Введение в Бутстреп

2021

Глава 9

Регрессионные модели

9.1 Введение

Регрессионные модели являются одними из самых полезных и наиболее часто используемых статистических методов. Они предлагают относительно простой анализ сложных ситуаций, когда мы пытаемся отсортировать влияние многих возможных объясняющих переменных на зависимую переменную. В главе 7 мы используем алгоритм одновыборочного бутстрепа, алгоритм для анализа точности регрессионного анализа данных холостирамина из таблицы 7.4. Здесь мы более критически смотрим на задачу регрессии. Рассмотрен общий бутстреп алгоритм, показанный на рис. 8.3, что приводит к несколько иному бутстреп анализу для задач регрессии.

9.2 Линейная регрессионная модель

Мы начнем с классической модели линейной регрессии, или линейной модели, восходящей к Лежандру и Гауссу в начале 19 века. Набор данных $\mathbf{x}$ для модели линейной регрессии состоит из n точек $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$, где каждый $\mathbf{x}_i$ сам по себе является парой, скажем

$$\mathbf{x}_i = (\mathbf{c}_i, y_i). \tag{9.1}$$

Здесь $\mathbf{c}_i$ — это $1 \times p$ вектор $\mathbf{c}_i = (c_{il}, c_{i2}, \cdots, c_{ip})$, называемый *вектором признаков* или *предиктором*, а y_i — действительное число, называемое *ответом*.

Пусть μ_i указывает условное ожидание i-го ответа y_i с учетом предиктора $\mathbf{c}_i$,

$$\mu_i = \mathrm{E}(y_i|\mathbf{c}_i) \quad (i = 1, 2, \cdots, n). \tag{9.2}$$

Ключевое предположение в линейной модели состоит в том, что μ_i представляет собой линейную комбинацию компонентов предиктора $\mathbf{c}_i$,

$$\mu_i = \mathbf{c}_i\boldsymbol{\beta} = \sum_{j=1}^{p} c_{ij}\beta_j. \tag{9.3}$$

Вектор параметров, или *параметр регрессии*, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^{\mathrm{T}}$ неизвестен, обычная цель регрессионного анализа состоит в том, чтобы вывести $\boldsymbol{\beta}$ из наблюдаемых данных $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$. В квадратичной регрессии (7.20) для данных холостирамина ответ y_i — это улучшение для i-го человека, признак $\mathbf{c}_i$ — это вектор $(1, z_i, z_i^2)$ и $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$. Примечание: «Линейность» в линейной регрессии относится к линейной форме математического ожидания (9.3). Нет никакого противоречия в том, что линейная модель (7.20) является квадратичной функцией z.

Вероятностная структура линейной модели обычно выражается как

$$y_i = \mathbf{c}_i\boldsymbol{\beta} + \varepsilon_i \quad \text{для} \quad i = 1, 2, \cdots, n. \tag{9.4}$$

Предполагается, что ошибка ε_i в (9.4) является случайной выборкой из неизвестного *распределения ошибок* F с математическим ожиданием 0,

Table 9.1. *The hormone data. Amount in milligrams of anti-inflammatory hormone remaining in 27 devices, after a certain number of hours of wear. The devices were sampled from 3 different manufacturing lots, called A, B, and C. Lot C looks like it had greater amounts of remaining hormone, but it also was worn the least number of hours. A regression analysis clarifies the situation.*

lot	hrs	amount	lot	hrs	amount	lot	hrs	amount
A	99	25.8	B	376	16.3	C	119	28.8
A	152	20.5	B	385	11.6	C	188	22.0
A	293	14.3	B	402	11.8	C	115	29.7
A	155	23.2	B	29	32.5	C	88	28.9
A	196	20.6	B	76	32.0	C	58	32.8
A	53	31.1	B	296	18.0	C	49	32.5
A	184	20.9	B	151	24.1	C	150	25.4
A	171	20.9	B	177	26.5	C	107	31.7
A	52	30.4	B	209	25.8	C	125	28.5
mean:	150.6	23.1		233.4	22.1		111.0	28.9

$$F \to (\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n) = \boldsymbol{\varepsilon} \quad [\mathrm{E}_F(\boldsymbol{\varepsilon}) = 0]. \tag{9.5}$$

Заметим, что (9.4), (9.5) влекут

$$\begin{aligned} \mathrm{E}(y_i|\mathbf{c}_i) &= \mathrm{E}(\mathbf{c}_i\boldsymbol{\beta} + \varepsilon_i|\mathbf{c}_i) = \mathrm{E}(\mathbf{c}_i\boldsymbol{\beta}|\mathbf{c}_i) + \mathrm{E}(\varepsilon_i|\mathbf{c}_i) \\ &= \mathbf{c}_i\boldsymbol{\beta}, \end{aligned} \tag{9.6}$$

что является предположением о линейности (9.3). Здесь мы использовали тот факт, что условное ожидание $\mathrm{E}(\varepsilon_i|\mathbf{c}_i)$ совпадает с безусловным ожиданием $\mathrm{E}(\varepsilon_i) = 0$, поскольку ε_i выбираются независимо от $\mathbf{c}_i$.

Мы хотим оценить вектор параметра регрессии $\boldsymbol{\beta}$ из наблюдаемых данных $(\mathbf{c}_1, y_1), (\mathbf{c}_2, y_2), \cdots, (\mathbf{c}_n, y_n)$. Пробное значение $\boldsymbol{\beta}$, скажем $\mathbf{b}$, дает *остаточную квадратичную ошибку*

$$\mathrm{RSE}(\mathbf{b}) = \sum_{i=1}^{n} (y_i - \mathbf{c}_i\mathbf{b})^2, \tag{9.7}$$

как в уравнении (7.21). Оценка *методом наименьших квадратов* $\boldsymbol{\beta}$ — это значение $\hat{\boldsymbol{\beta}}$ из $\mathbf{b}$, которое минимизирует $\mathrm{RSE}(\mathbf{b})$,

$$\mathrm{RSE}(\hat{\boldsymbol{\beta}}) = \min_{\mathbf{b}}[\mathrm{RSE}(\mathbf{b})]. \tag{9.8}$$

Пусть $\mathbf{C}$ — матрица размера $n \times p$ с i-й строкой $\mathbf{c}_i$ (design matrix), а $\mathbf{y}$ — вектор $(y_1, y_2, \cdots, y_n)^{\mathrm{T}}$. Тогда оценка методом наименьших квадратов является решением следующего уравнения

$$\mathbf{C}^{\mathrm{T}}\mathbf{C}\hat{\boldsymbol{\beta}} = \mathbf{C}^{\mathrm{T}}\mathbf{y} \tag{9.9}$$

и задается формулой

$$\hat{\boldsymbol{\beta}} = (\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{y}. \tag{9.10}$$

9.3 Пример: данные по гормонам

В таблице 9.1 показан небольшой набор данных, который является подходящим для регрессионного анализа. Медицинское устройство для непрерывной доставки противовоспалительного гормона было протестировано на 27 пациентах. Переменная ответа y_i — это количество гормона, оставшееся в устройстве после ношения,

$$y_i = \text{ оставшееся количество гормона в устройстве}i,\ i = 1, 2, \cdots, 27.$$

Есть две переменные–предикторы,

$$z_i = \text{ количество часов ношения i-го устройства}$$

и

$$L_i = \text{ производственная партия устройства i.}$$

Тестируемые устройства были случайным образом выбраны из трех различных производственных партий, названных A, B и C.

Левая часть рисунка 9.1 представляет собой диаграмму рассеяния 27 точек $(z_i, y_i) = (\text{часы}_i, \text{число}_i)$ с символом L_i, используемым в качестве графического сивола. Мы видим, что более длительное время ношения приводит к меньшему количеству оставшегося гормона, как и следовало ожидать. Мы можем количественно оценить это наблюдение с помощью регрессионного анализа.

Рассмотрим модель, в которой математическое ожидание y является линейной функцией z,

$$\mu_i = \mathrm{E}(y_i|z_i) = \beta_0 + \beta_1 z_i \quad i = 1, 2, \cdots, 27. \tag{9.11}$$

Эта модель игнорирует L_i: он имеет форму (9.3) с векторами признаков размерности $p = 2$,

$$\mathbf{c}_i = (1, z_i). \tag{9.12}$$

Вектор неизвестных параметров $\boldsymbol{\beta}$ был помечен (β_0, β_1) вместо β_1, β_2, так что индексы соответствуют степеням z, как в (7.20). Нормальные уравнения (9.10) дают оценку наименьших квадратов

$$\hat{\boldsymbol{\beta}} = (34.17, -0.0574)^{\mathrm{T}}. \tag{9.13}$$

Линия регрессии, оцененная методом наименьших квадратов

$$\hat{\mu}_i = \mathbf{c}_i\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 z_i \tag{9.14}$$

изображена на правой части рисунка 9.1. Среди всех возможных линий, которые можно было нарисовать, эта линия минимизирует сумму квадратов 27 вертикальных расстояний от точек до линии.

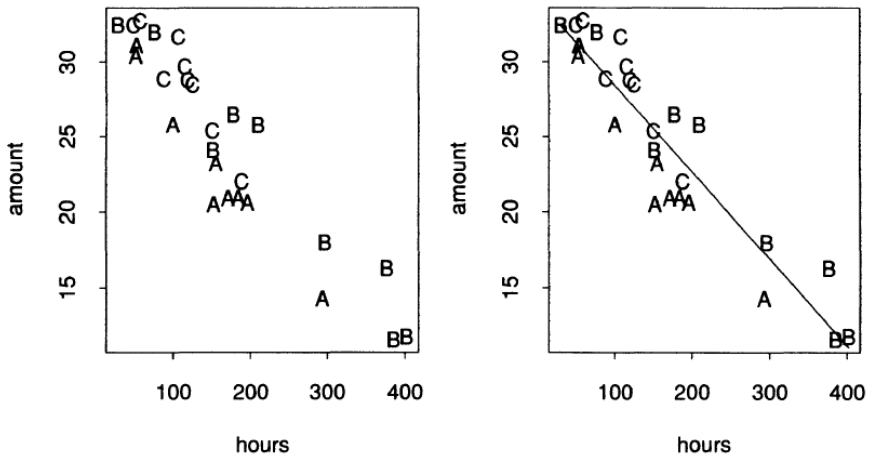

Figure 9.1. *Scatterplot of the hormone data points* $(z_i, y_i) = (hours_i, amount_i)$*, labeled by lot. It is clear that longer hours of wear result in lower amounts of remaining hormone. The right panel shows the least-squares regression of* y_i *on* z_i *:* $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 z_i$*, where* $\hat{\beta} = (34.17, -.0574)$.

Насколько точен оценочный вектор параметров $\hat{\boldsymbol{\beta}}$? Ответ дает чрезвычайно полезная формула, также восходящая к Лежандру и Гауссу. Пусть $\mathbf{G}$ — матрица скалярных произведений $p \times p$,

$$\mathbf{G} = \mathbf{C}^{\mathrm{T}}\mathbf{C}, \tag{9.15}$$

матрица с элементом $g_{hj} = \sum_{i=1}^{n} c_{ih}c_{ij}$ в строке h, столбце j. Пусть σ_F^2 будет дисперсией ошибок в модели (9.4),

$$\sigma_F^2 = \mathrm{var}_F(\varepsilon). \tag{9.16}$$

Тогда стандартная ошибка j-го компонента $\hat{\boldsymbol{\beta}}$, квадратного корня из его дисперсии, равна

$$\mathrm{se}(\hat{\beta}_j) = \sigma_F\sqrt{G^{jj}} \tag{9.17}$$

где G^{jj} — j-й диагональный элемент обратной матрицы $\mathbf{G}^{-1}$.

Последняя формула является обобщением формулы (5.4) для стандартной ошибки выборочного среднего, $\mathrm{se}_F(\bar{x}) = \sigma_F/\sqrt{n}$, см. задачу 9.1. На практике σ_F оценивается по формуле, аналогичной (5.11),

$$\hat{\sigma}_F = \{\sum_{i=1}^{n}(y_i - \mathbf{c}_i\hat{\boldsymbol{\beta}})^2/n\}^{1/2} = \{\mathrm{RSE}(\hat{\boldsymbol{\beta}})/n\}^{1/2} \tag{9.18}$$

или версией $\hat{\sigma}_F$ с корректированным смещением,

Table 9.2. *Results of fitting model (9.11) to the hormone data*

	Estimate	$\widehat{\text{se}}$	$\overline{\text{se}}$
$\hat{\beta}_0$	34.17	.83	.87
$\hat{\beta}_1$	-.0574	.0043	.0045

Table 9.3. *Results of fitting model (9.21) to the hormone data.*

	Estimate	$\widehat{\text{se}}$	$\overline{\text{se}}$
$\hat{\beta}_A$	32.13	.69	.75
$\hat{\beta}_B$	36.11	.89	.97
$\hat{\beta}_C$	35.60	.60	.66
$\hat{\beta}_1$	-.0601	.0032	.0035

$$\bar{\sigma}_F = \{\text{RSE}(\hat{\boldsymbol{\beta}})/(n-p)\}^{1/2}. \tag{9.19}$$

Соответствующие оценочные стандартные ошибки для компонентов $\hat{\boldsymbol{\beta}}$ равны

$$\widehat{\text{se}}(\hat{\beta}_j) = \hat{\sigma}_F\sqrt{G^{jj}} \text{ или } \overline{\text{se}}(\hat{\beta}_j) = \hat{\sigma}_F\sqrt{G^{jj}}. \tag{9.20}$$

Связь между $\widehat{\text{se}}(\hat{\beta}_j)$ и $\overline{\text{se}}(\hat{\beta}_j)$ такая же, как между формулами (5.12) и (2.2) для среднего.

Большинство программ для линейной регрессии с библиотеками обычно выдают результат $\overline{\text{se}}(\hat{\beta}_j)$ вместе с оценкой $\hat{\beta}_j$ методом наименьших квадратов. Применение такой программы к модели (9.11) для данных по гормону дает результаты в таблице 9.2.

Глядя на правую часть рисунка 9.1, большинство точек для партии A лежат ниже подобранной линии регрессии, в то время как большинство точек для партий B и C лежат выше этой линии. Это говорит о неточности модели (9.11). Если бы модель была точной, можно было бы ожидать, что примерно половина каждой партии будет лежать выше, а половина ниже установленной линии. Выражаясь обычной терминологией, похоже, что в данных присутствует эффект партии.

В нашу линейную модель легко включить эффект партии. Мы предполагаем, что условное математическое ожидание y при заданных L и z имеет вид

$$\text{E}(y|L,z) = \beta_L + \beta_1 z. \tag{9.21}$$

Здесь B_L равно одному из трех возможных значений: $\beta_A, \beta_B, \beta_C$, в зависимости от партии устройства. Это похоже на модель (9.11), за исключением того, что (9.21) допускает разные точки пересечения для каждой партии,

а не одну точку пересечения β_0 из (9.11). Анализ модели (9.21) методом наименьших квадратов дал результаты в таблице 9.3.

Обратите внимание, что $\hat{\beta}_A$ на несколько стандартных ошибок меньше чем $\hat{\beta}_B$ и $\hat{\beta}_C$, что указывает на то, что устройства в партии A содержат значительно меньше гормона.

9.4 Применение бутстрепа

Пока ни один из расчетов не требует бутстрепа. Однако полезно выполнить бутстреп-анализ для модели линейной регрессии. Оказывается, оценки стандартной ошибки бустрепа такие же, как $\hat{\text{se}}(\hat{\beta}_j)$, (9.20). Убедившись, что бутстреп дает разумные ответы в случае, который мы можем проанализировать математически, мы можем продолжить применять бутстреп к более общим моделям регрессии, которые не имеют математического решения: где функция регрессии нелинейна по параметрам $\boldsymbol{\beta}$, и где мы используем методы подбора, отличные от метода наименьших квадратов.

Вероятностная модель $P \to \mathbf{x}$ для линейной регрессии, как описано в (9.4), (9.5), состоит из двух компонентов:

$$P = (\boldsymbol{\beta}, F), \tag{9.22}$$

где $\boldsymbol{\beta}$ — вектор параметров коэффициентов регрессии, а F — распределение вероятностей ошибок. Общий алгоритм бустрепа на рис. 8.3 требует, чтобы мы оценили P. У нас уже есть доступная $\hat{\boldsymbol{\beta}}$, оценка методом наименьших квадратов для $\boldsymbol{\beta}$. Как мы можем оценить F? Если бы $\boldsymbol{\beta}$ было известно, мы могли бы вычислить ошибки $\varepsilon_i = y_i - \mathbf{c}_i\boldsymbol{\beta}$ для $i = 1, 2, \cdots, n$ и оценить F по их эмпирическому распределению. Мы не знаем $\boldsymbol{\beta}$, но можем использовать $\hat{\boldsymbol{\beta}}$ для вычисления аппроксимации ошибок

$$\hat{\varepsilon}_i = y_i - \mathbf{c}_i\hat{\boldsymbol{\beta}}, \text{ для } i = 1, 2, \cdots, n. \tag{9.23}$$

($\hat{\varepsilon}_i$ также называют *остатками*.) Очевидная оценка F — это эмпирическое распределение $\hat{\varepsilon}_i$,

$$\hat{F}: \text{ вероятность } 1/n \text{ для } \hat{\varepsilon}_i \text{ при } i = 1, 2, \cdots, n. \tag{9.24}$$

Обычно $\hat{F}$ будет иметь ожидание 0, как требуется в (9.5), см. задачу 9.5.

Имея в руках $\hat{P} = (\hat{\boldsymbol{\beta}}, \hat{F})$, мы знаем, как рассчитать наборы данных бутстрепа для модели линейной регрессии: $\hat{P} \to \mathbf{x}^*$ должно означать то же самое, что и $P \to \mathbf{x}$, вероятностный механизм (9.4), (9.5), дающий фактический набор данных $\mathbf{x}$. Чтобы сгенерировать $\mathbf{x}^*$, мы сначала выбираем случайную выборку бустреп ошибок

$$\hat{F} \to (\varepsilon_1^*, \varepsilon_2^*, \cdots, \varepsilon_n^*) = \boldsymbol{\varepsilon}^*. \tag{9.25}$$

Каждый ε_i^* равен любому из n значений $\hat{\varepsilon}_j$ с вероятностью $1/n$. Затем бустреп ответы y_i^* генерируются согласно (9.4),

$$y_i^* = \mathbf{c}_i\hat{\boldsymbol{\beta}} + \varepsilon_i^* \text{ для } i = 1, 2, \cdots, n. \tag{9.26}$$

Читатель должен убедиться, что (9.24), (9.25), (9.26) то же самое, что (9.4), (9.5), за исключением того, что $\hat{P} = (\hat{\boldsymbol{\beta}}, \hat{F})$ заменяет $P = (\boldsymbol{\beta}, F)$. Обратите внимание, что $\hat{\boldsymbol{\beta}}$ — фиксированная величина в (9.26), имеющая одинаковые значения для всех i.

Бутстреп набор данных $\mathbf{x}^*$ равен $(\mathbf{x}_1^*, \mathbf{x}_2^*, \cdots, \mathbf{x}_n^*)$, где $\mathbf{x}_i^* = (\mathbf{c}_i, y_i^*)$. Может показаться странным, что векторы признаков $\mathbf{c}_i$ для бутстреп данных такие же, как и для фактических данных. Это происходит потому, что мы рассматриваем $\mathbf{c}_i$ как фиксированные величины, а не как случайные. (Во всех наших примерах размер выборки n трактовался одинаково.) Этот момент дополнительно обсуждается ниже.

Бустреп оценка $\hat{\boldsymbol{\beta}}^*$ методом наименьших квадратов является минимизатором квадратичной остаточной ошибки для бустреп данных,

$$\sum_{i=1}^{n}(y_i^* - \mathbf{c}_i\hat{\boldsymbol{\beta}}^*)^2 = \min_{\mathbf{b}}(y_i^* - \mathbf{c}_i\mathbf{b})^2. \tag{9.27}$$

Нормальные уравнения (9.10), примененные к бутстреп данным, дают

$$\hat{\boldsymbol{\beta}}^* = (\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{y}^*. \tag{9.28}$$

В этом случае нам не нужны симуляции Монте–Карло, чтобы вычислить бутстреп стандартные ошибки для компонентов $\hat{\boldsymbol{\beta}}^*$. Несложный расчет дает выражение в явной форме для $\mathrm{se}_{\hat{F}}(\hat{\boldsymbol{\beta}}_j^*) = \hat{\mathrm{se}}_\infty(\hat{\beta}_j)$, идеальной оценки бутстреп стандартной ошибки:

$$\begin{aligned}\mathrm{var}(\hat{\boldsymbol{\beta}}^*) &= (\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{T}}\mathrm{var}(\mathbf{y}^*)\mathbf{C}(\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1} \\ &= \hat{\sigma}_F^2(\mathbf{C}^{\mathrm{T}}\mathbf{C})^{-1},\end{aligned} \tag{9.29}$$

поскольку $\mathrm{var}(\mathbf{y}^*) = \hat{\sigma}_F^2\mathbf{I}$, где $\mathbf{I}$ — единичная матрица. Следовательно

$$\hat{\mathrm{se}}_\infty(\hat{\beta}_j) = \hat{\sigma}_F\sqrt{G^{jj}}. \tag{9.30}$$

Другими словами, бутстреп оценка стандартной ошибки для $\hat{\beta}_j$ такая же, как и обычная оценка $\hat{\mathrm{se}}(\hat{\beta}_j)$, (9.20).

9.5 Бутстреп пары против бутстреп остатков

Читатель, возможно, заметил интересный факт: теперь у нас есть два разных способа бутстреп регрессионной модели. Метод, описанный в главе 7, выбирал пары $\mathbf{x}_i = (\mathbf{c}_i, y_i)$, так что бутстреп набор данных $\mathbf{x}^*$ имел форму

$$\mathbf{x}^* = \{(\mathbf{c}_{i_1}, y_{i_1}), (\mathbf{c}_{i_2}, y_{i_2}), \cdots, (\mathbf{c}_{i_n}, y_{i_n})\}, \tag{9.31}$$

для $i_1, i_2, \cdots, i_n$ в случайной выборке целых чисел от 1 до n. Обсуждаемый в этой главе метод (9.24), (9.25), (9.26) можно назвать «бутстрепом остатков». Он создает бутстреп наборы данных в форме

$$\mathbf{x}^* = \{(\mathbf{c}_1, \mathbf{c}_1\hat{\boldsymbol{\beta}} + \hat{\varepsilon}_{i_1}), (\mathbf{c}_2, \mathbf{c}_2\hat{\boldsymbol{\beta}} + \hat{\varepsilon}_{i_2}), \cdots, (\mathbf{c}_n, \mathbf{c}_n\hat{\boldsymbol{\beta}} + \hat{\varepsilon}_{i_n})\}. \tag{9.32}$$

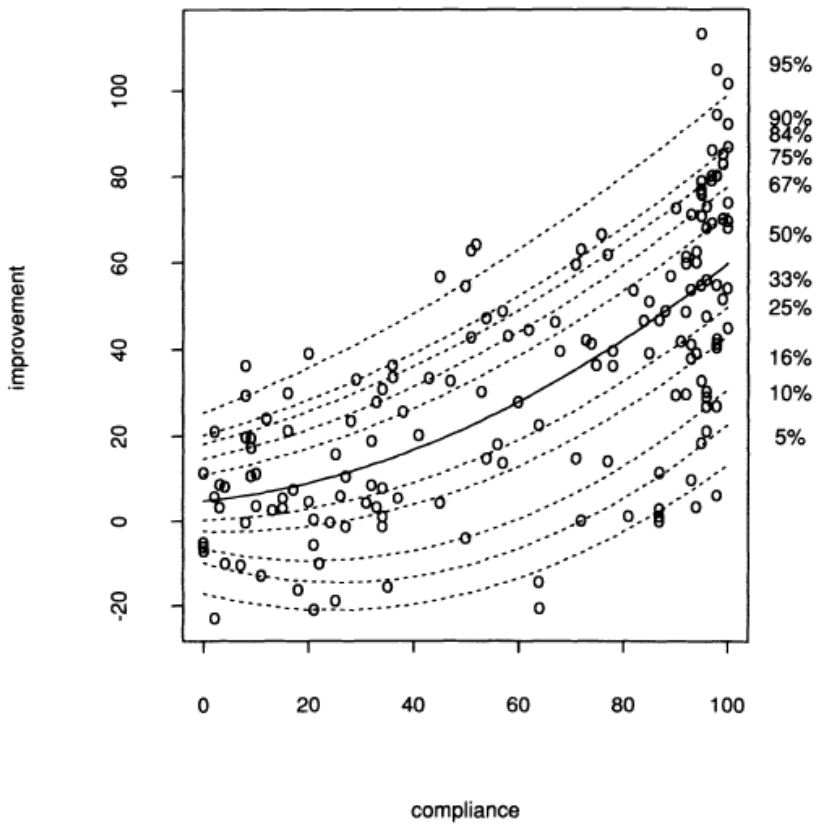

Figure 9.2. *Regression percentiles for the cholostyramine data of Figure 7.5; for example the curve labeled "75%" approximates the conditional 75th percentile of the Improvement y given the Compliance z, plotted as a function of z. The percentile curves are twice as far apart at $z = 100$ as at $z = 0$. The linear regression model (9.4), (9.5) can't be correct for this data set. (Regression percentiles calculated using asymmetric maximum likelihood, Efron, 1991.)*

Какой бутстреп метод лучше? Ответ зависит от того, насколько мы доверяем модели линейной регрессии (9.4). Эта модель говорит, что ошибка между y_i и его средним значением $\mu_i = \mathbf{c}_i\boldsymbol{\beta}$ не зависит от $\mathbf{c}_i$; он имеет одинаковое распределение «F» независимо от $\mathbf{c}_i$. Это сильное предположение, которое может оказаться неверным, даже если модель математического ожидания $\mu_i = \mathbf{c}_i\boldsymbol{\beta}$ верна. Это не соответствует данным холостирамина на рис. 7.4.

На рисунке 9.2 показаны *процентили регрессии* для данных холостирамина. Например, кривая, обозначенная «75%», аппроксимирует условный 75-й процентиль улучшения y как функцию соответствия z. Вблизи любого заданного значения z около 75% нанесенных на график точек лежат ниже кривой. Модель (9.4), (9.5) предсказывает, что эти кривые будут находиться на одинаковом расстоянии друг от друга для всех значений z. Вместо этого кривые расходятся по мере увеличения z, будучи вдвое дальше друг от друга при $z = 100$, чем при $z = 0$. Другими словами, ошибки ε_i в (9.4) стремятся быть вдвое больше при $z = 100$, чем при $z = 0$.

Бутстреп пары менее чувствительны к предположениям, чем бутстреп остатки. Стандартная оценка ошибки, полученная с помощью бутстреп пар (9.31), дает разумные ответы, даже если (9.4), (9.5) полностью неверны. Единственное предположение, стоящее за (9.31), состоит в том, что исходные пары $\mathbf{x}_i = (\mathbf{c}_i, y_i)$ были случайным образом выбраны из некоторого распределения F, где F — распределение на $(p+1)$–мерных векторах $(\mathbf{c}, y)$. Даже если (9.4), (9.5) верны, нет ничего плохого в бустреп парах, как показано в (9.31); можно показать, что ответ (9.31) приближается к ответу (9.32) по мере увеличения числа пар n. Простая модель для данных гормонов (9.12) была повторно проанализирована бутстреп-парами. $B = 800$ бутстреп репликаций дали

$$\hat{\mathrm{se}}_{800}(\hat{\beta}_0) = 0.77 \quad \hat{\mathrm{se}}_{800}(\hat{\beta}_1) = 0.0045, \tag{9.33}$$

что не сильно отличается от таблицы 9.2.

Можно привести и обратный аргумент. Модель (9.4), (9.5) не обязательно должна выполняться идеально, чтобы бутстреп остатков, как в (9.32), давал разумные результаты. Более того, различия в распределении ошибок, как и в данных холостирамина, могут быть включены в модель (9.4), (9.5), что приведет к более подходящей версии бутстреп остатков; см. модель (9.42). Возможно, наиболее важным моментом здесь является то, что бутстреп не является однозначно определенной концепцией. Рисунок 8.3 может быть реализован по-разному для одной и той же задачи, в зависимости от того, как интерпретируется вероятностная модель $P \to \mathbf{x}$.

Когда мы осуществляем бустреп остатков, бутстреп наборы данных $\mathbf{x}^* = \{(\mathbf{c}_1, y_1^*), (\mathbf{c}_2, y_2^*), \cdots, (\mathbf{c}_n, y_n^*)\}$ имеют вектора признаков $\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_n$ в точности такие же, как и для фактического набора данных $\mathbf{x}$. Это кажется неестественным для данных по гормонам, где $\mathbf{c}_i$ включает z_i, затраченное количество часов, которое является такой же случайной величиной, как и переменная ответа y_i, оставшееся количество гормона.

Даже когда признаки генерируются случайным образом, есть причины проводить анализ так, как если бы они были фиксированными. Коэффициенты регрессии имеют большую стандартную ошибку, когда признаки имеют меньшее стандартное отклонение. Рассматривая признаки как фиксированные константы, мы получаем стандартную ошибку, которая отражает точность, связанную с выборкой фактически наблюдаемых признаков. Однако, как показывает (9.33), разница между $\mathbf{c}_i$ фиксированной и $\mathbf{c}_i$ случайной обычно не сильно влияет на оценку стандартной ошибки.

9.6 Пример: данные о выживаемости клеток

Бывают ситуации в регрессии, когда признаки более естественно считать фиксированными, а не случайными. Данные по выживаемости клеток в таблице 9.4 показывают такую ситуацию. Радиолог провел эксперимент с 14 бактериальными пластинами. Пластинки подвергали воздействию различных доз радиации и измеряли долю выживших клеток. Как и следовало ожидать, более высокие дозы приводят к меньшей выживаемости. Знак вопроса после ответа на пластине 13 отражает некоторую неуверенность в этом результате, выраженную исследователем.

Исследователя интересовал регрессионный анализ с переменной–предиктором.

$$\text{доза}_i = z_i \quad i = 1, 2, \cdots, 14 \tag{34}$$

и переменная ответа

$$\log\left(\text{пропорция выживания}\right)_i = y_i \quad i = 1, 2, \cdots 14. \tag{9.35}$$

Были доступны две различные теоретические модели радиационного поражения, одна из которых предсказывала линейную регрессию,

$$\mu_i = \mathrm{E}(y_i|z_i) = \beta_1 z_i, \tag{9.36}$$

и другая квадратичная регрессия,

$$\mu_i = \mathrm{E}(y_i|z_i) = \beta_1 z_i + \beta_2 z_i^2. \tag{9.37}$$

В (9.36) или (9.37) нет пересекающих членов β_0, потому что мы знаем, что нулевая доза дает коэффициент выживаемости 1, $y = \log(1) = 0$.

В таблице 9.5 показаны оценки методом наименьших квадратов $(\hat{\beta}_1, \hat{\beta}_2)$ и их оценочные стандартные ошибки $\bar{\mathrm{se}}(\hat{\beta}_j)$, (9.20). Представлены два анализа методом наименьших квадратов, один с данными для всех 14 пластин, другой за исключением сомнительной пластины 13. В обоих анализах оцененный коэффициент квадратичной регрессии $\hat{\beta}_2$ является положительным. Является ли это отличие значимым? Другими словами, можем ли мы заключить, что $\hat{\beta}_2$ останется положительным, если будет исследовано гораздо больше пластин? Отношение $\hat{\beta}_2/\hat{\mathrm{se}}(\hat{\beta}_2)$ помогает ответить на этот вопрос. Отношение составляет 2.46 для анализа, основанного на всех 14 пластинах, что обычно считается убедительным доказательством того, что $\hat{\beta}_2$ значительно больше нуля. Если верить этому результату, то квадратичная модель (9.37) сильно предпочтительнее модели (9.36), которая имеет $\beta_2 = 0$.

Table 9.4. *The Cell Survival data. Fourteen cell plates were exposed to different levels of radiation. The observed response was the proportion of cells which survived the radiation exposure. The response in plate 13 was considered somewhat uncertain by the investigator.*

plate number	dose (rads/100)	survive prop.	log.surv prop.
1	1.175	0.44000	-0.821
2	1.175	0.55000	-0.598
3	2.350	0.16000	-1.833
4	2.350	0.13000	-2.040
5	4.700	0.04000	-3.219
6	4.700	0.01960	-3.219
7	4.700	0.06120	-2.794
8	7.050	0.00500	-5.298
9	7.050	0.00320	-5.745
10	9.400	0.00110	-6.812
11	9.400	0.00015	-8.805
12	9.400	0.00019	-8.568
13	14.100	0.00700?	-4.962?
14	14.100	0.00006	-9.721

Однако удаление сомнительной пластины 13 из анализа снижает $\hat{\beta}_2/\hat{\text{se}}(\hat{\beta}_j)$ только до 0.95, что является несущественным результатом. Вывод заключается не в том, что β_2 *обязательно* равен нулю, а в том, что он легко может быть равен нулю: если $\beta_2 = 0$, и если $(\beta_2)\doteq 0.0091$, как в строке 2 таблицы 9.5, то это вовсе не удивительно, что значение $\hat{\beta}_2$ такое же большое или больше наблюдаемого значения 0.0086. У нас нет убедительных доказательств для отказа от линейной модели в пользу квадратичной модели.

Статистика — это наука о сборе информации по крупицам с целью получения высокоинформативных сложных результатов. Статистики настораживаются, когда видят, что одна точка данных, особенно подозрительная, доминирует в ответе на важный вопрос. Действительная критика регрессии по методу наименьших квадратов состоит в том, что одна удаленная точка, такая как пластина 13, может иметь слишком большое влияние на подобранную кривую регрессии. Это показано на рисунке 9.3, который строит кривую регрессии методом наименьших квадратов как с данными из пластины 13, так и без них. Мощный эффект точки «?» очевиден. Даже если бы исследователь не подвергал сомнению достоверность пластинки 13, мы бы предпочли, чтобы наши подогнанные кривые не зависели так сильно от отдельных точек данных.

Table 9.5. *Estimated regression coefficients and standard errors for the quadratic model (9.37) applied to the cell survival data. Least squares estimates (9.10) were obtained using all 14 plates (line 1), and also excluding plate 13 (line 2). Estimated standard errors for lines 1 and 2 are* $\overline{se}(\hat{\beta}_j)$, *(9.20). The estimated standard errors for the least median of squares regression (all 14 plates), line 3, were obtained from a bootstrap analysis,* $B = 400$. *The quadratic coefficient looks significantly nonzero in line 1, but not in lines 2 or 3. Line 4 gives the standard errors for the least median of squares estimate, based on resampling residuals from model (9.42).*

	$\hat{\beta}_1$	$(\widehat{\text{se}})$	$\hat{\beta}_2$	$(\widehat{\text{se}})$	$\hat{\beta}_2/\widehat{\text{se}}$
1. Least Squares, 14 plates	-1.05	(.159)	.0341	(.0143)	2.46
2. Least Squares, 13 plates	-0.86	(.094)	.0086	(.0091)	0.95
3. Least Median of Squares	-0.83	(.272)	.0114	(.0362)	0.32
4. (Resampling residuals)		(.141)		(.0160)	

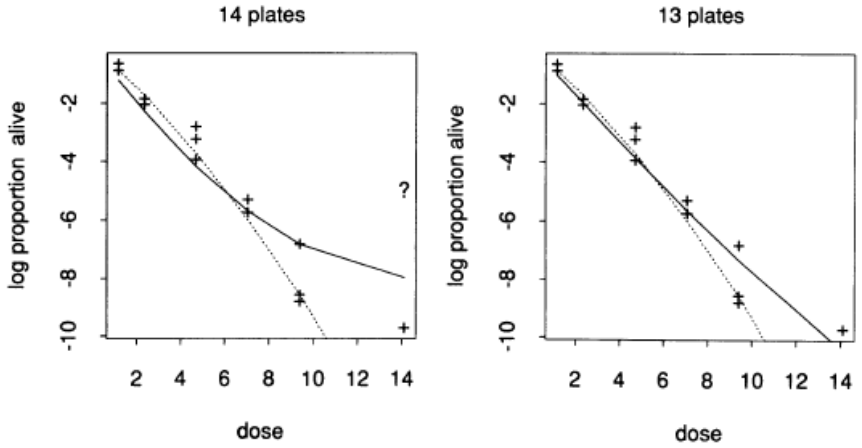

Figure 9.3. *Scatterplot of the cell survival data; solid line is the quadratic regression* $\hat{\beta}_1 z + \hat{\beta}_2 z^2$ *obtained by least-squares. Dashed line is quadratic regression fit by method of least median of squares (LMS). Left panel: all 14 plates; Right panel: thirteen plates, excluding the questionable result from plate 13. Plate 13, marked"?"in the left panel, has a large effect on the fitted least-squares curve. The questionable point has no effect on the LMS curve.*

9.7 Наименьшая медиана квадратов

Наименьшая медиана квадратов регрессии, сокращенно LMS, является менее чувствительным методом подбора, чем метод наименьших квадратов. Единственное различие между методом наименьших квадратов и LMS — это выбор критерия соответствия. Чтобы motivate критерий, давайте разделим остаточную квадратичную ошибку (9.7) на размер выборки, получив среднеквадратичные остатки

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{c}_i\mathbf{b})^2. \tag{9.38}$$

Минимизация (9.38), очевидно, то же самое, что минимизация (9.7). Средние выборок чувствительны к влияющим значениям, а медианы — нет. Следовательно, чтобы сделать (9.38) менее чувствительным, мы можем заменить среднее значение на медиану, получив *медианноквадратичные остатки*

$$\mathrm{MSE}(\mathbf{b}) = \text{медиана}(y_i - \mathbf{c}_i\mathbf{b})^2. \tag{9.39}$$

Оценка LMS $\boldsymbol{\beta}$ — это значение $\hat{\boldsymbol{\beta}}$, минимизирующее MSR($\mathbf{b}$),

$$\mathrm{MSR}(\hat{\boldsymbol{\beta}}) = \min_{\mathbf{b}}[\mathrm{MSR}(\mathbf{b})]. \tag{9.40}$$

Обратите внимание, что разница между методом наименьших квадратов и LMS заключается не в выборе модели, которая остается (9.3), а в том, как мы измеряем расхождения между моделью и наблюдаемыми данными. MSR($\mathbf{b}$) менее чувствителен, чем RSE($\mathbf{b}$), к удаленным точкам данных. Это можно увидеть на рис. 9.3, где, по-видимому, очень мало различий между квадратичной LMS-аппроксимацией с точкой «?» или без нее. На самом деле разницы нет. Расчетные коэффициенты регрессии равны $(\hat{\beta}_1, \hat{\beta}_2) = (-0.81, 0.0088)$ в обоих случаях.

Можно показать, что *разбивка* оценки LMS составляет примерно 50%. Разбивка оценщика — это наименьшая часть данных, которая может произвольно сильно повлиять на ее значение. Другими словами, оценщик имеет разбивку α, если по крайней мере $m = \alpha \cdot n$ точек данных должны быть «плохими», прежде чем он разобьет. Высокая разбивка — это хорошо, при этом 50% — это наибольшее значение, которое имеет смысл (если $\alpha > 50\%$, неясно, какие из них являются хорошими, а какие плохими). Например, среднее значение выборки имеет разбивку $1/n$, поскольку, изменяя только одно значение данных, мы можем заставить среднее значение выборки принимать любое значение. Медиана выборки имеет разбивку 50%, что отражает тот факт, что она менее чувствительна к отдельным значениям. Оценщик регрессии методом наименьших квадратов наследует чувствительность среднего и имеет разбивку $1/n$, в то время как оценщик наименьших средних квадратов, как и медиана, имеет разбивку примерно 50%. Точное определение разбивки дано в задаче 9.9.

Насколько точны LMS оценки $\hat{\beta}_1, \hat{\beta}_2$? Нет четкой формулы, подобной (9.20) для стандартных ошибок LMS. (Нет четкой формулы для самих оценок LMS. Они вычисляются с использованием алгоритма выборки с возвращением: см. задачу 9.8.) Стандартные ошибки в таблице 9.5 были получены методами бутстрепа. Стандартные ошибки в строке 3 основаны на парах повторной выборки, как в разделе 7.3. Бустреп набор данных был создан в форме $\mathbf{x}^* = ((\mathbf{c}_1^*, y_1^*), (\mathbf{c}_2^*, y_2^*), \cdots, (\mathbf{c}_n^*, y_n^*))$, как в (9.31), где $\mathbf{c}_i = (z_i, z_i^2)$. После генерации $\mathbf{x}^*$ была получена репликация бутстрепа $\hat{\boldsymbol{\beta}}^*$ для вектора регрессии LMS как минимизатор медианноквадратичных остатков для бустреп данных, то есть минимизатор над $\mathbf{b}$ для

$$\text{медиана}(y_i^* - \mathbf{c}_i^* \mathbf{b})^2 \tag{9.41}$$

$B = 400$ репликаций бустрепа дают оценочные стандартные ошибки в строке 3 таблицы 9.5. Обратите внимание, что $\hat{\beta}_2$ не намного больше нуля.

Признаками в данных выживаемости клеток были фиксированные числа, установленные исследователем: она выбрала дозы

$$1.175, 1.175, 2.35, \cdots, 14.100,$$

чтобы провести хороший эксперимент по различению линейной и квадратичной моделей радиационной выживаемости. Это заставляет нас больше интересоваться бустреп остатками (9.32), нежели бустреп парами. Тогда бустреп наборы данных $\mathbf{x}^*$ будут иметь те же векторы признаков $\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_n$, которые исследователь намеренно использовал в эксперименте.

Модель (9.4), (9.5) не совсем подходит для данных о выживаемости клеток. Глядя на рисунок 9.3, мы видим, что зависимая переменная y_i более рассеянна при больших значениях z. Это похоже на ситуацию с холостирамином на рис. 9.2, за исключением того, что у нас недостаточно точек для построения хороших процентилей регрессии. Грубо говоря, мы будем предполагать, что ошибки линейной модели линейно возрастают с дозой z. Это равносильно замене (9.4) на

$$y_i = \mathbf{c}_i \boldsymbol{\beta} + z_i \varepsilon_i \text{ для } i = 1, 2, \cdots, 14. \tag{9.42}$$

Мы по-прежнему предполагаем, что $(\varepsilon_1, \varepsilon_2, \cdots, \varepsilon n)$ — случайная выборка из некоторого распределения F, (9.5). Для модели квадратичной регрессии $\mathbf{c}_i = (z_i, z_i^2)$.

Модель вероятности для (9.42), как и раньше, равна $P = (\boldsymbol{\beta}, F)$; $\boldsymbol{\beta}$ было оценено при помощи LMS, $\hat{\boldsymbol{\beta}} = (-0.83, 0.0114)$. Затем F было оценено с помощью $\hat{F}$, эмпирического распределения величин $(y_i - \mathbf{c}_i \hat{\boldsymbol{\beta}}/z_i)$, $i = 1, 2, \cdots, 14$.

Строка 4 таблицы 9.5 сообщает о бутстреп стандартных ошибках для оценок наименьших медиан квадратов $\hat{\beta}_1$ и $\hat{\beta}_2$, полученных из $B = 200$ бутстреп репликаций, с бутстреп остатками в модели (9.42). Стандартные ошибки заметно меньше, чем при бустрепе пар. (Но недостаточно мал, чтобы сделать $\hat{\beta}_2$ существенно/значительно/значимо отличным от нуля.) Стандартные ошибки в строке 4 следует рассматривать с осторожностью, поскольку данные модели (9.42) лишь делают слабое предположение. Самым

важным в представлении модели было проиллюстрировать, как бутстреп остатков может быть выполнен в ситуациях, более сложных, чем (9.4).

9.8 Библиографические примечания

Регрессия обсуждается в большинстве текстов по элементарной статистике, и есть много книг, посвященных этой теме, в том числе Draper and Smith (1981) и Weisberg (1980). Бутстреп регрессионных моделей обсуждается на более глубоком математическом уровне в работах Фридмана (1981), Шорака (1982), Бикеля и Фридмана (1983), Вебера (1984), Ву (1986) и Шао (1988). Фридман и Петерс (1984), Петерс и Фридман (1984a, 1984b) рассмотрели некоторые практические аспекты. Rousseeuw (1984) вводит оценку наименьшей медианы квадратов. Эфрон (1991) обсуждает оценку процентилей регрессии.