Ф. Мостеллер
Дж. Тьюки

АНАЛИЗ ДАННЫХ И РЕГРЕССИЯ

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. Адлера. — 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*
Художн 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

МАТЕМАТИКО-СТАТИСТИЧЕСКИЕ МЕТОДЫ ЗА РУБЕЖОМ

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; Гласс 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 (≈ 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:

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

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.

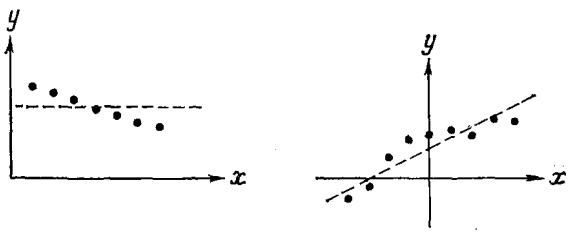
В этом случае мы *ни в коей мере* не претендуем на то, что получившаяся кривая имеет ту же форму, что и регрессионная кривая, которую можно было бы построить, имей мы неограниченные данные. Построенная таким образом кривая — не более чем аппроксимация.

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

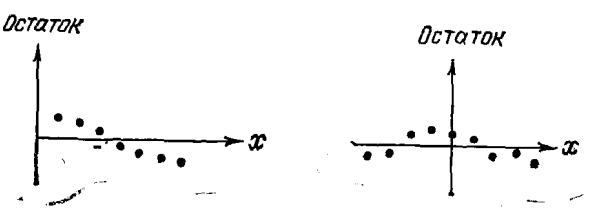
значительно расширяет область применимости регрессионных методов, так как на практике мы часто располагаем скромным набором экспериментальных данных, значительно меньшим, чем требуется для метода сглаживания, и совершенно недостаточным для эффективного применения первой концепции (здесь требуются тысячи 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-