371, 4307, 5308, 6243	9,2; 10,0; 7,6; 8,5; 10,3; 8,8
13,0	10,7	213°, 212°, 216°, 212°	3438, 4374, 5374, 6309	13,7; 13,5; 14,4; 13,5
17,6	15,3	230°, 225°, 234°, 230°, 234°, 230°	1053, 2440, 3053, 4440, 5440, 6375	17,2; 16,3; 18,0; 17,2; 18,0; 17,2
20,9	19,2	249°, 246°, 250°, 246°, 252°, 246°	1119, 2054, 3119, 4054, 5055, 6441	20,7; 20,2; 20,8; 20,2; 21,2; 20,2
22,1	22,5	271°, 266°, 272°, 267°, 276°, 273°	1186, 2121, 3186, 4121, 5122, 6057	22,0; 22,2; 22,0; 22,2; 21,8; 21,9
21,0	21,6	291°, 296°, 298°	2187, 4187, 6124	20,6; 20,2; 20,0
18,6	19,8	302°, 300°, 306°	1253, 3253, 5189	19,6; 19,8; 19,1
16,1	17,4	339°, 327°, 339°, 332°, 338°, 324°	1321, 2255, 3321, 4256, 5256, 6190	15,5; 16,7; 15,5; 16,2; 15,6; 16,0
14,7	15,4	351°	6257	15,3
(14,4)	(14,6)			

* = (**) ПЛЮС (***) из таблицы илл. 16.5.2=14+6 (θ−310°)+4 cos 2 (θ−75°), при θ=средний угол.

** Среднее между соседними значениями в столбце (соответствует $\hat{y}$ в точках θ = 0°, 20°, . . .).

Иллюстрация 16.5.6

Значения y (из таблицы илл. 16.5.2), $\hat{y}$ и $y - \hat{y}$ для отдельных абсолютных времен (из таблицы илл. 16.5.5); значения $q_{\hat{y}}$ и $Q_{\hat{y}}$

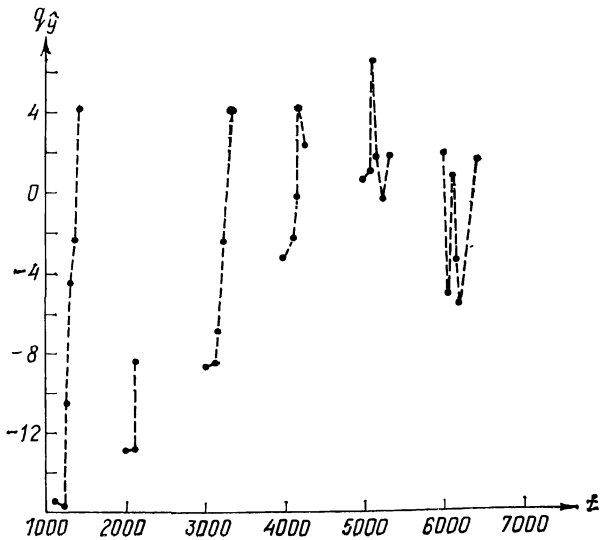
t абсолютное время	y уровень	$\hat{y}$	$y - \hat{y}$	$q_{\hat{y}}$	$\hat{y} - \hat{y}_{med}$	$Q_{\hat{y}}$
1037	22	15,2	6,8			
1053	5	17,2	—12,2			
1103	23	14,7	8,3			
1119	5	20,7	—15,7	—15,7	5,5	—2,8
1169	21	12,9	8,1			
1186	6	22,0	—16,0	—16,0	6,8	—2,4
1235	18	10,8	7,2			
1253	9	19,6	—10,6	—10,6	4,4	—2,4
1301	13	8,2	4,8	—4,8	—7,0	—0,7
1321	13	15,5	—2,5			
1368	8	5,6	2,4	—2,4	—9,6	—0,2
1422	22	14,8	—14,8			
1436	5	9,2	—4,2	4,2	—6,0	0,7
2038	22	14,3	7,7			
2054	7	20,2	—13,2	—13,2	5,0	—2,6
2104	20	15,1	4,9			
2121	9	22,2	—13,2	—13,2	7,0	—1,9
2170	15	12,3	2,7			
2187	12	20,6	—8,6	—8,6	5,4	—1,4
2235	10	10,1	—0,1			
2255	15	16,7	—1,7			
2357	22	14,7	7,3			
2372	4	10,0	—6,0			
2423	24	15,3	8,7			
2440	6	16,3	—10,3			
3037	20	15,2	4,8			
3053	9	18,0	—9,0	—9,0	2,8	—3,2
3103	18	14,2	3,8			
3119	12	20,8	—8,8	8,8	5,6	—1,6
3168	14	15,1	—1,1			
3186	15	22,0	—7,0	—7,0	6,8	—1,0
3234	9	12,3	—3,3			
3253	17	19,8	—2,8	—2,8	4,6	—0,6
3321	18	15,5	2,5			
3371	4	7,6	—3,6	3,6	—7,6	0,5
3422	21	14,9	6,1			
3438	7	13,7	—6,7			
4037	14	15,2	—1,2			
4054	17	20,2	—3,2	—3,2	5,0	—0,6
4103	10	14,6	—4,6			
4121	20	22,2	—2,2	—2,2	7,0	—0,3
4169	6	12,6	—6,6			
4187	20	20,2	—0,2	—0,2	5,0	0
4237	3	7,2	—4,2	4,2	—8,0	0,5
4256	19	16,2	2,8			
4307	6	8,5	—2,5	2,5	—6,7	0,4
4356	18	14,7	3,3			
4374	11	13,5	—2,5			
4423	15	15,2	—0,2			
4440	15	17,2	—2,2			
5037	8	15,2	—7,2			
5055	22	21,2	0,8	0,8	6,0	0,1

Продолжение

t абсолютное время	y уровень	$\hat{y}$	$y-\hat{y}$	$q_{\hat{y}}$	$\hat{y}-\hat{y}_{med}$	$Q_{\hat{y}}$
5104	5	15,1	—10,1			
5122	23	21,8	1,2	1,2	6,6	0,2
5171	3	9,5	—6,5	6,5	—5,7	1,1
5189	21	19,1	1,9	1,9	3,9	0,5
5241	6	5,7	0,3	—0,3	—9,5	0
5256	18	15,6	2,4			
5308	11	10,3	0,7			
5356	12	14,7	—2,7			
5374	16	14,4	1,6			
5423	9	15,2	—6,2			
5440	20	18,0	2,0	2,0	2,8	0,7
6039	4	15,1	11,1			
6057	24	21,9	2,1	2,1	6,7	0,3
6107	4	8,8	4,8	—4,8	—6,4	—0,8
6124	21	20,0	1,0	1,0	4,8	0,2
6176	8	5,2	3,2	—3,2	—10,0	—3,2
6190	16	16,0	0			
6243	14	8,8	5,2	—5,2	—6,4	0,8
6257	11	15,3	—4,3			
6291	9	14,6	—5,6			
6309	19	13,5	5,5			
6357	6	14,9	—8,9			
6375	21	17,2	3,8			
6424	5	15,3	—10,3			
6441	22	20,2	1,8	1,8	5,0	0,4

Иллюстрация 16.5.6.

$q_{\hat{y}}$ из таблицы илл. 16.5.6 в зависимости от абсолютного времени t. Пунктиром соединены значения для дней одного месяца



● ЗАДАНИЯ ДЛЯ УПРАЖНЕНИЙ

Все задания упорядочены по главам и параграфам, а внутри параграфов пронумерованы. Так, 7.3.4 обозначает 4-е задание в параграфе 7.3.

Надстрочный индекс «к», например $2.2.4^{к}$, означает, что это задание стоит использовать для коллективного решения.

Звездочкой помечены задания повышенной трудности.

Буква В, как в 16.В.4, соответствует заданию для внеаудиторной работы.

А буквой С, как в 14.С.1, помечены те задания, которые непосредственно связаны с другими главами.

Иллюстрации для упражнений, например илл. 1 для упр. 1.5.3, приводятся рядом с самими заданиями. Кроме того, у нас есть еще несколько массивов данных в разделе, озаглавленном «Приложение для упражнений», который следует сразу за данным разделом.

ГЛАВА 12

12-1. Сравните и разграничьте понятия «соответствие», «независимость» и «причинность».

12-2. Каковы три вида идей, необходимых для обоснования концепции причинности?

12-3. Разграничьте математическое и статистическое представление о «зависимости».

12-4. На илл. 1 к упр. 12-4 даны среднегодовые уровни воды в озере Виктория-Ньянза (x) относительно некоторого фиксированного значения и числа солнечных пятен (y) за 1902—1921 гг. Нанесите эти данные на график и постройте линейную регрессию y от x. Предложите механизм, который мог бы объяснить наблюдаемое соответствие между x и y.

Иллюстрация 1 к упражнению 12—4

Среднегодовые уровни озера Виктория-Ньянза и числа солнечных пятен, 1902—1921 гг.

Год	x	y	Год	x	y
1902	—10	5	1912	—11	4
1903	13	24	1913	—3	1
1904	18	42	1914	—2	10
1905	15	63	1915	4	47
1906	29	54	1916	15	57
1907	21	62	1917	35	104
1908	10	49	1918	27	81
1909	8	44	1919	8	64
1910	1	19	1920	3	38
1911	—7	6	1921	—5	25

Источник. Shaw Napier, Sir. Manual of Meteorology, Vol. 1: Meteorology in history; Cambridge University Press, London, 1942, p. 284. Публикуется с разрешения издателя.

12.1.1. Каково первое значение слова «регрессия»?

12.1.2. Опишите несколько ситуаций, где стоит предпочесть среднему арифметическому какую-нибудь другую свертку.

12.1.3. Каково второе значение слова «регрессия»?

12.1.4. Каковы некоторые преимущества и недостатки каждого вида регрессии?

12.1.5. Постройте график зависимости времени послеоперационного лечения в клинике (y) и физического состояния в момент выписки (x) для данных о больных, оперировавшихся по поводу грыжи (данные содержатся в табл. 8 из приложения для упражнений в конце книги).

а. Найдите средний y для каждого физического состояния и нанесите эти средние на график. В каком смысле здесь можно говорить о регрессии? Хороша ли на Ваш взгляд такая свертка?

б. Повторите задание пункта а) для медиан и сравните результаты. Каким образом можно было бы использовать другие процентные точки для усовершенствования регрессии как инструмента свертки данных?

12.1.6. По точкам на графике из предыдущего упражнения проведите прямую. Каково уравнение этой линии? Интерпретируйте его. Каковы достоинства и недостатки данного вида свертки в сравнении с тем, что приведен в упр. 12.1.б?

12.1.7. На илл. 1 к упр. 12.1.7 приведено число айсбергов, наблюдавшихся к югу от Ньюфаундленда (x) и к югу от Большой отмели* (y) помесячно за весь 1920 г.

а. Постройте график y—x и проведите по точкам прямую. Напишите ее уравнение.

б. Найдите остатки и нанесите их на график в зависимости от x. Что Вы можете сказать об ошибках предсказания y как функции от x?

в. Как соотносятся с этой задачей понятия о «соответствии», «зависимости» и «причинности»?

Иллюстрация 1 к упражнению 12.1.7

Число айсбергов, наблюдавшихся помесячно к югу от Ньюфаундленда (x) и к югу от Большой отмели (y) за 1920 г.

Месяц	1	2	3	4	5	6	7	8	9	10	11	12
x	3	10	36	83	130	68	25	13	9	4	3	2
y	0	1	4	9	18	13	3	2	1	0	0	0

Источник. S h a w Napier, Sir. Manual of Meteorology. Vol. 2, p. 407, Cambridge University Press, London, 1942. Публикуется с разрешения издателя.

12.2.1. Перечислите пять различных применений регрессии.

12.2.2. Почему при построении регрессии для исключения нас не заботят коэффициенты? А как о них можно заботиться?

12.2.3. Почему важно помнить, что часто много разных групп факторов предсказывают примерно одинаково хорошо?

12.2.4. Регрессия как средство локального усреднения представлена на илл. 1 к упр. 12.2.4 для максимального числа детей, которых хотели бы иметь женщины, по данным национальных опросов (в США) за 1955, 1960 и 1965 гг. Меняются ли намерения женщин от 1955 к 1965 г.? Постройте регрессию по 14 точкам желаемого числа детей по годам опросов (закодированных:— 1,0, + 1; без свободного члена) и найдите ожидаемые числа детей после исключения временного тренда.

* Ньюфаундлендская банка. — Прим ред.

Иллюстрации к упражнению 12.2.4

Максимальное желаемое число детей

Возраст опрашиваемых	Год опроса		
	1955	1960	1965
20—24	3,20	3,07	3,26
25—29	3,32	3,46	3,53
30—34	3,32	3,49	3,58
35—39	3,16	3,35	3,59
40—44	—	3,46	3,54

Источник. Ryder N. B. and Westoff C. F. Reproduction in the United States 1965. Princeton University Press, Princeton, N. J. 1971, p. 42.

12.2.5. Постройте график зависимости послеоперационного времени пребывания в клинике (y) от возраста пациентов (x) для данных об операциях грыжи (табл. 8 из приложения для упражнений).

а. Проведите по точкам прямую и интерпретируйте ее уравнение. Найдите время пребывания, скорректированное на возраст.

б. Как Вы думаете, разумно ли пользоваться простой линейной регрессией для коррекции на возраст в диапазоне возрастов от 2 до 80 лет, как в этих данных?

12.2.6. Постройте подходящий график для зависимости затрат на исследования в ВВС США (y) от времени в годах (x) по данным о военных исследованиях США (табл. 10 из приложения для упражнений).

а. Проведите по точкам прямую и используйте ее уравнение для получения *подходящих «поправок на линейный тренд»*.

б. Нанесите на график «скорректированный y» по годам. Наблюдается ли разница по времени? Какого рода функция должна работать лучше?

12.3.1. Что означает $x_{2;1}$?

12.3.2. Что такое $y_{;1}$ — на словах и в виде формулы?

12.3.3. Что такое $x_{2;1}$?

12.3.4. Уточните, что такое $y_{;12}$ и $x_{1;25}$?

12.3.5. Подберите зависимость x_2 на $y_{;1}$, которая была бы эквивалентна (алгебраически) зависимости $x_{2;1}$ на $y_{;1}$. В чем преимущество последнего метода?

12.3.6. В параграфе 12.2 мы узнаем, что часто разные наборы факторов могут давать практически одинаковые модели. Объясните, почему мы часто обнаруживаем, что совсем разные наборы коэффициентов для одних и тех же факторов могут успешно приводить к практически одинаковым моделям?

12.3.7. Воспользуйтесь графическим методом из этого параграфа для построения зависимости возраста пациентов (x_1) (сначала отдельно) и физического состояния (вместе) (x_2) на время послеоперационного долечивания в клинике (y) по данным об операциях грыжи (табл. 8 из приложения для упражнений). (Если Вы уже решили упр. 12.2.5, то первый этап для Вас позади.) Сравните уравнения

$$y = b_0 + b_1 x_1$$

и

$$y = c_0 + c_1 x_1 + c_2 x_2.$$

12.3.8. Постройте зависимость накоплений от доходов и процентных ставок (долговременных *Aaa* по Муди), пользуясь набором данных об экономике США (табл. 4 из приложения для упражнений). Сохраните результаты, они еще пригодятся.

12.3.9. а. Какая из переменных (из данных Коулмена) оценивает устную речь учеников лучше всех? (Табл. 7 из приложения для упражнений.)

б. Какая еще переменная кажется наилучшей для работы вместе с той, что Вы уже выбрали? Подтвердится ли это ожидание?

12.4.1. Если мы хотим скорректировать y по z и располагаем несколькими мерами z, то почему бы нам не включить их все сразу в регрессию? (Мы еще вернемся к этому вопросу в параграфах 12.6 и 13.7, а также в гл. 15, включая и обсуждение композиций, использующих веса, выбранные экспертами.)

12.4.2. Как узнать, будут ли два несравнимых носителя коллинеарны? (См., например, упр. 12—4.) Значит, коллинеарность часто не вытекает из природы переменных.

12.4.3. Модель «наименьших квадратов» для y и x_1, x_2, x_3, x_4, x_5 из данных Коулмена (табл. 7 из приложения для упражнений) дается формулой

$$y = 19{,}9 - 1{,}79x_1 + 0{,}0432x_2 + 0{,}556x_3 + 1{,}11x_4 - 1{,}79x_5.$$

а. Попадаются ли какие-нибудь неожиданные коэффициенты? Почему они все-таки не неожиданны для данного генератора?

б. Какие переменные должны быть коллинеарными? Почему должны? Почему не должны? Замените группу коллинеарных переменных одним носителем и пересчитайте модель, пользуясь урезанным генератором. Подумайте, как сравнить остатки для этой модели с остатками для модели полного генератора.

12.4.4. Для экономических данных (табл. 4 из приложения для упражнений) положим: y — накопления; x_1 — доходы; x_2 — долговременные процентные ставки (*Aaa* по Муди); x_3 — долговременные процентные ставки (*Bbb* по Муди); x_4 — текущие процентные ставки.

Сравните модели:

а. $y = b_0 + b_1x_1 + b_2x_2$;

б. $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$;

в. $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$.

12.4.5. Очень давно экономисты нашли эмпирическое соотношение, которое мы выразили так:

$$\text{инфляция} = a + b\ (\text{безработица}).$$

Оно известно как «кривая Филлипса». В экономических данных (табл. 4 из приложения для упражнений) положим: y — инфляция (если P_t — индекс розничных цен в году с номером t, то $y = \dfrac{(P_t - P_{t-1})}{P_{t-1}}$); x_1 — общее число безработных; x_2 — безработные мужчины старше 20 лет; x_3 — безработные женщины старше 20 лет.

а. Для выделенного Вам носителя x_i и пары носителей (x_i, x_j) постройте регрессии

$$y = b_0 + b_i x_i$$

и

$$y = c_0 + c_i x_i + c_j x_j.$$

б. Сравните коэффициенты Ваших моделей.

12.4.6. Сравните результаты из упр. 12.4.5 с моделью вида

$$y = d_0 + d_1x_1 + d_2x_2 + d_3x_3.$$

12.4.7. Падение атмосферного давления сопряжено с плохой погодой. В данных табл. 6 из приложения для упражнений положим: y = (давление на метеостанции) — (вчерашнее давление); x_1 — относительная влажность; x_2 — туман; x_3 — осадки.

а. Для выделенных Вам носителя (x_i) и пары (x_i, x_j) постройте регрессии

$$y = b_0 + b_i x_i$$

и

$$y = c_0 + c_i x_i + c_j x_j.$$

б. Сравните коэффициенты этих моделей.

12.4.8. Сравните результаты упр. 12.4.7 с моделью вида

$$y = d_0 + d_1x_1 + d_2x_2 + d_3x_3.$$

12.4.9. Какие переменные в данных о древних военных конфликтах (табл. 12 из приложения для упражнений) кажутся коллинеарными? Набросайте план этапов, которые должны привести Вас к получению регрессионной модели для оценки продолжительности войн в месяцах.

12.5.1. Почему часто оказывается возможным, что два полиномиальных уравнения с совершенно различными коэффициентами одинаково хорошо описывают данные? Что из этого следует относительно экстраполяции за область варьирования данных?

12.5.2. Что случится, если в множественную регрессию включить данные с коллинеарными носителями? А что же надо делать?

12.5.3. Выясните, совпадают ли пожелания женщин в упр. 12.2.4. Если брать одинаковые поправки на годы опросов, то что Вы могли бы сказать о том: а) есть ли различия в возрастных группах женщин (1916—1920, 1921—1925, ..., 1941—1945) с точки зрения их пожеланий; б) остается ли тенденция в чаяниях для разных возрастов после исключения данных о возрастах?

12.5.4.* Найдите корреляцию между X и X^2, если X распределен равномерно на отрезке $0—A$.

12.5.5. Почему модель

$$a^* + b^* x + c\,(x - x_0)^2$$

может иметь преимущество перед моделью

$$a + bx + cx^2,$$

несмотря на то, что они математически эквивалентны?

12.5.6. Служит ли упр. 13.3.4 о предсказании сегодняшней температуры по вчерашним данным примером полной зависимости?

12.6.1. Что надо делать для управления эффектом переменной, которая прямо не измеряется?

12.6.2. (Продолжение упр. 12.6.1.) Какие предположения при этом необходимы?

12.6.3. Что такое «инструментальная переменная»?

12.6.4. Как сравнить разбросы $y - (b_{yu}/b_{xu})\,x$ и $y - b_{yx}x$? Зачем это нужно?

12.6.5. (Продолжение упр. 12.3.8.) Пусть мы теперь хотим воздействовать на инфляцию. Из «кривой Филлипса» (инфляция $= b_0 + b_1$ (безработица)) следует, что безработица может служить удобной мерой инфляции. Для более непосредственного воздействия мы могли бы взять индекс розничных цен (ИРЦ): если ИРЦ для года t равен P_t, то наша мера инфляции (I) была бы равна: $I = (P_t - P_{t-1})/P_{t-1}$. Используя I как инструментальную переменную, скорректируйте регрессию в упр. 12.3.8 для получения «истинной» инфляции. Что случилось с коэффициентами перед доходами и перед процентными ставками?

12.6.6. (Продолжение упр. 12.6.5.) Какие из предпосылок этого метода могут оказаться нарушенными?

12.6.7. (Продолжение упр. 12.6.5.) Постройте регрессию для остатков от накоплений из упр. 12.3.8 в зависимости от I, скорректированного на доходы и процентные ставки. Как сравнить Ваш результат с тем, что получилось в упр. 12.6.5?

12.6.8. (Продолжение упр. 12.4.4.) Вот обычная модель, которой пользуются экономисты:

накопления $= b_0 + b_1$ (доходы) $+ b_2$ (процентные ставки) $+ b_3$ (инфляция за последний год).

Для корректировки регрессии из упр. 12.4.4 на инфляцию за последний год мы можем воспользоваться как инструментальной переменной инфляцией, выраженной в индексе розничных цен. Эта инфляция I_t для года t равна:

$$I_t = (\text{ИРЦ}_t - \text{ИРЦ}_{t-1})/\text{ИРЦ}_{t-1}.$$

Кривая Филлипса

$$\text{инфляция} = c_0 + c_1\,(\text{безработица})$$

предполагает, что безработица служит вторичной мерой инфляции.

а. Скорректируйте регрессию на «истинную» инфляцию за последний год. Обсудите состоятельность предпосылок этого метода.

б. Скорректируйте регрессию только по I_{t-1}. Как сравнивать результаты?

12.6.9. В каких случаях данные для США за 1955—1974 гг. поддержат или отвергнут экономические модели, бегло описанные в упр. 12.6.8?

12.7.1. Когда масштабы для x и y можно выбирать независимо, всегда удается получить более или менее линейную модель, а вот когда практически мы не знаем, чего хотим, то как нам выбирать масштаб?

12.7.2. Почему мы пользуемся этими правилами, когда физические ограничения задачи не влияют на выбор?

12.7.3. Какое правило «старше» и почему?

12.7.4. Нойес (Noyes) с соавторами (Journ. chim. phys. 6: 505 (1908) и Zeits. phys. Chem. 70: 350 (1910)) дали эквивалентные сопротивления для разных температур водных растворов некоторых веществ при различных концентрациях, которые приведены ниже:

Концентрация	KCl при 18°C	KCl при 100°C	KCl при 306°C	NaCl при 18°C	NaCl при 100°C	NaCl при 306°C	HCl при 18°C	HCl при 100°C	HNO_3 при 306°C
0,0005	128,1		1044	107,5	355	1003	375	835	374,0
0,002	126,3	393	1008	105,4	349	955	373,6	826	371,2
0,01	122,4	377	910	102,0	335,5	860	368,1	807	365,0
0,08	113,5	341,5	720	93,5	301	680	353,0	762	353,7
0,1	112,0	336		92,0	290		350,6	754	

а. Для этих веществ и температур постройте две модели:

эквивалентное сопротивление $= a^* + b^*$ (концентрация)

и

эквивалентное сопротивление $= a^{**} + b^{**} \sqrt{\text{концентрация}}$.

Какое приближение лучше?

б. Как Вы стали бы оценивать эквивалентное сопротивление для очень малых концентраций?

в. Как построить границы по t-распределению Стьюдента для Вашего ответа?

12.7.5. Клеменс (Klemenc) и Реми (Remi) приводят (Mon. Chem. 44, 307 (1924)) вязкости некоторых смесей: а) водорода и пропана и б) водорода и NO. Найдите разумную модель для данного Вам случая. Вот округленные данные вида а) (% пропана; вязкость): (0; 86), (3,1; 89), (7,8; 94), (8,9; 95), (15; 97), (22,2; 96), (32,7; 92), (51,8; 87), (69,8; 81), (80,4; 77), (100; 75); б) (% NO; вязкость): (0; 85), (19,8; 142), (23,0; 145), (28,4; 147), (45,1; 160), (70,4; 172), (85,0; 175), (100; 180).

12.7.6. В илл. 1 к упр. 12.7.6 приводятся длина (x_1), площадь сечения (x_2) и объем (y) пяти кирпичей. Соотношение $y = x_1 x_2^{\frac{2}{3}}$ выполняется строго. Как хороша будет регрессия $y = b_0 + b_1 x_1 + b_2 x_2$? Какие особенности этого набора данных делают хорошую модель возможной?

Иллюстрация 1 к упражнению 12.7.6

	Размеры кирпичей				
x_1	10	10	10	11	9
x_2	5	6	4	5	5
y	250	360	160	275	225

12.8.1. Какие преимущества сулит первый этап анализа большого массива данных, основанный на подвыборках?

12.8.2. Какие из этих преимуществ сохраняются, даже если у нас есть высокоскоростная вычислительная машина?

ГЛАВА 13

13.1.1. Что дает тождество

$$117 - 3x + 2x^2 = 109 + 5x + 2(x - 2)^2$$

для интерпретации коэффициента при x?

13.1.2. Пусть мы построили регрессию

(инфляция) $= b_0 + b_1$ (доля безработных) $+ b_2$ (инфляция за предыдущий год) $+$
$+ b_3$ (процентная ставка).

Что не верно в утверждении: «влияние процентной ставки на инфляцию определяется коэффициентом b_3»? Как следовало бы описать коэффициент b_3?

13.1.3. Что такое «генератор»?

13.1.4. Каковы главные факторы, ответственные за величину коэффициента b_1 в модели $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3$?

13.1.5. Предположим, что какому-то химику понадобилось выяснить влияние галогенов: хлора (Cl), брома (Br) и иода (I) на точки кипения некоторых алкилгалогенидов. Любая алкильная группа (скажем, C_4H_9) может сочетаться с любым галогеном (здесь Cl, Br и I), образуя галогениды (C_4H_9Cl, C_4H_9Br и C_4H_9I). Мы же хотим выяснить, как различаются точки кипения. Их значения (в °C) представлены в илл. 1 к упр. 13.1.5 вместе со значениями молекулярных весов алкильных групп. Положим, y — точка кипения; x_1 — молекулярный вес алкильной группы

$$x_2 = \begin{cases} 1 & \text{для хлора,} \\ 0 & \text{для любого галогена, кроме хлора;} \end{cases}$$

и пусть x_3 и x_4 — аналогично задают бром и иод.

а. Постройте графически или как-нибудь иначе регрессионную модель

$$y = b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4.$$

(Простой график y—x_i даст вполне приличные оценки всех параметров.)

б. Что собой представляют «эффекты галогенов»?

Иллюстрация к упражнению 13.1.5

Молекулярные веса алкильных групп и точки кипения их галогенидов

Алкильная группа	Молекулярный вес	Галоген		
		Cl	Br	I
C_2H_5	29	12,5	38	72
n—C_3H_7	43	47	71	102
n—C_4H_9	57	78,5	102	130
n—C_5H_{11}	71	108	130	157
n—C_6H_{13}	85	134	156	180
n—C_7H_{15}	99	160	180	204
n—C_8H_{17}	113	185	202	225,5

13.1.6. Другой химик, быть может, предпочтет молекулярный вес всего соединения. Тогда x_5 — молекулярный вес алкилгалогенида. Поскольку атомные веса для Cl, Br и I соответственно равны 35,5; 80 и 127, мы имеем

$$x_5 = x_1 + 35{,}5x_2 + 80x_3 + 127x_4.$$

а. Перестройте модель так:

$$y = c_5x_5 + c_2x_2 + c_3x_3 + c_4x_4.$$

б. Что же такое теперь «эффекты галогенов»?

13.1.7. При ответе на следующие вопросы воспользуйтесь данными о задолженностях муниципалитетов ряда городов (табл. 9 из приложения для упражнений).

а. Подберите регрессию $y = b_0 + b_2x_2$.

б. Подберите регрессию $y = c_0 + c_1x_1 + c_2x_2$ и сравните ее коэффициенты с теми, что получились в пункте а).

в. Подберите регрессию $y = d_0 + d_1x_1$. Согласуются ли результаты с Вашими интуитивными ожиданиями, основанными на результатах пунктов а и б?

13.1.8. В табл. 5 из приложения для упражнений положим

y = E (расходы на обучение в расчете на одного ученика бесплатной средней школы);

x_1 = SBG (размер штатной субсидии на одну школу в штате Массачусетс);

x_2 = W (стоимость имущества, подлежащего налогообложению, на одного ученика бесплатной средней школы).

а. Найдите регрессии $y = b_0 + b_1x_1$ и $y = c_0 + c_2x_2$.

б. Найдите $y = d_0 + d_1x_1 + d_2x_2$ и сравните с тем, что было выше.

13.2.1. Что мы имеем в виду под $x_{1;25}$, или $y_{;1}$, или $y_{.12}$, или $x_{2.1}$?

13.2.2. Пусть мы хотим строить регрессию y от x_1 и x_2 поэтапно. В качестве первого шага подбирается для x_1 регрессия к y и выделяется $y_{;1}$. Что Вы посоветовали бы теперь сделать, чтобы получить модель для y с двумя носителями x_1 и x_2? Среди возможностей, которые Вы видите, какая предпочтительнее? Почему?

а. Постройте модель для x_2 по x_1, выделяя $x_{2;1}$. Постройте модель $y_{;1}$ по $x_{2;1}$.

б. Постройте модель для y по x_2.

в. Постройте модель для x_2 по x_1, выделяя $x_{2;1}$.
Постройте модель для y по $x_{2;1}$.

г. Постройте модель для $y_{;1}$ по x_2.

13.2.3. а. Как следует понимать коэффициенты в упр. 12.3.7?

б. Как следует понимать коэффициенты в упр. 12.3.8?

13.2.4. В данных об операциях по поводу грыжи (табл. 8 из приложения для упражнений) положим: y — время послеоперационного долечивания в клинике; x_1 — возраст; x_2 — физическое состояние.

а. Сопровождайте метод наименьших квадратов на каждом шаге графическим анализом при построении модели $y = b_0 + b_1x_1 + b_2x_2$, причем сначала введите x_1, а затем x_2.

б. Сравните оценки метода наименьших квадратов для коэффициентов b_0, b_1 и b_2 с теми, что получились графически в упр. 12.3.7. Сравните и сами модели.

13.3.1. Мы наблюдали эффект слабых вариаций регрессионных моделей, обусловленный округлением. Причем все точки имеют шансы подвергнуться округлению на одну и ту же величину. Что может случиться, если выкинуть важный фактор? Станет ли модель ошибочной?

13.3.2. В примере 5 из гл. 13 мы установили, что существует «наилучшая» модель метода наименьших квадратов, но ее коэффициенты остались неопределенными. Какими же они могут быть? Что отсюда следует в связи с использованием моделей и интерпретацией их коэффициентов?

13.3.3. Обратите внимание на то, что в регрессионных моделях из упр. 13.1.5 нет постоянных членов. Если бы мы положили $x_0 \equiv 1$ для постоянного члена, то какие из переменных среди x_0, x_1, x_2, x_3 и x_4 оказались бы строго коллинеарными?

13.3.4. Представьте себе, что нам нужно предсказать температуру для 2—30 апреля по данным о погоде из табл. 6 приложения для упражнений. Один из возможных подходов — это исключение из данных тренда, поскольку зима заканчивается, и использование температуры в качестве носителя, поскольку тепло и холод сменяют друг друга последние несколько дней.

а. Для осуществления этой программы воспользуйтесь шаговой регрессией и исключите из данных первый эффект. К чему это привело?

б. Исчезнет ли эта проблема, если мы попытаемся предсказать температуры только для дождливых дней? Да или нет? Почему?

13.3.5. Выясните, как точность зависимостей, найденных в упр. 13.32 и 13.3.3, влияет на интерпретацию эффектов, которые мы пытались измерить.

13.3.6. На илл. 1 к упр. 13.3.6 представлены значения $y = -1 + x + 0{,}5x^3$, округленные до двух десятичных знаков. Значения x заимствованы из илл. 13.3.1 (с. 70). Постройте регрессию $y = b_0 + b_1x + b_2x^2$ и обсудите результаты.

Иллюстрация 1 к упражнению 13.3.6

Округленные значения $y = -1 + x + 0{,}5\ x^3$

x	0,9	1,0	1,1	1,2	1,3	1,4	1,5
y	0,26	0,50	0,77	1,06	1,40	1,77	2,19

13.3.7. Повторите упр. 13.3.6 для значений y, округленных до одного десятичного знака.

13.3.8. Постройте по данным из упр. 13.3.6 регрессию $y = b_0 + b_1x + b_2x^2 + b_3x^3$ и обсудите результаты.

13.3.9. Повторите упр. 13.3.8 при округлении y до одного знака после запятой.

13.3.10. а. Найдите оценки метода наименьших квадратов для регрессии $y = b_0 + b_1x_1 + b_2x_2$ по данным из илл. 12.3.1 (с. 35).

б. Добавьте «ошибки округления» величиной $\pm$ 2 для x_1, $\pm$ 500 для x_2 и $\pm$ 100 для y. Используйте таблицу случайных чисел. Сравните коэффициенты модели

$$y^* = b_0^* + b_1^*\ x_1^* + b_2^*\ x_2^*.$$

в. Как изменилась теперь модель?

13.3.11. Повторите упр. 13.3.10 с иными «ошибками округления»: $\pm$ 1 для x_1, $\pm$ 2000 для x_2 и $\pm$ 300 для y. Как Вы думаете, насколько ошибки этих переменных обусловлены ошибкой измерения, и насколько — тем, что мы измеряем не совсем то, что имеем в виду измерить, а нечто в какой-то мере отличающееся?

13.4.1. Пусть мы построили регрессию с k носителями

$$y = b_0 + b_1x_1 + \ldots + b_kx_k$$

и пусть теперь переходим от этих k носителей к их линейным комбинациям z_1, z_2, ..., z_k, причем так, чтобы генератор не изменился. Если построить новую регрессию

$$y = c_0 + c_1z_1 + \ldots + c_kz_k,$$

то изменятся ли коэффициенты? А изменится ли сама модель?

13.5.1. Как Вы понимаете выражение «x_1 служит заменителем x_2»?

13.5.2. Предположим, некий исследователь находит, что результаты общей оценки знаний строго коррелируют с числом лет обучения. Горя желанием определить предметы, наиболее важные при обучении, исследователь ввел новые переменные x_1, ..., x_k для вклада «цикла гуманитарных дисциплин», «цикла статистических курсов» и т. п. Но новая регрессия

$$y = b_0 + b_1x_1 + \ldots + b_kx_k$$

не дала коэффициентов, существенно отличающихся от нуля. Что же случилось?

13.5.3. Пусть нас интересует регрессия «числа обводок за сезон» (T) от роста (H) (в дюймах) и веса W (в фунтах) защитников в профессиональном футболе. Удалось получить модель $T = b_0 + 0{,}50\,W - 0{,}10H$. Значит ли это, что «коротышкам» играть легче? Что вообще Вы могли бы сказать об относительной важности роста и веса?

13.5.4. Положим, что y — величина муниципальной задолженности, x_1 — стоимость участков под застройку и x_5 — число студентов колледжей, отнесенное к населению в данных о городских долгах (табл. 9 из приложения для упражнений).

а. Постройте следующие регрессии: $y = a_0 + a_1x_1$, $y = b_0 + b_5x_5$ и $y = c_0 + c_1x_1 + c_5x_5$.

б. Влияют ли доли студентов на увеличение долгов своих муниципалитетов? Как бы Вы объяснили полученные результаты?

13.5.5. Подумайте, какие переменные из данных об экономике (табл. 4 из приложения для упражнений) могли бы служить заменой для а) каких-то невключенных или неизмеримых переменных, б) для других переменных из этого же набора данных.

13.5.6. Выполните упр. 13.5.5 для данных Коулмена (табл. 7 из приложения для упражнений).

13.5.7. Выполните упр. 13.5.5 для данных об операциях по поводу грыжи (табл. 8 из приложения для упражнений).

13.6.1. Можете ли Вы отличить «эффект x_i на y, когда все остальные x_s остаются постоянными» от «эффекта x_i, скорректированного на все остальные иксы, на y, скорректированный на все остальные иксы»? Почему да или почему нет?

13.6.2. Если какие-то стратегические причины сдвигают значение x_j, то почему подстановка этого нового значения в старую регрессионную модель часто оказывается несостоятельной? Почему обычно не удается воспользоваться старой регрессией и для всех новых значений $x_1, x_2, \ldots, x_k$?

13.7.1. Чем применение регрессии в физических науках часто отличается от других ее приложений?

13.7.2. Почему в регрессионном анализе нет гарантии управления основными переменными из их наблюдаемого множества? Приведите пример, где это действительно может случиться из той области приложений, которая Вам близка.

13.7.3. Объясните, почему множественная регрессия с коррелированными носителями может привести к плохо определенным индивидуальным коэффициентам даже тогда, когда модель в целом хорошо согласуется с имеющимися данными?

13.7.4. Найдите и сравните остатки $(v—\hat{v})$ для данных о размерах яиц из илл. 13.7.1 (с. 74), если:

а) $\hat{v} = x_2 + 2x_3$;

б) $\hat{v} = 0{,}320 + 0{,}728x_2 + 1{,}812x_3$.

13.7.5. Если у нас есть несколько переменных, измеряющих разные аспекты одного и того же (вроде «домашнего климата»), то почему неразумно «запихивать» их все порознь в одну модель? Как Вы бы воспользовались такой многоаспектной информацией?

13.7.6. Для экономических данных из табл. 4 приложения для упражнений положим: y — накопления; x_1 — безработица (всего); x_2 — процентные ставки *Aaa* по Муди; x_3 — текущие процентные ставки; x_4 — индекс розничных цен; x_5 — (год — 1965).

а. Постройте регрессию y по всем x и попытайтесь интерпретировать найденное уравнение.

б. Какие носители действуют как заменители и для каких других носителей в этой регрессии? для переменных, не входящих в эту модель? Помните, что каждый из этих носителей растет во времени.

в. Смогли ли Вы угадать эффект изменения процентных ставок, пользуясь этим уравнением?

13.7.7. Сделайте упр. 13.7.6 по данным об обучении (табл. 5 из приложения для упражнений) для y (расходы на одного ребенка, обучающегося в бесплатной средней школе) и со всеми прочими факторами из этой модели. Стоит ли взять в качестве некоторых носителей логарифмы переменных? Как по всем носителям наиболее достоверно отличить богатые города от бедных?

13.7.8. Почему в упр. 13.7.6 и 13.7.7 получились уравнения регрессии, не имеющие преимуществ в точности или достоверности из-за управления столь многими важными переменными?

13.8.1. Опишите какую-нибудь задачу, переменные которой нельзя разделить даже при сколь угодно больших объемах данных.

13.8.2. Объясните, как могло случиться, что коэффициенты b_1 и b_2 определены очень плохо, зато их сумма $b_1 + b_2$ найдена очень точно.

13.8.3. Каков общий подход к преодолению трудностей, подобных той, что приведена в упр. 13.8.2?

13.8.4. Постройте регрессию y по x_1, x_2 и году для данных илл. 12.3.1 (с. 35). Как сделать методы из этого параграфа полезными для нейтрализации трудности, получения устойчивых коэффициентов?

13.8.5. Воспользуйтесь методами этого параграфа чтобы а) улучшить понимание модели и б) упростить уравнение регрессии из упр. 13.7.6.

13.8.6. Выполните упр. 13.8.5, используя уравнение регрессии из упр. 13.7.7.

13.8.7. Как методы этого параграфа могут помочь обнаружению того, что а) несколько носителей — меры практически одного и того же? б) два носителя совершенно различны, но строго следят друг за другом в исследуемой выборке? в) три или больше носителей — меры разных смесей двух переменных?

ГЛАВА 14

14.1.1. Докажите, что оценка метода наименьших квадратов для параметра α в модели $y = \alpha + \beta x$ есть $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

14.1.2. Докажите, что

$$\operatorname{var}(\hat{\alpha}) = \frac{\sigma^2 \Sigma x^2}{n\Sigma (x - \bar{x})^2}.$$

Проверьте эту формулу в частном случае, когда всякий x может принимать значения либо 0, либо 1 (не обязательно все одновременно).

14.1.3. Для модели $\mu + \beta (x - \bar{x})$, полученной методом наименьших квадратов, докажите, что наличие или отсутствие в модели коэффициента μ не повлияет на оценку β и наоборот.

14.1.4. Пусть мы подобрали модель $y = \alpha + \gamma x$ методом наименьших квадратов с равными весами. Чему равна дисперсия для γ? Можно ли переписать $\alpha + \gamma x^2$ так, чтобы оба коэффициента можно было найти по отдельности? Что тогда станет с дисперсиями оцениваемых коэффициентов?

14.1.5. Пусть мы получили модель $y = \alpha + \beta x$ и теперь хотим переписать ее так, чтобы α и новый коэффициент δ можно было получить раздельно. Выразите условия, обеспечивающие эту операцию, в терминах дисперсий и ковариаций α и δ. Найдите новое уравнение, удовлетворяющее этим условиям.

14.1.6. Пусть мы построили модель $y = \varepsilon \cos\theta + \eta \sin\theta$ для значений θ, взятых с нерегулярным шагом. Попробуйте переписать эту модель с коэффициентами λ и η, такими, что λ и η можно оценить раздельно. Что такое «естественные» условия? Удовлетворяет ли им Ваше переписанное уравнение?

14.1.7. Пусть мы нашли модель $y = K_A e^{At} + K_B e^{Bt}$, где A и B — известные константы. Сможете ли Вы переписать ее с помощью коэффициентов K_A и, допустим, τ_B, обеспечивающих раздельную оценку? Как? А с помощью коэффициентов K_B и, скажем, τ_A, допускающих раздельную оценку? Как? Чему равна дисперсия K_A в первом случае и дисперсия K_B во втором? А какими они были сначала?

14.1.8. (Используйте упр. 14.1.7.) Пусть B настолько близко к A, что

$$e^{(B-A)t} \approx 1 + (B - A)\, t$$

оказывается вполне удовлетворительным приближением.
Чему примерно равна дисперсия K_A? Что случилось бы, если B было бы очень близко к A?

14.2.1. Что такое «балансир»?

14.2.2. Почему множество из единиц служит балансиром при получении уравнения $y = \alpha + \beta x$ методом наименьших квадратов?

14.2.3. Докажите, что при подборе модели $y = \beta x$ методом наименьших квадратов коэффициенты при x служат балансирами. Только ли они служат здесь балансирами?

14.2.4. Докажите, что в построении $y = \alpha + \beta x$ методом наименьших квадратов величины $\{x(i) - \bar{x}\}$ будут балансирами, даже если α и β линейно зависимы.

14.2.5. Почему для подбора методом наименьших квадратов модели $y = \alpha + \beta x$, если α и β линейно независимы, необходимо иметь два балансира?

14.2.6. Пусть $\hat{\beta}$ будет оценкой метода наименьших квадратов для коэффициента β в модели $\beta\ (x - \bar{x})$. Пусть $\hat{\hat{\beta}}$ будет какой-нибудь другой оценкой β, найденной без обращения к методу наименьших квадратов, и пусть $h_i = x_i - \bar{x}$. Докажите, что если для заданного набора данных $\Sigma\hat{y}_i h_i > \Sigma\hat{\hat{y}}_i h_i$, то и $\hat{\beta} > \hat{\hat{\beta}}$ для тех же данных.

14.2.7. Пусть мы строим модель $y = \beta x$, в которой значения y — целые числа, а значения x даются в «осьмушках» (долях дюйма или ценах акций?). Какой балансир обеспечит Вам наибольшую простоту счета? (Под простотой мы здесь разумеем отсутствие дробей и не слишком большие числа.)

14.2.8. Пусть x принимает одно из следующих значений: 0, 1, 2, 3, 4, 5 и пусть мы строим модель $y = \alpha + \beta x$. Какие два из балансиров, обсуждаемых в тексте, кажутся самыми подходящими? Почему именно они?

14.2.9. Пусть мы ищем модель $y = \alpha + \beta x + \gamma x^2$ и в качестве балансиров предлагаются выражения $1 + x + 3x^3$, $1 + x + 5x^2$, $1 + x + 11x^2$, $1 + x + 17x^2$, $1 + x + 29x^2$ и $2 + x + 1023x^2$. Сколько из них нам надо выбрать? (Назовем их число k.) Какие подмножества сочетания из 6 по k мы можем выбирать? Каков наипростейший выбор? А наш выбор похож на него? Предложите ли Вы лучший выбор (не обязательно из этих 6 кандидатов)?

14.2.10. Что случилось бы, если бы мы попытались найти модель $y = \alpha + \beta x + \gamma x^2 + \sigma\,(3x^2 - 17x + 12)$? Сколькими балансирами пришлось бы воспользоваться? Как бы мы их нашли?

14.2.11. Пусть мы хотим построить модель $y = \alpha e^{\beta x} + \gamma$ и у нас есть начальное приближение β_0 для β. Как в этом случае выглядит «естественная» аппроксимация для $\alpha e^{\beta x} + \gamma$? А каковы соответствующие ей балансиры? Можно ли сохранить тот же самый набор балансиров, если мы имеем более чем одно значение для β_0? Почему? Почему нет?

14.3.1. Что значит подогнать модель $y = \Sigma_j \beta_j x_j$?

14.3.2. Каково *значение* c, такое, что балансир $\{1 + cx\ (i)\}$ настроен на α и не настроен на β при получении модели $y = \alpha + \beta x$ методом наименьших квадратов?

14.3.3. Что такое «уловитель» для подбираемого коэффициента?

14.3.4. Покажите, что $[x\ (j) - \bar{x}]/[\Sigma\ (x\ (i) - \bar{x})^2]$ есть уловитель для $\hat{\beta}$ в модели $y = \mu + \beta\ (x - \bar{x})$.

14.3.5. Пусть мы подбираем модель $y = \alpha + \beta x$. Что будет служить уловителем для β? А для α? Ну а для μ (в выражении $\mu + \beta\ (x - \bar{x})$)?

14.3.6. Пусть x принимает значения 0, 1, 3, 6 и 10. Какие численные формы надо придать уловителю, чтобы мы могли получить этот результат непосредственно?

14.3.7. Пусть (1) значения x симметричны относительно нуля и (2) мы строим модель

$$\alpha + \beta x^2 + \gamma x^4 + \delta x^6 + \varepsilon x^7 + \eta x^8 + \lambda x^{10}.$$

Каков уловитель для ε? Как его выразить арифметически, если x равен ± 2, ± 5, ± 8, ± 9 либо, наконец, ± 10?

14.3.8. Каким был бы последний ответ, если бы x принимал значения ± 2 (по 7 раз), ± 5 (по 5 раз), ± 8 (по 2 раза), ± 9 (по 2 раза) и ± 10 (по одному разу)?

14.3.9. Пусть $\{c_1\ (i)\}$ — уловитель для α_1, когда мы ищем модель

$$\alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3,$$

и $\{c_2\ (i)\}$ — уловитель для α_2, а $\{c_3\ (i)\}$ — для α_3 в той же модели. Что будет, если мы найдем не всю модель, а только $\alpha_1 x_1 + \alpha_3 x_3$?

14.3.10. Если $\{d_1\ (i)\}$ — новый уловитель для α_1, а $\{d_3\ (i)\}$ — для α_3 в последней модели (см. упр. 14.3.9), то как найти значения d, зная c_1, c_2 и c_3?

14.3.11. (Только для тех, кто имеет выход на ЭВМ.) В справочнике County and City Data Book (1962) среди прочего приводятся данные об x_{203} (общее население в 1960 г.), x_{213} (% родившихся вне США), x_{223} (% окончивших менее чем

5 классов школы) и x_{233} (% мужчин в численности рабочей силы для городов, не входящих в объединения и имеющих население 25 000 и более). Часть этих данных для Пенсильвании представлена на илл. 1 к упр. 14.3.11. Пусть мы строим модель

$$\alpha_0 x_{203} + \alpha_1 x_{213} + \alpha_2 x_{223} + \alpha_3 x_{233}.$$

Каковы уловители для α_0, α_1, α_2 и α_3? (Ответ дайте в численном виде, приведя одно значение для каждого из 15 городов.)

Иллюстрация к упражнению 14.3.11

Данные для больших необъединенных городов штата Пенсильвания

	x_{203}	x_{213}	x_{223}	x_{233}
Абингдон	55831	5,7	2,3	70,0
Бристоль	59298	3,4	2,3	74,6
Челтенхем	35990	7,8	2,6	69,4
Фолс	29082	2,5	1,8	74,2
Хейверфорд	54019	6,7	2,8	71,3
Хемпфилд	29704	3,8	7,8	74,6
Лоуэр-Мерион	59420	7,3	2,9	64,6
Мидлтаун	26894	3,9	1,1	77,5
Миллкрик	28441	2,9	3,8	72,5
Маунт-Лебанон	35361	3,7	1,1	72,5
Пенн-Хиллз	51512	4,3	2,9	75,3
Ридли	35738	4,0	3,9	73,5
Росс	25952	4,0	1,8	73,6
Спрингфилд	26733	4,7	1,8	73,7
Аппер-Дарби	93158	6,7	2,8	66,5

14.3.12. (Компьютер не требуется.) В том же справочнике County and City Data Book (1962) содержатся аналогичные сведения о 13 городах штата Нью-Йорк. Если бы мы снова искали модель

$$\alpha_0 x_{203} + \alpha_1 x_{213} + \alpha_2 x_{223} + \alpha_3 x_{233}$$

по этим 13 точкам, то что мы могли бы ожидать для уловителей в штате Нью-Йорк, зная их для штата Пенсильвания? А что на сей счет можно сказать о 12 городах из Калифорнии?

14.3.13. (Компьютер нужен.) В данных из упр. 14.3.11 вычеркните одну указанную Вам строчку и повторите все вычисления. В каких соотношениях, по Вашему мнению, окажутся уловители для исходного и данного множеств? А как получилось на самом деле?

14.3.14. (Классная контрольная работа; воспользуйтесь результатами упр. 14.3.13.) Соберите вместе результаты, полученные при отбрасывании различных городов, (1) исследуйте их совместно, (2) обсудите результаты этого исследования.

14.4.1. Определите подбор методом наименьших квадратов модели $y = \Sigma_j \beta_j x_j$ в терминах остаточной суммы квадратов и балансиров.

14.4.2. Набор балансиров $\{x_1, \ldots, x_k\}$ дает модель метода наименьших квадратов вида $y = \Sigma_j \beta_j x_j$. Единствен ли он? Если нет, то дайте аналогичный набор.

14.4.3. Почему выражение $c_1 x_1 + \ldots + c_k x_k$ будет служить балансиром, когда методом наименьших квадратов подгоняется модель $y = \Sigma_j \beta_j x_j$?

14.4.4. Найдите уловитель для $\hat{\beta}$-оценки метода наименьших квадратов параметра β в модели $y = \alpha + \beta x$.

14.4.5. Воспользуйтесь концепцией балансиров, чтобы доказать, что $\Sigma\, (\hat{y} - y)\, \hat{y} = 0$, где $\hat{y}$ — предсказанное значение отклика в модели $y = \Sigma \beta_j x_j$ с балансирами, которые представляют собой линейные комбинации носителей $x_1, \ldots, x_k$.

14.4.6. Найдите оценки метода наименьших квадратов для параметров β_1 и β_2 из модели $y = \beta_1 x_1 + \beta_2 x_2$, полагая, что $(\Sigma x_1^2)(\Sigma x_2^2) - (\Sigma x_1 x_2)^2 \neq 0$.

14.4.7. (Только для пользователей ЭВМ.) Отталкиваясь от представления суммы квадратичных отклонений в виде

$$\Sigma (y - \widehat{\beta_1} x_1 - \widehat{\beta_2} x_2 - \widehat{\beta_3} x_3 - \ldots - \widehat{\beta_k} x_k)^2$$

и пользуясь ЭВМ, найдите набор условий, таких, чтобы $\widehat{\beta}_1$, $\widehat{\beta}_2$, ..., $\widehat{\beta}_k$ соответствовали минимуму этой суммы без ограничений. Теперь переформулируйте Ваши утверждения так, чтобы они говорили что-нибудь о балансирах. Приложите результат к содержимому параграфа 14.4 (с. 82).

14.4.8. В каких случаях обычный метод наименьших квадратов даст единственное решение? А в каких нет?

14.4.9. Почему $\Sigma (y - \hat{y})^2$ не может быть отрицательной?

14.5.1. Почему $x_{1 \cdot 2 \ldots k}$ будет балансиром при подборе уравнения $y = \Sigma_{j=1}^{k} \beta_j x_j$ методом наименьших квадратов?

14.5.2. Будет ли $x_{1 \cdot 2 \ldots k}$ уловителем для β_1?

14.5.3. Укажите несколько уловителей для β_1.

14.5.4. Почему $\Sigma x_{1 \cdot 2 \ldots k}^2 \leqslant \Sigma x_1^2 \cdot$ меньшее?

14.5.5. Пусть мы строим модель $y = \beta_0 + \beta_1 x_1 + \beta_2 \sin x$, где x по одному разу принимает следующие значения:

$$x = \pm 1 \times 10^{-k}, \pm 2 \times 10^{-k}, \pm 3 \times 10^{-k} \text{ и } \pm 4 \times 10^{-k}.$$

Найдите, воспользовавшись приближением $\sin x \approx x - \frac{1}{6} x^3$, величину $x_{3 \cdot 12}$, приняв $\sin x \equiv x_3$ (для соответствующего приближения), $x \equiv x_2$ и $1 \equiv x_1$. Как сравнить $\Sigma (x_{3 \cdot 12})^2$ с $\Sigma (x_3)^2$? Как Вы считаете, насколько большим должно было бы быть число k, чтобы обеспечить построение приличной модели? Как бы Вы справились с вычислениями при относительно больших k?

14.5.6. Пусть мы, как и в последнем примере, строим модель $y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2$, причем $\Sigma x_0 x_1 = 0$ и $\Sigma x_0 x_2 = 0$. Если $\Sigma (x_{3 \cdot 12})^2$ очень мала по сравнению с $\Sigma (x_3)^2$, то что Вы можете сказать о величине $\Sigma (x_{2 \cdot 13})^2$ по сравнению с $\Sigma (x_2)^2$? Почему?

14.5.7. (Используйте результаты упр. 14.5.6.) Пусть мы строим модель $y = \beta_0 x_0 + \gamma_1 x_1 + \gamma_2 x_{3 \cdot 12}$ при тех же условиях, что и в упр. 14.5.6. Какова будет дисперсия $\hat{\gamma}_1$ в сравнении с дисперсией $\hat{\beta}_1$? Почему? Что бы это значило?

14.5.8. Пусть мы подгоняем модель $y = \beta_0 + \beta_1 x + \beta_2 \sin x + \beta_3 \operatorname{tg} x$, где x принимает по одному разу следующие значения:

$$x = \pm 1 \times 10^{-k}, \pm 2 \times 10^{-k}, \pm 3 \times 10^{-k} \text{ и } \pm 4 \times 10^{-k}.$$

Воспользуйтесь приближениями

$$\sin x \approx x - \frac{1}{6} x^3 + \frac{1}{120} x^5, \quad \operatorname{tg} x \approx x + \frac{1}{3} x^3 + \frac{2}{15} x^5,$$

чтобы упростить запись $\beta_0 + \beta_1 x + \beta_2 \sin x + \beta_3 \operatorname{tg} x$, придав ей вид $\beta_0 + \gamma_1 x + \gamma_3 x^3 + \gamma_5 x^5$. Как это сделать наиболее естественным и доступным образом? Как Вы могли бы воспользоваться этим с целью получения верхних границ для $\Sigma (x_{2 \cdot 013})^2$ и $\Sigma (x_{3 \cdot 012})^2$? Что говорят эти границы о $\operatorname{var} \hat{\beta}_2$ и $\operatorname{var} \hat{\beta}_3$?

14.5.9. (Продолжение упр. 14.5.8.) Как сравнить $\operatorname{var} \hat{\beta}_2$ и $\operatorname{var} \hat{\beta}_3$ из упр. 14.5.8 с а) $\operatorname{var} \hat{\delta}_2$ из модели $y = \delta_0 + \delta_2 \sin x$ и б) с $\operatorname{var} \hat{\eta}_3$ из модели $y = \eta_0 + \eta_3 \operatorname{tg} x$? Прокомментируйте и обсудите.

14.5.10. Пусть мы подбираем модель $y = \beta_0 + \beta_1 x + \beta_2 \sin x$, где $y \equiv \operatorname{tg} x$, а x принимает по одному разу следующие значения: $\pm 1 \times 10^{-k}$, $\pm 2 \times 10^{-k}$, $\pm 3 \times 10^{-k}$ и $\pm 4 \times 10^{-k}$. Постройте эту модель, воспользовавшись приближениями из упр. 14.5.8. Что представляет собой $x_{2 \cdot 01}$ (для $x_0 = 1$, $x_1 = x$, $x_2 = \sin x$) для этого приближения? А что такое $y_{\cdot 01}$? Постройте график $(x_{2 \cdot 01}, y_{\cdot 01})$ по точкам. Что Вы скажете на счет $\hat{\beta}_2$?

14.5.11. (Продолжение упр. 14.5.10.) Давайте положим, что $y = \operatorname{tg} x +$ малые случайные ошибки. Как велики могут быть эти «малые» случайные

ошибки, чтобы знак оценки $\hat{\beta}_2$ определился разумно? А для любого k? Для $k = 5$? Для $k = 10$?

14.5.12. Один химик, строя модель $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_6 x_6$ нашел следующие значения (x_6, $x_{6 \cdot 543210}$; $y_{\cdot 543210}$): (1; 0,002; 0,0007), (4; — 0,001; 0,0001), (9; 0,001; — 0,0002), (16; 0,003; — 0,0004), (25; — 0,004; — 0,0002). Постройте графики: а) зависимости $y_{\cdot 543210}$ от x_6 и б) $y_{\cdot 543210}$ от $x_{6 \cdot 543210}$. Что Вы скажете, сопоставляя эти графики? Как Вы думаете, хорошо ли определен знак $\hat{\beta}_6$? А если бы модель имела вид $y_{\cdot 012345} = \gamma_6 x_6$, то был бы тогда знак $\hat{\gamma}_6$ верным? Отчего все это происходит? Что бы Вы посоветовали химику?

14.5.13. (Только для пользователей ЭВМ.) Вернитесь к данным из упр. 14.3.11 и положите $y = x_{203}$, $x_0 = 1$, $x_1 = x_{213}$, $x_2 = x_{223}$, $x = x_{233}$. Теперь найдите значения $x_{0 \cdot 123}$, $x_{1 \cdot 023}$, $x_{2 \cdot 013}$, $x_{3 \cdot 012}$ и $y_{\cdot 0123}$. Сделайте 4 графика зависимостей $y_{\cdot 0123}$ от каждого из 4 найденных x. Рассмотрите все точки, которые Вам покажутся важными. Что говорят Вам эти графики о том, какие из городов оказывают относительно большее влияние на какие из коэффициентов: $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ и $\hat{\beta}_3$ (в предположении, что модель имеет вид $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$)?

14.6.1. Что такое взвешенный метод наименьших квадратов для модели $y = \Sigma \beta_j x_j$?

14.6.2. Покажите, что дело сводится к обычному методу наименьших квадратов, когда

$$w_{ij} = \begin{cases} 1 & i = j, \\ 0 & i \neq j. \end{cases}$$

14.6.3. Если в модели $y = \beta x$ точность каждого измерения y_i пропорциональна $1/x_i$ и подходящими весами для i-го наблюдения будут $1/x_i$, то покажите, что оценкой взвешенных наименьших квадратов для β выступает $\overline{y}/\overline{x}$.

14.6.4. Покажите в условиях упр. 14.6.3 (но для точности каждого измерения y_i, пропорциональной $1/x_i^2$), что взвешенная оценка наименьших квадратов параметра β равна:

$$\frac{\Sigma_{i=1}^{n} \dfrac{y_i}{x_i}}{n}.$$

14.6.5. Некий астроном строит модель $y = \beta^x$ с весами по следующим данным (w, x, y): (4; 0; 0,13), (9; 1; 0,27), (4; 2; 0,43); (1; 3; 0,69), (16; 4; 0,91), (25; 5; 1,32), (9; 6; 1,50); (1; 8; 2,03). Выпишите эквивалентное множество данных с равными весами. Затем постройте график.

14.6.6. (Продолжение упр. 14.6.5.) На графике из упр. 14.6.5 подберите подходящее значение B для β и изобразите остатки для $y = Bx$ в простейшем случае (считая, что модель строится без весов). О чем они говорят Вам? Есть ли хотя бы одна точка, которая кажется отклонившейся от прямой? Сколько раз надо было бы поменять веса точек, чтобы уложить их более или менее на прямую?

14.6.7. (Только для пользователей ЭВМ.) Снова вернитесь к данным из упр. 14.3.7 и положите $w = x_{203}$, $y = x_{233}$, $x_1 = x_{213}$, $x_2 = x_{223}$, модель $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

14.6.8. (Для всех.) По данным, полученным в упр. 14.6.7, постройте графики зависимостей $\sqrt{wy}$ от $\sqrt{wx_1}$ и $\sqrt{wx_2}$. Каковы Ваши выводы о богатстве городов?

14.6.9. (Продолжение упр. 14.6.8.) Постройте «на глаз» оценку C_1 для γ_1 из модели $y = \gamma_0 + \gamma_1 x_1$ и сделайте график зависимости $\sqrt{w}(y - C_1 x)$ от $\sqrt{wx^2}$. Каковы Ваши заключения?

14.6.10. Однажды некий исследователь построил модель $y = \beta_0 + \beta_1 x$ по 47 точкам со следующими весами: для 32 точек — 1, для пяти — 3, для пяти — 10, для одной — 100, для двух — 1000 и для двух — 1 000 000. Если бы Вы знали пары (x, y), то как Вы смогли бы приблизить эту модель более просто?

14.6.11. В примере со 100 точками (3 сомнительные), который обсуждался в тексте (с. 101), положим, что x согласованы с y, что искомая модель имеет вид $y = \beta x$ и что у тех трех точек, которые заслуживают малых весов, иксы лежат

где-то вблизи их ожидаемых значений. Сколь большого уменьшения дисперсии $\hat{\beta}$ мы ожидаем после исключения этих сомнительных точек (с правильными весами)? Заметно ли сократится при этом доверительный интервал для $\hat{\beta}$? Почему? Почему нет?

14.6.12. (Продолжение упр. 14.6.11.) А что, если значения для трех «сомнительных» точек примерно в 100 раз отличаются от всех прочих x? Как тогда ответить на вопросы из упр. 14.6.11? А как сразу догадаться об этом?

14.7.1. Для тех же самых 10 постоянных значений и 1 переменного постройте кривую влияния для серединного среднего, определяемого как среднее той половины значений, которая приходится на середину распределения. Причем найдите это среднее сначала как середину между пятью значениями из одиннадцати с каждого конца, а затем — как середину пяти с половиной точек (т. е. пяти точек с единичными весами и каждой следующей с весом 1/4). Сколь сильно различаются эти кривые? А в общем случае? А в самом крайнем?

14.7.2. Повторите упр. 14.7.1 для двух других усеченных средних: а) среднее середины по трем точкам, б) среднее середины семи точек.

14.7.3. Повторите упр. 14.7.1 для усечения:

$$\frac{\text{нижний квартиль} + 2\ (\text{медиана}) + \text{верхний квартиль}}{4}.$$

(Помните, что объем выборки 11; квартили лежат в 3,5 — на полпути между третьим и четвертым значениями с каждой стороны.)

14.7.4. Повторите упр. 14.7.1 для просто и дважды винзоризованных средних. Причем в просто винзоризованных средних а) наибольшее значение передвигается до совпадения со вторым по величине (в дважды винзоризованных средних два наибольших совмещаются с третьим) и б) наименьшее значение перемещается до совпадения со вторым с этого конца (а в дважды винзоризованных средних два наименьших сливаются с третьим).

14.7.5. (Воспользуйтесь упр. 14.7.1, 14.7.2, 14.7.3 и 14.7.4.) Кривые влияния для мер положения легко классифицируются в шесть категорий в соответствии с тем, а) ведут ли они в середину или нет и б) уменьшают ли они неопределенность или нет (с возвратом или без возврата). Их общие свойства сведены в таблицу на илл. 1 к упр. 14.7.5. Сделайте копию этой таблицы 2 × 3 с теми же градациями и заполните ее, вписав в соответствующие клетки среднее арифметическое, медиану, бивес-среднее (два варианта), четыре усеченные средние (просто и с винзоризацией). Выясните, что показывают Ваши результаты относительно того, какой индикатор положения стоит брать в каких случаях.

Иллюстрация 1 к упражнению 14.7.5

Характеристика оценок положения

	Неопределенность растет	Тенденция к росту без возвращения	Тенденция к росту с возвращением
Не ведет в середину	Не надежно; эффективность не очень велика для любого разумного распределения	Надежно при умеренном разбросе «хвостов»; эффективность как в ←	Очень надежно; эффективность не может быть очень большой для любого разумного распределения
Ведет в середину	Не надежно; эффективность очень велика для некоторых разумных распределений, но не для всех	Надежно при умеренном разбросе «хвостов»; эффективность очень велика для некоторых и может быть велика для целого ряда других распределений	Очень надежно; эффективность очень велика для некоторых и может быть велика для целого ряда других распределений

14.8.1. Почему желательно, чтобы метод наименьших квадратов был итеративным?

14.8.2. Проведите анализ, аналогичный тому, что дан в примере 1 на с. 93, но для весов

$$w(u) = \begin{cases} 1 - |u| & |u| < 1, \\ 0 & |u| \geqslant 1. \end{cases}$$

14.8.3. Один аналитик, подбирая модель

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

попытался найти бивес-приближение итеративно. Его результаты приведены на илл. 1 к упр. 14.8.3. Как Вы думаете, что случилось во время счета?

Иллюстрация 1 к упражнению 14.8.3

Бивес-приближение, получаемое итеративно

Приближения	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\beta}_3$	$\widehat{S}$
Нулевое	0	0	0	0	7,5
1-е	3,1	2,4	1,9	7,3	3,0
2-е	2,9	2,5	2,0	7,2	2,8
3-е	4,9	2,5	2,1	7,1	4,0
4-е	3,0	2,5	2,1	7,1	2,9
5-е	2,9	2,5	2,1	7,1	2,7
6-е	2,9	2,5	2,1	7,1	4,7
7-е	3,0	2,4	2,2	7,0	2,9
8-е	3,0	2,5	2,1	7,1	2,8
9-е	3,9	2,5	2,1	7,1	2,8

14.8.4. Постройте сначала график зависимости весов от u, получаемых путем простого шагового взвешивания при $c = 5$, а затем такой же график для биквадратных весов при $c = 5$ (умножьте последний на константу, чтобы он выглядел лучше первого!). Как наилучшим образом сравнить эти два варианта?

14.8.5. Повторите предыдущее упражнение для (несколько более сложного) шагового взвешивания при $c = 9$ и биквадратного взвешивания для $c = 9$.

14.8.6. Что может случиться, если некоторая пара (x, y) передвинется вперед или назад вдоль итераций по ошибке в шаговом взвешивании?

14.8.7. (Для пользователей ЭВМ.) Вернитесь к данным из упр. 14.3.11. На этот раз положите $x = x_{203}$ и $y = x_{233}$ и найдите модель $y = \beta x$ сначала методом наименьших квадратов, а затем бивес-итерацией при $c = 4$ начиная с $\beta = 0$. Обсудите различия. (Если надо, постройте график остатков.)

14.8.8. (Для пользователей ЭВМ; продолжение упр. 14.8.7.) Проделайте то же самое бивес-итерациями начиная с $\beta = 0$ для $c = 6$, 8 и 10. Сравните все пять результатов и обсудите их различия. (Стройте графики остатков, если надо!)

14.8.9. (Для пользователей ЭВМ; продолжение упр. 14.8.7 и 14.8.8.) Проделайте то же самое бивес-итерациями, но начиная с $\beta = -0{,}0002$ и полагая $c = 4$ и $c = 10$. Сравните с предыдущими результатами и обсудите. (Стройте графики остатков, если они Вам помогают.)

14.9.1. Сравните достоинства и недостатки метода наименьших квадратов и метода наименьших абсолютных отклонений (модулей).

14.9.2. В чем состоит неблагоприятная особенность метода наименьших абсолютных отклонений?

14.9.3. Пусть множество балансиров $w_l x_1, \ldots, w_l x_k$ для l-й итерации взвешенного метода наименьших квадратов при построении модели $y = \Sigma_{j=1}^{k} \beta_j x_j$

сходятся при $l \to \infty$ к $w_{\text{кон}} x_1, \ldots, w_{\text{кон}} x_k$. Покажите, что эти последние балансиры дают модель $y = \Sigma_{j=1}^{k} \beta_j x_j$, построенную методом наименьших абсолютных отклонений.

14.9.4.* Если у нас есть одни y без x, то какого рода модель даст метод наименьших абсолютных отклонений? В свете результатов упр. 14.8.5 чего следует опасаться при использовании наименьших абсолютных отклонений?

14.9.5.* Вернитесь к примеру из параграфа 14.8. с 10 фиксированными и одной подвижной точками и установите, какое значение $\hat{\hat{x}}$ минимизирует $\Sigma\, \Psi(x_i - \hat{\hat{x}})$, где $\Psi(u)$ равно u^2 или $k|u|$, когда $|u| \leqslant$ или $\geqslant k$ для множества значений 11-й точки (а можно ли воспользоваться итеративным взвешиванием?). Какая кривая влияния будет соответствовать такому способу получения модели?

14.9.6.* (Только для пользователей ЭВМ; город надо выделить каждому, кто возьмется делать это упражнение.) Вернитесь к данным из упр. 14.3.11 и возьмите $x = x_{203}$, $y = x_{233}$, $c = 9$, а начальное приближение $\beta = 0$. Сделайте линейные модели методом наименьших модулей а) для данных, как они есть, б) для выделенного Вам города замените y значениями ± 10, ± 8, ± 6, ± 4, ± 2 (а y для 11 других городов оставьте без изменения). Найдите подходящую кривую влияния для $\hat{\beta}$. Сравните данные для разных городов в классе.

14.9.7.* (Только для пользователей ЭВМ; продолжение упр. 14.9.7.) Повторите упр. 14.9.6 для $y = x_{233}$, $x_1 = 1$, $x_2 = x_{223}$, $x_3 = x_{213}$, $W = x_{203}$ (вес первого рода) при старых значениях c, начальных β и с изменениями в одном y.

14.10.1. Один исследователь получил (или хотел получить) модель $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$. Все значения коэффициентов β получились явно большими, много бóльшими, чем ожидалось; так что вместо того, чтобы думать о графике зависимости, например, $y_{\cdot 023} = y_{\cdot 0123} + \hat{\beta}_1 x_{1 \cdot 023}$ от $x_{1 \cdot 023}$, пришлось думать о графике $y_{\cdot 0123}$ (окончательные остатки) от $x_{1 \cdot 023}$. Для 237 точек вида (y, x_1, x_2, x_3) исследователь нашел следующие соотношения:

Число точек в данных	Величина $y_{\cdot 0123}$	Величины			
		$x_{0 \cdot 123}$	$x_{1 \cdot 023}$	$x_{2 \cdot 013}$	$x_{3 \cdot 012}$
211	мала	мала	мала	мала	мала
13	мала	мала	велика	мала	мала
10	мала	мала	мала	велика	мала
1	мала	мала	мала	мала	велика
1	мала	велика	мала	мала	мала
1	велика	велика	велика	велика	мала

Какие вопросы здесь следует выяснить? На каких экспериментальных точках следует прежде всего сосредоточить внимание? Какой из коэффициентов $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$ Вы считаете возможным отбросить? Какие экспериментальные точки Вы считали бы нужным исключить? Какие из коэффициентов $\hat{\beta}$ было бы лучше изменить, чем оставить так, как есть?

14.10.2. Одна исследовательница тщательно изучала основания для описания множества из n экспериментальных точек моделью

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5,$$

где для всех x выполняются соотношения $n/10 \leqslant x_i^2 \leqslant 10n$. Однако еще до получения всех y она нашла, что

$$\Sigma\,(x_{0 \cdot 12345})^2 = 0{,}27n,\ \Sigma\,(x_{1 \cdot 02345})^2 = 0{,}38n,\ \Sigma\,(x_{2 \cdot 01345})^2 = 0{,}0003n,$$

$$\Sigma\,(x_{3 \cdot 01245})^2 = 0{,}79n,\ \Sigma\,(x_{4 \cdot 01235})^2 = 0{,}0001n,\ \Sigma\,(x_{5 \cdot 01234})^2 = 1{,}23n.$$

Она огорчилась из-за двух очень маленьких сумм квадратов и была удивлена, что на ее долю выпали такие трудности. Чем Вы могли бы ее утешить? (Во-первых, относительно трудностей; во-вторых, насчет того, как их нужно было бы преодолеть.) Будет ли здесь что-нибудь зависеть от значений y?

14.10.3. Спортивный комментатор располагает сведениями о 1279 баскетболистах. Он обозначил через y среднюю продолжительность жизни, x_1 — длину ног (дюймы), x_2 — длину рук (дюймы), x_3 — рост (дюймы), x_4 — вес. Исследуя x он нашел:

$$\Sigma (x_{1 \cdot 234})^2 = 1279 \text{ (0,01 дюйма)}^2, \quad \Sigma (x_{2 \cdot 134})^2 = 1279 \text{ (0,008 дюйма)}^2,$$

$$\Sigma (x_{3 \cdot 124})^2 = 1279 \text{ (0,012 дюйма)}^2, \quad \Sigma (x_{4 \cdot 123})^2 = 1279 \text{ (13,2 фунта)}^2.$$

Он полагает, что мог бы отбросить один или несколько факторов. Прав ли он? Почему? Почему нет? Если бы надо было отбросить один или несколько факторов, то какие из них следовало бы выбрать? Помогут ли приведенные суммы узнать, какие из факторов стоит попытаться выкинуть в первую очередь? И с каких надо было бы начать?

14.10.4. Другой спортивный исследователь собрал сведения о 534 теннисистах. В данном случае в качестве y могла выступать любая измеримая характеристика, x_1 — средняя скорость (миль/ч) в первой подаче в момент пересечения сетки, x_2 — % первых подач, x_3 — средняя скорость на второй подаче, x_4 — время бега на 40 ярдов, x_5 — время в марафоне (26 миль, несколько часов). При анализе x было найдено:

$$\Sigma (x_{1 \cdot 2345})^2 = 534 \text{ (0,07 миль/ч)}^2, \quad \Sigma (x_{2 \cdot 1345})^2 = 534 \text{ (11 \%)}^2,$$

$$\Sigma (x_{3 \cdot 1245})^2 = 534 \text{ (17 миль/ч)}^2, \quad \Sigma (x_{4 \cdot 1235})^2 = 534 \text{ (0,47 с)}^2,$$

$$\Sigma (x_{5 \cdot 1234})^2 = 534 \text{ (32 мин)}^2.$$

Как Вы думаете, получатся ли разумные ответы? Почему да? Почему нет?

14.10.5. Пусть мы рассматриваем лишь два фактора x_1 и x_2. Если $x_0 = 1$, а r — коэффициент парной корреляции Пирсона (произведение моментов) между x_1 и x_2, то докажите, что

$$\Sigma (x_{1 \cdot 02})^2 = (1 - r^2) \Sigma (x_{1 \cdot 0})^2 \text{ и } \Sigma (x_{2 \cdot 01})^2 = (1 - r^2) \Sigma (x_{2 \cdot 0})^2.$$

Напомним, что $r^2 = (\Sigma (x_{1 \cdot 0} x_{2 \cdot 0}))^2 / (\Sigma (x_{1 \cdot 0})^2) (\Sigma (x_{2 \cdot 0})^2)]$. Что это значит для

$$\frac{\Sigma (x_{1 \cdot 02})^2}{\Sigma (x_{1 \cdot 0})^2} \text{ и } \frac{\Sigma (x_{2 \cdot 01})^2}{\Sigma (x_{2 \cdot 0})^2}?$$

14.10.6. (Продолжение упр. 14.10.5.) А как насчет

$$\frac{\Sigma (x_{1 \cdot 2})^2}{\Sigma (x_1)^2} \text{ и } \frac{\Sigma (x_{2 \cdot 1})^2}{\Sigma (x_2)^2}?$$

14.10.7. (Продолжение упр. 14.10.6.) Что можно сказать о

$$\frac{\Sigma (x_{1 \cdot 2345})^2}{\Sigma (x_{1 \cdot 345})^2} \text{ и } \frac{\Sigma (x_{2 \cdot 1345})^2}{\Sigma (x_{2 \cdot 345})^2}?$$

А как дальше? Почему? Почему нет?

14.10.8. Могут ли сосуществовать следующие соотношения:

$$\frac{\Sigma (x_{1 \cdot 234})^2}{\Sigma (x_1)^2} = 10^{-11}, \quad \frac{\Sigma (x_{2 \cdot 134})^2}{\Sigma (x_2)^2} = 10^{-2},$$

$$\frac{\Sigma (x_{3 \cdot 124})^2}{\Sigma (x_3)^2} = 10^{-1}, \quad \frac{\Sigma (x_{4 \cdot 123})^2}{\Sigma (x_4)^2} = 1?$$

Почему да? Почему нет?

Об упражнениях 14.С. Эти упражнения, относящиеся к параграфу 10.7, нуждаются в методах из параграфов 14.7, 14.8 и 14.9.

14.С.1. Составьте свою собственную выборку из городов (взятых из упр. 3.4.6). Включите в нее три переменные из табл. 1 приложения для упражнений, среди которых двумя должны быть 320 — медианный семейный доход и 327 — % окончивших колледж, и проведите все вычисления $\hat{c}o\hat{b}$ — $\hat{b}a\hat{r}$ из параграфа 10.7 (вып. 1).

14.С.2. Проделайте предыдущие упражнения для других трех переменных, среди которых обязательно должны быть 331 — % живущих в тех же домах, что и в 1955 г., и 358 — переселившиеся между 1958 и 1960 гг.

14.С.3. Снова повторите то же самое для переменных x_1, x_5 и y из табл. 7 приложения для упражнений.

14.С.4. И еще раз то же задание, теперь для x_3, x_4 и x_5 из табл. 9 приложения для упражнений.

ГЛАВА 15

15—1. Что Вы понимаете под «гибкой» регрессией?

15—2. Почему гибкая регрессия имеет смысл только тогда, когда нам не надо интерпретировать коэффициенты? Какие улучшения кажутся на первый взгляд снимающими эту трудность? Это происходит в действительности или нет?

15.1.1. Назовите пять «идеальных условий», которые мы выдвигали для гибкой регрессии. Если бы мы беспокоились об их состоятельности, то что бы мы сделали для проверки этого?

15.1.2. Объясните, что такое «генератор».

15.1.3. Что мы минимизируем?

15.1.4. Когда есть много альтернативных генераторов, некоторые из них имеют почти такое же минимизируемое значение, как и наименьшее среди них. Как же с этим поступить на практике?

15.1.5. Что представляет собой метод PRESS?

15.1.6. Что такое $s^2/(n-k)$ Энскамби (Тьюки)?

15.1.7. Что такое C_p Маллоуза?

15.1.8. Как использовать на практике величины из упр. 15.1.5 — 15.1.7?

15.2.1. Что такое «шаговая» регрессия? Почему мы часто нуждаемся в ней больше, чем в попытках перебрать все подмножества носителей?

15.2.2. Что представляют собой шаги «включения» и «исключения»? Зачем нам нужны исключения?

15.2.3. Как получаемые шаг за шагом модели и их остатки говорят нам о том, что происходит во время шаговой процедуры? Что означает, если коэффициент при интересующем нас носителе непрестанно меняется с каждым новым шагом?

15.2.4. Как можно построить устойчивую процедуру?

15.2.5. В гл. 13 («Беды регрессионных коэффициентов») мы предостерегали против таких интерпретаций коэффициентов, которые казались Макдональду и Уорду вполне подходящими. Почему могут не срабатывать такие приемы на некоторых разновидностях данных?

15.2.6. Пусть мы хотим предсказать расходы на обучение одного школьника по данным об обучении из табл. 5 в приложении для упражнений. Там приведен перечень из шести носителей. Для пяти из них, которые Вам выделены, проделайте следующее.

а. Сколько генераторов размера $k=2$ существует? Выберите среди них тот, который Вы считаете наилучшим.

б. Используйте для выделения «наилучшего» носителя шаговую процедуру, чтобы быть уверенным в том, что каждый носитель имеет свой шанс попасть в генератор. Как соотносятся результаты пунктов а) и б)?

15.2.7. Найдите все 10 регрессий для генераторов с $k=2$ из упр. 15.2.6. Сколько из них близки к оптимальным? Попало Ваше оптимальное решение в пункт а) или в пункт б)?

15.2.8. Повторите упр. 15.2.6 для генераторов размера $k=3$.

15.2.9. Найдите все 10 регрессий для генераторов с $k = 3$ из упр. 15.2.8. Сколько среди них близких к оптимальным? Оказалось ли Ваше оптимальное решение в пункте а) или б)?

15.2.10. Воспользуйтесь C_p Маллоуза для управления выбором в упр. 15.2.6, а затем шаговой процедурой для получения «наилучшего» генератора. Как сравнить Ваш ответ с результатами четырех предыдущих упражнений?

15.2.11—15.2.15. Проделайте упр. 15.2.6—15.2.10 для своих собственных данных.

15.3.1. Каковы преимущества методов всех подмножеств по сравнению с гибкой регрессией?

15.3.2. Можно ли извлечь выгоду из перебора генераторов в какой-нибудь особой последовательности? В какой именно?

15.3.3. Как благодаря работам Дэниеля и Вуда или Фернивала и Уилсона расширились возможности приложения этого метода?

15.3.4. Как можно построить устойчивую версию методов всех подмножеств?

15.3.5. (Громоздкие вычисления.) Пусть в экономических данных (табл. 4 из приложения для упражнений) y — log личного потребления, x_1 — log дохода, x_2 — долговременные процентные ставки (*Aaa* по Муди) и x_3 — текущие процентные ставки. Постройте регрессию y со всеми возможными генераторами выбранных x и обсудите результаты. Какую интерпретацию можно дать регрессионным коэффициентам?

15.4.1. Как бы Вы разделили большинство носителей на 3 категории по важности для построения модели?

15.4.2. Каковы основные этапы такого анализа?

15.4.3. Почему, когда исключаются ключевые носители, не приходится ожидать, что отклонения $(y_i - \hat{y}_i)$ будут независимы или некоррелированы?

15.4.4. Для предсказания дневной температуры в данных о погоде в г. Бостоне (табл. 6 из приложения для упражнений) используется список из 6 носителей. Могут быть важными и другие переменные, такие, например, как дата, вчерашняя температура или колебания давления по сведениям метеостанции и пр. Составьте список носителей, которые могут представлять интерес при предсказании температуры (y). Выберите среди них два носителя x_1 и x_2 и постройте для них устойчивую модель с помощью метода из параграфа 12.3. Каковы Ваши значения $y_{;12}$ и $x_{2;1}$? (Не включайте носителей, связанных с осадками, давлением, относительной влажностью или туманом, поскольку они понадобятся в других задачах.)

15.4.5. Рассортируйте данные о древних военных конфликтах (табл. 12 из приложения для упражнений) по трем категориям в зависимости от степени их важности. Отберите не более двух ключевых носителей и сосчитайте остатки для данных о продолжительности войн в месяцах в качестве y.

15.4.6. (Громоздкие вычисления.) Это первое в серии упражнений для предсказания инфляции по экономическим данным (табл. 4 из приложения для упражнений). (Определение инфляции в терминах индекса розничных цен см. в упр. 12.6.5.) Выберите 1 или 2 из ключевых носителей и разделите остальные на «интересные» и «для позднейшей угадайки». У Вас может возникнуть желание ввести новые носители, вроде «прошлогодней инфляции» или «колебаний валового национального продукта (ВНП)».

15.5.1. В чем состоят две платы за использование в регрессии многих носителей?

15.5.2. Каковы интуитивные мотивы для перевода «футбольных» носителей (модифицированные носители в параграфе 15.5) в логарифмическую шкалу?

15.5.3. Если факторы комбинируются после тщательного исследования данных, то оценка остаточной дисперсии будет, видимо, меньше, чем после объединения на основе суждений профессионалов, сделанных *априори*. Объясните, почему это не противоречит утверждению из параграфа 12.5, где говорится, что дисперсия $\hat{y}$ будет больше.

15.5.4. Почему линейные комбинации часто чувствительны к способу комбинирования носителей? В каких ситуациях могут оказаться более предпочтительными нелинейные комбинации?

15.5.5. (Продолжение упр. 15.4.4.) Постройте содержательный компонент для зависимости остатков от температуры в зависимости от меры «влажности», выражаемой через осадки, давление, относительную влажность и туман. Введите этот компонент в модель.

15.5.6. (Продолжение упр. 15.4.5.) Постройте содержательные компоненты для некоторых категорий носителей по данным о древних военных конфликтах (табл. 12 из приложения для упражнений). Обсудите, почему мы вынуждены комбинировать носители в этой задаче? Постройте зависимость этих компонентов от остатков для месяцев войн из упр. 15.4.5.

15.5.7. (Продолжение упр. 15.4.6.) Постройте содержательные компоненты для «процентных ставок» и «безработицы» и введите их в модель.

15.6.1. Что такое «главные компоненты»? Обсудите наиболее важные различия в анализе при использовании главного компонента $0{,}5x_1 - 2x_2 + x_5$ и содержательного компонента $0{,}5x_1 - 2x_2 + x_5$.

15.6.2. Можно ли всегда пользоваться главными компонентами вместо *априорных* содержательных компонент? Какие вопросы возникают по поводу а) данных или б) экспертов, вырабатывающих компоненты *априори*?

15.6.3. В чем различие между $\hat{\text{var}}\,(x_j)$ и содержательной мерой дисперсии x_j в параграфе 15.6?

15.6.4. Какие четыре категории носителей удобно исследовать методом главных компонент? Как их сравнить с тремя категориями из параграфа 15.4?

15.6.5. Почему Вы считаете, что носители, будучи превращенными в главные компоненты, становятся более пригодными для интерпретации? Многое ли мы при этом рискуем потерять?

15.6.6. (Продолжение упр. 15.4.4.) Какими будут главные компоненты для осадков, давления, относительной влажности и тумана? Как наиболее важные из них сравнить с Вашими содержательными компонентами? А как сравнить модели?

15.6.7. (Продолжение упр. 15.4.5.) Сравните главные компоненты с некоторыми из Ваших содержательных компонент. Если у Вас получилось несколько главных компонент, то как бы Вы прореагировали на то, что эта модель оказалась явно лучше, чем модель с содержательными компонентами?

15.6.8. (Продолжение упр. 15.5.7.) Как главные компоненты сравнить с содержательными для «процентных ставок»? Что Вы предпочитаете?

15.7.1. Обсудите коэффициенты, получаемые в регрессионном анализе в свете замечаний этого параграфа.

15.9.1. Каковы роли прошлых знаний и текущего анализа?

15.9.2. Оставаясь в рамках линейной регрессии, постройте зависимость y от x_1, x_2, x_3, x_4, x_5, алгебраически эквивалентную зависимости тех же факторов от $(y - 17x_1 - 5x_3 - 3x_5)$. Что мы выгадываем, пользуясь последним приближением?

15.9.3. Пусть мы занимались анализом погоды в г. Бостоне в апреле 1976 г. Как Вы могли бы при этом воспользоваться аналогичными результатами за 1975 г. (табл. 6 из приложения для упражнений) с целью усиления Вашего анализа? Как бы получше употребить здесь идеи параграфа 15.8?

ГЛАВА 16

16.1.1. Почему в регрессионный анализ всегда стоит включать исследование остатков?

16.1.2. Почему построение зависимости $(y - \hat{y})$ от y — это плохая идея? Какая идея хороша?

16.1.3. Если мы обнаруживаем какую-нибудь закономерность в картине остатков, то существуют по крайней мере два подхода к усовершенствованию модели. Каковы они?

16.1.4. Какие функции называются выпуклыми вверх? А какие — вогнутыми вверх?

16.1.5. В чем состоят различные пути исследования $(y-\hat{y})$ по отношению к $\hat{y}$?

16.1.6. Что бы Вы выбрали из возможностей, представленных в упр. 16.1.5, будь у Вас 10, 40, 100 или 1000 точек?

16.2.1. Что Вы думаете о «преобразовании» x как альтернативе более общему термину «функция» от x?

16.2.2. Что такое свертка?

16.2.3. В чем разница между носителем и фактором?

16.2.4. Когда два разных набора факторов дают эффект при одном и том же генераторе, можно получить идентичные модели. Почему же тогда встает проблема выбора переменных?

16.2.5. Ограничиваетесь ли Вы применением переменных только в их ныне наиболее распространенной форме? Что Вы теряете, вводя новые преобразования? А что можете выгадать?

16.3.1. Когда строится график зависимости $(y-\hat{y})$ от $y_{ст}$, то возникают ли проблемы, если $y_{ст}$ преобразован?

16.3.2. Что такое $x_{корр}$?

16.3.3. Что можно сделать, когда нам не хватает данных для суждения о зависимости $(y-\hat{y})$ от $t_{ст}$?

16.3.4. Для каких целей имело бы смысл строить график зависимости остатков от $t_{ст}$, а не от $x_{корр}$, даже если $x_{корр}$ уже есть?

16.3.5. Как Вы сглаживаете остатки? Когда это может быть полезно?

16.3.6. Если график $(y-\hat{y})$ от $x_{корр}$ показывает, что $t_{ст}$ (или эквивалентно $x_{корр}$) стоило бы включить в регрессию, то как это можно проще всего сделать?

16.3.7. *Полнота регистрации смертей.* Для некоторых популяций имеет место тождество

$$\text{смертность} = \text{рождаемость} - \text{прирост}.$$

Если мы захотим определить в последовательности популяций ту часть популяции, которая старше x лет, то мы получим последовательность отношений

$$\text{смертность } (x) = \text{рождаемость } (x) - \text{прирост } (x),$$

где

$$\text{смертность } (x) = \frac{N \text{ людей, умерших после } x \text{ лет}}{N \text{ людей, живущих после } x \text{ лет}};$$

$$\text{рождаемость } (x) = \frac{N \text{ людей, ,,дотянувших'' до } x \text{ лет}}{N \text{ людей, живущих после } x \text{ лет}};$$

$$\text{прирост } (x) = \frac{\text{дополнительное } N \text{ людей, старше } x \text{ лет}}{N \text{ людей, живущих после } x \text{ лет}}.$$

Демографы часто пользуются понятием «стабильной популяции», т. е. такой, в которой каждому из равных отрезков времени соответствует один и тот же прирост, так что прирост (x) не зависит от x. Стабильные популяции получаются в результате длительных периодов неизменных рождаемости и смертности.
При одном обследовании сельскохозяйственных районов Китая, проведенном в 1930 г., возникло подозрение, что рождаемость и смертность не должны меняться год от года, но что сообщался лишь некоторый процент c от общего числа имевших место смертей, причем этот процент не зависел от возраста.

а. Как это должно сказаться на последовательности отношений, приведенной выше? Как приспособить линейную регрессию для получения оценки c. В илл. 1 к упр. 16.3.7 даны рождаемость и смертность для переживших x лет мужчин из этого обследования. Какова Ваша оценка «полноты регистрации» c?

б. Похоже ли, что это стабильная популяция? Почему? Почему нет? Действительно ли полнота регистрации не зависит от возраста? Почему? Почему нет? Как Вы предпочитаете строить зависимость остатков — от возраста, или от рождаемости? Почему? Почему нет? Скажут ли Вам оба графика больше, чем один? Почему? Почему нет?

Иллюстрация 1 к упражнению 16.3.7.

Сообщенные данные о рождениях и смертях (в тыс.)

Возраст	Рождаемость	Смертность
5	29,5	16,2
10	30,1	15,8
15	31,8	16,8
20	33,6	18,2
25	36,7	19,6
30	40,3	21,8
35	45,3	24,9
40	54,7	29,0
45	63,4	33,8
50	77,9	41,6
55	91,4	52,1
60	120,8	64,6
65	143,5	82,8
70	169,4	111,8
75	231,6	129,4

Источник. Данные приведены с разрешения M. A. Stoto. Дискуссию смотри в: Barclay G. W., Coale A. J., Stoto M. A. and Trussel T. J. (1976). A reassessment of the demography of traditional rural China.— Population Index,42, 606—635.

16.3.8. Пусть y — величина задолженности; x_3 — прирост населения; x_4 — общий чистый долг для данных о муниципальных долгах (табл. 9 из приложения для упражнений).

а. Постройте регрессии y по x_3, y по x_4 и y по x_3 и x_4.

б. Воспользуйтесь графиками остатков для диагностики того, что случилось, и для оценки качества модели $y = b_0 + b_3x_3 + b_4x_4$.

в. Как можно было бы продолжить Ваш анализ?

16.4.1. Почему, когда в регрессионную модель добавляется новая переменная $t_{\text{нов}}$, часто никакого улучшения не получается, тогда как $x_{\text{корр}}$ существенно улучшает модель?

16.4.2. (Продолжение упр. 16.3.8.) Воспользуйтесь графиками остатков для добавления наилучшей новой переменной, включаемой в Вашу наилучшую модель из упр. 16.3.8. Почему графики остатков в этой ситуации помогут, несмотря на то, что любые нормальные шаговые алгоритмы скорее всего потерпят провал?

16.4.3. (Продолжение упр. 16.4.2.) Завершите построение модели для задолженностей по муниципальным данным (табл. 9 из приложения для упражнений). Подумайте, почему Вы пришли именно к такому приближению, и обменяйтесь мнениями о Вашей модели.

16.5.1. Почему мы рассматриваем мультипликативные модели вида $\hat{y}_{\text{med}} + (\hat{y} - \hat{y}_{\text{med}})(1 + u)$ вместо того, чтобы брать $\hat{y}(1 + u)$?

16.5.2. Что такое $Q_{\hat{y}}$? Зачем мы им пользуемся?

16.5.3. Что такое $q_{\hat{y}}$? Как его сравнить с $Q_{\hat{y}}$?

16.5.4. Почему мы строим график остатков, вместо того чтобы просто искать новую модель?

16.5.5. Как мы судим о том, стоит ли вводить в модель произведение носителей?

16.7.1. (Продолжение упр. 15.6.8.) Постройте модель, показывающую, как меняются экономические переменные с ростом инфляции.

а. «Запаздывающие» переменные могут обеспечить предсказание инфляции в наступающем году. Сколь хорошим можно сделать это предсказание, если не учитывать информации о значениях других переменных модели в наступающем году? Какие риски связаны с подобным предсказанием?

б. Можно ли считать коэффициенты Вашей регрессии интерпретируемыми? Для тех факторов, которыми можно манипулировать, что бы Вы предложили в качестве стратегии их изменения, направленного на уменьшение инфляции?

в. Обсудите, как изменения, предложенные в пункте б, могут сказаться на значениях других носителей.

16.B.1. (Самостоятельная работа.) Перестройте анализ из упр. 9.1.10, воспользовавшись регрессией для исследования зависимости между давлением и температурой на протяжении всего года. Выясните, действительно ли «месяц»—это неподходящий носитель для регрессии, если не прибегать к преобразованиям (и почему)? Один из возможных подходов к началу рассмотрения этой задачи состоит в том, чтобы расчленить ее на 12 меньших задач, а затем объединить результаты анализов (ср. параграфы 12.7, 12.8, 15.8 и 15.9).

16.B.2. (Самостоятельная работа.) Проанализируйте данные по трансплантации (пересадке) сердца (табл. 11 из приложения для упражнений). Для хирурга наиболее интересны следующие вопросы.

а. Какие факторы кажутся важными при длительной операции?

б. Как лучше избежать ошибочных предсказаний длительности операции? Каков риск неудачи?

в. Что бы Вы могли сказать врачу насчет соотношения между принятием плохой оценки времени операции и увеличением времени ожидания? Служит ли возраст фактором?

г. Каковы риски, если воспользоваться изменением стратегии, рекомендованным в пункте в? Как бы Вы объяснили это врачам на примере?

д. Нуждаются ли эти данные в коррекции на временной тренд?

16.B.3. (Самостоятельная работа.) Проанализируйте данные о военных исследованиях (табл. 10 из приложения для упражнений). Как сравнить имеющиеся суммы с требуемыми? Что потребуется в будущем году? Как меняется этот процесс во времени? Как за представленный период времени затраты перераспределялись между тремя родами вооруженных сил?

16.B.4. (Самостоятельная работа.) (Продолжение упр. 15.6.7, 15.4.5 и др.) Завершите анализ данных о древних военных конфликтах (табл. 12 из приложения для упражнений). Каков Ваш вывод об эффективности военной силы как сдерживающей силы в истории?

16.B.5. (Самостоятельная работа.) *Демографический сдвиг.* В 1888 г. в Швейцарии начался период, известный как «демографический сдвиг», т. е. рождаемость начала падать с высокого до низкого уровня, сохраняющегося до сегодняшнего дня. Если мы обозначим изменение «I_g» общей нормализованной меры рождаемости, то его можно попробовать соотнести с изменениями различных социально-экономических показателей. Значения I_g и пяти таких показателей, собранных Ф. ван де Валле (Francine van de Walle) в книге о демографическом сдвиге в Швейцарии, представлены на илл. 1 к упр. 16.B.5 для 47 франкоязычных кантонов за 1888 г. Проанализируйте эти данные. Какие переменные кажутся наиболее важными? Обсудите «статическую» природу этого подхода к «динамической» задаче.

Иллюстрация 1 к упражнению 16.B.5

Рождаемость швейцарцев и социально-экономические показатели

№ кантона	I_g	x_1	x_2	x_3	x_4	x_5
1	0,802	0,170	0,15	0,12	9,96	0,222
2	0,831	0,451	0,06	0,09	84,84	0,222
3	0,925	0,397	0,05	0,05	93,40	0,202
4	0,858	0,365	0,12	0,07	33,77	0,203
5	0,769	0,435	0,17	0,15	5,16	0,206
6	0,761	0,353	0,09	0,07	90,57	0,266
7	0,838	0,702	0,16	0,07	92,85	0,236
8	0,924	0,678	0,14	0,08	97,16	0,249
9	0,824	0,533	0,12	0,07	97,67	0,210
10	0,829	0,452	0,16	0,13	91,38	0,244
11	0,871	0,645	0,14	0,06	98,61	0,245
12	0,641	0,620	0,21	0,12	8,52	0,165
13	0,669	0,675	0,14	0,07	2,27	0,191
14	0,689	0,607	0,19	0,12	4,43	0,227
15	0,617	0,693	0,22	0,05	2,28	0,187
16	0,683	0,726	0,18	0,02	24,20	0,212
17	0,717	0,340	0,17	0,08	3,30	0,200
18	0,557	0,194	0,26	0,28	12,11	0,202
19	0,543	0,152	0,31	0,20	2,15	0,108
20	0,651	0,730	0,19	0,09	2,84	0,200
21	0,655	0,598	0,22	0,10	5,23	0,180
22	0,650	0,551	0,14	0,03	4,52	0,224
23	0,566	0,509	0,22	0,12	15,14	0,167
24	0,574	0,541	0,20	0,06	4,20	0,153
25	0,725	0,712	0,12	0,01	2,40	0,210
26	0,742	0,581	0,14	0,08	5,23	0,238
27	0,720	0,635	0,06	0,03	2,56	0,180
28	0,605	0,608	0,16	0,10	7,72	0,163
29	0,583	0,268	0,25	0,19	18,46	0,209
30	0,654	0,495	0,15	0,08	6,10	0,225
31	0,755	0,859	0,03	0,02	99,71	0,151
32	0,693	0,849	0,07	0,06	99,68	0,198
33	0,773	0,897	0,05	0,02	100,00	0,183
34	0,705	0,782	0,12	0,06	98,96	0,194
35	0,794	0,649	0,07	0,03	98,22	0,202
36	0,650	0,759	0,09	0,09	99,06	0,178
37	0,922	0,846	0,03	0,03	99,46	0,163
38	0,793	0,631	0,13	0,13	96,83	0,181
39	0,704	0,384	0,26	0,12	5,62	0,203
40	0,657	0,077	0,29	0,11	13,79	0,205
41	0,727	0,167	0,22	0,13	11,22	0,189
42	0,644	0,176	0,35	0,32	16,92	0,230
43	0,776	0,376	0,15	0,07	4,97	0,200
44	0,676	0,187	0,25	0,07	8,65	0,195
45	0,350	0,012	0,37	0,53	42,34	0,180
46	0,447	0,466	0,16	0,29	50,43	0,182
47	0,428	0,277	0,22	0,29	58,33	0,193

Определения:

x_1 — доля населения, постоянно занимающегося сельским хозяйством;

x_2 — доля «призывников», получивших наивысшую оценку на армейских испытаниях;

x_3 — доля населения, которая продолжала обучение после начальной школы;
x_4 — доля населения, исповедующая католицизм;
x_5 — детская смертность: доля детей, не доживших до года, от общего числа родившихся живыми.

Источник. Данные используются с разрешения Francine van de Walle. Отдел изучения народонаселения Принстонского университета, 1976. Неопубликованные данные.

ПРИЛОЖЕНИЕ ДЛЯ УПРАЖНЕНИЙ

(Продолжение. Начало см. в вып. 1)

Перечень упражнений, в которых используются приводимые ниже данные

4. Экономика: 12.3.8, 12.4.4, 12.4.5, 12.4.6, 12.6.5—12.6.9; 13.2.3, 13.5.5, 13.7.7, 13.7.9, 13.8.6; 15.3.5, 15.4.6, 15.5.7, 15.6.8; 16.7.1.
5. Обучение: 13.1.8, 13.7.8, 13.7.9, 13.8.7; 15.2.6—15.2.10.
6. Метеорология: 12.4.7, 12.4.8, 12.5.6; 13.3.4; 15.4.4, 15.5.5, 15.6.6, 15.9.3.
7. Данные Коулмена: 12.3.9, 12.4.3; 13.5.6; 14.С.3.
8. Операции по поводу грыжи: 12.1.5, 12.1.6, 12.2.5, 12.3.7, 13.2.3; 13.2.4, 13.5.7.
9. Муниципальная задолженность: 13.1.7, 13.5.4; 14.С.4; 16.3.8, 16.4.2, 16.4.3.
10. Военные исследования: 12.2.6; 16.В.3.
11. Трансплантация сердца: 16.В.2.
12. Древние военные конфликты: 12.4.9; 15.4.5, 15.5.6, 15.6.7; 15.В.4.
13. Самоубийства (резерв).

Таблица 4

Экономические данные из экономического доклада президента США, февраль 1975 (выбраны данные за 1955—1974 гг).

Год	Валовой нац. продукт	Индекс розн. цен*	Безработные			Процентные ставки			Личные (в млн.)		
			всего	мужчины старше 20 лет	женщины старше 20 лет	долговременные — по Муди		текущие	потребление	накопления	доходы
						Aaa	Bbb				
1955	438,0	69,0	4,4	3,8	4,4	3,06	3,53	1,89	39,6	34,1	310,9
1956	446,1	70,0	4,1	3,4	4,2	3,36	3,88	2,77	38,9	36,0	333,0
1957	452,5	72,5	4,3	3,6	4,1	3,89	4,71	3,12	40,8	34,1	351,1
1958	447,3	74,5	6,8	6,2	6,1	3,79	4,73	2,15	37,9	33,0	361,2
1959	475,9	75,1	5,5	4,7	5,2	4,38	5,05	3,36	44,3	34,7	383,5
1960	487,7	76,3	5,5	4,7	5,1	4,41	5,19	3,53	45,3	32,3	401,0
1961	497,2	77,0	6,7	5,7	6,3	4,35	5,08	3,00	44,2	31,7	416,8
1962	529,8	77,9	5,5	4,6	5,4	4,33	5,02	3,00	49,5	37,1	442,6
1963	551,0	78,8	5,7	4,5	5,4	4,26	4,86	3,23	53,9	39,1	465,5
1964	581,0	79,9	5,2	3,9	5,2	4,40	4,83	3,55	59,2	49,5	497,5
1965	617,8	81,3	4,5	3,2	4,5	4,49	4,87	4,04	66,3	55,4	538,9
1966	658,1	83,6	3,8	2,5	3,8	5,13	5,67	4,50	70,8	65,1	587,2
1967	675,2	86,0	3,8	2,3	4,2	5,51	6,23	4,19	73,1	65,0	629,3
1968	706,6	89,6	3,6	2,2	3,8	6,18	6,94	5,17	84,0	68,3	688,9
1969	725,6	94,4	3,5	2,1	3,7	7,03	7,81	5,87	90,8	60,6	750,9
1970	722,5	100,0	4,9	3,5	4,8	8,04	9,11	5,95	91,3	76,2	808,3
1971	746,3	104,3	5,9	4,4	5,7	7,39	8,56	4,88	103,9	87,4	864,0
1972	792,5	107,7	5,6	4,0	5,4	7,21	8,16	4,50	118,4	97,9	944,9
1973	839,2	114,4	4,9	3,2	4,8	7,44	8,24	6,44	130,3	120,2	1055,0
1974	821,1	127,0	5,6	3,8	5,5	8,57	9,50	7,83	127,8	×	1050,4

* Индекс розничных цен относится к 1970 г.

Таблица 5

Расходы на обучение

Следующие данные были собраны экономистом М. С. Фельдштейном (M. S. Feldstein). Подробности в работе [«Wealth, neutrality, and local choice in public education» by Martin S. Feldstein. The American Economic Review, 65, № 1, March 1975, p. 75—89]. Данные воспроизводятся с разрешения автора и Американской экономической ассоциации.

Переменные:

MFI — медианный семейный доход;
SBG — штатные субсидии на одну школу;
FG — федеральные субсидии;
RES — % местных налогов на недвижимость;
PSC — число учащихся бесплатных школ, на душу населения;
EEP — расходы на обучение одного школьника;
TVP — размер подлежащей налогообложению недвижимости, на одного школьника;
P — цена образования относительно города, имеющего 100 долларов на эту цель.

Города, чьи школы частично поддерживаются из бюджета штата, должны вносить менее одного доллара на каждый доллар, полученный для покрытия расходов по обучению.

MFI	SBG	FG	RES	PSC	EEP	TVP	P
9890	181	308	68	0,180	873	19560	100
12247	65	84	82	0,180	864	34311	100
10904	1	102	71	0,185	874	24418	75
11292	1	84	76	0,208	758	26824	74
13030	1	99	80	0,256	841	28044	82
10377	146	111	64	0,212	719	17453	100
9815	47	146	45	0,103	1184	48012	100
9738	150	149	60	0,197	622	15950	100
14958	66	155	90	0,273	1008	32105	100
12516	1	101	76	0,211	842	25743	74
10086	68	117	35	0,175	949	50993	100
9881	1	220	70	0,294	859	39036	87
11982	1	158	71	0,248	903	19296	70
11550	1	128	70	0,232	782	16941	63
9750	96	144	62	0,200	894	26419	100
9756	1	196	29	0,181	861	22569	70
11031	113	101	85	0,256	871	25478	100
11645	162	106	60	0,220	574	18759	100
11278	1	164	79	0,539	854	22904	75
10871	1	70	55	0,228	850	23858	73

Продолжение

MFI	SBG	FG	RES	PSC	EEP	TVP	P
17558	112	136	88	0,299	1027	24570	100
9739	1	90	59	0,167	742	31847	80
11020	1	126	65	0,246	772	16347	66
10665	1	91	66	0,205	805	24561	75
12424	1	56	84	0,227	761	24857	77
9638	156	119	73	0,226	674	17915	100
12580	1	63	50	0,300	745	17616	68
12656	91	321	81	0,215	720	27058	100
13144	1	122	80	0,273	844	26744	82
9992	130	316	69	0,200	858	20570	100
8924	1	162	50	0,197	878	19348	67
12629	1	107	75	0,286	783	14790	63
12606	1	114	68	0,228	867	24950	74
10621	215	88	67	0,294	760	10848	100
11629	157	185	68	0,207	749	21630	100
9510	60	83	67	0,523	917	40645	100
11094	142	108	68	0,189	954	25794	100
13434	1	93	88	0,244	763	23991	74
10067	66	79	88	0,178	810	52291	100
11541	1	125	85	0,240	815	26512	82
14805	1	172	84	0,302	948	19009	71
9594	1	78	50	0,153	761	22433	67
9752	131	65	60	0,239	634	18598	100
12281	1	105	83	0,231	795	26689	77
9957	162	137	67	0,150	751	22710	100
12412	1	89	82	0,222	808	30002	83
9802	1	136	51	0,230	786	25084	74
9418	1	188	70	0,146	769	36782	81
12837	1	169	56	0,224	903	22132	74
11631	1	105	55	0,240	688	29349	82
9279	1	140	69	0,250	822	12658	61
11685	1	92	90	0,177	731	23856	71
10038	1	125	44	0,165	955	20115	66

Таблица 6

Данные о погоде в г. Бостоне в апреле 1975 г.

Эти данные взяты из сообщений метеостанции и пункта сбора первичной местной климатологической информации Национального бюро погоды (США), отдел прогнозов, Бостон, Массачусетс.

Все данные об осадках округлялись до 0,01 дюйма. «Солнечная погода» измерялась с помощью фотоэлектрических батарей числом часов, превыщающих пороговый уровень яркости. Отсчеты атмосферного давления и относительной влажности не усреднялись; показания снимались между 12 ч 50 мин и 1 ч 00 мин пополудни. Туман кодировался нулем (0), если он отсутствовал, 1 — если он был умеренным и 2 — при видимости менее 1/4 мили.

Дата	Средн. t	Осадки (дюймы)	Средн. скорость ветра	Солнечная погода (часы)	Средн. облачность (в десятых)	Давление (дюймы)*	Относит. влажность*	Туман**
1	38	0	7,4	4,1	8	29,82	57	0
2	40	0	8,6	9,0	8	30,00	55	0
3	44	1,40	21,5	0,0	10	29,02	93	1
4	35	0,23	17,0	0,0	10	29,10	85	2
5	36	0,13	14,8	2,2	10	29,49	82	1
6	41	0,01	14,6	2,2	10	29,52	58	0
7	41	0,01	15,1	5,6	9	29,62	46	0
8	39	0,04	12,4	10,1	6	29,76	47	1
9	40	0	13,3	5,1	9	29,85	43	0
10	43	0	8,6	13,0	1	29,81	35	0
11	42	0	9,8	13,2	0	29,82	42	0
12	40	0	8,6	11,2	8	29,85	68	0
13	42	0	13,0	11,5	4	29,98	30	0
14	48	0	11,6	13,3	0	30,13	25	0
15	44	0	7,5	10,0	9	30,00	52	0
16	48	0	13,6	9,9	10	29,85	45	0
17	52	0	11,0	10,8	5	29,77	40	0
18	54	0,01	11,9	8,5	6	29,91	25	1
19	61	0,12	16,2	1,2	10	29,45	63	1
20	53	0	21,4	8,1	5	29,62	30	0
21	45	0	15,8	11,8	3	29,98	17	0
22	47	0	10,1	13,7	1	30,32	25	0
23	51	0	11,9	12,6	2	30,27	28	0
24	54	0,28	6,9	0,0	10	29,84	72	2
25	52	0,04	7,2	0,0	10	29,89	74	2
26	45	0,16	12,9	3,7	8	29,86	71	1
27	44	0,01	14,5	2,8	10	29,92	31	1
28	44	0	9,4	8,8	7	29,81	34	0
29	43	0	8,3	14,0	2	30,11	46	0
30	54	0	7,7	14,0	1	30,17	34	0

* Между 12 ч. 50 мин. и 14 ч. 00 мин. пополудни.

** 0—отсутствует; 1—умеренный туман; 2—видимость $\leqslant$ 1/4 мили.

Таблица 7

Случайная выборка из данных отчета Коулмена по 20 школам штатов Средней Атлантики и Новой Англии

Переменные:

y — средняя оценка за устную речь (за все 6 классов);

x_1 — оплата школьного персонала в расчете на одного ученика;

x_2 — % отцов-белых у шестиклассников;

x_3 — социально-экономическое положение, складывающееся из средних (для шестиклассников): размера семей, полноты семей, образования отцов, образования матерей, процента отцов-белых и размера квартиры;
x_4 — средняя школьная оценка за устную речь;
x_5 — средний образовательный уровень матерей шестиклассников (единице соответствуют 2 класса школы).

№ школы	x_1	x_2	x_3	x_4	x_5	y
1	3,83	28,87	7,20	26,60	6,19	37,01
2	2,89	20,10	—11,71	24,40	5,17	26,51
3	2,86	69,05	12,32	25,70	7,04	36,51
4	2,92	65,40	14,28	25,70	7,10	40,70
5	3,06	29,59	6,31	25,40	6,15	37,10
6	2,07	44,82	6,16	21,60	6,41	33,90
7	2,52	77,37	12,70	24,90	6,86	41,80
8	2,45	24,67	—0,17	25,01	5,78	33,40
9	3,13	65,01	9,85	26,60	6,51	41,01
10	2,44	9,99	—0,05	28,01	5,57	37,20
11	2,09	12,20	—12,86	23,51	5,62	23,30
12	2,52	22,55	0,92	23,60	5,34	35,20
13	2,22	14,30	4,77	24,51	5,80	34,90
14	2,67	31,79	—0,96	25,80	6,19	33,10
15	2,71	11,60	—16,04	25,20	5,62	22,70
16	3,14	68,47	10,62	25,01	6,94	39,70
17	3,54	42,64	2,66	25,01	6,33	31,80
18	2,52	16,70	—10,99	24,80	6,01	31,70
19	2,68	86,27	15,03	25,51	7,51	43,10
20	2,37	76,73	12,77	24,51	6,96	41,01

Таблица 8

Данные об операциях по поводу грыжи

В следующей таблице представлены сведения о 32 пациентах, подвергшихся операции по удалению грыжи (без летального исхода). При этом измерялись:

LEAVE (условия жизнеобеспечения в операционной):
1) обычное восстановление,
2) использование блока интенсивной терапии с вечера накануне операции,
3) использование блока интенсивной терапии; требуется умеренное внимание,
4) использование блока интенсивной терапии, необходимо усиленное внимание.

NURSE (уровень требуемого обслуживания медицинскими сестрами и санитарками на протяжении первой недели после операции):
1) усиленный,
2) высокий,
3) средний,
4) облегченный.

LOS (время долечивания в клинике после операции, в днях).

Переменные, описывающие предоперационное состояние пациентов, без расшифровки:

PSTAT (физическое состояние с учетом обстоятельств, связанных с предстоящей операцией) по шкале от 1 до 5, где 1 соответствует отличному здоровью, а 5 — очень плохому.

BUILD (телосложение пациента):
1) истощенный,
2) худой,

3) нормальный, 4) толстый, 5) тучный.

CARDIAC или RESP (предоперационные осложнения):
1) нет, 3) средние,
2) незначительные, 4) тяжелые.

Пациент	Возраст	Пол	PSTAT	BUILD	CARDIAC	RESP.	LEAVE	LOS	NURSE
1	78	М	2	3	1	1	2	9	3
2	60	М	2	3	2	2	2	4	—
3	68	М	2	3	1	1	1	7	4
4	62	М	3	5	3	1	1	35	3
5	76	М	3	4	3	2	2	9	4
6	76	М	1	3	1	1	1	7	—
7	64	М	1	2	1	2	1	5	—
8	74	Ж	2	3	2	2	1	16	3
9	68	М	3	4	2	1	1	7	—
10	79	Ж	2	2	1	1	2	11	3
11	80	Ж	3	4	4	1	1	4	—
12	48	М	1	3	1	1	1	9	3
13	35	Ж	1	4	1	2	1	2	—
14	58	М	1	3	1	2	1	4	—
15	40	М	1	4	1	1	1	3	—
16	19	М	1	3	1	1	1	4	—
17	79	М	3	2	3	3	3	3	—
18	51	М	1	3	1	1	1	5	—
19	57	М	2	3	2	1	1	8	3
20	51	М	3	3	3	2	1	8	4
21	48	М	1	3	1	1	1	3	—
22	48	М	1	3	1	1	1	5	—
23	66	М	1	3	1	1	1	8	4
24	71	М	2	3	2	2	2	2	—
25	75	Ж	3	1	3	1	2	7	—
26	02	Ж	1	3	1	1	1	0	—
27	65	Ж	2	3	1	1	2	16	3
28	42	Ж	2	3	1	1	2	3	—
29	54	М	2	2	2	2	2	2	—
30	43	М	1	2	1	1	1	3	—
31	04	М	2	2	2	1	1	3	—
32	52	М	1	3	1	1	1	8	3

Источник. McPeek B. and Gilbert J. P. of the Harvard Anesthesia Center. Данные воспроизводятся с их разрешения.

Таблица 9

Данные о муниципальных задолженностях для 20 городов

Переменные:
y — величина задолженности;
x_1 — стоимость земли под застройку (в числе 1000-долларовых акций);
x_2 — срок платежей по векселям (в сотнях месяцев);
x_3 — прирост населения (в 100000 чел.);
x_4 — чистый общий долг;
x_5 — отношение числа учащихся колледжей ко всему населению.

№	Город	x_1	x_2	x_3	x_4	x_5	y
1	Бирмингем	30	1,81	3,61	0,280	1,03	335
2	Окснард	10	1,93	0,29	0,012	0,00	365
3	Салинас	30	2,79	0,24	0,023	4,29	315
4	Данбери	15	1,81	0,40	0,036	2,38	325
5	Нью-Хейвен	15	1,87	1,65	0,186	6,54	283
6	Норуолк	40	2,17	0,59	0,145	0,15	300
7	Новый Орлеан	15	2,34	6,40	0,710	1,91	327
8	Балтимор	10	1,85	9,74	1,827	2,24	290
9	Детройт	10	2,09	19,25	1,703	1,81	317
10	Сент-Луис	55	2,03	8,73	0,533	3,09	273
11	Клифтон	5	2,37	0,81	0,088	0,00	356
12	Нью-Йорк	5	2,33	82,00	20,720	2,25	314
13	Норт-Хемпстед	35	1,93	2,05	0,075	2,10	345
14	Талса	25	2,53	2,54	0,193	2,57	315
15	Филадельфия	80	2,14	22,00	3,861	2,36	305
16	Мемфис	90	1,93	4,53	0,465	1,55	285
17	Хопуэлл	15	2,16	0,22	0,023	0,00	350
18	Норфолк	10	1,90	2,99	0,356	0,47	320
19	Мадисон	100	1,93	1,17	0,205	21,09	270
20	Южн. Милуоки	25	1,81	0,17	0,027	0,00	305

Таблица 10

Военные исследования и разработки (в млн. дол.)

Приводимая таблица содержит суммы, требуемые вооруженными силами США на исследования и разработки по годам с 1953 по 1973, и соответствующие суммы, ассигнованные на эти цели конгрессом США (в млн. дол.).

Год	Армия		Флот		Авиация	
	требуемые	полученные	требуемые	полученные	требуемые	полученные
1953	450,0	440,0	75,7	70,0	525,0	525,0
1954	475,0	345,0	74,9	58,6	537,0	440,0
1955	355,0	345,0	61,0	419,9	431,0	418,1
1956	333,0	333,0	439,2	439,2	570,0	570,0
1957	410,0	410,0	477,0	492,0	610,0	710,0
1958	400,0	400,0	505,0	505,0	661,0	661,0
1959	471,0	498,7	641,0	821,2	719,0	743,0
1960	1046,5	1035,7	970,9	1015,9	750,0	1159,9
1961	1041,7	1041,2	1169,0	1218,6	1334,0	1552,9
1962	1130,4	1203,2	1267,0	1301,5	1637,0	2403,2
1963	1329,0	1319,5	1474,0	1475,9	3439,0	3632,1
1964	1474,6	1390,2	1578,4	1530,5	3627,9	3458,7
1965	1401,5	1344,1	1456,3	1377,5	3210,9	3117,3
1966	1442,7	1410,6	1478,1	1444,2	3153,9	3109,4
1967	1522,2	1531,9	1752,5	1762,4	3058,1	3116,8
1968	1544,0	1514,2	1863,9	1826,5	3293,6	3251,2
1969	1661,9	1522,6	2146,4	2141,3	3364,7	3570,3
1970	1849,5	1596,8	2211,5	2186,4	3561,2	3060,6
1971	1717,9	1618,2	2197,3	2165,1	2909,7	2762,1
1972	1951,5	1839,5	2431,4	2372,3	3017,0	2912,9
1973	2122,7	1829,0	2813,8	2545,3	3262,2	3122,5

Источник. Capra James R. Analysis of data describing Congressional responses to DOD budget requests (Ph. D. thesis). Naval Postgraduate School, Monterey, California, June 1974.

Таблица 11

Трансплантация сердца

После того как пациент включен в Стенфордскую программу, ему подбирается сердце донора с соответствующей группой крови.

Мы отобрали для представления здесь только тех пациентов, наблюдение за которыми велось вплоть до их смерти, поскольку это позволяет избежать трудностей, связанных с анализом «цензурированных» данных (о тех пациентах, которые живы, мы знаем лишь, что время их послеоперационной жизни будет

ПВ	Отторжение? Да=+	ТТ5	В	ВО	КВ
15		1,11	54,3	0	6
3		1,66	40,4	35	123
624	+	1,32	51	50	244
46	+	0,61	42,5	11	235
127		0,36	48	25	253
64	+	1,89	54,6	16	279
1350	+	0,87	54,1	36	300
280	+	1,12	49,5	27	327
23		2,05	56,9	19	325
10	+	2,76	55,3	17	412
1024	+	1,13	43,4	7	405
39	+	1,38	42,8	11	454
730	+	0,96	58,4	2	469
136	+	1,62	52	82	565
1		0,47	54,2	70	594
836	+	1,58	45	15	612
60	+	0,69	64,5	16	623
54	+	2,09	49	45	870
47	+	0,87	61,5	18	864
44		0,0	36,2	0	1101
994	+	0,81	48,6	1	1107
51	+	1,38	47,2	20	1149
253	+	1,08	48,8	31	1210
51	+	1,51	52,5	9	1326
322	+	1,82	48,1	20	1382
65	+	0,66	49,1	2	1420
551		0,12	48,9	32	1525
66	+	1,12	51,3	11	1538
65	+	1,68	45,2		1561
25	+	1,68	53	4	1634
63	+	2,16	56,4	26	1742
12		0,61	29,2	4	1727
29	+	1,08	54	66	1862
48		3,05	53,4	31	1882
297	+	0,60	42,6	36	1893
50	+	2,25	46,4	59	1966
68	+	1,33	51,4	138	2060
26		0,82	52,5	159	2082
161	+	1,2	43,8	3	2087

Источник. Stanford Heart Transplantation Program. Более подробные данные описаны в работе: Miller Rupert G., Jr. Least-squares regression with censored data. — Biometrika, 1976, 63, 449—464. Воспроизведено с разрешения автора и журнала.

больше, чем то, что они уже прожили, а это приводит к неполным или «цензурированным» данным). Приведенные цифры охватывают пациентов, умерших после пересадки сердца, в период с января 1968 г. по апрель 1974 г.

Изучались следующие переменные:

постоперационное время (ПВ) — число дней жизни пациента после операции;
возраст (В) — число лет в момент операции;
время ожидания (ВО) — число дней с момента включения пациента в программу до дня операции;
календарное время (КВ) — число дней от 1 января 1968 г. до дня операции;
тест Т5 на совместимость (ТТ5) — мера степени совместимости тканей донора и реципиента.

Таблица 12

Древние военные конфликты; служит ли милитаризация сдерживанию?

Эти данные содержат примеры воздействия более высоких цивилизаций[1] на историю. Для адекватного описания этого воздействия остается определить роль всех нижеследующих носителей. Для каждой цивилизации случайным образом выбиралось десятилетие в заданном столетии. Данные о Швейцарской конфедерации были добавлены с тем, чтобы они представляли страны с республиканским государственным строем.

Описание переменных (без тех подробностей, которые можно восстановить как функции других переменных):

	Столбец	
Война	1	месяцы войны во втором из выбранных десятилетий;
Территория	2	территориальные приобретения рассматриваемого государства (% исходной площади);
Обороняемая площадь	3	территория, требующая обороны;
Численность	4	число офицеров в армии;
Мобильность	5	число командиров, всадников или кораблей;
Качество	6	дисциплинированная армия, по свидетельству историков;
Фортификации	7	укрепление границы с противником;
Престиж	8	*Всеобщее* согласие, усиливающее армию, равную армии противника;
Соседи	9	те, кто имеют общую границу с противником;
Преграды	10	более $^2/_3$ общей границы — естественные преграды;
Главный город	11	главный город рассматриваемого государства в пределах 300 миль от границы;
Военная добыча	12	добыча одной стороны за счет другой;
Культура	13	общий культурный обмен;
Торговля	14	общий торговый обмен.

[1]Общество, располагающее по крайней мере одним городом с фундаментальными постройками и населением не менее 25 000 человек на территории радиусом 6 км, в котором широко распространена литература, переводимая на другие языки.

Историческая общность	Начало десятилетия	Государство	Противник	Грубые оценки	
				месяцы войны	терр. приобретения, %
Китай	125 до н. э.	династия Раняя Хань	гунны	105	3,6
	25 до н. э.			0	0
	776 н. э.	династия Тан	тибетцы (кяны)	51	0
	1076	династия Сун	тангуты	34	0
	1376	династия Мин	сев. Юань	1	6
Мусульмане	776	Аббасиды	Византия	72	0,01
Русские	1476	Московия	Новгород	6	25,2
Греки — римляне	225 до н. э.	Рим	Карфаген	34	—19,9
	25 до н. э.		Парфия	0	1,5
	176 н. э.		маркоманны и квады	34	0
	376		вестготы	66	0
	576	Византия	Персия	120	0
Западная католическая церковь	1276	Франция	Англия	0	—1,3
	1376	Англия	Франция	90	—25,6
	1576	Испания	Нидерланды	120	—2,27
	1676	Франция	То же	46	4
	1776	Англия	Франция	67	—39
Швейцарцы	1376	Швейцарская конфедерация	Кибург	17	13
	1476	То же	Бургундия	12	11,5
	1576	швейцарцы-протестанты	швейцарцы-католики	0	0

* Для носителей используются такие кодовые обозначения: Р — данный признак имеет

Источник. Naroll R., Bullough V. L. and Naroll F. (1974). New York, p. 377, Appendix C. Данные воспроизводятся с разрешения авторов и

Табл. 12 (продолжение)

Переменные (носители)											
военные						географические			культурные		
оборонямая территория	численность	мобильность	качество	фортификация	престиж	соседи	естественные преграды	главный город	добыча	культурный обмен	торговля
A	P	A	P	P	A	P	P	A	P	A	0
A	P	A	P	A	P	P	P	A	P	A	0
P	P	A	A	P	A	P	P	P	P	P	0
A	P	0	0	P	P	P	P	A	P	A	A
A	P	A	A	A	P	P	P	A	A	P	A
A	P	0	P	P	P	P	P	A	P	A	0
A	P	P	P	A	A	P	A	P	P	P	P
A	P	A	A	A	A	P	P	P	P	A	0
P	P	A	P	A	P	P	P	A	A	P	P
P	A	P	P	P	P	P	P	A	P	P	P
P	A	A	P	0	P	P	A	P	P	P	P
0	0	A	P	P	0	P	P	A	A	A	A
P	P	A	A	P	P	P	A	P	P	P	P
0	A	A	P	P	A	P	A	P	A	P	P
P	P	P	P	A	P	A	P	A	A	A	A
A	P	A	P	P	P	A	A	P	P	P	P
P	P	A	A	A	P	A	P	P	A	P	P
P	P	A	P	P	P	P	A	P	A	P	P
A	A	A	P	P	A	A	P	0	A	P	P
P	A	A	A	0	A	P	P	0	A	P	0

место; A — данный признак отсутствует; 0 — нет сведений.

Military Deterrence in History. State University of New York Press, Albany, издательства.

Таблица 13

Самоубийства и смертность

А. Доля самоубийц среди мужчин

Страна	Возраст, лет					
	15—24	25—34	35—44	45—54	55—64	65—74
Болгария	8,3	7,9	10,9	19,1	25,2	40,2
Финляндия	15,5	35,2	53,2	62,2	68,6	68,9
Венгрия	30,5	44,1	56,0	65,2	70,7	76,4
Израиль	4,8	10,2	11,4	16,4	18,2	23,0
Голландия	4,1	7,5	8,4	13,9	18,8	24,5
Швеция	13,8	28,3	40,8	51,0	50,5	47,4
США	9,9	17,3	22,4	28,4	36,0	35,7

Источник. Hamermesh D. S. and Goss N. M. (1974). An Economic Theory of Suicide.—Journal of Political Economy, 82, p. 83. Данные публикуются с разрешения авторов и издателя.

Сопоставьте с частью Б.

Табл. 13 (продолжение)

Б. Смертность в семи странах за 1934 или 1965 г.

Страна	Пол	Население (в тыс. чел.) возраст, лет						Число умерших (в тыс. чел.) возраст, лет					
		15—24	25—34	35—44	45—54	55—64	65—74	15—24	25—34	35—44	45—54	55—64	65—74
Болгария	М	661,0	609	650,6	413,4	419,3	212,7	758	306	1620	2602	6283	8528
	Ж	650,0	605,2	647,7	441,8	421,6	249,7	334	521	1050	1783	3184	7743
Финляндия	М	428,3	304,5	287,4	233,3	200,1	995,1	514	638	1366	2698	5338	5958
	Ж	413,5	293,1	305,2	279,2	253,3	155,6	185	234	543	1207	2767	5562
Венгрия	М	781	677	713	534	546	310	874	1182	2215	3830	10415	15209
	Ж	766	712	774	606	621	417	413	607	1499	2749	6998	14085
Израиль	М	174,7	132,3	128,0	119,1	91,6	41,4	191	146	251	644	1448	1716
	Ж	163,9	136,3	134,1	120,6	83,1	41,5	90	100	223	561	1074	1449
Голландия	М	1074,8	812,0	753,4	641,2	538,6	348,6	376	819	1607	3891	2023	13961
	Ж	1021,0	769,7	766,7	671,6	595,1	413,1	414	322	1020	2272	4831	10352
Швеция	М	624,6	466,9	511,9	526,1	457,0	291,2	629	585	1154	2968	6401	1136
	Ж	595,0	449,3	505,6	520,4	477,0	338,9	260	285	763	1794	3948	8505
США	М	14 467	10 866	11 923	11 601	8011	5146	22 406	21 553	45 831	102 426	183 919	256 867
	Ж	14 591	11 246	12 438	11 132	8664	5259	9081	12 097	29 079	58 279	99 227	174 175

Источник. Keyfitz N. and Flieger W. (1968). World Population. University of Chicago Press, p. 162, 222 276, 310, 372, 424, 508.

● УКАЗАТЕЛЬ ПЕРЕВОДА ТЕРМИНОВ*

Абсолютное отклонение	absolute deviation
~ ые значения (модули) отклонений	~ values of deviations
абсцисса; горизонтальная ось	horizontal axis
автоматическая корректировка смещения	automatic bias adjustment
~ программа оптимизации	~ scheme of optimization
аддитивность	additivity
анализ, аддитивный	PLUS analysis
~, альтернативный	alternative ~
~, графический	exploratory plotting
~ данных	data analysis
~, двухфакторный	two-factor ~
~, дисперсионный (ANOVA)	analysis of variance
~ ~, многомерный	~~~, complex
~ ~, полный	~~~, complete
~, итеративный; повторный, циклический	reanalysis
~, латентно-структурный	latent structure analysis
~, многомерный	multiple analysis
~, последовательный	successive analysis
~, причинный	causal ~
~, регрессионный	regression ~
~ ~, множественный	~~, multiple
~, статистический	statistical analysis
~, факторный	factor ~
анкета [при обследовании], вопросник	questionnaire
анкетный тест [с письменными ответами]	paper-and-pencil test
аномалия; аномальные значения	anomalous value
арифметическое среднее	arithmetic average
асимметрия	skewness
асимметрично	skew
аттестационная (экспертная) оценка	assessment
аукционная стоимость	valuation
аффинное преобразование	affine transformation
Балансир	matcher
~ы, линейно-независимые	~s, lineary independent
баланс	balance
балл; число очков	score; scoring grade
~, нормализованный (ранг)	~, normal; rankit

* Термины приводятся в порядке русского алфавита. Знак ~ (тильда) заменяет слово, стоящее на соответствующем месте в предыдущем термине. Синонимы — в круглых скобках. Слова с близкими значениями — через точку с запятой. Пояснения — в квадратных скобках. Английские эквиваленты даны без артиклей. Запятая, как правило, указывает на инверсию порядка слов.

бесконечность	infinity
бесконтрольно	out of control
бета-функция	beta function
бивес	biweight
бивзвешивание	biweighting
биквадратный вес	bisquare weight
биостатистик	biostatistics
блок; однородная группа	batch
бугор; горб	hump
бугорчатый; ломаный; комковатый	humpy
быстродействующая ЭВМ	high-speed computer
Вариации, источники	sourses of variation
~, мера	measure of ~
вариация; изменчивость	variation
~, выборочная,	~, sampling
вверх, выпуклость	upward hollow
вес	weight
~, априорный	~, a priori
~, матричный	~, matrix
веса, выравнивающие	flattened weights
~, неравные	unequal ~
~, последовательные	successive ~
весами, с равными	equally weighted
~~, метод наименьших квадратов	~~ least squares
весов, матрица	matrix of weights
взаимосвязь; корреляция	correlation
взвешенный метод наименьших квадратов	weighted least squares
взвешенная оценка	~ estimate
~ разность	~ difference
~ сумма	~ sum
взвешивание, шаговое	stepweighting
вектор	vector
величина; количество	magnitude; quantity
~, обратная	reciprocal
~ свернутости	size of plurality
~, случайная	random variable
величины, независимые нормально распределенные	independently normally distributed values
~ ~, случайные	independent random variables
вероятность	probability
вероятности, плотность	~, density
~, распределение	~ distribution
вид; форма; очертание	shape
вклад	contribution
~, линейный	~, linear
внешний заменитель	external proxie
внешняя (дополнительная) неопределенность	external uncertainty; supplementary uncertainty
~ точка	outer point
внутренний заменитель (заместитель)	proxie within fit
внутренняя неопределенность	internal uncertainty
~ стандартная ошибка	~ standard error
~ точка	inter point
возведение в степень со сдвигом	started powers
воздействие	handle
возмущение	perturbation
волнистость; неровность	wiggliness
вращение	rotation
вращения, эллипсоид	ellipsoid of revolution

временные ряды	time-series
всплеск	wedge shape
вторичная статистика	secondary statistic
выборка	sample
~ из выборки	subsample
~, контрольная	test sample
~, не вполне случайная	judgment ~
~ независимых наблюдений	sample of independent observations
~, повторная	~, replicate
~, простая случайная	~, random, simple
~, расслоенная	~, stratified
~, случайная	~, random
выборки взаимопроникающие	interpenetrating samples
~, объем	sample size
выборочное обследование	~ survey
выборок, теория	theory of sampling
~ ~ больших	large-sample theory
выборочное среднее	sample mean
выборочный размах	~ range
выброс	outlier
вывод надежный; фидуциальный	fiducial inference
~ статистический	statistical ~
~ формальный	formal ~
выпуклости, правило	bulging rule
выпуклость	hollowness
~ вниз (вогнутость)	downward hollow
выражение	expression
высокочастотный шум	high-frequency noise
вытянутый (удлиненный) хвост	long tail
вычисления, однократные (не итеративные)	once-through calculation
вычислительный метод шаговой регрессии	computational method for stepwise regression
выявление закономерностей; распознавание образов	looking for patterns
Гауссова эффективность	Gaussian efficiency
Гауссово (нормальное) отклонение	~ deviate
генератор [моделей]	stock
геометрическое среднее	geometric mean
генеральная совокупность	population; universe
генеральное среднее	population mean
генеральной совокупности, квадратичная ошибка	~ standard deviation
~ ~ распределение (теоретическое распределение)	parent distribution
гистограмма	histogram
главный член	leading term
гладкая (сглаженная) регрессия	smooth regression
гладкость; сглаженность	smoothness
Гольбаха, числа	Goldbach counts
горизонталь [ная линия];	horizontal line
~ прямая (параллельная абсциссе)	~~, straight
градус; отметка; уровень	grade
граница, доверительная	confidence limit
~, резкая	abrupt terminus
границы, доверительные	confidence limits
~ ~, разумные	~~ sensible
грубое приближение к линии регрессии	crude regression line
группа; кластер	group; cluster
~, компактная	clump

Данные, входные (исходные)	input data
~, дополнительные	additional ~ ; further ~
~, необработанные (сырье)	raw ~
~, скудные	sparse ~
~, массив; файл	file
данных, набор	collection of data; data set
~~ исходных	original set of data
~, порядок	order of data
двусторонний 95%-ный доверительный предел (граница)	two-sided 95% confidence limit
~ 5%-ный уровень	~~ 5% level
«дважды слепое» исследование	«double blind» study
делящийся на (кратный)	devisible by
демографическая статистика	vital statistics
десятичный логарифм	common logarithm
детерминантные уравнения метода главных компонент	determinantal equations of method of principal components
детерминированная система	deterministic system
~ функция	functional relationship
диаграмма рассеяния (поле корреляции)	scatter diagram
диаметр	diameter
дискретность	discreteness
дискретные случайные величины с неупорядоченной областью значений	discrete unordered variables
дискриминант	discriminant
дискриминантная функция, линейная	linear discriminant function
дискриминация	discrimination
дисперсий, отношение	variance ratio
~ — ковариаций, матрица	~ — covariance matrix
дисперсия	variance
~, асимптотическая	~, asymptotic
~, кажущаяся	~, apparent
~, общая	~, common
~, остаточная	~, residual
~ разности	~ of difference
добавочная переменная (фактор)	extra variable
доверительные границы, нормированные (концы доверительного интервала)	standard confidence points
доверительный интервал (граница)	confidence interval
доказательства, статистические	statistical arguments
доля брака за долгое время, средняя	long-run average proportion of defective
~ (соотношение), накопленная	cumulative proportion
~ «успехов» в совокупности	proportion of successes in population
~ ~; примерная (приблизительная)	crude success rate
дополнительная переменная (фактор)	auxiliary variable
дополнительный член	additional term
допуск	allowance
достоверность (надежность)	reliability
~ оценки	reality of assessment
достоверный	trustworthy
дрейф (тренд)	trend
дробь	fraction
Единица; объект; элемент	unit
единичное изменение (скачок)	~ change
единственное решение	unique solution

Зависимая переменная (отклик)	response variable; dependent variable
зависимость	association; dependence
~, линейная	linear dependence
~ ~, приближенная	~~, approximate
~ ~, точная	~~, exact
~ монотонная	monotone relationship
~ однозначная (детерминированная)	exclusive dependence
зависимый, линейно	linearity dependent
задача, многокритериальная (с многими откликами)	multiple-response problem
~, однопараметрическая регрессионная	one-parameter regression problem
зазубренный (зубчатый)	jagged
закон, математический	mathematical law
~, эмпирический (эмпирическое правило)	empirical ~
замкнутая система	closed system
знак	sign; cutoff
~, десятичный	decimal; decimal place
~ минус	minus sign
~ плюс	plus ~
знаков, критерий	sign test
знаменатель	denominator
значение	value
~, «истинное»	~, true
~, критическое	~, critical
~, ожидаемое (математическое ожидание)	~, expected
~, особое (особая точка)	~, letter
~, отрицательное	~, negative
~, положительное	~, positive
~~переменной (фактора)	~~ of variable
~ среднее	~, average (mean)
~ типичное	~, typical
~ центральное	~, central
значений, набор	set of values
значения, упорядоченные	ordered values
значимости, критерий (проверка)	significance test; test of significance
~, уровень	confidence level
~ ~, принятый	conventional level of significance
зеркальное отражение	anti-image
Иерархия	hierarchy
извив кривой [особенность поведения]	bump on curve
изгиб	bend
изменчивость [качественная]	variability
~, наблюдаемая естественная	~, observing natural
~, совместная	covariability
измерение	measurement
~я, ошибка	~ error
~ ~, систематическая (смещение прибора)	error of measurement, systematic
~, ошибки	errors of measurement
~, точность	measurement precision
инвариантность	invariance
индикатор	indicator
~, качественный	~, qualitative
~, количественный	~, quantitative
индикации, устойчивость	stability of indication
индикация (указание)	indication
инструментальная переменная	instrumental variable

интервал класса	class interval
интерквартильный размах	interquartile range
интерполяция	interpolation
исключение (элиминирование)	elimination; exclusion
искусственные («рукотворные») эксперименты	manmade experiments
источники отклонений	sourses of deviation
«истинная» регрессия	true regression
исходный носитель	raw carrier
итеративная модификация метода наименьших квадратов	iteratively modified least squares
итеративный линейный метод взвешенных наименьших квадратов	~ weighted linear least squares
итоги	amounts and counts
Калибровка	calibration
каноническая корреляция	canonical correlation
касательная (тангенс)	tangent
квадрат, полный	complete square
~ ухудшения модели, средний	mean-square failure of fit
квадратов, среднее	average of squares
~, сумма	sum of ~
~~ для предсказания	~~~, prediction
~~ остатков	~~~, residuals
~~ отклонений	~~~ of deviations
~~, средняя	~~~~~, average
квадратичная ошибка (стандартное отклонение)	standard deviation
~~ выборочного среднего	~~ of sample mean
~ форма	quadratic form
~~, неотрицательно определенная	~~, nonnegative
~ функция (полином второго порядка); квадратичное (параболическое) приближение	parabolic fit
квадратичный член	quadratic term
квадратный корень	square root
квантиль	quantile
квантовать	quantize
квартиль	quartile; hinge
клин; клиновидная форма	wedge shape
ковариация	covariance
~, нулевая	~, zero
количественное соотношение	quantitative relationship
количественный метод	~ method
коллекция (собрание)	collection
коллинеарность	collinearity
комбинации наблюдений, линейные	linear combinations of observations
компонент (составная часть)	constituent
~, главный	principal component
~, смысловой	judgement ~
~ы, ортогональные	orthogonal ~s
компьютер (ЭВМ)	computer
константа; постоянная; общий член	common; constant
контроль качества (управление качеством)	quality control
концепция неопределенности (размытое понятие)	vague concept
~ точности	exactness
координат, начало	origin
~, система	coordinate system
корректировка; коррекция	adjustment

коэффициент пропорциональности (постоянный множитель)	constant multiplier
~ уравнения регрессии (регрессионный коэффициент)	regression coefficient
косвенная нормализация	indirect standardizing
краевой эффект	end effect
кривая	curve
~ влияния	~, influence
~, линеаризуемая	~, straightening
~, регрессионная (линия регрессии)	~, regression
~, сглаженная	~, smooth
критерий-отношение	test ratio
~ перепроверки	cross-validation criterion
~ подгонки	fitting criteria
критерия, мощность	power of test
кубический полином	cubic polinomial
Латентная (скрытая) переменная	latent variable
лестница переформулировок (последовательность преобразований)	ladder of re-expressions
~ статистик (оценок)	staircase of statistics
линейная коррекция	linear adjustment
~ модель	~ model
~~, полная	~~, full
~ регрессия (простая регрессия)	simple regression
линейное ограничение	~ constraint
~ преобразование	~ transformation
линейность	linearity
линейный эффект фактора	linear effect of variable
линия, сплошная	heavy line
логарифмирование со сдвигом	started logs
логит (логарифмическая единица)	logit
локализация, точная	exact location
локальное среднее	local average
локальный максимум	~ maxima
~ минимум	~ minima
луч (полупрямая)	half-line
Макроразброс	macro-variability
малая добавка (капля; чуточка)	smidgen
малость	smallness
масштаб; масштабный множитель	scale; scale factor
математическое ожидание (генеральное среднее)	population mean
матричный язык	matrix language
матрица, обратная	inverse matrix
машинная программа	computer program
медиана	median
~ абсолютного отклонения от медианы	~ absolute deviation from median
~ совокупности (теоретическая медиана)	~, population
медианный ранг	~ rank
медианы, место	depth of median
~ скользящие (текущие)	running medians
межквартильный размах	interquartile range
мера положения	measure of location
~ разброса	~~ spread
~~остатков	~~~ of residuals
~ связности	~~ association
~ эффективности	~~ efficiency

метод включения	forward (selection) procedure
~ всех подмножеств	all-subset technique
~ главных компонент	principal components
~ исключения	backward (elimination) procedure; backward technique
~ многократной проверки	multiple-control method
~ наименьших квадратов (МНК)	least squares
~~, обычный	~~, ordinary
~ наименьших модулей (абсолютных отклонений; первых степеней)	~-absolutes; least absolute deviation; least first powers
~~*p*-х степеней	~ *p*-th powers
~ «свободный» от распределений	distribution-free procedure
~ сглаживания; сглаживание	smoothing procedure
~, статистический [техника]	statistical technology
~а наименьших квадратов, решение	least-square solution
механизм	mechanism
~, причинный (причинно-следственный)	~, causal
мешающая переменная	interfering variable
микроразличие (различие «в малом»)	micro-difference
минимизируемое (то, что минимизируется)	minimand
«мишура»	forget
множество факторов, полное	complete set of variables
~ упорядоченных пар	set of ordered pairs
модель	model; fit; form
~, неизвестная	unknown fit
~ распределения	distribution model
моделирование, оценивающее	estimating fitness
монотонность	monotonicity
«мощная опора»	splitting stem
мощность	strength
мультипликативный эффект	multiplicative effect
Наблюдателя, уравнение	personal equation
наблюдение	observation
~, повторное	repeated measurement
~ с большим разбросом	high-variability observation
надежности коэффициент	reliability coefficient
наклон (угол наклона; угловой коэффициент)	slope
наличие соответствия	presence of association
настроен на	tune to
неадекватность (степень неадекватности модели)	lack of fit
невзвешенное (простое) среднее	simple mean
невзвешенность	unweighting
независимая ошибка	independent error
~переменная (фактор)	~ variable
~~ в уравнении для предсказания (предиктор)	predictor
независимо распределено	distributed independently
независимости, отсутствие	failure of independence
нелинейная модель	nonlinear model
~ регрессия	~ regression
нелинейность	nonlinearity
нелинейный генератор	nonlinear stock
~ метод наименьших квадратов	~ least squares
ненадежность (отсутствие надежности)	unreliability; lack of reliability

ненадежность среднего	unreleability of mean
неопределенности, количественный показатель	uncertainty, expressing amount of
~ свертки, оценка	~ of numerical summary, assessing
неопределенность (недостоверность; ненадежность)	~ ; indeterminacy
~ оценки коэффициента	~ of estimate of coefficient
~ числовой свертки	~~ numerical summary
неотрицательность	nonnegativity
непараметрический метод	non-parametric procedure
неподходящий фактор	unrelated variable
непренебрежим	nonnegligible
непрерывная категория [с непрерывно меняющимся признаком]	broad category
~ совокупность	continuous population
неравенство	inequality
нерегулярная функция	irregular function
нерегулярность	irregularity
нерасчлененная опора	unsplit stem
неслучайный	nonrandom
несмещенно относительно среднего	unbiased on average
несмещенность	unbiasedness
несовместные	incompatible
нестабильность (неустойчивость)	instability
неформальный вывод	informal inference
нецентрированные результаты	uncentred form
нечетное число	odd number
низкочастотный шум	low-frequency noise
нормальное отклонение	normal deviation
~~, нормированное (гауссово)	~ deviate, standard
~~, случайное	~~, random
нормально распределенные ошибки	normally distributed errors
нормальность	normality
нормирование	standardizing
нормит (нормальная единица)	normit
носители, мультиколлинеарные	multicollinear carriers
носитель [информации]	carrier
~, перестроенный	~, rearranging
~, спрямляющий	~, straightened
нулевая гипотеза (нуль-гипотеза)	null hypothesis
«нуль-ситуация»	~-situation
Облако точек	point-cloud
обоснованность	validation
обработка	treatment
обследование	survey
~, пробное (предварительное)	~, pilot
обратно пропорционально	inverse proportion
общая причина	common-cause
общее среднее	grand mean
объединение популяций (пул)	pool of populations
объединенная сумма квадратов	pooled sum of squares
ограничение	restraint
огромное отклонение	huge deviation
однородность	homogeneous
окрестность	neighborhood
округлость	roundness
оперативная характеристика	operating characteristic
описательная статистика	descriptive statistics
«опора и консоль»	stem-and-leaf
«опора» с переменной толщиной	change in stem width

«опоры», толщина	stem width
опрашиваемый; респондент	respondent
определитель, ненулевой (невырожденный)	nonzero determinant
определяющий (основной; фундаментальный) фактор	fundamental variable
опрос общественного мнения	public-opinion polling
оптимизация	optimization
опыт, управляемый	controlled trial
ортогонализация	orthogonalization
остатки (отклонения от избранной линии регрессии)	departures from observed regression line
остаток	residual
остаточная переменная	~ variable
осцилляция	oscillation
отклик	response variable
отклонение	deviation
~, среднее	~, mean
~, стандартное	~, standard
~я остатков	~s of residuals
~, стандартного, оценка	estimate of standard deviation
относительная эффективность	relative efficiency
отношение (связь)	relation
отражение	image
отрезок на оси ординат, отсекаемый прямой (свободный член)	intercept
~ прямой	straight-line segment
отрицательная связь, сильная	strong negative relation
отрицательно	negative
отстроен от	tune out
отсутствие соответствия	lack of association
оценка	estimate
~, бивес-	~, biweight
~дисперсии коэффициента регрессии	estimated variance of regression coefficient
~ достоинств	appreciation
~ как формула	estimator
~ количества	evaluation
~ корня из квадратичной ошибки	estimated root-mean-square error
~, несмещенная	unbiased estimate
~ свертки (выборочное значение свертки)	sample summary value
~ стоимостная (ценовая)	appraisal
~, условная	conventional estimate
~ ценностная (назначаемая; аукционная)	valuation
оценки, точность	precision of assessment
оцениваемое	estimand
оцениваемый параметр (подлежащий оценке)	~, parameter being
оценивания, точность	~, precision of
оцениватель	estimator
ошибка	error
~ округления	~, rounding
~, систематическая (смещение)	~, systematic
~ углового коэффициента	~ in slope
Парное произведение	cross-product
пассивный эксперимент	observational study
первичная статистика	primary statistic
переменная (фактор)	variable

переменная, «подставная»	variable, proxy
перепроверка	cross-validation
перепись	census
переподгонка (повторный подбор)	refitting
перестановка	permutation
переформулировка (преобразование)	re-expression
переодичности, степень	degree of periodicity
перпендикуляр	perpendicular
план [эксперимента]	design
планируемый эксперимент, рандомизированный	randomized controlled trials
плато	flat spot
плоскость	plane
~ регрессии	~, regression
плоскостность; ровность	flatness
плотность	density
~ распределения	~ function
поверхность	surface
подбор линии методом наименьших квадратов (подбор МНК-линии)	fit least-squares line
~, шаговый	fitting by stages
~ функции	~ of function
подвыборка	subsumple
подгенератор	costock
подгонка	fit
подгонки, степень	~, degree of
подсчет оценок параметров	determination
показатель степени	exponent
полином	polinomial
~ высокого порядка	~, high-degree
полиномиальная модель	~ fit
положение (расположение; сдвиг)	location
положительная связь, сильная	strong positive relation
положительно	positive
~ полуопределенная матрица	~ semidefinite matrix
полуось	semiaxe
полуразмах (полусумма крайних значений выборки)	midrange
полушаг	half-step
порядковая статистика	order statistics
последовательность длины n, случайная	random sequence of length n
~ чисел, упорядоченная	array of numbers
последовательный подход	sequential approach
правдоподобные рассуждения	probability statement
правило (прием) эмпирическое	rule of thumb
правильная (с постоянным шагом) решетка	grid of equally spaced points
правильность [в смысле меры отклонения от истинного или точного значения] (точность)	accuracy
предсказание (представление)	prediction; fit
~, смещенное	~, biased
~, статистическое	~, statistical
предсказательная сила	worth straightening
предсказательный (экстраполирующий)	straightening
преобразование	re-expression; transformation
~, пробное	~ , tentative
~, тригонометрическое	angular transformation

прецизионность (точность; разрядность машинной сетки)	precision
приближение	fit
~, хорошее	~, good
причина	cause
причинность (каузальность)	causation
причинный подход	causal approach
пробит	probit
прогиб	sag
прогноз (прогнозирование)	forecast [ing]; prediction
программа для ЭВМ, статистическая	statistical computer program
программы наблюдений и эксперимента, план	design in observational program and experiments
проективный тест [психологический]	projective test
пропуск (пустая, незаполненная ячейка)	empty cell
простое число	prime
простота интерпретации коэффициентов	simplicity of interpretation of coefficients
пространство, трехмерное	three-dimentional space
процесс сбора данных	data-gathering process
процент (степень) свернутости	plurality; folded %
процентиль (процентная точка)	percentage point; percentage
прямая [линия]	straight line
прямой, подбор	~~, fitting
~, проходящей через начало координат, уравнение	~~ through origin, equation of
прямоугольник	rectangle
псевдозначение	pseudo-value
псевдонуль	pseudo-zero
пунктир [ная линия]	dotted line
Равномерно (с равным шагом) расположенные	uniformly spread
разброс (рассеивание)	scatter
~ данных (точек) относительно прямой	spread of data around straight line
~~ для биномиального распределения	binomial variation of counts
разбросанные хвосты	straggling tails
размах (различие выборочных средних; разброс)	spread
различие, огромное	large-scale difference
размерность	dimention
размытость (расплывчатость)	fuzziness
разности, счетные	balance
рандомизация	randomness
распределение	distribution
~, апостериорное	~, posterior
~, асимметричное выборочное	~, unsymmetrical sampling
~, безусловное (маргинальное)	~, marginal
~, биномиальное	~, binomial
~, выборочное	~, sampling
~~ рангового коэффициента корреляции	~~ of rank — correlation coefficient
~, загрязненное	~, contaminated
~, идеальное (теоретическое)	~, complete
~ индивидуальных значений	~ of individual values
~, колоколообразное	~, bell-shaped
~ Коши	~, Cauchy
~, кумулятивное логистическое	~, cumulative logistic

распределение, кумулятивное логистическое, обратное	distribution, cumulative logistic, inverse
~, логистическое	~, logistic
~, нормальное (гауссово)	~, normal
~ ~, кумулятивное	~~, cumulative
~~, обратное	~~~, inverse
~, нормированное нормальное	~, standard normal
~, прямоугольное (равномерное)	~, rectangular
~, равномерное	~, uniform
~ с бесконечным стандартным отклонением	~ with infinite standard deviation
~ ~ «поджатым» хвостом	~, squeezed-tailed
~~ разбросанными хвостами («длиннохвостое»)	~, straggling-tailed
~ симметричное треугольное	~, symmetric triangular
~, треугольное	~, triangular
~, условное	~, conditional
~, фидуциальное	~, fiducial
~ хи-квадрат (χ^2)	~, $-\chi^2$
~, эмпирическое	~, empirical
распределений, смесь нормальных	~s, mixture of normal
распределенные случайные числа, равномерно	uniform random numbers
расслоение (стратификация)	stratification
рассеяние	dispersion
растянутые хвосты	stretched tails
расчет приближенный	rough calculation
~ прикидочный	crude ~
регрессии, уравнение	regression equation
регрессионная функция	~ function
регрессия	regression
~, гибкая (управляемая)	~, guided
~ для измерения	~-as-measurement
~, структурная	~, structural
~ y по x	~ of y on x
результаты, центрированные	centered form
робастность (устойчивость)	robustness
~ (~) к предпосылкам	~ of validity
~ (~) ~ эффективности	~~ efficiency
ранкит	rankit
ряд, степенной	power serie
Свертка (свертывание)	folding
~ [процесс]	summarize
~ [результат] (сводка)	summary
~, графическая	~, graphical
~, общая	~, grand
~, статистическая	summarize statistics
~, числовая	summary, numerical
свертывание	summarization; folding
~ корнем	folded roots (froots)
~ логарифмом	~ logs (flogs)
свободный член уравнения регрессии (точка пересечения прямой с осью абсцисс)	level of line
связь, умеренная (относительно слабая)	modest relation
сглаживание	smoothing
~ текущее по 3 соседним точкам	~, 3R
сдвиг	start
середина интервала класса	midpoint of class interval

сетка, параллельная диагонали	parallel diagonal bars
сечение, поперечное	cross section
сильно смещенный	severely biased
симметричное усечение	symmetrical trimming
синусоидальный характер	sinusoidal character
систематическая составляющая	systematic variation
«складной нож»	jackknife
скользящее (текущее) среднее	running means
следствие	consequence
сложная система	complex system
слой	layer; stratum
случайное число	random digit
случайных чисел, таблица	~-number table
смещения	bias
совокупность, бесконечная	population, infinite
~, дихотомическая	~, dichotomous
~, изучаемая (искомая; требуемая)	~, target
~, эталонная (эталон)	~, standard
согласие	agreement
соотношение	relationship
сопоставление (взаимодействие)	comparison value
сопротивляемость (стойкость; устойчивость)	resistence
серединное (усеченное) среднее	midmean
среднее	average; mean
~ по столбцу	~, column
~, усеченное (урезанное; цензурированное)	trimmed mean
средняя по большой выборке сумма квадратов отклонений оценок	long-run average sum of squares of deviations of estimates
состоятельность (непротиворечивость)	consistency
состязание (конкуренция)	competition
стабильность (устойчивость)	stability
стабилизировать	held constant
степени свободы (число степеней свободы)	degrees of freedom
степень связи	degree of relation
стиснутый концами	squeezed-in tails
структура	structure
сумма парных произведений отклонений	sum of cross-products of deviations
счетная дробь [правильная] (отношение счетных сумм; счетная доля)	counted fraction
счетные суммы	amounts and counts
Таблица с двумя входами (таблица сопряженности 2 × 2)	two-way table
~, сканирующая	break table
~ значений хи-квадрат	chi-square table
теория чисел	number theory
теснота связи	closeness of relation
точка	point
~ «возврата» (локальный экстремум)	~, turning
точно	precise
точность	exactness
тренд	trend
третичная статистика	tertiary statistics
трудности, шкала	difficulty, scale of
трудность	difficulty
тяжести, центр	center of gravity

Угол	corner
уловитель	catcher
упорядоченные ярлыки	grades-ordered labels
уравнение множественной регрессии	multiple regression equation
~ свободного члена	intercept equation
~ углового коэффициента	slope ~
уравновешенная сумма наблюдаемых значений	matcher-weighted sum of observed values
~ ~ предсказанных ~	~—~~~fitted values of response variable
уровни группирования (группировки)	levels of grouping
усреднение	averaging
устойчивость	stability
Функция правдоподобия	likelihood function
~ потерь	loss ~
Характеристика [логарифма]	characteristic
хвост	tail
хорда	chord
Целое число	integer
центрированный свободный член уравнения прямой	centered intercept
центроид (центр «тяжести»)	centroid
Частные производные по коэффициентам	partial derivations with respect to coefficients
четное число	even number
чувствительность	responsiveness
~ результатов	sensitive of results
Эксперимент	experiment
~ в естественных условиях (натурный)	~, natural
экстраполяция	extrapolation
эксцесс	kurtosis
эффект	effect
~ фактора	~ of variable
эффективность	effectiveness
Яйцевидный	ovoid
ячейка	cell

● ОГЛАВЛЕНИЕ ВЫПУСКА 1

● ОГЛАВЛЕНИЕ

Мостеллер Ф., Тьюки Дж.

М84 Анализ данных и регрессия: В 2-х вып. Вып. 2 / Пер. с англ. Б. Л. Розовского; Под ред. и с предисл. Ю. П. Аллера. — М.: Финансы и статистика, 1982. — 239 с., ил. — (Математико-статистические методы за рубежом).

В пер.: 1 р. 90 к.

В книге исследуются проблемы границ применимости статистических методов к анализу реального мира, проблемы качества статистических выводов — что в них существенно и что несущественно. Под этим углом зрения рассматриваются основные статистические методы, предлагаются новые подходы. Второй выпуск посвящен главным образом проблемам регрессионного анализа.

Для статистиков, экономистов, демографов. Полезна студентам старших курсов по этим специальностям.

М $\frac{0702000000-165}{010(01)-82}$ 40—82

ББК 22.172
517.8

Ф. Мостеллер, Дж. Тьюки

АНАЛИЗ ДАННЫХ И РЕГРЕССИЯ

Рекомендована к изданию редколлегией серии
5 июня 1979 г.

Зав. редакцией *А. В. Павлюков*
Редактор *К. М. Чижевская*
Мл. редактор *И. Н. Горина*
Техн. редакторы *К. К. Букалова, Г. А. Полякова*
Корректоры *Г. В. Хлопцева, З. С. Кандыба*
Худож. редактор *Э. А. Смирнов*

ИБ № 941

Сдано в набор 22.04.82 Подписано в печать 4.10 82.
Формат 60×90$^1/_{16}$. Бум. тип. № 2. Гарнитура «Литературная». Печать высокая. П л. 15,0. Усл. п. л. 15,0. Усл. кр.-отт. 15,0. Уч.-изд. л. 16,20. Тираж 6500 экз. Заказ 935. Цена 1 р. 90 к.

Издательство «Финансы и статистика», Москва, ул. Чернышевского, 7

Московская типография № 4 Союзполиграфпрома при Государственном комитете СССР по делам издательств, полиграфии и книжной торговли
129041, Москва, Б. Переяславская ул., д. 46