Норман Р. Дрейпер

Прикладной регрессионный анализ

Том 2

Москва
«Книга по Требованию»

УДК 51
ББК 22.1
Н83

Норман Р. Дрейпер
Н83 Прикладной регрессионный анализ: Том 2 / Норман Р. Дрейпер – М.: Книга по Требованию, 2021. – 350 с.

ISBN 978-5-458-25572-1

Работа американских ученых посвящена регрессионному анализу, применяемому во всех отраслях народного хозяйства н научных исследованиях. Второе издание (1-е изд. перевода 1973 г. вышло в одной книге) значительно переработано и дополнено новыми алгоритмами и сравнением их достоинств. Кн. 1 содержит классическое описание модели линейной регрессии, включая описание алгоритмов для ЭВМ. В кн. 2 приводится описание модели, нелинейной по параметрам регрессии, обширная библиография и приложения.Для специалистов статистиков, экономистов, социологов, научных работников.

ISBN 978-5-458-25572-1

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

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

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

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

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

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

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

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

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

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

Ю. АДЛЕР, В. ГОРСКИЙ

Глава 6. ● ВЫБОР «НАИЛУЧШЕГО» УРАВНЕНИЯ РЕГРЕССИИ

6.0. Введение

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

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

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

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

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

Некоторые предостережения относительно использования данных пассивного эксперимента

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

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

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

6.1. МЕТОД ВСЕХ ВОЗМОЖНЫХ РЕГРЕССИЙ

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

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

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

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

Статистика R^2

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

---

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

Серия В включает все двухфакторные уравнения (модель $E(Y) = \beta_0 + \beta_i X_i + \beta_j X_j$).

Серия Г состоит из всех трехфакторных уравнений (модель строится аналогично).

Серия Д — из всех уравнений с четырьмя факторами.

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

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

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

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

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

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

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

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

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

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

Остаточный средний квадрат s^2

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

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

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

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

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

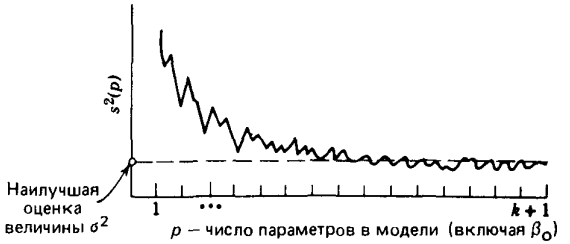

Рис. 6.1. Переподгонка, показывающая типичную стабилизацию s^2

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

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

Критерий Маллоуза C_p

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

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

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

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

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

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

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

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

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

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

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

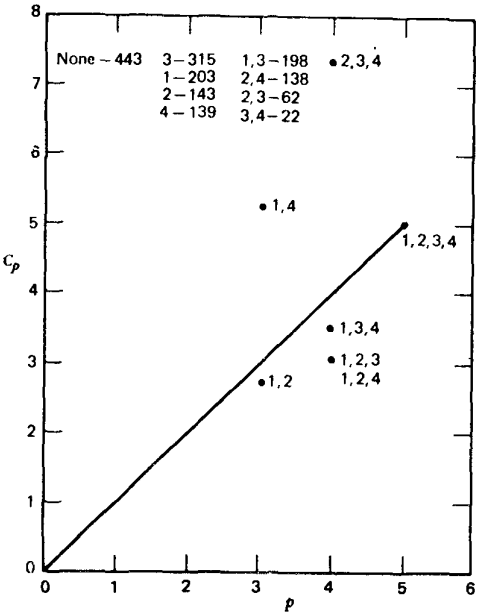

Рис. 6.3. График статистики C_p для данных Хальда

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

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

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