8

Linear Regression Models

Regression models are used to describe how one or perhaps a few *response* variables depend on other *explanatory* variables. The idea of regression is at the core of much statistical modelling, because the question 'what happens to y when x varies?' is central to many investigations. It is often required to predict or control future responses by changing the other variables, or to gain an understanding of the relation between them. There is usually a single response, treated as random. Often there are many explanatory variables, which are treated as non-stochastic. The simplest models involve linear dependence and are described in this chapter, while Chapter 9 deals with more structured situations in which the explanatory variables have been chosen by the experimenter according to a design. Chapter 10 describes some of the many extensions of regression to nonlinear dependence. Throughout we simplify our previous notation by using y to represent both the response variable and the value it takes; no confusion should arise thereby.

8.1 Introduction

If we denote the response by y and the explanatory variables by x, our concern is how changes in x affect y. In Section 5.1, for example, the key question was how the annual maximum sea level in Venice depended on the passage of time. We fitted the straight-line regression model

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, \ldots, n,$$

where we took y_j to be the jth annual maximum sea level and x_j to be the year in which this occurred. The parameters β_0 and β_1 represent a baseline maximum sea level and the annual rate at which sea level increases, while ε_j is a random variable that represents the difference between the underlying level, $\beta_0 + \beta_1 x_j$, and the value observed, y_j.

An immediate generalization is to increase the number of explanatory variables, setting

$$y_j = \beta_1 x_{j1} + \cdots + \beta_p x_{jp} + \varepsilon_j = x_j^{\mathrm{T}}\beta + \varepsilon_j,$$

where $x_j^{\mathrm{T}} = (x_{j1}, \ldots, x_{jp})$ is a $1 \times p$ vector of explanatory variables associated with the jth response, β is a $p \times 1$ vector of unknown parameters and ε_j is an unobserved error accounting for the discrepancy between the observed response y_j and $x_j^{\mathrm{T}}\beta$. In matrix notation,

$$y = X\beta + \varepsilon, \tag{8.1}$$

where y is the $n \times 1$ vector whose jth element is y_j, X is an $n \times p$ matrix whose jth row is x_j^{T}, and ε is the $n \times 1$ vector whose jth element is ε_j. The data on which the investigation is to be based are y and X, and the aim is to disentangle systematic changes in y due to variation in X from the haphazard scatter added by the errors ε. Model (8.1) is known as a *linear regression model* with *design matrix* X.

Example 8.1 (Straight-line regression) For the straight-line regression model, (8.1) becomes

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

so X is an $n \times 2$ matrix and β a 2×1 vector of parameters. ■

Example 8.2 (Polynomial regression) Suppose that the response is a polynomial function of a single covariate,

$$y_j = \beta_0 + \beta_1 x_j + \cdots + \beta_{p-1} x_j^{p-1} + \varepsilon_j.$$

For example, we might wish to fit a quadratic or cubic trend in the Venice sea level data, in which case we would have $p = 3$ or $p = 4$ respectively. Then

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{p-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^{p-1} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

where X has dimension $n \times p$. ■

A key point is that (8.1) is linear in the parameters β. Polynomial regression can be written in form (8.1) because of its linearity, not in x, but in β.

Case	x_1	x_2	x_3	x_4	y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

Table 8.1 Cement data (Woods *et al.*, 1932): y is heat evolved in calories per gram of cement, and x_1, x_2, x_3, and x_4 are percentage weight of clinkers, with x_1, $3CaO.Al_2O_3$, x_2, $3CaO.SiO_2$, x_3, $4CaO.Al_2O_3.Fe_2O_3$, and x_4, $2CaO.SiO_2$.

Example 8.3 (Cement data) Table 8.1 contains data on the relationship between the heat evolved in the setting of cement and its chemical composition. Data on heat evolved, y, for each of $n = 13$ independent samples are available, and for each sample the percentage weight in clinkers of four chemicals, x_1, $3CaO.Al_2O_3$, x_2, $3CaO.SiO_2$, x_3, $4CaO.Al_2O_3.Fe_2O_3$, and x_4, $2CaO.SiO_2$, is recorded.

Figure 8.1 shows that although the response y depends on each of the covariates $x_1, \ldots, x_4$, the degrees and directions of the dependences differ.

In this case we might fit the model

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \beta_3 x_{3j} + \beta_4 x_{4j} + \varepsilon_j,$$

where Figure 8.1 suggests that β_1 and β_2 are positive, and that β_3 and β_4 are negative. The design matrix has dimension 13×5, and is

$$X = \begin{pmatrix} 1 & 7 & 26 & 6 & 60 \\ 1 & 1 & 29 & 15 & 52 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 10 & 68 & 8 & 12 \end{pmatrix};$$

the vectors y and ε have dimension 13×1 and β has dimension 5×1. ■

In the examples above the explanatory variables consist of numerical quantities, sometimes called *covariates*. *Dummy variables* that represent whether or not an effect is applied can also appear in the design matrix.

Example 8.4 (Cycling data) Norman Miller of the University of Wisconsin wanted to see how seat height, tyre pressure and the use of a dynamo affected the time taken to ride his bicycle up a hill. He decided to collect data

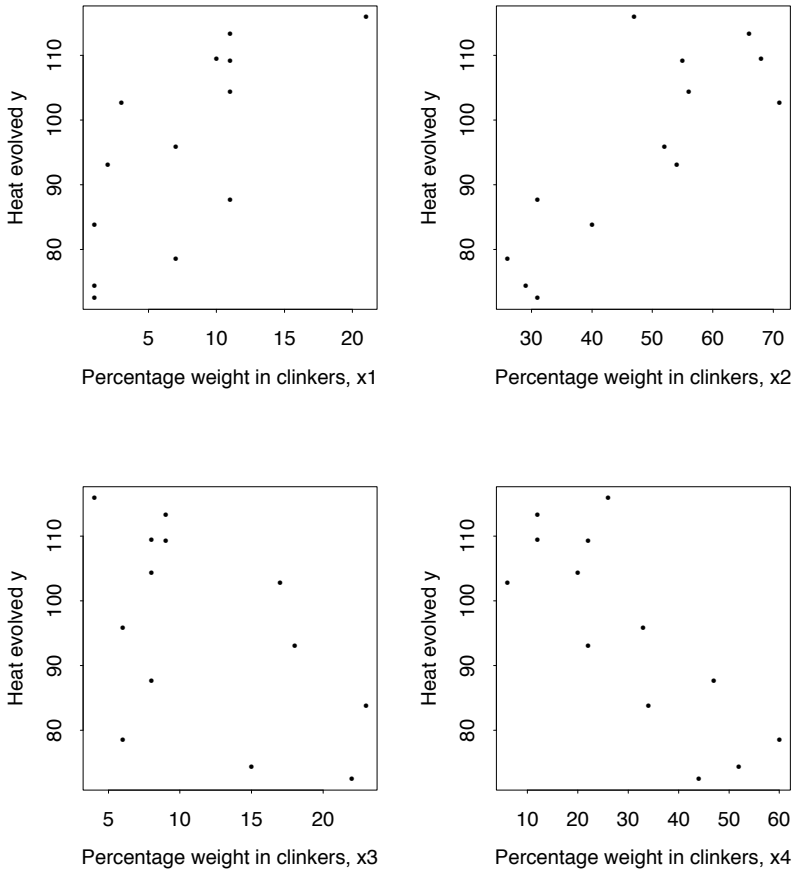


Figure 8.1 Plots of cement data. The variables are heat evolved in calories per gram, y, percentage weight in clinkers of x_1, $3CaO.Al_2O_3$, x_2, $3CaO.SiO_2$, x_3, $4CaO.Al_2O_3.Fe_2O_3$, and x_4, $2CaO.SiO_2$.

at each combination of two seat heights, 26 and 30 inches from the centre of the crank, two tyre pressures, 40 and 55 pounds per square inch (psi) and with the dynamo on and off, giving eight combinations in all. The times were expected to be quite variable, and in order to get more accurate results he decided to make two timings for each combination. He wrote each of the eight combinations on two pieces of card, and then drew the sixteen from a box in a random order. He planned to make four widely separated runs up the hill on each of four days, first adjusting his bicycle to the setups on the successive pieces of card, but bad weather forced him to cancel the last run on the first day; he made five on the third day to make up for this. Table 8.2 gives timings obtained with his wristwatch.

The lower part of Table 8.2 shows how average time depends on experimental setup. There is a large reduction in the average time when the seat

Setup	Day	Run	Seat height (inches)	Dynamo	Tyre pressure (psi)	Time (secs)
1	3	2	−	−	−	51
2	4	1	−	−	−	54
3	2	2	+	−	−	41
4	2	3	+	−	−	43
5	3	3	−	+	−	54
6	2	1	−	+	−	60
7	3	1	+	+	−	44
8	4	3	+	+	−	43
9	1	1	−	−	+	50
10	4	4	−	−	+	48
11	3	5	+	−	+	39
12	4	2	+	−	+	39
13	3	4	−	+	+	53
14	1	3	−	+	+	51
15	1	2	+	+	+	41
16	2	4	+	+	+	44

Table 8.2 Data and experimental setup for bicycle experiment (Box *et al.*, 1978, pp. 368–372). The lower part of the table shows the average times for each of the eight combinations of settings of seat height, tyre pressure, and dynamo, and the average times for the eight observations at each setting, considered separately.

	Seat height (inches from centre of crank)	Dynamo	Tyre pressure (psi)
−	26	Off	40
+	30	On	55

Dynamo	Tyre pressure low		Tyre pressure high	
	Seat low	Seat high	Seat low	Seat high
Off	52.5	42.0	49.0	39.0
On	57.0	43.5	52.0	42.5

Dynamo		Tyre pressure		Seat	
Off	On	Low	High	Low	High
45.63	48.75	48.75	45.63	52.63	41.75

is raised and smaller reductions when the tyre pressure is increased and the dynamo is off.

The quantities that are varied in this experiment — seat height, tyre pres-

sure, and the state of the dynamo — are known as *factors*. Each takes two possible values, known as *levels*. Here there are two types of factors: quantitative and qualitative. The two levels of seat height and tyre pressure are quantitative — other values might have been chosen, and more than two levels could have been used — but the dynamo factor has only two possible levels and is qualitative.

An experiment like this, in which data are collected at each combination of a number of factors, is known as a *factorial experiment*. Such designs and their variants are widely used; see Section 9.2.4. In this case an experimental setup with three factors each having two levels is applied twice: the design consists of two replicates of a 2^3 factorial experiment.

One linear model for the data in Table 8.2 is that at the lower seat height, with the dynamo off, and the lower tyre pressure, the mean time is μ, and the three factors act separately, changing the mean time by α_1, α_2, and α_3 respectively. This corresponds to the linear regression model

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{16} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \end{pmatrix}.$$

Table 8.2 suggests that $\mu \doteq 52.5$, that $\alpha_1 < 0$, $\alpha_2 > 0$, and $\alpha_3 < 0$. The baseline time is μ, which corresponds to the mean time at the lower level of all three factors, and the overall average time is $\overline{y} = \mu + \frac{1}{2}\alpha_1 + \frac{1}{2}\alpha_2 + \frac{1}{2}\alpha_3 + \overline{\varepsilon}$, where $\overline{\varepsilon}$ is the average of the unobserved errors.

A different formulation of the model would take the overall mean time as

the baseline, leading to

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \\ y_{11} \\ y_{12} \\ y_{13} \\ y_{14} \\ y_{15} \\ y_{16} \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \\ \varepsilon_{10} \\ \varepsilon_{11} \\ \varepsilon_{12} \\ \varepsilon_{13} \\ \varepsilon_{14} \\ \varepsilon_{15} \\ \varepsilon_{16} \end{pmatrix}. \tag{8.2}$$

In (8.2) the effect of increasing seat height from 26 to 30 inches is $2\beta_1$, the effect of switching the dynamo on is $2\beta_2$, and the effect of increasing tyre pressure is $2\beta_3$. As each column of the design matrix apart from the first has sum zero, the overall average time in this parametrization is $\beta_0 + \overline{\varepsilon}$. Although the parameter β_0 is related to the overall mean, it does not correspond to a combination of factors that can be applied to the bicycle — how can the dynamo be half on? Despite this, we shall see below that (8.2) is convenient for some purposes. ■

Often it is better to apply a linear model to transformed data than to the original observations.

Example 8.5 (Multiplicative model) Suppose that the data consist of times to failure that depend on positive covariates x_1 and x_2 according to

$$y = \gamma_0 x_1^{\gamma_1} x_2^{\gamma_2} \eta,$$

where η is a positive random variable. Then

$$\log y = \log \gamma_0 + \gamma_1 \log x_1 + \gamma_2 \log x_2 + \log \eta,$$

which is linear in $\log \gamma_0$, γ_1, and γ_2. The variance of the transformed response $\log y$ does not depend on its mean, whereas y has variance proportional to the square of its mean, so in addition to achieving linearity, the transformation equalizes the variances. ■

Exercises 8.1

1 Which of the following can be written as linear regression models, (i) as they are, (ii) when a single parameter is held fixed, (iii) after transformation? For those that can be so written, give the response variable and the form of the design matrix.
(a) $y = \beta_0 + \beta_1/x + \beta_2/x^2 + \varepsilon$;
(b) $y = \beta_0/(1 + \beta_1 x) + \varepsilon$;
(c) $y = 1/(\beta_0 + \beta_1 x + \varepsilon)$;
(d) $y = \beta_0 + \beta_1 x^{\beta_2} + \varepsilon$;
(e) $y = \beta_0 + \beta_1 x_1^{\beta_2} + \beta_3 x_2^{\beta_4} + \varepsilon$;

2 Data are available on the weights of two groups of three rats at the beginning of a fortnight, x, and at its end, y. During the fortnight, one group was fed normally and the other group was fed a growth inhibitor. Consider a linear model for the weights,

$$y_{jg} = \alpha_g + \beta_g x_{jg} + \varepsilon_{jg}, \quad j = 1, \ldots, 3, \quad g = 1, 2.$$

(a) Write down the design matrix for the model above.
(b) The model is to be reparametrized in such a way that it can be specialized to (i) two parallel lines for the two groups, (ii) two lines with the same intercept, (iii) one common line for both groups, just by setting parameters to zero. Give one design matrix which can be made to correspond to (i), (ii), and (iii), just by dropping columns.

8.2 Normal Linear Model

8.2.1 Estimation

Suppose that the errors ε_j in (8.1) are independent normal random variables, with means zero and variances σ^2. Then the responses y_j are independent normal random variables with means $x_j^{\mathrm{T}}\beta$ and variances σ^2, and (8.1) is the *normal linear model*. The likelihood for β and σ^2 is

$$L(\beta, \sigma^2) = \prod_{j=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y_j - x_j^{\mathrm{T}}\beta)^2\right\},$$

and the log likelihood is

$$\ell(\beta, \sigma^2) \equiv -\tfrac{1}{2}\left\{n \log \sigma^2 + \frac{1}{\sigma^2}\sum_{j=1}^{n}(y_j - x_j^{\mathrm{T}}\beta)^2\right\}.$$

Whatever the value of σ^2, the log likelihood is maximized with respect to β at the value that minimizes the *sum of squares*

$$SS(\beta) = \sum_{j=1}^{n}(y_j - x_j^{\mathrm{T}}\beta)^2 = (y - X\beta)^{\mathrm{T}}(y - X\beta). \tag{8.3}$$

We obtain the maximum likelihood estimate of β by solving simultaneously

the equations

$$\frac{\partial SS(\beta)}{\partial \beta_r} = 2\sum_{j=1}^{n} x_{jr}(y_j - \beta^{\mathrm{T}} x_j) = 0, \quad r = 1, \ldots, p.$$

In matrix form these amount to the *normal equations*

$$X^{\mathrm{T}}(y - X\beta) = 0, \tag{8.4}$$

which imply that the estimate satisfies $(X^{\mathrm{T}}X)\beta = X^{\mathrm{T}}y$. Provided the $p \times p$ matrix $X^{\mathrm{T}}X$ is of full rank it is invertible, and the *least squares estimator* of β is

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y.$$

The maximum likelihood estimator of σ^2 may be obtained from the profile likelihood for σ^2,

$$\ell_{\mathrm{p}}(\sigma^2) = \max_{\beta} \ell(\beta, \sigma^2) = -\tfrac{1}{2}\left\{ n \log \sigma^2 + \frac{1}{\sigma^2}(y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta}) \right\}, \tag{8.5}$$

and it follows by differentiation that the maximum likelihood estimator of σ^2 is

$$\widehat{\sigma}^2 = n^{-1}(y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta}) = n^{-1}\sum_{j=1}^{n}(y_j - x_j^{\mathrm{T}}\widehat{\beta})^2.$$

We shall see below that $\widehat{\sigma}^2$ is biased and that an unbiased estimator of σ^2 is

$$S^2 = \frac{1}{n-p}(y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta}) = \frac{1}{n-p}\sum_{j=1}^{n}(y_j - x_j^{\mathrm{T}}\widehat{\beta})^2.$$

Example 8.6 (Straight-line regression) We write the straight-line regression model (5.3) in matrix form as

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 - \overline{x} \\ 1 & x_2 - \overline{x} \\ \vdots & \vdots \\ 1 & x_n - \overline{x} \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

The least squares estimates are

$$\begin{aligned} \widehat{\beta} = \begin{pmatrix} \widehat{\gamma}_0 \\ \widehat{\gamma}_1 \end{pmatrix} &= \begin{pmatrix} n & \sum(x_j - \overline{x}) \\ \sum(x_j - \overline{x}) & \sum(x_j - \overline{x})^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_j \\ \sum(x_j - \overline{x})y_j \end{pmatrix} \\ &= \begin{pmatrix} n^{-1} & 0 \\ 0 & \frac{1}{\sum(x_j - \overline{x})^2} \end{pmatrix} \begin{pmatrix} \sum y_j \\ \sum(x_j - \overline{x})y_j \end{pmatrix} \\ &= \begin{pmatrix} \overline{y} \\ \frac{\sum(x_j - \overline{x})y_j}{\sum(x_j - \overline{x})^2} \end{pmatrix}. \end{aligned}$$

If all the x_j are equal, $X^{\mathrm{T}}X$ is not invertible, and $\widehat{\gamma}_1$ is undetermined: any value is possible.

The unbiased estimator of σ^2 is

$$\frac{1}{n-2}\sum_{j=1}^{n}\left\{y_j - \overline{y} - (x_j - \overline{x})\frac{\sum(x_k - \overline{x})y_k}{\sum(x_k - \overline{x})^2}\right\}^2.$$

■

Example 8.7 (Surveying a triangle) Suppose that we want to estimate the angles α, β, and γ (radians) of a triangle ABC based on a single independent measurement of the angle at each corner. Although there are three angles, their sum is the constant $\alpha+\beta+\gamma=\pi$, and so just two of them vary independently. In terms of α and β, we have $y_A = \alpha + \varepsilon_A$, $y_B = \beta + \varepsilon_B$, and $y_C = \pi - \alpha - \beta + \varepsilon_C$, and this gives the linear model

$$\begin{pmatrix} y_A \\ y_B \\ y_C - \pi \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -1 & -1 \end{pmatrix}\begin{pmatrix}\alpha \\ \beta\end{pmatrix} + \begin{pmatrix}\varepsilon_A \\ \varepsilon_B \\ \varepsilon_C\end{pmatrix}.$$

Hence

$$\begin{pmatrix}\widehat{\alpha} \\ \widehat{\beta}\end{pmatrix} = \tfrac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2\end{pmatrix}\begin{pmatrix}\pi + y_A - y_C \\ \pi + y_B - y_C\end{pmatrix} = \tfrac{1}{3}\begin{pmatrix}\pi + 2y_A - y_B - y_C \\ \pi + 2y_B - y_A - y_C\end{pmatrix}.$$

It is straightforward to show that $s^2 = (y_A + y_B + y_C - \pi)^2/3$. ■

The sum of squares $SS(\beta)$ plays a central role. Its minimum value,

$$SS(\widehat{\beta}) = \sum_{j=1}^{n}(y_j - x_j^{\mathrm{T}}\widehat{\beta})^2 = (y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta}),$$

is called the *residual sum of squares* because it is the residual squared discrepancy between the observations, y, and the *fitted values*, $\widehat{y} = X\widehat{\beta}$. The vector $\widehat{y}$ is the linear combination of the columns of X that best accounts for the variation in y, in the sense of minimizing the squared distance between them. Note that

$$\widehat{y} = X\widehat{\beta} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = Hy,$$

say, where the *hat matrix* $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ "puts hats" on y. Evidently H is a projection matrix; see Section 8.2.2.

The unobservable error $\varepsilon_j = y_j - x_j^{\mathrm{T}}\beta$ is estimated by the jth *residual* $e_j = y_j - \widehat{y}_j = y_j - x_j^{\mathrm{T}}\widehat{\beta}$. In vector terms,

Sometimes e_j is called a *raw residual*.

$$e = y - X\widehat{\beta} = y - Hy = (I_n - H)y,$$

where I_n is the $n \times n$ identity matrix.

Table 8.3 Data from bicycle experiment, together with fitted values $\widehat{y}$, raw residuals e, standardized residuals, r, deletion residuals r', leverages h and Cook distances C.

Setup	Seat height	Dynamo	Tyre pressure	Time y	$\widehat{y}$	e	r	r'	h	C
1	–1	–1	–1	51	52.62	–1.625	–0.99	–0.99	0.25	0.08
2	–1	–1	–1	54	52.62	1.375	–0.84	0.83	0.25	0.06
3	1	–1	–1	41	41.75	–0.750	–0.46	–0.44	0.25	0.02
4	1	–1	–1	43	41.75	1.250	0.76	0.75	0.25	0.05
5	–1	1	–1	54	55.75	–1.750	–1.06	–1.07	0.25	0.09
6	–1	1	–1	60	55.75	4.250	2.59	3.72	0.25	0.56
7	1	1	–1	44	44.87	–0.875	–0.53	–0.52	0.25	0.02
8	1	1	–1	43	44.87	–1.875	–1.14	–1.16	0.25	0.11
9	–1	–1	1	50	49.50	0.500	0.30	0.29	0.25	0.01
10	–1	–1	1	48	49.50	–1.500	–0.91	–0.91	0.25	0.07
11	1	–1	1	39	38.62	0.375	0.23	0.22	0.25	0.00
12	1	–1	1	39	38.62	0.375	0.23	0.22	0.25	0.00
13	–1	1	1	53	52.62	0.375	0.23	0.22	0.25	0.00
14	–1	1	1	51	52.62	–1.625	–0.99	–0.99	0.25	0.08
15	1	1	1	41	41.75	–0.750	–0.46	–0.44	0.25	0.02
16	1	1	1	44	41.75	2.250	1.37	1.43	0.25	0.16

Example 8.8 (Cycling data) For model (8.2) we find that

$$(X^{\mathrm{T}}X)^{-1} = \tfrac{1}{16} I_4,$$

so the least squares estimates $(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ are

$$\tfrac{1}{16}\begin{pmatrix} y_1+y_2+y_3+y_4+y_5+y_6+y_7+y_8+y_9+y_{10}+y_{11}+y_{12}+y_{13}+y_{14}+y_{15}+y_{16} \\ -y_1-y_2+y_3+y_4-y_5-y_6+y_7+y_8-y_9-y_{10}+y_{11}+y_{12}-y_{13}-y_{14}+y_{15}+y_{16} \\ -y_1-y_2-y_3-y_4+y_5+y_6+y_7+y_8-y_9-y_{10}-y_{11}-y_{12}+y_{13}+y_{14}+y_{15}+y_{16} \\ -y_1-y_2-y_3-y_4-y_5-y_6-y_7-y_8+y_9+y_{10}+y_{11}+y_{12}+y_{13}+y_{14}+y_{15}+y_{16} \end{pmatrix} = \begin{pmatrix} 47.19 \\ -5.437 \\ 1.563 \\ -1.563 \end{pmatrix}.$$

Thus the overall average time is 47.19 seconds, putting the seat at height 30 inches rather than 26 inches changes the time by an average of $2 \times (-5.437) = -10.87$ seconds, putting the dynamo on rather than off changes the time by an average of $2 \times 1.563 = 3.13$ seconds, and increasing the tyre pressure from 40 to 55 psi changes the time by –3.13 seconds. The largest effect is due to increasing the seat height. The model suggests that the fastest time is obtained with no dynamo, a high seat and tyres at 55 psi.

The residual sum of squares for this model is 43.25 seconds squared, the overall sum of squares is $\sum y_j^2 = 36221$ seconds squared, and therefore the sum of squares explained by the model is $36221 - 43.25 = 36177.75$ seconds squared; this is the amount of variation removed when $X\beta$ is fitted.

The fitted values are $\widehat{y} = X\widehat{\beta}$, giving $\widehat{y}_1 = \widehat{\beta}_0 - \widehat{\beta}_1 - \widehat{\beta}_2 - \widehat{\beta}_3 = 52.625$, $e_1 = y_1 - \widehat{y}_1 = 51 - 52.625 = -1.625$, and so forth. Table 8.3 gives the data, fitted values, residuals and quantities discussed in Examples 8.22 and 8.27. ■

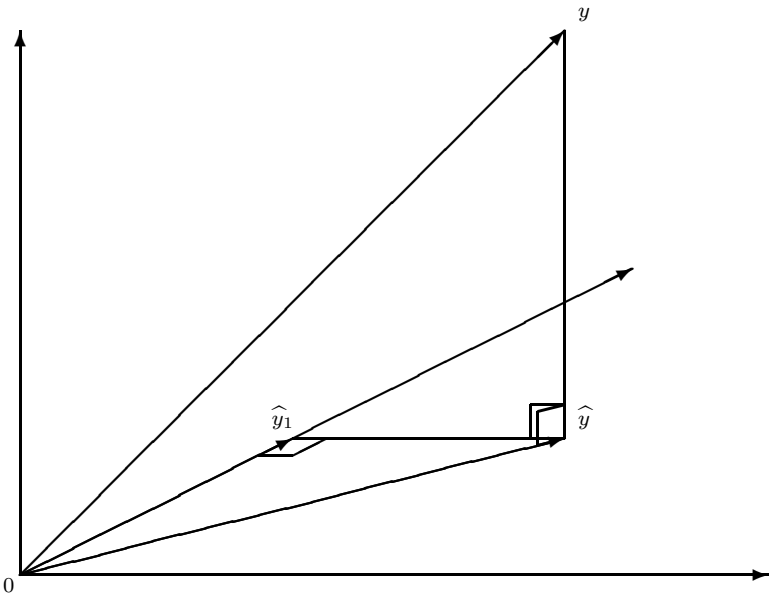


Figure 8.2 The geometry of least squares estimation. The space spanned by all three axes represents the n-dimensional observation space in which y lies. The horizontal plane through O represents the p-dimensional space in which the linear combination $X\beta$ lies, and estimation by least squares amounts to minimizing the squared distance $(y - X\beta)^{\mathrm{T}}(y - X\beta)$. In the figure the value of $X\beta$ that gives the minimum lies vertically below y, which corresponds to orthogonal projection of y into the p-dimensional subspace spanned by the columns of X; the fitted value $\widehat{y} = Hy$ is the point closest to y in that subspace, and the projection matrix is $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. The vector of residuals $e = y - \widehat{y}$ is orthogonal to the fitted value $\widehat{y}$. The line $x = z = 0$ represents the space spanned by the columns of the reduced model matrix X_1, with corresponding fitted value $\widehat{y}_1$. The orthogonality of $\widehat{y}_1$, $\widehat{y} - \widehat{y}_1$, and $y - \widehat{y}$ implies that when the data are normal the corresponding sums of squares are independent.

8.2.2 Geometrical interpretation

Figure 8.2 shows the geometry of least squares. The n-dimensional vector space inhabited by the observation vector y is represented by the space spanned by all three axes, and the p-dimensional subspace in which $X\beta$ lies is represented by the horizontal plane through the origin. The least squares estimate $\widehat{\beta}$ minimizes $(y - X\beta)^{\mathrm{T}}(y - X\beta)$, which is the squared distance between $X\beta$ and y. We see that $(y - X\beta)^{\mathrm{T}}(y - X\beta)$ is minimized when the vector $y - X\beta$ is orthogonal to the horizontal plane spanned by the columns of X, so that for any column x of X we have $x^{\mathrm{T}}(y - X\beta) = 0$. Equivalently the normal equations $X^{\mathrm{T}}(y - X\beta) = 0$ hold, and provided $X^{\mathrm{T}}X$ is invertible we obtain $\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$. The fitted value $\widehat{y} = X\widehat{\beta} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = Hy$ is the orthogonal projection of y onto the plane spanned by the columns of X, and the matrix representing that projection is H. Notice that $\widehat{y}$ is unique whether or not $X^{\mathrm{T}}X$ is invertible.

Figure 8.2 shows that the vector of residuals, $e = y - \widehat{y} = (I_n - H)y$, and the vector of fitted values, $\widehat{y} = Hy$, are orthogonal. To see this algebraically, note that

$$\widehat{y}^{\mathrm{T}}e = y^{\mathrm{T}}H^{\mathrm{T}}(I_n - H)y = y^{\mathrm{T}}(H - H)y = 0, \tag{8.6}$$

because $H^{\mathrm{T}} = H$ and $HH = H$, that is, the projection matrix H is symmetric and idempotent (Exercise 8.2.5). The close link between orthogonality and independence for normally distributed vectors means that (8.6) has important consequences, as we shall see in Section 8.3. For now, notice that (8.6) implies

that

$$y^{\mathrm{T}}y = (y - \widehat{y} + \widehat{y})^{\mathrm{T}}(y - \widehat{y} + \widehat{y}) = (e + \widehat{y})^{\mathrm{T}}(e + \widehat{y}) = e^{\mathrm{T}}e + \widehat{y}^{\mathrm{T}}\widehat{y}, \qquad (8.7)$$

as is clear from Figure 8.2 by Pythagoras' theorem. That is, the overall sum of squares of the data, $\sum y_j^2 = y^{\mathrm{T}}y$, equals the sum of the residual sum of squares, $SS(\widehat{\beta}) = \sum(y_j - \widehat{y}_j)^2 = e^{\mathrm{T}}e$, and the sum of squares for the fitted model, $\sum \widehat{y}_j^2 = \widehat{y}^{\mathrm{T}}\widehat{y}$.

Such decompositions are central to analysis of variance, discussed below.

8.2.3 Likelihood quantities

Chapter 4 shows how the observed and expected information matrices play a central role in likelihood inference, by providing approximate variances for maximum likelihood estimates. To obtain these matrices for the normal linear model, note that the log likelihood has second derivatives

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial\beta_r\partial\beta_s} &= -\frac{1}{\sigma^2}\sum_{j=1}^{n} x_{jr}x_{js}, \quad \frac{\partial^2 \ell}{\partial\beta_r\partial\sigma^2} = \frac{1}{\sigma^4}\sum_{j=1}^{n} x_{jr}(y_j - x_j^{\mathrm{T}}\beta), \\
\frac{\partial^2 \ell}{\partial(\sigma^2)^2} &= -\tfrac{1}{2}\left\{-\frac{1}{\sigma^4} + \frac{2}{\sigma^6}\sum_{j=1}^{n}(y_j - x_j^{\mathrm{T}}\beta)^2\right\}, \quad r, s = 1, \ldots, p.
\end{aligned}$$

Thus elements of the expected information matrix are

$$\mathrm{E}\left(-\frac{\partial^2 \ell}{\partial\beta_r\partial\beta_s}\right) = \frac{1}{\sigma^2}\sum_{j=1}^{n} x_{jr}x_{js}, \quad \mathrm{E}\left(-\frac{\partial^2 \ell}{\partial\beta_r\partial\sigma^2}\right) = 0, \quad \mathrm{E}\left\{-\frac{\partial^2 \ell}{\partial(\sigma^2)^2}\right\} = \frac{n}{2\sigma^4},$$

or in matrix form

$$I(\beta, \sigma^2) = \begin{pmatrix} \sigma^{-2}X^{\mathrm{T}}X & 0 \\ 0 & \frac{1}{2}n\sigma^{-4} \end{pmatrix}, \quad I(\beta, \sigma^2)^{-1} = \begin{pmatrix} \sigma^2(X^{\mathrm{T}}X)^{-1} & 0 \\ 0 & 2\sigma^4/n \end{pmatrix}.$$

Provided that X has rank p, the matrices $I(\beta, \sigma^2)$ and $J(\widehat{\beta}, \widehat{\sigma}^2)$ are positive definite (Exercise 8.2.7).

Under mild regularity conditions on the design matrix and the errors, the general theory of likelihood estimation implies that the asymptotic distribution of $\widehat{\beta}$ and σ^2 is normal with means β and σ^2, and covariance matrix given by $I(\beta, \sigma^2)^{-1}$, the block diagonal structure of which implies that $\widehat{\beta}$ and $\widehat{\sigma}^2$ are asymptotically independent. We shall see in the next section that stronger results are true: when the errors are normal the estimates $\widehat{\beta}$ have an exact normal distribution and are independent of $\widehat{\sigma}^2$ for every value of n, while $\widehat{\sigma}^2$ has a distribution proportional to χ^2_{n-p} provided that $n > p$.

The quantities $\widehat{\beta}$ and $SS(\widehat{\beta})$ are minimal sufficient statistics for β and σ^2 (Problem 8.7).

Example 8.9 (Two-sample model) Suppose that we have two groups of normal data, the first with mean β_0,

$$y_{0j} = \beta_0 + \varepsilon_{0j}, \quad j = 1, \ldots, n_0,$$

and the second with mean $\beta_0 + \beta_1$,

$$y_{1j} = \beta_0 + \beta_1 + \varepsilon_{1j}, \quad j = 1, \ldots, n_1,$$

where the ε_{gj} are independent with means zero and variances σ^2. The matrix form of this model is

$$\begin{pmatrix} y_{01} \\ \vdots \\ y_{0n_0} \\ y_{11} \\ \vdots \\ y_{1n_1} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} + \begin{pmatrix} \varepsilon_{01} \\ \vdots \\ \varepsilon_{0n_0} \\ \varepsilon_{11} \\ \vdots \\ \varepsilon_{1n_1} \end{pmatrix}.$$

The estimator of β is $\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$, that is,

$$\begin{aligned} \begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \end{pmatrix} &= \begin{pmatrix} n_0 + n_1 & n_1 \\ n_1 & n_1 \end{pmatrix}^{-1} \begin{pmatrix} n_0\overline{y}_{0\cdot} + n_1\overline{y}_{1\cdot} \\ n_1\overline{y}_{1\cdot} \end{pmatrix} \\ &= \begin{pmatrix} n_0^{-1} & -n_0^{-1} \\ -n_0^{-1} & n_0^{-1} + n_1^{-1} \end{pmatrix} \begin{pmatrix} n_0\overline{y}_{0\cdot} + n_1\overline{y}_{1\cdot} \\ n_1\overline{y}_{1\cdot} \end{pmatrix} \\ &= \begin{pmatrix} \overline{y}_{0\cdot} \\ \overline{y}_{1\cdot} - \overline{y}_{0\cdot} \end{pmatrix}, \end{aligned}$$

where $\overline{y}_{0\cdot} = n_0^{-1}\sum y_{0j}$ and $\overline{y}_{1\cdot} = n_1^{-1}\sum y_{1j}$ are the group averages. One can verify directly that the elements of $\sigma^2(X^{\mathrm{T}}X)^{-1}$ give the variances and covariance of the least squares estimators.

In this example the fitted values are $\widehat{\beta}_0 = \overline{y}_{0\cdot}$ for the first group and $\widehat{\beta}_0 + \widehat{\beta}_1 = \overline{y}_{1\cdot}$ for the second group, and the unbiased estimator of σ^2 is

$$S^2 = \frac{1}{n_0 + n_1 - 2}\left\{\sum_{j=1}^{n_0}(y_{0j} - \overline{y}_{0\cdot})^2 + \sum_{j=1}^{n_1}(y_{1j} - \overline{y}_{1\cdot})^2\right\}.$$

A minimal sufficient statistic for $(\beta_0, \beta_1, \sigma^2)$ is $(\overline{y}_{0\cdot}, \overline{y}_{1\cdot}, s^2)$. ■

Example 8.10 (Maize data) The discussion in Example 1.1 suggests that a model of matched pairs better describes the experimental setup for the maize data than the two-sample model of Example 8.9. We parametrize the matched pair model so that the jth pair of observations is

$$y_{1j} = \beta_j - \beta_0 + \varepsilon_{1j}, \quad y_{2j} = \beta_j + \beta_0 + \varepsilon_{2j}, \quad j = 1, \ldots, m,$$

where we assume that the ε_{ji} are independent normal random variables with

means zero and variances σ^2. We have $m = 15$. The average difference between the heights of the crossed and self-fertilized plants in a pair is $2\beta_0$, and the mean height of the pair is β_j. The matrix form of this model is

$$\begin{pmatrix} y_{11} \\ y_{21} \\ y_{12} \\ y_{22} \\ \vdots \\ y_{1m} \\ y_{2m} \end{pmatrix} = \begin{pmatrix} -1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ -1 & 0 & 1 & \cdots & 0 \\ 1 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ -1 & 0 & 0 & \cdots & 1 \\ 1 & 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{21} \\ \varepsilon_{12} \\ \varepsilon_{22} \\ \vdots \\ \varepsilon_{1m} \\ \varepsilon_{2m} \end{pmatrix},$$

so β has dimension $(m+1) \times 1$ and $X^{\mathrm{T}}X = \mathrm{diag}(2m, 2, \ldots, 2)$ has dimension $(m+1) \times (m+1)$.

We see that

$$\begin{aligned} \widehat{\beta}_0 &= (y_{21} - y_{11} + y_{22} - y_{12} + \cdots + y_{2m} - y_{1m})/(2m), \\ \widehat{\beta}_j &= \tfrac{1}{2}(y_{1j} + y_{2j}), \quad j = 1, \ldots, m, \end{aligned}$$

and that the estimators are independent. The unbiased estimator of σ^2 is

$$S^2 = \frac{1}{2m - (m+1)} \sum_{j=1}^{m} \left\{ (y_{1j} - \widehat{\beta}_j + \widehat{\beta}_0)^2 + (y_{2j} - \widehat{\beta}_j - \widehat{\beta}_0)^2 \right\},$$

which can be written as $\{2(m-1)\}^{-1} \sum (d_j - \overline{d})^2$, where $d_j = y_{2j} - y_{1j}$ is the difference between the heights of the crossed and self-fertilized plants in the jth pair, and $\overline{d} = m^{-1} \sum d_j$ is their average. Note that $\widehat{\beta}_0$ equals $\frac{1}{2}\overline{d}$. ■

Likelihood ratio statistic

The likelihood ratio statistic is a standard tool for comparing nested models. In the context of the normal linear model, let

$$y = X\beta + \varepsilon = \begin{pmatrix} X_1 & X_2 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon,$$

where X_1 is an $n \times q$ matrix, X_2 is an $n \times (p-q)$ matrix, $q < p$, and β_1 and β_2 are vectors of parameters of lengths q and $p-q$. Suppose that we wish to compare this with the simpler model in which $\beta_2 = 0$, so the mean of y depends only on X_1. Under the more general model the maximum likelihood estimators of β and σ^2 are $\widehat{\beta}$ and $\widehat{\sigma}^2 = n^{-1} SS(\widehat{\beta})$, where $SS(\beta) = (y - X\beta)^{\mathrm{T}}(y - X\beta)$, and it follows from (8.5) that the maximized log likelihood is

$$\ell_{\mathrm{p}}(\widehat{\sigma}^2) = -\tfrac{1}{2} \left\{ n \log SS(\widehat{\beta}) + n - n \log n \right\},$$

where $\ell_{\mathrm{p}}(\sigma^2) = \max_\beta \ell(\beta, \sigma^2)$ is the profile log likelihood for σ^2. When $\beta_2 = 0$, the maximum likelihood estimator of σ^2 is

$$\widehat{\sigma}_0^2 = n^{-1} SS(\widehat{\beta}_1) = n^{-1}(y - X_1\widehat{\beta}_1)^{\mathrm{T}}(y - X_1\widehat{\beta}_1),$$

where $\widehat{\beta}_1$ is the estimator of β_1 when $\beta_2 = 0$. Hence the likelihood ratio statistic for comparison of the models is

$$\begin{aligned} 2\left\{\ell_{\mathrm{p}}(\widehat{\sigma}^2) - \ell_{\mathrm{p}}(\widehat{\sigma}_0^2)\right\} &= n\log\left\{SS(\widehat{\beta})/SS(\widehat{\beta}_1)\right\} \\ &= n\log\left[1 + \frac{p-q}{n-p}\frac{\left\{SS(\widehat{\beta}_1) - SS(\widehat{\beta})\right\}/(p-q)}{SS(\widehat{\beta})/(n-p)}\right] \\ &= n\log\left(1 + \frac{p-q}{n-p}F\right), \end{aligned} \tag{8.8}$$

say. Here $F \geq 0$, with equality only if the two sums of squares are equal. This event can occur only if the columns of X_2 are linearly dependent on those of X_1. If not, the results of Section 4.5.2 imply that the likelihood ratio statistic has an approximate χ^2 distribution, but as it is a monotonic function of F, large values of (8.8) correspond to large values of F. We shall see in Section 8.5 that the exact distribution of F is known and can be used to compare nested models, with no need for approximations.

It is instructive to express F explicitly in terms of the least squares estimators. As (8.8) is a likelihood ratio statistic for testing $\beta_2 = 0$, it is invariant to 1–1 reparametrizations that leave β_2 fixed, and we write $\mathrm{E}(y)$ as

$$\begin{aligned} X_1\beta_1 + X_2\beta_2 &= X_1\beta_1 + H_1X_2\beta_2 + (I - H_1)X_2\beta_2 \\ &= X_1\left\{\beta_1 + (X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}X_2\beta_2\right\} + Z_2\beta_2 \\ &= X_1\lambda + Z_2\psi, \end{aligned}$$

say, where $H_1 = X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}$ is the projection matrix for X_1, $Z_2 = (I - H_1)X_2$ is the matrix of residuals from regression of the columns of X_2 on those of X_1, and the new parameters are λ and $\psi = \beta_2$. Note that

$$X_1^{\mathrm{T}}Z_2 = X_1^{\mathrm{T}}\left\{I - X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}\right\}X_2 = 0,$$

and that H_1 is idempotent. In this new parametrization the parameter estimates are

$$\begin{pmatrix}\widehat{\lambda}\\ \widehat{\psi}\end{pmatrix} = \begin{pmatrix}X_1^{\mathrm{T}}X_1 & X_1^{\mathrm{T}}Z_2\\ Z_2^{\mathrm{T}}X_1 & Z_2^{\mathrm{T}}Z_2^{\mathrm{T}}\end{pmatrix}^{-1}\begin{pmatrix}X_1^{\mathrm{T}}\\ Z_2^{\mathrm{T}}\end{pmatrix}y = \begin{pmatrix}(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}y\\ (Z_2^{\mathrm{T}}Z_2)^{-1}Z_2^{\mathrm{T}}y\end{pmatrix},$$

while if $\psi = \beta_2 = 0$, the least squares estimate of λ remains $\widehat{\lambda}$. Consequently

$$\begin{aligned} SS(\widehat{\beta}) &= (y - X_1\widehat{\lambda} - Z_2\widehat{\psi})^{\mathrm{T}}(y - X_1\widehat{\lambda} - Z_2\widehat{\psi}) \\ &= (y - X_1\widehat{\lambda})^{\mathrm{T}}(y - X_1\widehat{\lambda}) - 2\widehat{\psi}^{\mathrm{T}}Z_2^{\mathrm{T}}(y - X_1\widehat{\lambda}) + \widehat{\psi}^{\mathrm{T}}Z_2^{\mathrm{T}}Z_2\widehat{\psi} \\ &= SS(\widehat{\beta}_1) - \widehat{\psi}^{\mathrm{T}}Z_2^{\mathrm{T}}Z_2\widehat{\psi}, \end{aligned}$$

since

$$\widehat{\psi}^{\mathrm{T}}Z_2^{\mathrm{T}}(y - X_1\widehat{\lambda}) = \widehat{\psi}^{\mathrm{T}}Z_2^{\mathrm{T}}y - \widehat{\psi}^{\mathrm{T}}Z_2^{\mathrm{T}}X_1\widehat{\lambda}$$

$$\begin{aligned} &= \widehat{\psi}^{\mathrm{T}}(Z_2^{\mathrm{T}}Z_2)(Z_2^{\mathrm{T}}Z_2)^{-1}Z_2^{\mathrm{T}}y \\ &= \widehat{\psi}^{\mathrm{T}}(Z_2^{\mathrm{T}}Z_2)\widehat{\psi}. \end{aligned}$$

Thus the F statistic in (8.8) may be written as

$$F = \frac{n-p}{p-q}\frac{\widehat{\beta}_2^{\mathrm{T}}X_2^{\mathrm{T}}(I-H_1)X_2\widehat{\beta}_2}{SS(\widehat{\beta})}$$

and this is large if $\widehat{\beta}_2$ differs greatly from zero.

If β_2 is scalar, then $p-q=1$, the matrix $Z_2^{\mathrm{T}}Z_2 = X_2^{\mathrm{T}}(I-H_1)X_2 = v_{pp}^{-1}$ is scalar, and $F = T^2$, where

$$T = \frac{\widehat{\beta}_2 - \beta_2}{(v_{pp}s^2)^{1/2}} \tag{8.9}$$

with $s^2 = SS(\widehat{\beta})/(n-p)$ and $\beta_2 = 0$. Thus F is a monotonic function of T^2. We shall see in Section 8.3.2 that T has a t_{n-p} distribution.

8.2.4 Weighted least squares

Suppose that a normal linear model applies but that the responses have unequal variances. If the variance of y_j is σ^2/w_j, where σ^2 is unknown but the w_j are known positive quantities giving the relative precisions of the y_j, the log likelihood can be written as

$$\ell(\beta,\sigma^2) \equiv -\tfrac{1}{2}\left\{n\log\sigma^2 + \frac{1}{\sigma^2}(y-X\beta)^{\mathrm{T}}W(y-X\beta)\right\},$$

where $W = \mathrm{diag}\{w_1,\ldots,w_n\}$ is known as the matrix of weights. Let $W^{1/2} = \mathrm{diag}\{w_1^{1/2},\ldots,w_n^{1/2}\}$, and set $y' = W^{1/2}y$ and $X' = W^{1/2}X$. Then the sum of squares may be written as $(y'-X'\beta)^{\mathrm{T}}(y'-X'\beta)$. As this has the same form as (8.3), the estimates of β and σ^2 are

$$\widehat{\beta} = (X'^{\mathrm{T}}X')^{-1}X'^{\mathrm{T}}y' = (X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}Wy, \tag{8.10}$$

and

$$\begin{aligned} s^2 &= (n-p)^{-1}y'^{\mathrm{T}}\{I - X'(X'^{\mathrm{T}}X')^{-1}X'^{\mathrm{T}}\}y' \\ &= (n-p)^{-1}y^{\mathrm{T}}\{W - WX(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W\}y. \end{aligned} \tag{8.11}$$

These are the *weighted least squares estimates*. This device of replacing y and X with $W^{1/2}y$ and $W^{1/2}X$ allows methods for unweighted least squares models to be applied when there are weights (Exercise 8.2.9).

Example 8.11 (Grouped data) Suppose that each y_j is an average of a random sample of m_j normal observations, each with mean $x_j^{\mathrm{T}}\beta$ and variance σ^2, and that the samples are independent of each other. Then y_j has mean

$x_j^{\mathrm{T}}\beta$ and variance σ^2/m_j, and the y_j are independent. The estimates of β and σ^2 are given by (8.10) and (8.11) with weights $w_j \equiv m_j$. ■

Weighted least squares can be extended to situations where the errors are correlated but the relative correlations are known, that is, $\mathrm{var}(y) = \sigma^2 W^{-1}$, where W is known but not necessarily diagonal. This is sometimes called *generalized least squares*. The corresponding least squares estimates of β and σ^2 are given by (8.10) and (8.11).

Weighted least squares turns out to be of central importance in fitting nonlinear models, and is used extensively in Chapter 10.

Exercises 8.2

1 Write down the linear model corresponding to a simple random sample $y_1, \ldots, y_n$ from the $N(\mu, \sigma^2)$ distribution, and find the design matrix. Verify that

$$\widehat{\mu} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = \overline{y}, \quad s^2 = SS(\widehat{\beta})/(n-p) = (n-1)^{-1}\sum(y_j - \overline{y})^2.$$

2 Verify the formula for s^2 given in Example 8.7, and show directly that its distribution is $\sigma^2\chi_1^2$.

3 The angles of the triangle ABC are measured with A and B each measured twice and C three times. All the measurements are independent and unbiased with common variance σ^2. Find the least squares estimates of the angles A and B based on the seven measurements and calculate the variance of these estimates.

4 In Example 8.10, show that the unbiased estimator of σ^2 is $\{2(m-1)\}^{-1}\sum(d_j - \overline{d})^2$.

Recall that: (i) if the matrix A is square, then $\mathrm{tr}(A) = \sum a_{ii}$; (ii) if A and B are conformable, then $\mathrm{tr}(AB) = \mathrm{tr}(BA)$; (iii) λ is an eigenvalue of the square matrix A if there exists a vector of unit length a such that $Aa = \lambda a$, and then a is an eigenvector of A; and (iv) a symmetric matrix A may be written as ELE^{T}, where L is a diagonal matrix of the eigenvalues of A, and the columns of E are the corresponding eigenvectors, having the property that $E^{\mathrm{T}} = E^{-1}$. If the matrix is symmetric and positive definite, then all its eigenvalues are real and positive.

5 Show that if the $n \times p$ design matrix X has rank p, the matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is symmetric and idempotent, that is, $H^{\mathrm{T}} = H$ and $H^2 = H$, and that $\mathrm{tr}(H) = p$. Show that $I_n - H$ is symmetric and idempotent also. By considering $H^2 a$, where a is an eigenvector of H, show that the eigenvalues of H equal zero or one. Prove also that H has rank p.
Give the elements of H for Examples 8.9 and 8.10.

6 In a linear model in which $n \to \infty$ in such a way that $\widehat{\beta} \xrightarrow{P} \beta$, show that $e_j \xrightarrow{P} \varepsilon_j$. Generalize this to any finite subset of the residuals e. Is this true for the entire vector e?
Let $y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$ with $x_1 = \cdots = x_k = 0$ and $x_{k+1} = \cdots = x_n = 1$. Is $\widehat{\beta}$ consistent if $n \to \infty$ and $k = 1$? If $k = m$, for some fixed m? If $k = n/2$? Which of the ε_j can be estimated consistently in each case?

7 Show that in a normal linear model in which X has rank p, the matrices $I(\beta, \sigma^2)$ and $J(\widehat{\beta}, \widehat{\sigma}^2)$ are positive definite.

8 (a) Consider the two design matrices for Example 8.4; call them X_1 and X_2. Find the 4×4 matrix A for which $X_1 = X_2 A$, and verify that it is invertible by finding its inverse.
(b) Consider the linear models $y = X_1\beta + \varepsilon$ and $y = X_2\gamma + \varepsilon$, where $X_1 = X_2 A$, $\gamma = A\beta$, and A is an invertible matrix. Show that the hat matrices, fitted values,

residuals, and sums of squares are the same for both models, and explain this in terms of the geometry of least squares.

9 (a) Consider a normal linear model $y = X\beta + \varepsilon$ where $\mathrm{var}(\varepsilon) = \sigma^2 W^{-1}$, and W is a known positive definite symmetric matrix. Show that a inverse square root matrix $W^{1/2}$ exists, and re-express the least squares problem in terms of $y_1 = W^{1/2}y$, $X_1 = W^{1/2}X$, and $\varepsilon_1 = W^{1/2}\varepsilon$. Show that $\mathrm{var}(\varepsilon) = \sigma^2 I_n$. Hence find the least squares estimates, hat matrix, and residual sum of squares for the weighted regression in terms of y, X, and W, and give the distributions of the least squares estimates of β and the residual sum of squares.
(b) Suppose that W depends on an unknown scalar parameter, ρ. Find the profile log likelihood for ρ, $\ell_{\mathrm{p}}(\rho) = \max_{\beta,\sigma^2} \ell(\beta, \sigma^2, \rho)$, and outline how to use a least squares package to give a confidence interval for ρ.

8.3 Normal Distribution Theory

8.3.1 Distributions of $\widehat{\beta}$ and s^2

The derivation of the least squares estimators in the previous section rests on the assumption that the errors satisfy the *second-order assumptions*

$$\mathrm{E}(\varepsilon_j) = 0, \quad \mathrm{var}(\varepsilon_j) = \sigma^2, \quad \mathrm{cov}(\varepsilon_j, \varepsilon_k) = 0, \quad j \neq k, \tag{8.12}$$

and in addition are normal variables. As they are uncorrelated, their normality implies they are independent. On setting $\varepsilon^{\mathrm{T}} = (\varepsilon_1, \ldots, \varepsilon_n)$, we have

$$\mathrm{E}(\varepsilon) = 0, \quad \mathrm{cov}(\varepsilon, \varepsilon) = \mathrm{E}(\varepsilon\varepsilon^{\mathrm{T}}) = \sigma^2 I_n,$$

where I_n is the $n \times n$ identity matrix. The least squares estimator equals

$$\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(X\beta + \varepsilon) = \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon,$$

which is a linear combination of normal variables, and therefore its distribution is normal. Its mean vector and covariance matrix are

$$\begin{aligned} \mathrm{E}(\widehat{\beta}) &= \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{E}(\varepsilon), \\ \mathrm{var}(\widehat{\beta}) &= \mathrm{cov}\{\beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon, \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon\} \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{cov}(\varepsilon, \varepsilon)X(X^{\mathrm{T}}X)^{-1}, \end{aligned}$$

so

$$\mathrm{E}(\widehat{\beta}) = \beta, \quad \mathrm{var}(\widehat{\beta}) = \sigma^2(X^{\mathrm{T}}X)^{-1}. \tag{8.13}$$

Therefore $\widehat{\beta}$ is normally distributed with mean and covariance matrix given by (8.13). We shall see below that the residual sum of squares has a chi-squared distribution, independent of $\widehat{\beta}$. Thus the key distributional results for the normal linear model are

$$\widehat{\beta} \sim N_p\left\{\beta, \sigma^2(X^{\mathrm{T}}X)^{-1}\right\} \quad \text{independent of} \quad SS(\widehat{\beta}) \sim \sigma^2\chi^2_{n-p}. \tag{8.14}$$

To show that the least squares estimator and residual sum of squares are independent, note that the residuals can be written as

$$e = (I_n - H)y = (I_n - H)(X\beta + \varepsilon) = (I_n - H)\varepsilon,$$

because $HX = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X = X$. Therefore the vector $e = (I_n - H)\varepsilon$ is a linear combination of normal random variables and is itself normally distributed, with mean and variance matrix

$$\begin{aligned} \mathrm{E}(e) &= \mathrm{E}\{(I_n - H)\varepsilon\} = 0, \\ \mathrm{var}(e) &= \mathrm{var}\,\{(I_n - H)\varepsilon\} = (I_n - H)\mathrm{var}(\varepsilon)(I_n - H)^{\mathrm{T}} = \sigma^2(I_n - H). \end{aligned} \tag{8.15}$$

The covariance between $\widehat{\beta}$ and e is

$$\begin{aligned} \mathrm{cov}(\widehat{\beta}, e) &= \mathrm{cov}\{\beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\varepsilon, (I_n - H)\varepsilon\} \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\mathrm{cov}(\varepsilon, \varepsilon)(I_n - H)^{\mathrm{T}} \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\sigma^2 I_n(I_n - H)^{\mathrm{T}} = 0. \end{aligned}$$

As both e and $\widehat{\beta}$ are normally distributed and their covariance matrix is zero, they are independent, which implies that $\widehat{\beta}$ and the residual sum of squares $SS(\widehat{\beta}) = e^{\mathrm{T}}e$ are independent.

The key to the distribution of $SS(\widehat{\beta})$ is the decomposition

$$\begin{aligned} \varepsilon^{\mathrm{T}}\varepsilon &= (y - X\beta)^{\mathrm{T}}(y - X\beta) \\ &= (y - X\widehat{\beta} + X\widehat{\beta} - X\beta)^{\mathrm{T}}(y - X\widehat{\beta} + X\widehat{\beta} - X\beta) \\ &= \left\{e + X(\widehat{\beta} - \beta)\right\}^{\mathrm{T}} \left\{e + X(\widehat{\beta} - \beta)\right\}, \end{aligned}$$

which leads to

$$\varepsilon^{\mathrm{T}}\varepsilon/\sigma^2 = e^{\mathrm{T}}e/\sigma^2 + (\widehat{\beta} - \beta)^{\mathrm{T}}X^{\mathrm{T}}X(\widehat{\beta} - \beta)/\sigma^2, \tag{8.16}$$

because $e^{\mathrm{T}}X = y^{\mathrm{T}}(I_n - H)X = 0$. The left-hand side of (8.16) is a sum of the n independent chi-squared variables ε_j^2/σ^2, so its distribution is χ_n^2; its moment-generating function is $(1 - 2t)^{-n/2}$, $t < \frac{1}{2}$. It follows from applying (3.23) to the normal distribution of $\widehat{\beta}$ in (8.14) that $(\widehat{\beta} - \beta)^{\mathrm{T}}X^{\mathrm{T}}X(\widehat{\beta} - \beta)/\sigma^2 \sim \chi_p^2$. On taking moment-generating functions of both sides of (8.16) we therefore obtain

$$(1 - 2t)^{-n/2} = \mathrm{E}\left\{\exp(te^{\mathrm{T}}e/\sigma^2)\right\} \times (1 - 2t)^{-p/2}, \quad t < \tfrac{1}{2},$$

because e and $\widehat{\beta}$ are independent. Therefore $e^{\mathrm{T}}e/\sigma^2$ has moment-generating function $(1 - 2t)^{-(n-p)/2}$, showing that its distribution is χ_{n-p}^2. We need only recall that $SS(\widehat{\beta}) = e^{\mathrm{T}}e$ to establish the remaining result in (8.14): under the normal linear model, we have $SS(\widehat{\beta})/\sigma^2 \sim \chi_{n-p}^2$.

As the distribution of $SS(\widehat{\beta})$ is $\sigma^2\chi^2_{n-p}$, its mean is $\mathrm{E}\{SS(\widehat{\beta})\} = (n-p)\sigma^2$, and its variance is $\mathrm{var}\{SS(\widehat{\beta})\} = 2(n-p)\sigma^4$. Thus

$$S^2 = \frac{1}{n-p}\sum_{j=1}^{n}(y_j - x_j\widehat{\beta})^2 = \frac{1}{n-p}SS(\widehat{\beta})$$

is an unbiased estimator of σ^2, whereas $\widehat{\sigma}^2 = SS(\widehat{\beta})/n$ is biased.

8.3.2 Confidence and prediction intervals

Confidence intervals for components of β are based on the distributions of $\widehat{\beta}$ and S^2. Under the normal linear model the rth element of $\widehat{\beta}$ satisfies

$$\widehat{\beta}_r \sim N(\beta_r, \sigma^2 v_{rr}),$$

where v_{rr} is the rth diagonal element of $(X^{\mathrm{T}}X)^{-1}$, and $\widehat{\beta}$ is independent of S^2, whose distribution is $(n-p)^{-1}\sigma^2\chi^2_{n-p}$. Therefore

$$T = \frac{\widehat{\beta}_r - \beta_r}{\sqrt{S^2 v_{rr}}} \sim t_{n-p},$$

which makes the connection with (8.9). A $(1-2\alpha)$ confidence interval for β_r is $\widehat{\beta}_r \pm s v_{rr}^{1/2} t_{n-p}(\alpha)$. When σ^2 is known, we replace s by σ and $t_{n-p}(\alpha)$ by the normal quantile z_α.

$t_\nu(\alpha)$ is the α quantile of the t_ν distribution.

Similar reasoning gives confidence intervals for linear functions of β. The maximum likelihood estimator of the linear function $x_+^{\mathrm{T}}\beta$ is $x_+^{\mathrm{T}}\widehat{\beta}$, which has a normal distribution with mean $x_+^{\mathrm{T}}\beta$ and variance

$$\mathrm{var}(x_+^{\mathrm{T}}\widehat{\beta}) = x_+^{\mathrm{T}}\mathrm{var}(\widehat{\beta})x_+ = \sigma^2 x_+^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_+.$$

As S^2 is independent of $\widehat{\beta}$, confidence regions for $x_+^{\mathrm{T}}\beta$ can be based on

$$\frac{x_+^{\mathrm{T}}\widehat{\beta} - x_+^{\mathrm{T}}\beta}{\left\{S^2 x_+^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_+\right\}^{1/2}} \sim t_{n-p}.$$

If σ^2 is known, the observed s is replaced in the confidence interval by σ and quantiles of the t distribution are replaced by those of the normal. Notice that the variance of a fitted value $\widehat{y}_j = x_j^{\mathrm{T}}\widehat{\beta}$ is $\sigma^2 x_j^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_j$, and this equals $\sigma^2 h_{jj}$, where h_{jj} is the jth diagonal element of the hat matrix H.

A confidence interval for a function of parameters is different from a *prediction interval* for a new observation, $y_+ = x_+^{\mathrm{T}}\beta + \varepsilon_+$. The presence of ε_+ would introduce uncertainty about y_+ even if β was known, and a prediction interval must take this into account. If ε_+ is normal with mean zero and variance σ^2, independent of the data from which $\widehat{\beta}$ is estimated, we have

$$\begin{aligned}\mathrm{E}(x_+^{\mathrm{T}}\widehat{\beta} + \varepsilon_+) &= x_+^{\mathrm{T}}\beta,\\ \mathrm{var}(x_+^{\mathrm{T}}\widehat{\beta} + \varepsilon_+) &= \mathrm{var}(x_+^{\mathrm{T}}\widehat{\beta}) + \mathrm{var}(\varepsilon_+) = \sigma^2\{x_+^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_+ + 1\}.\end{aligned}$$

When σ^2 is unknown, therefore, a prediction interval for y_+ can be based on

$$\frac{y_+ - x_+^{\mathrm{T}}\widehat{\beta}}{\left[S^2\left\{1 + x_+^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_+\right\}\right]^{1/2}} \sim t_{n-p},$$

with the appropriate changes if σ^2 is known.

Example 8.12 (Cycling data) The covariance matrix for the parameter estimates in Example 8.8 is $\frac{\sigma^2}{16}I_4$. As the residual sum of squares is $SS(\widehat{\beta}) = 43.25$, $n = 16$ and $p = 4$, an estimate of σ^2 is $s^2 = 43.25/12 = 3.604$ on 12 degrees of freedom, and each estimate $\widehat{\beta}_r$ has standard error $(s^2/16)^{1/2} = 0.475$.

A 0.95 confidence interval for the true value of β_1 is $\widehat{\beta}_1 \pm st_{12}(0.025)/4$, and this is $-5.437 \pm 0.475 \times 2.18 = (-6.47, -4.40)$ seconds, clear evidence that the time is shorter when the seat is higher. The change due to the effect of tyre pressure is $2\widehat{\beta}_3$ seconds, for which the standard error is $2 \times s/4 = 0.95$ seconds.

A 0.95 prediction interval for a further timing y_+ made with all three factors set at their higher levels would be $41.75 \pm (1 + \frac{4}{16})^{1/2} st_{12}(0.025)$, which is $(39.49, 46.01)$. The variability introduced by ε_+ forms the bulk of the variability of y_+, whose variance is five times that of the fitted value. ■

Example 8.13 (Maize data) Consider the two-sample model applied to the data in Table 1.1. If we assume that the heights of the cross-fertilized plants form a random sample with means $\beta_0 + \beta_1$, and that the heights of the self-fertilized plants form a random sample with height β_0, and that both have variance σ^2, the results of Example 8.9 establish that the estimates are

$$\widehat{\beta}_0 = \overline{y}_{0\cdot} = 140.6, \quad \widehat{\beta}_1 = \overline{y}_{1\cdot} - \overline{y}_{0\cdot} = 161.53 - 140.6 = 20.93,$$

that the unbiased estimate of σ^2 is $s^2 = 553.19$, and that the estimated variance of $\widehat{\beta}_1$ is $s^2(n_0^{-1} + n_1^{-1}) = 73.78$. As s^2 has 28 degrees of freedom, a 0.95 confidence interval for β_1 has limits

$$\widehat{\beta}_1 \pm s(n_0^{-1} + n_1^{-1})^{1/2} t_{28}(0.025) = 20.93 \pm 73.78^{1/2} \times 2.048 = 3.34, 38.52.$$

This does not contain zero, and is evidence that the crossed plants are significantly taller than self-fertilized plants.

For the matched pairs model of Example 8.10, there are $m = 15$ pairs, with $\widehat{\beta}_0 = 10.48$ and $s^2 = 712.36$, on $2m - (m+1) = 14$ degrees of freedom. A 0.95 confidence interval for β_0 based on this model has limits

$$\widehat{\beta}_0 \pm \{s^2/(2m)\}^{1/2} t_{14}(0.025) = 10.48 \pm (712.36/30)^{1/2} \times 2.154 = 0.00, 20.96.$$

The corresponding interval for the height increase for crossed plants is an interval for $2\beta_0$, that is, $(0.00, 41.91)$. This is wider than the interval for the

two-sample model, and just contains the value zero, giving evidence that there may be no increase due to cross-fertilization. The increase in interval width has two causes. First, the estimate of σ^2 for the matched pairs model equals 712.36, which is larger than the value 553.19 for the two-sample model. Second, there are only 14 degrees of freedom for the matched pairs estimate of variance, and $|t_{14}(0.025)| > |t_{28}(0.025)|$, which slightly inflates the matched pairs confidence interval relative to the interval from the matched analysis. ■

Exercises 8.3

1 The following table gives the parameter estimates, standard errors and correlations, when the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ is fitted to the cement data of Example 8.3. The residual sum of squares is 48.11.

	Estimate	SE	Correlations of Estimates		
(Intercept)	48.19	3.913	(Intercept)	x1	x2
x1	1.70	0.205	x1 -0.736		
x2	0.66	0.044	x2 -0.416	-0.203	
x3	0.25	0.185	x3 -0.828	0.822	-0.089

On the assumption that this normal linear model applies, compute 0.95 confidence intervals for β_0, β_1, β_2, and β_3, and test the hypothesis that $\beta_3 = 0$. Compute a 0.90 confidence interval for $\beta_2 - \beta_3$.

2 Let $\widehat{\beta}$ be a least squares estimator, and suppose that $\varepsilon_+ \sim N(0, \sigma^2)$ independent of $\widehat{\beta}$. Verify that $\mathrm{var}(x_+^{\mathrm{T}}\widehat{\beta}) = \sigma^2 x_+^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_+$ and that $\mathrm{var}(x_+^{\mathrm{T}}\widehat{\beta} + \varepsilon_+) = \sigma^2\{1 + x_+^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_+\}$. Assuming that a normal linear model is suitable for the cycling data, calculate a 0.90 confidence interval for the mean time to cycle up the hill when the three factors are at their lowest levels. Obtain also a 0.90 prediction interval for a future observation made with that setup.

8.4 Least Squares and Robustness

In Section 8.2.1 we established that $\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ is the maximum likelihood estimator of the regression parameter β under the assumption of normal responses. The model is a linear exponential family with complete minimal sufficient statistic $(\widehat{\beta}, S^2)$, and it follows that these are the unique minimum variance unbiased estimators of (β, σ^2). It is natural to ask to what optimality properties hold more generally. We shall see below that $\widehat{\beta}$ has minimum variance among all estimators linear in the responses y, under assumptions on the mean and variance structure of y alone. Thus the least squares estimator retains optimality properties even without full distributional assumptions. This has important generalizations, as we shall see in Section 10.6.

Suppose that the second-order assumptions (8.12) hold, but that the errors are not necessarily normal. Thus, although uncorrelated, they may be dependent. Then $\mathrm{E}(y) = X\beta$ and $\mathrm{var}(y) = \sigma^2 I_n$. Let $\tilde{\beta}$ denote any unbiased

estimator of β that is linear in y. Then a $p \times n$ matrix A exists such that $\tilde{\beta} = Ay$, and unbiasedness implies that $\mathrm{E}(\tilde{\beta}) = AX\beta = \beta$ for any parameter vector β; this entails $AX = I_p$. Now

The $n \times n$ hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ is symmetric and idempotent and hence so is $I_n - H$.

$$\begin{aligned} \mathrm{var}(\tilde{\beta}) - \mathrm{var}(\widehat{\beta}) &= A\sigma^2 I_n A^{\mathrm{T}} - \sigma^2 (X^{\mathrm{T}}X)^{-1} \\ &= \sigma^2 \left\{ AA^{\mathrm{T}} - AX(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}A^{\mathrm{T}} \right\} \\ &= \sigma^2 A(I_n - H)A^{\mathrm{T}} \\ &= \sigma^2 A(I_n - H)(I_n - H)^{\mathrm{T}}A^{\mathrm{T}} \end{aligned}$$

and this $p \times p$ matrix is positive semidefinite. Thus $\widehat{\beta}$ has smallest variance in finite samples among all linear unbiased estimators of β, provided that the second-order assumptions hold. This result, the *Gauss–Markov* theorem, gives further support for using $\widehat{\beta}$ if a linear estimator of β is sought, though of course nonlinear estimators may have smaller variance.

Johann Carl Friedrich Gauss (1777–1855) was born and educated in Brunswick. He studied in Göttingen and obtained a doctorate from the University of Helmstedt. His first book, published at the age of 24, contained the largest advance in geometry since the Greeks. He became director of the Göttingen observatory and invented least squares estimation for the combination of astronomical observations, though his statistical work was not published until much later. He also wrote treatises on theoretical astronomy, surveying, terrestial magnetism, infinite series, integration, number theory, and differential geometry.

Example 8.14 (Student t density) Suppose that $y = X\beta + \sigma\varepsilon$, where the ε_j are independent and have the Student t density (3.11) with ν degrees of freedom. Now $\mathrm{var}(\varepsilon_j)$ is finite and equals $\nu/(\nu-2)$ provided $\nu > 2$, and then the least squares estimator has variance matrix $\sigma^2\nu/(\nu-2) \times (X^{\mathrm{T}}X)^{-1}$.

How much efficiency is lost by using least squares rather than maximum likelihood estimation for β? To see this we must compute the expected information matrix, which gives the inverse variance of the maximum likelihood estimator. The log likelihood assuming ν and σ^2 known is

$$\ell(\beta) \equiv -\frac{\nu+1}{2} \sum_{j=1}^{n} \log\left\{1 + (y_j - x_j^{\mathrm{T}}\beta)^2/(\nu\sigma^2)\right\},$$

and differentiation with respect to β gives

$$\begin{aligned} \frac{\partial \ell(\beta)}{\partial \beta} &= \frac{\nu+1}{\nu\sigma^2} \sum_{j=1}^{n} \frac{y_j - x_j^{\mathrm{T}}\beta}{1 + (y_j - x_j^{\mathrm{T}}\beta)^2/(\nu\sigma^2)} x_j, \\ -\frac{\partial^2 \ell(\beta)}{\partial \beta \partial \beta^{\mathrm{T}}} &= \frac{\nu+1}{\nu\sigma^2} \sum_{j=1}^{n} \frac{1 - (y_j - x_j^{\mathrm{T}}\beta)^2/(\nu\sigma^2)}{\left\{1 + (y_j - x_j^{\mathrm{T}}\beta)^2/(\nu\sigma^2)\right\}^2} x_j x_j^{\mathrm{T}}. \end{aligned}$$

Now $\mathrm{E}\{(1 + \varepsilon^2/\nu)^{-r}\} = (\nu + 2r - 2)\cdots\nu/\{(\nu + 2r - 1)\cdots(\nu + 1)\}$, so the expected information for β is $\sigma^{-2}(\nu+1)/(\nu+3) \times X^{\mathrm{T}}X$. Thus the maximum likelihood estimator is a nonlinear function of y with large-sample variance matrix $\sigma^2(\nu+3)/(\nu+1) \times (X^{\mathrm{T}}X)^{-1}$. It follows that the least squares estimator has asymptotic relative efficiency $(\nu-2)(\nu+3)/\{\nu(\nu+1)\}$, independent of the design matrix, β, or σ^2. As $\nu \to \infty$, the efficiency tends to one; for $\nu = 5$, 10, and 20 it equals 0.8, 0.95, and 0.99. Maximum likelihood estimation of β barely improves on least squares for a wide range of ν, because the t density is close to normal unless ν is small. ∎

M-estimation

The least squares estimators have strong optimality properties, but because they are linear in y, they are sensitive to outliers. When data are too extensive to be carefully inspected or when bad data are present, robust or resistant estimators are more appropriate. One approach to constructing them is to replace the sum of squares with a function $\sum \rho\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\}$ that down-weights extreme values of $(y_j - x_j^{\mathrm{T}}\beta)/\sigma$. The resulting estimators are called *M-estimators* because they are maximum-likelihood-like: the function ρ takes the place of a negative log likelihood. They may also be defined as the solutions of the $p \times 1$ estimating equation (Section 7.2)

$$\sigma^{-1} \sum_{j=1}^{n} x_j \rho'\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\} = 0, \tag{8.17}$$

where $\rho'(u) = d\rho(u)/du$, which extends the least squares estimating equation

$$X^{\mathrm{T}}(y - X\beta) = \sum_{j=1}^{n} x_j (y_j - x_j^{\mathrm{T}}\beta) = 0. \tag{8.18}$$

Many functions $\rho(u)$ have been proposed. Setting $\rho(u) = u^2/2$ gives least squares. Other possibilities include $\rho(u) = |u|$, $\rho(u) = \nu \log(1 + u^2/\nu)/2$, and

$$\rho(u) = \begin{cases} u^2, & \text{if } |u| < c, \\ c(2|u| - c), & \text{otherwise,} \end{cases}$$

corresponding to the median, a t_ν density, and a Huber estimator (Example 7.19). These have the drawback that large outliers are not downweighted to zero. This can be achieved with a redescending function such as the biweight,

$$\rho'(u) = u \max\left[\left\{1 - (u/c')^2\right\}^2, 0\right];$$

taking $c' = 4.865$ gives asymptotic efficiency 0.95 for normal data.

Notice that $\sum \rho\{(y_j - x_j^{\mathrm{T}}\beta)/\sigma\}$ has second derivative $\sigma^{-2} \sum x_j x_j^{\mathrm{T}} g'(y_j - x_j^{\mathrm{T}}\beta)$, whose expectation is of form $\sigma^{-2} X^{\mathrm{T}} X \times \mathrm{E}\{g'(\varepsilon)\}$ under a model in which $y_j = x_j^{\mathrm{T}}\beta + \sigma\varepsilon_j$ and the ε_j are independent and identically distributed with zero mean and unit variance. The ideas of Section 7.2 imply that the M-estimator has asymptotic variance

$$\sigma^2 (X^{\mathrm{T}} X)^{-1} \times \mathrm{E}\left\{g(\varepsilon)^2\right\} / \mathrm{E}\left\{g'(\varepsilon)\right\},$$

so its efficiency relative to least squares is simply $\mathrm{E}\{g'(\varepsilon)\}/\mathrm{E}\left\{g(\varepsilon)^2\right\}$. The Huber estimator for regression has efficiencies given by the right panel of Figure 7.4, for instance.

Equation (8.17) may be solved using iterative versions of least squares described in Section 10.2.2, though these may fail to converge if ρ is not convex.

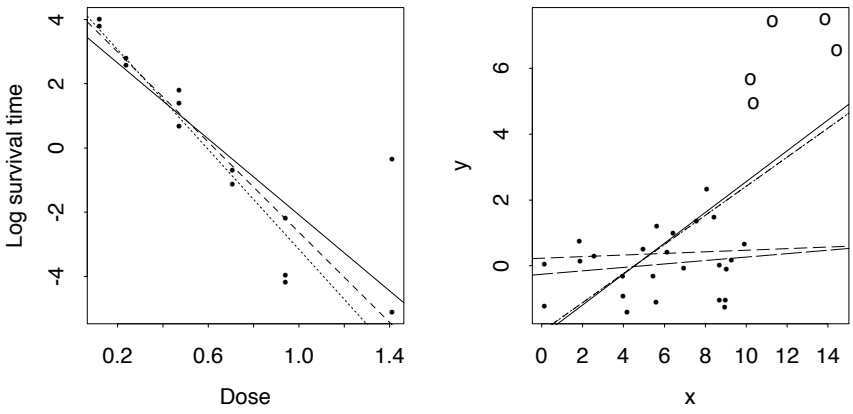


Figure 8.3 Data for which least squares estimation fails. Left: log survival proportions for rats given doses of radiation, with lines fitted by least squares with (solid) and without (dots) the outlier, and a Huber M-estimate for the entire data (dashes) (Efron, 1988). Right: simulated data with a batch of outliers (circles), and fits by least squares to all data (solid), least squares to good data only (large dash), Huber (dot-dash), biweight (dashes), and least trimmed squares (medium dash). The Huber and biweight fits are the same to plotting accuracy.

In practice σ too must be estimated, by the median absolute deviation of the residuals $y_j - x_j^{\mathrm{T}}\widehat{\beta}$ at each iteration, or using an M-estimator of scale.

Initial values for these fits can be found by a highly resistant procedure such as *least trimmed squares*, whereby β is chosen to minimize $\sum_{i=1}^{q}(y_j - x_j^{\mathrm{T}}\beta)^2_{(i)}$; this is the sum of the smallest $q = \lfloor n/2 \rfloor + \lfloor (p+1)/2 \rfloor$ squared residuals, found by a Monte Carlo search. Highly resistant procedures do not usually provide standard errors, which can be obtained by a data-based simulation procedure such as the bootstrap; see the bibliographic notes.

Example 8.15 (Survival data) The left panel of Figure 8.3 shows data on batches of rats given doses of radiation. They are well fit by a straight line, apart from an apparent outlier, which strongly affects the least squares fit — note what the pattern of residuals will be. The least squares estimates of slope and its standard error with and without the outlier are -5.91 (1.05) and -7.79 (0.59), while Huber estimation gives -7.02 (0.46). Downweighting the outlier using the robust estimator gives a result intermediate between keeping it and deleting it.

This sample is small and the outlier sticks out, so robust methods are not really needed. They are more valuable for larger more complex data sets where visualization is difficult and outliers non-obvious. ■

Example 8.16 (Simulated data) To illustrate and compare some robust estimators, we generated sets of 25 standard normal observations y with a single covariate x, and then added k outliers with mean 6, having the t_5 distribution. The right panel of Figure 8.3 shows one of these datasets, with $k = 5$. We then computed five estimates of slope, from least squares applied

Table 8.4 Bias (standard deviation) of estimators of slope in sample of 25 good data and k outliers, estimated from 200 replications.

k	Least squares		M-estimation		Least trimmed squares
	No outliers	With outliers	Huber	Biweight	
1	0.00 (0.07)	0.17 (0.06)	0.07 (0.07)	0.01 (0.07)	−0.01 (0.13)
2	0.00 (0.07)	0.26 (0.06)	0.13 (0.07)	0.02 (0.09)	0.01 (0.14)
5	0.00 (0.07)	0.41 (0.05)	0.38 (0.06)	0.19 (0.19)	0.01 (0.14)
10	0.00 (0.06)	0.48 (0.04)	0.48 (0.04)	0.46 (0.12)	0.05 (0.20)

with and without the outliers, from Huber and biweight M-estimators having efficiency 0.95 at the normal model, and from least trimmed squares. Table 8.4 shows the bias and standard deviation of the slope estimators for various k, computed from 200 replicate data sets.

Inclusion of just one outlier ruins the least squares estimator, which is the benchmark when outliers are excluded. The biweight gives the better of the M-estimators, but with $k \geq 5$ it is badly biased. The M-estimators perform as badly as least squares when contamination is high. Least trimmed squares is least biased overall, but is very inefficient even for $k = 1$. This suggests that a good practical data analysis strategy is to use an initial least trimmed squares fit to identify and delete outliers, and then apply M-estimation to the remaining data. ■

Misspecified variance

Outliers are just one of many possible problems in regression. Suppose that although $\mathrm{E}(y) = X\beta$, the variance is $\mathrm{var}(y) = V$ rather than the assumed $\sigma^2 I_n$. Then $\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ has variance

$$(X^{\mathrm{T}}X)^{-1}(X^{\mathrm{T}}VX)(X^{\mathrm{T}}X)^{-1}. \tag{8.19}$$

If $V = \sigma^2 I_n$, then $\mathrm{var}(\widehat{\beta}) = \sigma^2(X^{\mathrm{T}}X)^{-1}$, which itself is the inverse Fisher information for β under the normal model. Thus if the variance of y is correctly supposed to equal $\sigma^2 I_n$, the least squares estimator attains the Cramér–Rao lower bound appropriate to normal responses, while (7.20) implies that $\mathrm{var}(\widehat{\beta})$ is inflated otherwise.

Most packages use the formula $\sigma^2(X^{\mathrm{T}}X)^{-1}$ and make no allowance for possible variance misspecification. If plots such as those described in Section 8.6 do not suggest a particular variance to be fitted using weighted least squares, the weights being $W = V^{-1}$, then it may be better to apply least squares but to base confidence intervals on an estimate of (8.19). One simple possibility is to replace V with $\widehat{V} = \mathrm{diag}\{r_1^2, \ldots, r_n^2\}$, where $r_j = (y_j - \widehat{y}_j)/(1 - h_{jj})$.

Exercises 8.4

1 Check the details of Example 8.14.

2 Show that $\widehat{\beta}$ and S^2 are unbiased estimators of β and σ^2 even when the errors are not normal, provided that the second-order assumptions are satisfied.

3 Consider a linear regression model (8.1) in which the errors ε_j are independently distributed with Laplace density

$$f(u;\sigma) = (2^{3/2}\sigma)^{-1} \exp\{-|u/(2^{1/2}\sigma)|\}, \quad -\infty < u < \infty, \sigma > 0.$$

Verify that this density has variance σ^2. Show that the maximum likelihood estimate of β is obtained by minimizing the L^1 norm $\sum |y_j - x_j^{\mathrm{T}}\beta|$ of $y - X\beta$. Show that if in fact the $\varepsilon_j \overset{\text{iid}}{\sim} N(0,\sigma^2)$, the asymptotic relative efficiency of the estimators relative to least squares estimators is $2/\pi$.

4 Consider a linear model $y_j = x_j\beta + \varepsilon_j$, $j = 1,\ldots,n$ in which the ε_j are uncorrelated and have means zero. Find the minimum variance linear unbiased estimators of the scalar β when (i) $\mathrm{var}(\varepsilon_j) = x_j\sigma^2$, and (ii) $\mathrm{var}(\varepsilon_j) = x_j^2\sigma^2$. Generalize your results to the situation where $\mathrm{var}(\varepsilon) = \sigma^2/w_j$, where the weights w_j are known but σ^2 is not.

5 Use (8.18) to establish that (7.20) takes form

$$(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}VX(X^{\mathrm{T}}X)^{-1} \geq \sigma^2(X^{\mathrm{T}}X)^{-1}$$

when $\mathrm{var}(y)$ is wrongly supposed equal to $\varepsilon^2 I_n$ instead of V.

8.5 Analysis of Variance

8.5.1 F statistics

In most regression models a key question is whether or not the explanatory variables affect the response. For example, in the bicycle data, we were concerned how the time to climb the hill depended on the seat height and other factors. Ockham's razor suggests that we use the simplest model we can. This poses the question: which explanatory variables are needed? To be concrete, suppose that we fit a normal linear model

$$y = X\beta + \varepsilon = (X_1, X_2)\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix} + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon, \tag{8.20}$$

where X_1 is an $n \times q$ matrix, X_2 is an $n \times (p-q)$ matrix, $q < p$, and β_1 and β_2 are vectors with respective lengths q and $p - q$. We suppose that X has rank p and X_1 has rank q. The explanatory variables X_2 are unnecessary if $\beta_2 = 0$, in which case the simpler model $y = X_1\beta_1 + \varepsilon$ holds. How can we detect this?

In Figure 8.2, let the line $x = 0$ in the horizontal plane through the origin represent the linear subspace spanned by the columns of X_1. The fitted value $\widehat{y}_1 = X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}y$ is the orthogonal projection of y onto this subspace.

The vector of residuals, $y - \widehat{y}_1 = \{I_n - X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}\}y$, resolves into the two orthogonal vectors $y - \widehat{y}$ and $\widehat{y} - \widehat{y}_1$; that is,

$$y - \widehat{y}_1 = (y - \widehat{y}) + (\widehat{y} - \widehat{y}_1),$$

where $(y - \widehat{y})^{\mathrm{T}}(\widehat{y} - \widehat{y}_1) = 0$. These vectors are the residual from the more complex model, $y - \widehat{y}$, and the change in fitted values when X_2 is added to the design matrix, $\widehat{y} - \widehat{y}_1$. As these vectors are orthogonal linear functions of the normally distributed vector y, they are independent. Pythagoras' theorem implies that

$$(y - \widehat{y}_1)^{\mathrm{T}}(y - \widehat{y}_1) = (y - \widehat{y})^{\mathrm{T}}(y - \widehat{y}) + (\widehat{y} - \widehat{y}_1)^{\mathrm{T}}(\widehat{y} - \widehat{y}_1),$$

or equivalently

$$SS(\widehat{\beta}_1) = SS(\widehat{\beta}) + \left\{SS(\widehat{\beta}_1) - SS(\widehat{\beta})\right\}. \tag{8.21}$$

Thus the residual sum of squares for the simpler model is the sum of two independently distributed parts: the residual sum of squares for the more elaborate model, $SS(\widehat{\beta})$, and the reduction in sum of squares when the columns of X_2 are added to the design matrix, $SS(\widehat{\beta}_1) - SS(\widehat{\beta})$.

If the submodel is correct, so too is the more elaborate model, because β_2 takes the particular value zero. In this case $SS(\widehat{\beta}_1)$ has a $\sigma^2\chi^2_{n-q}$ distribution, and $SS(\widehat{\beta})$ has a $\sigma^2\chi^2_{n-p}$ distribution. Since $SS(\widehat{\beta}_1) - SS(\widehat{\beta})$ is independent of $SS(\widehat{\beta})$, (8.21) implies that when $\beta_2 = 0$, $SS(\widehat{\beta}_1) - SS(\widehat{\beta})$ has a $\sigma^2\chi^2_{p-q}$ distribution, and that

$$F = \frac{\{SS(\widehat{\beta}_1) - SS(\widehat{\beta})\}/(p-q)}{SS(\widehat{\beta})/(n-p)} \quad \sim \quad F_{p-q,n-p};$$

recall (8.8). If β_2 is non-zero, the reduction in sum of squares due to including the columns of X_2 in the design matrix will be larger on average than if $\beta_2 = 0$. Thus if $\beta_2 \neq 0$, F will tend to be large relative to the $F_{p-q,n-p}$ distribution. We can therefore test the adequacy of the simpler model using the statistic F, large values of which suggest that $\beta_2 \neq 0$.

Exercise 8.5.3 gives the algebraic equivalent of the geometric argument above. As we saw in Section 8.2.3, F arises from the likelihood ratio statistic for comparison of the two models. When X_2 consists of a single covariate, β_2 is scalar, and tests and confidence intervals for it may be obtained by fitting the more elaborate model (8.20) and calculating $T = (\widehat{\beta}_2 - \beta_2)/(sv_{rr}^{1/2})$. Here s^2 is the estimate of σ^2 from the more elaborate model, and the null distribution of T is t_{n-p}. In this situation there is a simple connection to F: when testing $\beta_2 = 0$, $F = T^2 = \widehat{\beta}_2^2/(s^2v_{rr})$.

Example 8.17 (Cement data) Suppose that we want to compare the models $y = \beta_0 + x_1\beta_1 + \varepsilon$ and $y = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + \varepsilon$. This

corresponds to asking if is there any effect on y of x_2, x_3, or x_4, after allowing for the effect of x_1. Here X_1 is a 13×2 matrix whose columns are a vector of ones and x_1, and X_2 is a 13×3 matrix whose columns are x_2, x_3, and x_4; both matrices have full rank.

For the full model $p = 5$ and the residual sum of squares is $SS(\widehat{\beta}) = 47.86$, and for the simpler model $q = 2$ and the residual sum of squares is $SS(\widehat{\beta}_1) = 1265.7$. Thus the reduction in sum of squares due to the columns of X_2 after fitting X_1 is $1265.7 - 47.86 = 1217.84$ on three degrees of freedom. To test whether this is a significant reduction, we compute

$$F = \frac{(1265.7 - 47.86)/(5-2)}{47.86/(13-5)} = 67.86,$$

$F_{\nu_1,\nu_2}(\alpha)$ is the α quantile of the F distribution with ν_1 and ν_2 degrees of freedom.

which would be consistent with an $F_{3,8}$ distribution if the simpler model was adequate. As F greatly exceeds $F_{3,8}(0.95) = 4.066$, there is strong evidence that there are effects of the added covariates.

Having established that adding extra covariates helps to explain the overall variation, it is natural to ask whether this is due to a subset of them rather than to all three. Is there a more informative decomposition of the sum of squares due to adding X_2? ■

8.5.2 Sums of squares

The interpretation of sums of squares is most useful if they can be decomposed into the reductions from successively adding different explanatory variables to the design matrix.

Suppose that we have a normal linear model

$$y = 1_n\beta_0 + X_1\beta_1 + X_2\beta_2 + \cdots + X_m\beta_m + \varepsilon, \tag{8.22}$$

where we call the matrices $1_n, X_1, X_2$, and so forth *terms*; the constant term 1_n is a column of n ones. Usually the simplest model that might be considered is $y = 1_n\beta_0 + \varepsilon$, in which case the fitted value is $\widehat{y}_0 = 1_n\overline{y}$, and the residual sum of squares is $SS_0 = \sum(y_j - \overline{y})^2$ with $\nu_0 = n - 1$ degrees of freedom.

We now consider the successive reductions in sum of squares due to adding the terms X_1, X_2, and so forth to the design matrix. Let $\widehat{y}_r$ be the fitted value when the terms $X_1, \ldots, X_r$ are included, and write

$$y - \widehat{y}_0 = (y - \widehat{y}_m) + (\widehat{y}_m - \widehat{y}_{m-1}) + \cdots + (\widehat{y}_1 - \widehat{y}_0).$$

This decomposition extends that leading to (8.21) and shown in Figure 8.2. The geometry of least squares implies that the quantities in parentheses on the right are mutually orthogonal. Pythagoras' theorem tells us that $(y - \widehat{y}_0)^{\mathrm{T}}(y - \widehat{y}_0)$ equals

$$(y - \widehat{y}_m)^{\mathrm{T}}(y - \widehat{y}_m) + (\widehat{y}_m - \widehat{y}_{m-1})^{\mathrm{T}}(\widehat{y}_m - \widehat{y}_{m-1}) + \cdots + (\widehat{y}_1 - \widehat{y}_0)^{\mathrm{T}}(\widehat{y}_1 - \widehat{y}_0),$$

Table 8.5 Analysis of variance table.

Terms	df	Residual sum of squares	Terms added	df	Reduction in sum of squares	Mean square
1_n	$n-1$	SS_0				
$1_n, X_1$	ν_1	SS_1	X_1	$n-1-\nu_1$	$SS_0 - SS_1$	$\frac{SS_0-SS_1}{n-1-\nu_1}$
$1_n, X_1, X_2$	ν_2	SS_2	X_2	$\nu_1-\nu_2$	$SS_1 - SS_2$	$\frac{SS_1-SS_2}{\nu_1-\nu_2}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$1_n, X_1, \ldots, X_m$	ν_m	SS_m	X_m	$\nu_{m-1}-\nu_m$	$SS_{m-1} - SS_m$	$\frac{SS_{m-1}-SS_m}{\nu_{m-1}-\nu_m}$

or equivalently

$$SS_0 = SS_m + (SS_{m-1} - SS_m) + \cdots + (SS_0 - SS_1), \tag{8.23}$$

where SS_r denotes the residual sum of squares that corresponds to the fitted value $\widehat{y}_r$, on ν_r degrees of freedom. In (8.23) the difference $SS_{r-1} - SS_r$ is the reduction in residual sum of squares due to adding the term X_r when the model already contains $1_n, X_1, \ldots, X_{r-1}$. As y is normal and the vectors $\widehat{y}_r - \widehat{y}_{r-1}$ and $y - \widehat{y}_m$ are all linear functions of the data, the geometry of least squares implies that SS_m and all the $SS_{r-1} - SS_r$ are mutually independent.

As more terms are successively added to the model, the degrees of freedom of the residual sums of squares decrease, that is, $\nu_0 \geq \nu_1 \geq \cdots \geq \nu_m$, with $\nu_r = \nu_{r+1}$ when the columns of X_{r+1} are a linear combination of the columns of the matrices $1_n, X_1, \ldots, X_r$. If $\nu_r = \nu_{r+1}$, $\widehat{y}_r = \widehat{y}_{r+1}$, and $SS_r = SS_{r+1}$. The term X_{r+1} is then redundant, because its inclusion does not change the fitted model.

Analysis of variance

The sums of squares can be laid out in an *analysis of variance table*. The prototype is Table 8.5. The residual sums of squares decrease as terms are added successively to the model. Often the three leftmost columns are omitted and their bottom row is placed under the right-hand columns; SS_m is used to compute the denominator for the F statistics for inclusion of X_1, X_2 and so forth, and these may be included also, as in the examples below.

Example 8.18 (Cement data) Table 8.6 gives the analysis of variance when the covariates x_1, x_2, x_3, and x_4 are successively included in the design matrix. There are very large reductions due to fitting x_1 and x_2, but those due to x_3 and x_4 are smaller. The F statistics for testing the effects of x_1 and x_2 are highly significant, but once x_1 and x_2 are included the F statistic for x_3 is not large compared to the $F_{1,8}$ distribution. A similar conclusion holds for x_4.

Table 8.6 Analysis of variance table for the cement data, showing reductions in overall sum of squares when terms are entered in the order given.

Term	df	Reduction in sum of squares	Mean square	F
x_1	1	1450.1	1450.1	242.5
x_2	1	1207.8	1207.8	202.0
x_3	1	9.79	9.79	1.64
x_4	1	0.25	0.25	0.04
Residual	8	47.86	5.98	

Table 8.7 Models for the means of the crossed and self-fertilized plants in the pth pot and jth pair for the maize data.

Terms	Crossed	Self-fertilized
1	μ	μ
1+Fertilization	$\mu + \alpha$	μ
1+Fertilization+Pot	$\mu + \alpha + \beta_p$	$\mu + \beta_p$
1+Fertilization+Pot+Pair	$\mu + \alpha + \beta_p + \gamma_j$	$\mu + \beta_p + \gamma_j$

Thus once x_1 and x_2 are included, x_3 and x_4 are unnecessary in accounting for the response variation. ■

Example 8.19 (Maize data) Consider models for the maize data with means as in Table 8.7. In order, these correspond to: no differences among pairs and no difference between cross-fertilization and self-fertilization; no differences among pairs but an effect of fertilization type; differences among the pots and an effect of fertilization type; and differences among the pots and among the pairs and an effect of fertilization type. Table 8.8 gives the analysis of variance when these models are fitted successively.

There are four pot parameters β_p, but the reduction in degrees of freedom when the pots term is included is three because although the corresponding 30×4 matrix has rank four, its columns sum to a column of ones. As the design matrix already contains a column of ones, including the four columns for the pots term increases the rank of the design matrix by only three. Likewise only 11 columns of the 30×15 matrix of terms for pairs increase the rank of a design matrix that already contains the overall mean and the pots term: the remaining four columns are linear combinations of those already present.

The residual sum of squares for the eventual model is 9972.5 on 14 degrees of freedom, so the denominator for F statistics is $9972.5/14 = 712.3$. The F

Table 8.8 Analysis of variance table for linear models fitted to the maize data.

Term	df	Reduction in sum of squares	Mean square	F
Fertilization	1	3286.5	3286.5	4.61
Pot	3	1053.6	351.2	0.49
Pair	11	4467.3	406.1	0.57
Residual	14	9972.5	712.3	

statistic for fertilization is just significant at the 5% level, but there seem to be no differences among pots or pairs. We can attribute to random variation the reduction in sum of squares when the pots and pairs terms are added, and obtain a better estimate of σ^2, namely

$$(9972.5 + 1053.6 + 4467.3)/(14 + 3 + 11) = 553.3$$

on 28 degrees of freedom. The F statistic for fertilization with this pooled estimate of σ^2 as denominator is 5.94 on 1 and 28 degrees of freedom and its significance level is 0.02, so the addition of the sums of squares for pots and pairs to the residual has resulted in a more sensitive analysis. ■

8.5.3 Orthogonality

The reduction in sum of squares when a term is added depends on the terms already in the model. This can obscure the interpretation of an analysis of variance, if a term that gives a large reduction early in a sequence of fits gives a small reduction if fitted later in the sequence instead.

Suppose that a normal linear model (8.22) applies. The reductions in sum of squares due to the terms X_r are unique only if the vector spaces spanned by the columns of the X_r are all mutually orthogonal, that is, $X_r^{\mathrm{T}} X_s = 0$ when $r \neq s$. Suppose that this is true, that in addition $X_r^{\mathrm{T}} 1_n = 0$, and that

$$y = 1_n\beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon. \tag{8.24}$$

Then the orthogonality of 1_n, X_1, and X_2 implies that the least squares estimators are

$$\begin{pmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} 1^{\mathrm{T}}1 & 0 & 0 \\ 0 & X_1^{\mathrm{T}}X_1 & 0 \\ 0 & 0 & X_2^{\mathrm{T}}X_2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & X_1 & X_2 \end{pmatrix}^{\mathrm{T}} y,$$

so that $\widehat{\beta}_0 = \overline{y}$, $\widehat{\beta}_1 = (X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}y$, and $\widehat{\beta}_2 = (X_2^{\mathrm{T}}X_2)^{-1}X_2^{\mathrm{T}}y$, with residual sum of squares

$$y^{\mathrm{T}}y - \widehat{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\widehat{\beta} = y^{\mathrm{T}}y - n\overline{y}^2 - \widehat{\beta}_1^{\mathrm{T}}X_1^{\mathrm{T}}X_1\widehat{\beta}_1 - \widehat{\beta}_2^{\mathrm{T}}X_2^{\mathrm{T}}X_2\widehat{\beta}_2. \tag{8.25}$$

For the simpler models

$$y = 1_n\beta_0 + \varepsilon, \quad y = 1_n\beta_0 + X_1\beta_1 + \varepsilon \quad y = 1_n\beta_0 + X_2\beta_2 + \varepsilon,$$

a similar calculation gives residual sums of squares

$$y^{\mathrm{T}}y - n\overline{y}^2, \quad y^{\mathrm{T}}y - n\overline{y}^2 - \widehat{\beta}_1^{\mathrm{T}}X_1^{\mathrm{T}}X_1\widehat{\beta}_1, \quad y^{\mathrm{T}}y - n\overline{y}^2 - \widehat{\beta}_2^{\mathrm{T}}X_2^{\mathrm{T}}X_2\widehat{\beta}_2,$$

and comparison with (8.25) shows that the reductions due to X_1 and X_2 are $\widehat{\beta}_1^{\mathrm{T}}X_1^{\mathrm{T}}X_1\widehat{\beta}_1$ and $\widehat{\beta}_2^{\mathrm{T}}X_2^{\mathrm{T}}X_2\widehat{\beta}_2$ whether or not the other has been included in the design matrix. Consequently the reductions in sums of squares due to X_1 and X_2 are unique. This argument readily extends to models with more than two mutually orthogonal terms X_r. In fact (8.24) has three, as we see by writing $1_n = X_0$.

Example 8.20 (Orthogonal polynomials) Consider a normal linear model with design matrix

$$X = (1_n, x_1, x_2, x_3, x_4) = \begin{pmatrix} 1 & -2 & 2 & -1 & 1 \\ 1 & -1 & -1 & 2 & -4 \\ 1 & 0 & -2 & 0 & 6 \\ 1 & 1 & -1 & -2 & -4 \\ 1 & 2 & 2 & 1 & 1 \end{pmatrix},$$

the last four columns of which correspond to linear, quadratic, cubic, and quartic polynomials in a covariate with five values equally spaced one unit apart. The columns of X are mutually orthogonal, and it follows that the reduction due to any of them does not depend on which of the others have already been fitted.

If the values had been equally-spaced but δ units apart, the model would be $y = 1_n\beta_0 + \delta x_1\beta_1 + \cdots + \delta^4 x_4\beta_4 + \varepsilon$, and the orthogonality of the terms would be unaffected. ■

The argument leading to (8.25) rarely applies directly, but it may do so if an overall mean, corresponding to a column of ones in the design matrix, is fitted first. Suppose that the matrices X_1 and X_2 in (8.24) are not mutually orthogonal and are not orthogonal to 1_n, but that we rewrite the model as

$$\begin{aligned} y &= 1_n(\beta_0 + \overline{x}_1^{\mathrm{T}}\beta_1 + \overline{x}_2^{\mathrm{T}}\beta_2) + (X_1 - 1_n\overline{x}_1^{\mathrm{T}})\beta_1 + (X_2 - 1_n\overline{x}_2^{\mathrm{T}})\beta_2 + \varepsilon \\ &= 1_n\gamma_0 + Z_1\beta_1 + Z_2\beta_2 + \varepsilon, \end{aligned}$$

say, where $\overline{x}_1^{\mathrm{T}}$ and $\overline{x}_2^{\mathrm{T}}$ are the averages of the rows of X_1 and X_2. Then Z_1 and Z_2 are centred and $Z_1^{\mathrm{T}}1_n = Z_2^{\mathrm{T}}1_n = 0$. This rearrangement of the model changes the intercept but leaves β_1 and β_2 unaffected. If the original matrices X_1 and X_2 are such that $Z_1^{\mathrm{T}}Z_2 = 0$, we can apply the argument leading to

(8.25) to our new model, to obtain the successive residual sums of squares

$$\begin{aligned} SS_0 &= y^{\mathrm{T}}y - n\overline{y}^2, \\ SS_1 &= y^{\mathrm{T}}y - n\overline{y}^2 - \widehat{\beta}_1^{\mathrm{T}} Z_1^{\mathrm{T}} Z_1 \widehat{\beta}_1, \\ SS_2 &= y^{\mathrm{T}}y - n\overline{y}^2 - \widehat{\beta}_1^{\mathrm{T}} Z_1^{\mathrm{T}} Z_1 \widehat{\beta}_1 - \widehat{\beta}_2^{\mathrm{T}} Z_2^{\mathrm{T}} Z_2 \widehat{\beta}_2, \end{aligned}$$

as the terms Z_1 and Z_2, or equivalently X_1 and X_2, are added to the design matrix. Since Z_1 is defined purely in terms of X_1 and 1_n, and Z_2 is defined purely in terms of X_2 and 1_n, the reduction in sum of squares due to adding X_1 after including the constant column 1_n in the design matrix is the same whether or not X_2 is present. Hence provided the constant is fitted first, the reductions in sum of squares due to X_1 and X_2 are independent of the order in which they are included. This argument extends to models with more than two X_r, provided that the centred matrices Z_r are mutually orthogonal.

Example 8.21 (3×2 layout) In a 3×2 layout with no interaction the observations and their means can be written

$$\begin{array}{cc} y_{11} & y_{12} \\ y_{21} & y_{22}\,, \\ y_{31} & y_{32} \end{array} \qquad \begin{array}{cc} \mu & \mu+\alpha \\ \mu+\delta_1 & \mu+\delta_1+\alpha\,. \\ \mu+\delta_2 & \mu+\delta_2+\alpha \end{array}$$

In terms of the parameter vector $(\mu, \alpha, \delta_2, \delta_3)^{\mathrm{T}}$, the design matrix is

$$X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix},$$

with X_1 the second column of X, and X_2 the third and fourth columns of X. Evidently X_1 and X_2 are not orthogonal and they are not orthogonal to the constant. On the other hand Z_1 and Z_2 in the corresponding centred matrix,

$$\begin{pmatrix} 1 & -\frac{1}{2} & -\frac{1}{3} & -\frac{1}{3} \\ 1 & \frac{1}{2} & -\frac{1}{3} & -\frac{1}{3} \\ 1 & -\frac{1}{2} & \frac{2}{3} & -\frac{1}{3} \\ 1 & \frac{1}{2} & \frac{2}{3} & -\frac{1}{3} \\ 1 & -\frac{1}{2} & -\frac{1}{3} & \frac{2}{3} \\ 1 & \frac{1}{2} & -\frac{1}{3} & \frac{2}{3} \end{pmatrix},$$

are orthogonal to the constant by construction and to each other because the design is balanced: δ_2 and δ_3 each occur equally often with α and without α. This balance has the consequence that provided that μ is fitted first, the reductions in sums of squares due to X_1 and X_2, or equivalently Z_1 and Z_2, are unique. ■

A designed experiment such as Example 8.21 can often be balanced, so that orthogonality is arranged, at least approximately, and the interpretation of its analysis of variance is relatively clear-cut. Even if the terms are not orthogonal, however, it may be possible to order them unambiguously. One example is polynomial dependence of y on x, where terms of increasing degree are added successively. Another example is when some terms represent classifications that are known to affect y but which are of secondary importance, and others correspond to the question of primary interest. For instance, it would be natural to assess the effects of different treatments on the incidence of heart disease after taking into account the effects of classifying variables such as age, sex, and previous medical history.

Exercises 8.5

1 Consider the cement data of Example 8.3, where $n = 13$. The residual sums of squares for all models that include an intercept are given below.

Model	SS	Model	SS	Model	SS
– – – –	2715.8	1 2 – –	57.9	1 2 3 –	48.11
1 – – –	1265.7	1 – 3 –	1227.1	1 2 – 4	47.97
– 2 – –	906.3	1 – – 4	74.8	1 – 3 4	50.84
– – 3 –	1939.4	– 2 3 –	415.4	– 2 3 4	73.81
– – – 4	883.9	– 2 – 4	868.9		
		– – 3 4	175.7	1 2 3 4	47.86

Compute the analysis of variance table when x_4, x_3, x_2, and x_1 are fitted in that order, and test which of them should be included in the model. Are your conclusions the same as in Example 8.18?

2 (a) Let A, B, C, and D represent $p\times p$, $p\times q$, $q\times q$, and $q\times p$ matrices respectively. Show that provided that the necessary inverses exist

$$(A+BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}.$$

(b) If the matrix A is partitioned as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

and the necessary inverses exist, show that the elements of the corresponding partition of A^{-1} are

$$\begin{aligned} A^{11} &= (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1}, \quad A^{22} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}, \\ A^{12} &= -A_{11}^{-1}A_{12}A^{22}, \quad A^{21} = -A_{22}^{-1}A_{21}A^{11}. \end{aligned}$$

3 In (8.20), suppose that X_1 and X_2 have ranks q and $p-q$ respectively, and define $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, $P = I_n - H$, $H_1 = X_1(X_1^{\mathrm{T}}X_1)^{-1}X_1^{\mathrm{T}}$ and $P_1 = I_n - H_1$. Let $\widehat{y} = Hy$, and $\widehat{y}_1 = H_1 y$.

Use the previous exercise.

Table 8.9 Sums of squares for models fitted to maize data.

Model	SS	Model	SS	Model	SS	Model	SS
— — —	18780	— Po —	17726	F Po —	14440	F — Pa	9972
F — —	15493	— — Pa	13259	— Po Pa	13259	F Po Pa	9972

(a) Show that $(y - \widehat{y})^{\mathrm{T}}(\widehat{y} - \widehat{y}_1) = 0$ if and only if $HH_1 = H_1$, and show that $H_1H = HH_1$. Give a geometrical interpretation of the equations $H_1H = HH_1 = H_1$.
(b) Show that

$$(X_1^{\mathrm{T}}P_2X_1)^{-1} = (X_1^{\mathrm{T}}X_1)^{-1} - H_1X_2(X_2^{\mathrm{T}}P_1X_2)^{-1}X_2^{\mathrm{T}}X_1(X_1^{\mathrm{T}}X_1)^{-1}.$$

(c) Show that

$$H = X_1(X_1^{\mathrm{T}}P_2X_1)^{-1}X_1^{\mathrm{T}} - H_1X_2(X_2^{\mathrm{T}}P_1X_2)^{-1}X_2^{\mathrm{T}} + X_2(X_2^{\mathrm{T}}P_1X_2)^{-1}X_2^{\mathrm{T}}P_1.$$

(d) Use (b) and (c) to show that $HH_1 = H_1$.

4 Under what two circumstances might one of the reductions in residual sum of squares $SS_r - SS_{r+1}$ in an analysis of variance table for a normal linear model equal zero? Does the more probable of these occur when the columns of either of the design matrices below are included successively in their models:

$$\text{(a)} \quad \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 \end{pmatrix}, \qquad \text{(b)} \quad \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}?$$

5 Suppose that the maize data consisted of three pots each containing two pairs of plants, 12 plants in all. Using the parametrization in Example 8.19, write out the 12×11 design matrix whose first two columns are terms for the overall mean and for cross-fertilization, whose next three columns are the pots term, and whose last six columns are the pairs term. Say what the degrees of freedom for the four models in Example 8.19 would then be, and hence give the degrees of freedom in the analysis of variance table.

6 The residual sums of squares in Example 8.19 are given in Table 8.9. For which of the terms are the reductions in residual sum of squares independent of the order of fitting? Explain why adding the `Pots` term to a model that already contains the `Pairs` term does not reduce the sum of squares, even if `Fertilization` is not included.

7 Verify that the columns of the design matrix in Example 8.20 are orthogonal. Use Gram–Schmidt orthogonalization to derive the corresponding matrices for two, three, and four observations.

8 Verify that 1_n, Z_1, and Z_2 in Example 8.21 are orthogonal. Show that if one of the rows of the original design matrix is missing, the Z_r are not orthogonal.

8.6 Model Checking

8.6.1 Residuals

Discrepancies between data and a regression model may be isolated or systematic, or both. One type of isolated discrepancy is when there are outliers: a few observations that are unusual relative to the rest. Systematic discrepancies arise, for example, when a transformation of the response or a covariate is needed, when correlated errors are supposed independent, or when a term is incorrectly omitted. There are many techniques for detecting such problems. Graphs are widely used, often supplemented by more formal methods that sharpen their interpretation.

The assumptions underlying the linear regression model (8.1) are:

- *linearity* — the response depends linearly on each explanatory variable and on the error, with no systematic dependence on any omitted terms;
- *constant variance* — the responses have equal variances, which in particular do not depend on the level of the response;
- *independence* — the errors are uncorrelated, and independent if normal; and sometimes
- *normality* — in the normal linear model the errors are normally distributed.

Many graphical methods for checking these assumptions are based on the raw residuals, $e = y - \widehat{y}$. These are estimates of the unobserved errors ε, with mean vector and variance matrix

$$\mathrm{E}(e) = 0, \quad \mathrm{var}(e) = \sigma^2(I_n - H),$$

where H is the hat matrix $X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. The covariance of two different residuals, e_j and e_k, equals $-\sigma^2 h_{jk}$, so in general the residuals are correlated.

A difficulty in direct comparison of the e_j is that their variances, $\sigma^2(1-h_{jj})$, are usually unequal. We therefore construct *standardized residuals*

$$r_j = \frac{e_j}{s(1-h_{jj})^{1/2}} = \frac{y_j - x_j^{\mathrm{T}}\widehat{\beta}}{s(1-h_{jj})^{1/2}}, \tag{8.26}$$

where $x_j^{\mathrm{T}}\widehat{\beta} = \widehat{y}_j$ is the jth fitted value and s^2 is the unbiased estimate of σ^2 based on the model. The r_j have means zero and approximately unit variances, and hence are comparable with standard normal variables.

The simplest check on linearity is to plot the response vector y against each column of the design matrix X. It is also useful to plot the standardized residuals r against each variable, whether or not it has been used in the model. Incorrect form of dependence on an explanatory variable, or omission of one, will show as a pattern in the corresponding plot. More formal techniques designed to detect wholesale nonlinearity are discussed below.

Constancy of variance is usually checked by a plot of the r_j or $|r_j|$ against fitted values. A common failure of this assumption occurs when the error variance increases with the level of the response; this shows as a trumpet-shaped plot. Since the raw residuals e and the fitted values $\widehat{y}$ are uncorrelated, we would expect random scatter if the model fitted adequately. This plot can also help to detect a nonlinear relation between the response and fitted value, as in Example 8.24 below.

Non-independence of the errors can be hard to detect and can have a serious effect on the standard errors of estimates, but serial correlation of time-ordered observations may show up in scatterplots of lagged r_j, or in their correlogram.

Assumptions about the distribution of the errors can be checked by probability plots of the r_j. In particular, normal scores plots are widely used.

Single outliers — maybe due to mistakes in data recording, transcription, or entry — are likely to show up on any of the plots described above, while multiple outliers may lead to *masking* where each outlier is concealed by the presence of others.

Example 8.22 (Cycling data) Figure 8.4 shows plots of the r_j for the model that includes effects of seat height, dynamo and tyre pressure. The top panels show the r_j plotted against the day on which the run took place, and the order of the run within each day. There is slight evidence of dependence on these, but we must beware of spurious patterns when there are only sixteen observations. To check whether these patterns might be genuine, we construct the F statistic for inclusion of factors corresponding to day and run after including seat height, dynamo, and tyre pressure in the model. Its value is 3.99, to be compared to $F_{7,5}(0.95) = 4.88$. Any evidence of differences among days and runs is weak, and we discount it.

The lower left panel of the figure shows residuals plotted against fitted values. There is a slight suggestion that the error variance increases as the fitted value does, but this is mostly due to the largest observation at the right of the plot.

The lower right panel of the figure shows a normal probability plot of the residuals. This is slightly upwardly curved, but not remarkably so in so small a set of data.

Inspection of Table 8.3 shows that the largest residual is for the sixth setup, of which the experimenter writes:

> Its comparison run (setup 5) was only 54 seconds. This is the largest amount of variation in the whole table. I suspect that the correct reading for setup 6 was 55 seconds, that is, I glanced at my watch and thought that it said 60 instead of 55 seconds. Since I am not sure, however, I have not changed it for the analysis. The conclusions would be the same in any case.

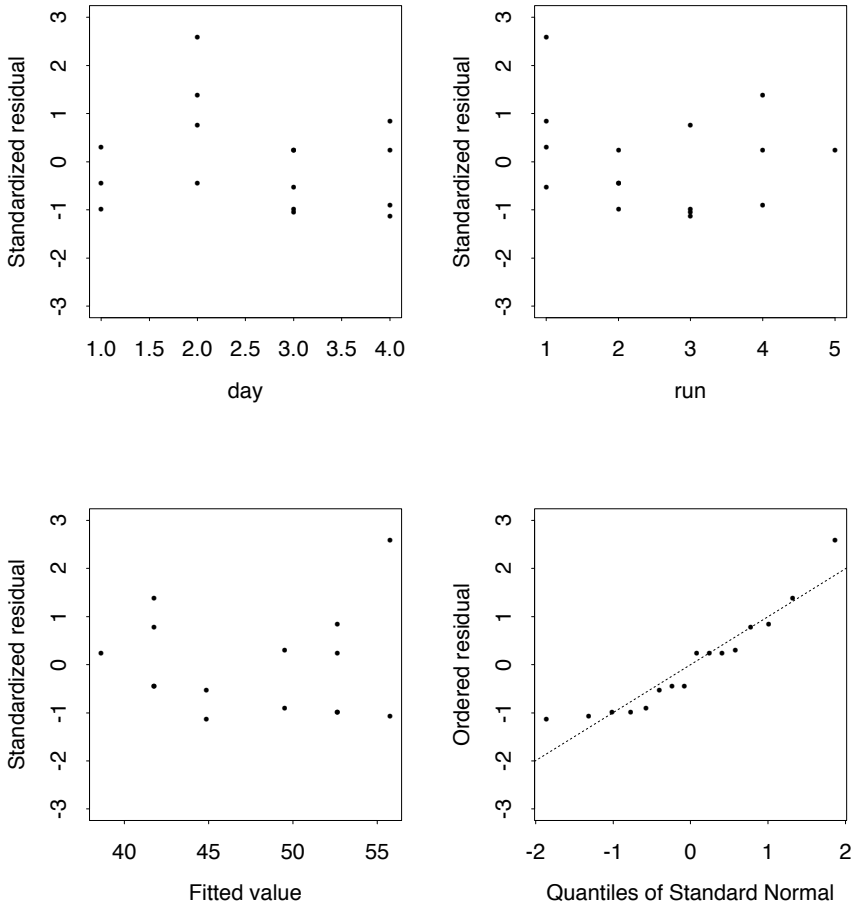


Figure 8.4
Residual plots for data on cycling up a hill. The panels showing residuals plotted against levels of day and run, and against fitted values, would show random variation if the model is adequate, as seems to be the case. The normal scores plot shows that the errors appear close to normal.

One reason that the conclusions would be unchanged is that a well-designed experiment like this is relatively robust to a single bad value.

To sum up: the linear model (8.2) seems to fit these data adequately. ■

8.6.2 Nonlinearity

Linearity is usually a convenient fiction for describing how a response depends on the explanatory variables, and there are many ways it can fail. For example, a linear model may be appropriate for a transformation of the original response, so that $a(y) = x^{\mathrm{T}}\beta + \varepsilon$ for some function $a(\cdot)$; then $y = a^{-1}(x^{\mathrm{T}}\beta + \varepsilon)$ and error is not additive on the original scale. Another possibility is that the response is a nonlinear function of $x^{\mathrm{T}}\beta$ but the error is additive, that is,

$y = b(x^{\mathrm{T}}\beta)+\varepsilon$ for some $b(\cdot)$. More generally we could put $a(y) = b(x^{\mathrm{T}}\beta)+c(\varepsilon)$ for fairly arbitrary functions $a(\cdot)$, $b(\cdot)$ and $c(\cdot)$. Such models can be fitted, but they are beyond our scope.

For a simpler approach, we consider parametric transformation of the response, in which we assume that for some family of transformations $a(\cdot)$ indexed by a parameter λ, there is a transformation such that $a(y) = x^{\mathrm{T}}\beta + \varepsilon$. In principle we might consider many possible transformations, but practical experience suggests that power and logarithmic transformations are among the most fruitful. The following example gives a general approach.

Example 8.23 (Box–Cox transformation) Suppose that a normal linear model applies not to y, but to

$$y^{(\lambda)} = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log y, & \lambda = 0. \end{cases}$$

As λ varies in the range $(-2, 2)$ this encompasses the inverse transformation ($\lambda = -1$), log ($\lambda = 0$), cube and square roots ($\lambda = \frac{1}{3}, \frac{1}{2}$), and the original scale ($\lambda = 1$), as well as the square transformation ($\lambda = 2$). We assume below that all the y_j are positive. If not, the transformation must be applied to $y_j + \xi$, with ξ chosen large enough to make all the $y_j + \xi$ positive.

Suggested by Box and Cox (1964). George E. P. Box (1919–) was educated at London University and has held posts in industry and at Princeton and the University of Wisconsin. He has made important contributions to robust and Bayesian statistics, experimental design, time series, and to industrial statistics. Sir David Roxbee Cox (1924–) was born in Birmingham and educated in Cambridge and Leeds. He has held posts at Imperial College London, Cambridge, and Oxford where he nows works. He has made highly influential contributions across the whole of statistical theory and methods. See DeGroot (1987a) and Reid (1994).

Now let $y^{(\lambda)}$ denote the $n \times 1$ vector of transformed responses, and assume that a normal linear model

$$y^{(\lambda)} = X\beta + \varepsilon$$

applies for some values of λ, β, and error variance σ^2. We assume that the design matrix contains a column of ones, so that using $y^{(\lambda)}$ rather than y^λ leaves the fit unchanged; it merely changes the intercept and rescales β.

To obtain the likelihood for β, σ^2, and λ, note that on taking into account the Jacobian of the transformation from $y^{(\lambda)}$ to y, the density of y_j is

$$f(y_j; \beta, \sigma^2, \lambda) = \frac{y_j^{\lambda-1}}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(y_j^{(\lambda)} - x_j^{\mathrm{T}}\beta)^2\right\}.$$

Consequently the log likelihood based on independent $y_1, \ldots, y_n$ is

$$\ell(\beta, \sigma^2, \lambda) \equiv -\tfrac{1}{2}\left\{n \log \sigma^2 + \frac{1}{\sigma^2}\sum_{j=1}^{n}(y_j^{(\lambda)} - x_j^{\mathrm{T}}\beta)^2\right\} + (\lambda - 1)\sum_{j=1}^{n} \log y_j.$$

If λ is regarded as fixed, the maximum likelihood estimates of β and σ^2 are $\widehat{\beta}_\lambda = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y^{(\lambda)}$ and $SS(\widehat{\beta}_\lambda)/n$, where $SS(\widehat{\beta}_\lambda)$ is the residual sum of squares for the regression of $y^{(\lambda)}$ on the columns of X. Thus the profile log likelihood for λ is

$$\ell_{\mathrm{p}}(\lambda) = \max_{\beta,\sigma^2} \ell(\beta, \sigma^2, \lambda) \equiv -\frac{n}{2}\left\{\log SS(\widehat{\beta}_\lambda) - \log g^{2(\lambda-1)}\right\},$$

where $g = (\prod y_j)^{1/n}$ is the geometric average of $y_1, \ldots, y_n$. Equivalently $\ell_{\mathrm{p}}(\lambda)$ equals $-\frac{1}{2}n \log SS_g(\widehat{\beta}_\lambda)$, where $SS_g(\widehat{\beta}_\lambda)$ is the residual sum of squares for the regression of $y^{(\lambda)}/g$ on the columns of X. Exercise 8.6.3 invites you to provide the details.

A plot of the profile log likelihood $\ell_{\mathrm{p}}(\lambda)$ summarizes the information concerning λ; a $(1-2\alpha)$ confidence interval is the set for which $\ell_{\mathrm{p}}(\lambda) \geq \ell_{\mathrm{p}}(\widehat{\lambda}) - \frac{1}{2}c_1(1-2\alpha)$. The exact maximum likelihood estimate of λ is rarely used, since a nearby value is usually more easily interpreted. ■

$c_\nu(\alpha)$ is the α quantile of the χ^2_ν distribution.

A different approach is to consider whether the model $y = b(x^{\mathrm{T}}\beta) + \varepsilon$ might apply. This cannot be linearized by a response transformation and if there is evidence that $b(\cdot)$ is substantially nonlinear but the variance is constant it may be necessary to fit a nonlinear normal model. The following example gives one method for detecting this sort of nonlinearity.

Example 8.24 (Non-additivity) Suppose that it is feared that $y = b(x^{\mathrm{T}}\beta) + \varepsilon$, where $b(\cdot)$ is a smooth nonlinear function. Taylor series expansion of $b(\cdot)$ about a typical value of $x^{\mathrm{T}}\beta$, η, say, gives

$$y \doteq b(\eta) + b'(\eta)(x^{\mathrm{T}}\beta - \eta) + \tfrac{1}{2}b''(\eta)(x^{\mathrm{T}}\beta - \eta)^2 + \varepsilon.$$

If the model contains a constant, so that $x^{\mathrm{T}}\beta = \beta_0 + x_1\beta_1 + \cdots$, then $y \doteq x^{\mathrm{T}}\gamma + \delta(x^{\mathrm{T}}\gamma)^2 + \varepsilon$, where γ is just a reparametrization of β, and $\delta \propto b''(\eta)$. A large value of δ corresponds to strong nonlinear dependence of y on $x^{\mathrm{T}}\beta$.

Let us fit the model $y = X\beta + \varepsilon$, giving fitted values $x_j^{\mathrm{T}}\widehat{\beta}$ and residual sum of squares $SS(\widehat{\beta})$. Then as $y - x^{\mathrm{T}}\gamma \doteq \delta(x^{\mathrm{T}}\gamma)^2 + \varepsilon$, non-additivity should show up as curvature in a plot of standardized residuals against fitted values.

A formal test for non-zero δ is based on refitting the model with the column $(x_j^{\mathrm{T}}\widehat{\beta})^2$ added to the design matrix. Although the residual sum of squares for this model, SS_δ, depends upon the fitted values for the previous fit, the F statistic for inclusion of $(x_j^{\mathrm{T}}\widehat{\beta})^2$,

$$\frac{SS(\widehat{\beta}) - SS_\delta}{SS_\delta/(n-p-1)}, \tag{8.27}$$

See Tukey (1949).

has an $F_{1,n-p-1}$ distribution; this is known as *Tukey's one degree of freedom for non-additivity.* ■

Covariates that are artificially created to help assess model fit, such as $(x_j^{\mathrm{T}}\widehat{\beta})^2$ in Example 8.24, are known as *constructed variables.*

Example 8.25 (Poisons data) Table 8.10 contains data from a completely randomized experiment on the survival times of 48 animals. The animals were divided at random into groups of size four, and then each group was given one of three poisons and one of four treatments. Thus there are two factors, one with three and the other with four levels. The lower part of Table 8.10 and the

Table 8.10 Poison data (Box and Cox, 1964). Survival times in 10-hour units of animals in a 3×4 factorial experiment with four replicates. The table underneath gives average (standard deviation) for the poison × treatment combinations.

Treatment	Poison 1	Poison 2	Poison 3
A	0.31, 0.45, 0.46, 0.43	0.36, 0.29, 0.40, 0.23	0.22, 0.21, 0.18, 0.23
B	0.82, 1.10, 0.88, 0.72	0.92, 0.61, 0.49, 1.24	0.30, 0.37, 0.38, 0.29
C	0.43, 0.45, 0.63, 0.76	0.44, 0.35, 0.31, 0.40	0.23, 0.25, 0.24, 0.22
D	0.45, 0.71, 0.66, 0.62	0.56, 1.02, 0.71, 0.38	0.30, 0.36, 0.31, 0.33

Treatment	Poison 1	Poison 2	Poison 3	Average
A	0.41 (0.07)	0.32 (0.08)	0.21 (0.02)	0.31
B	0.88 (0.16)	0.82 (0.34)	0.34 (0.05)	0.68
C	0.57 (0.16)	0.38 (0.06)	0.24 (0.01)	0.39
D	0.61 (0.11)	0.67 (0.27)	0.33 (0.03)	0.53
Average	0.62	0.55	0.28	0.48

upper panels of Figure 8.5 both show strong effects of treatment and poison: poison 3 is most potent, and treatments B and D are more efficacious than A and C. There is also evidence that the response variance depends on the mean: the standard deviations are smaller for poison × treatment combinations with smaller average response.

One model for these data is

$$y_{tpj} = \mu + \alpha_t + \beta_p + \varepsilon_{tpj}, \quad t = 1, 2, 3, 4,\ p = 1, 2, 3,\ j = 1, 2, 3, 4. \quad (8.28)$$

Here μ represents a baseline average response in the absence of treatments or poisons, α_t represents the effect of the tth treatment, β_p the effect of the pth poison and ε_{tpj} is the unobserved error for the jth replicate given the tth treatment and pth poison. We assess the fit of (8.28) initially through the plot of standardized residuals against fitted values in the upper left panel of Figure 8.6, which shows a striking increase of error variance with the mean response. The model underpredicts for the lowest responses, where $r_j > 0$ and therefore $y_j > \widehat{y}_j$, and overpredicts for the middle responses, where the residuals are mostly negative. Following Example 8.24, this suggests that the poison and treatment effects are not additive. The neighbouring panel shows that the errors are somewhat positively skewed relative to the normal distribution. The model fits the data poorly, not owing to a few bad observations, but in a systematic way, as was also suggested by the lower left panel of Figure 8.5.

Ignoring for a moment the nonconstancy of variance, we explore whether

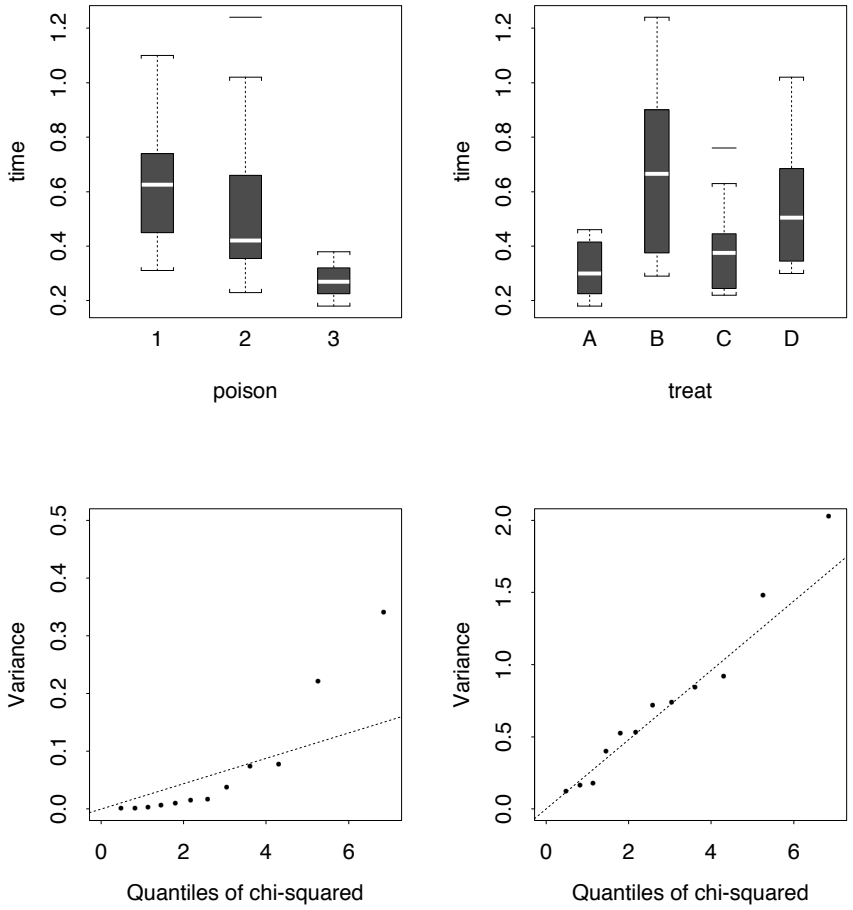


Figure 8.5 Poison data. The upper panels show how the responses depend on the factor levels. The lower left panel shows a χ^2_3 probability plots of the $3s^2_{pt}$, where s^2_{pt} is the sample variance of the four replicates y_{ptj} given the pth poison and tth treatment. The lower right panel shows the same plot for the y^{-1}_{ptj}.

the explanatory variables act additively. The F statistic for non-additivity, (8.27), equals 14.03. This is large compared with the 0.95 quantile of the $F_{1,41}$ distribution and gives strong evidence of non-additivity.

The lower right panel of Figure 8.6 shows the profile log likelihood for the transformation parameter, λ. There is strong evidence that the original scale ($\lambda = 1$) is poor; log transformation ($\lambda = 0$) also seems inappropriate. The most readily interpretable value of λ in the 95% confidence interval seems to be -1, corresponding to fitting a linear model to the inverse response $1/y$. This can be interpreted in terms of the rate of dying, whose units are time^{-1}. The lower left panel of the figure suggests that the evidence for non-additivity has gone, and that the inverse transformation has roughly equalized the error

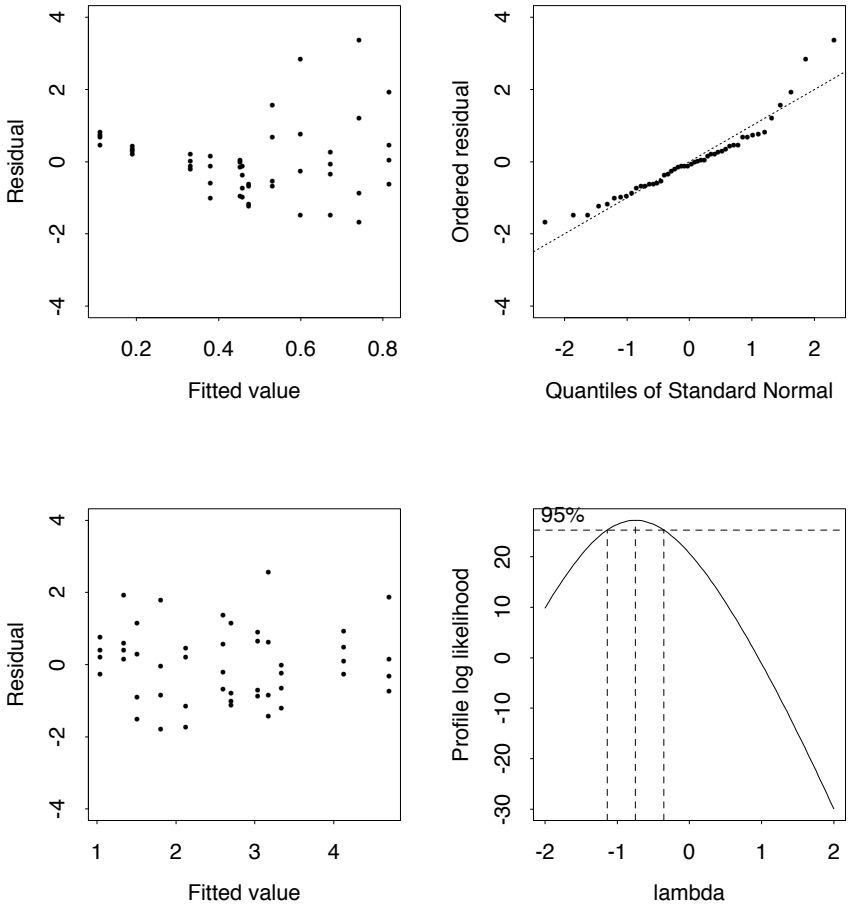


Figure 8.6
Diagnostic plots for the two-way layout model for the poisons data. The upper left panel a plot of standardized residuals for the fit of the two-way layout model to the original data against the fitted value, while its neighbour shows the normal probability plot of these residuals. The lower right panel shows the profile log likelihood for the Box–Cox parameter λ and suggests that a linear model should be fitted to the inverse response, $1/y$. The lower left panel shows the residuals for the two-way layout model with response $1/y$ plotted against its fitted values; this does not display the non-linearity and systematic increase of variance of the panel above.

variances. A probability plot shows that the residuals on this scale are close to normal.

To sum up, the model $y^{-1} = \mu + \alpha_t + \beta_p + \varepsilon_{tpj}$ seems to fit the data adequately, and has a direct interpretation as a linear model for the effect of poisons and treatments on the speed of dying.

We return to these data in Examples 9.6 and 9.8. ■

8.6.3 Leverage, influence, and case deletion

We call the explanatory and response variables (x_j, y_j) the jth *case*. We have already seen how an odd y_j can arise, but there can also be effects due to unusual explanatory variables. To see how, recall that $\mathrm{var}(y_j - x_j^{\mathrm{T}}\widehat{\beta}) = \sigma^2(1 -$

h_{jj}), and notice that if h_{jj} is close to one the jth fitted value must lie very close to y_j itself. Indeed, if $h_{jj} = 1$, the model is constrained so that $x_j^{\mathrm{T}}\widehat{\beta} = y_j$. This is undesirable because in effect a degree of freedom, the equivalent of one parameter, is used to fit one response value exactly. The effect on $\widehat{\beta}$ could be catastrophic if y_j were outlying.

The quantity h_{jj} is called the *leverage* of the jth case. Other things being equal, the argument above suggests that low leverage is good. But $\mathrm{tr}(H) = \sum h_{jj} = p$ (Exercise 8.2.5), so the average leverage cannot be reduced below p/n. Approximate equalization of leverage is one attribute of good design. In the factorial experiment in Table 8.3, for example, $h_{jj} = \frac{1}{4}$ for each case. A general guideline is that cases for which $h_{jj} > 2p/n$ deserve closer inspection; it may be worthwhile to repeat an analysis without them in order to assess their effect on both the values and the precision of the estimates. In itself, however, high leverage is not sufficient reason to delete a case, which if not outlying may be very informative.

Example 8.26 (Straight-line regression) The matrix formulation of

$$y_j = \gamma_0 + (x_j - \overline{x})\gamma_1 + \varepsilon_j, \quad j = 1, \ldots, n,$$

is given in Example 8.6, and it is easily deduced that the jth leverage is

$$h_{jj} = \frac{1}{n} + \frac{(x_j - \overline{x})^2}{\sum_k (x_k - \overline{x})^2}.$$

When the constant is dropped the leverage is $(x_j - \overline{x})^2 / \sum_k (x_k - \overline{x})^2$, and when the covariate x_j is dropped the leverage is n^{-1}. Thus h_{jj} can be interpreted as a sum of contributions for each parameter. As the contribution corresponding to γ_1 is quadratic in $x_j - \overline{x}$, responses with large values of $|x_j - \overline{x}|$ will strongly affect the slope of the fitted line. All the responses have equal weight in estimating the intercept. These effects do not depend on the response values and depend purely on the design matrix. ■

Having seen that an individual case may substantially affect least squares estimates, it is natural to ask how to measure this. One overall *influence measure* for the jth case is *Cook's distance*, defined as

After Cook (1977). R. Dennis Cook is a professor of statistics at the University of Minnesota.

$$C_j = \frac{1}{ps^2}(\widehat{y} - \widehat{y}_{-j})^{\mathrm{T}}(\widehat{y} - \widehat{y}_{-j}),$$

where $\widehat{y}_{-j} = X\widehat{\beta}_{-j}$, and subscript $-j$ denotes a quantity calculated with the jth case deleted from the model. Cook's distance measures the overall change in the fitted values when the jth case is deleted from the model, standardized by the dimension of β and the estimate of σ^2. It can be revealing to refit a model without the cases whose values of C_j are largest.

To gain some insight into C_j, note that the least squares estimate of β calculated without the jth case is

$$\widehat{\beta}_{-j} = (X^{\mathrm{T}}X - x_j x_j^{\mathrm{T}})^{-1}(X^{\mathrm{T}}y - x_j y_j).$$

Some linear algebra shows that

$$\widehat{\beta}_{-j} = \widehat{\beta} - (X^{\mathrm{T}}X)^{-1}x_j \frac{y_j - \widehat{y}_j}{1 - h_{jj}}, \tag{8.29}$$

and it follows that (Exercise 8.6.5)

$$C_j = \frac{r_j^2 h_{jj}}{p(1 - h_{jj})}, \tag{8.30}$$

where r_j is the standardized residual. Therefore large values of C_j arise if a case has high leverage or a large standardized residual, or both. A plot of C_j against $h_{jj}/(1 - h_{jj})$ helps to distinguish between these possibilities. A crude rule is that as a residual with $|r_j| > 2$ or a case with leverage $h_{jj} > 2p/n$ deserve attention, a value of C_j greater than $8/(n - 2p)$ is worth a closer look. It is possible for the model to depend on a case whose Cook's distance is zero (Exercise 8.6.6), however, and there is no substitute for careful inspection of the data, residuals, and leverages.

As an observation with a large standardized residual can have a big effect on a fitted model, it is natural to ask whether an outlier is more easily detected by comparing y_j with its predicted value based on the other observations, $x_j^{\mathrm{T}}\widehat{\beta}_{-j}$. After all, if the model is correct and y_j is not an outlier, we expect that $\mathrm{E}(\widehat{\beta}) = \mathrm{E}(\widehat{\beta}_{-j}) = x_j^{\mathrm{T}}\beta$, although of course $\widehat{\beta}_{-j}$ will be a less precise estimate of β than $\widehat{\beta}$. On the other hand, an outlying response y_j does not affect $x_j^{\mathrm{T}}\widehat{\beta}_{-j}$, so any discrepancy between them should be more obvious. There is a close connection to the idea of cross-validation. Now (8.29) implies that

$$y_k - x_k^{\mathrm{T}}\widehat{\beta}_{-j} = y_k - \widehat{y}_k + x_k^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_j \frac{y_j - \widehat{y}_j}{1 - h_{jj}},$$

and since $x_k^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_j = h_{jk}$, we find that $\mathrm{var}(y_j - x_j^{\mathrm{T}}\widehat{\beta}_{-j}) = \sigma^2/(1 - h_{jj})$. This suggests that *deletion residuals* be defined as

$$r'_j = \frac{y_j - x_j^{\mathrm{T}}\widehat{\beta}_{-j}}{\mathrm{var}(y_j - x_j^{\mathrm{T}}\widehat{\beta}_{-j})^{1/2}} = \frac{y_j - \widehat{y}_{-j,j}}{s_{-j}(1 - h_{jj})^{1/2}},$$

where $\widehat{y}_{-j,j}$ is the jth element of the vector $\widehat{y}_{-j}$ and the estimate of σ^2 based on the data with the jth case deleted equals

$$s_{-j}^2 = \frac{1}{n - 1 - p}\left[(y - \widehat{y}_{-j})^{\mathrm{T}}(y - \widehat{y}_{-j}) - \left\{y_j - \widehat{y}_j + \frac{h_{jj}(y_j - \widehat{y}_j)}{1 - h_{jj}}\right\}^2\right].$$

Table 8.11
Simulated data and case diagnostics.

Case	x_1	x_2	y	$\widehat{y}$	r	r'	h	C
1	0.02	–6.31	0.95	0.41	1.16	1.20	0.88	3.28
2	0.36	0.39	0.44	0.53	–0.08	–0.07	0.13	0.00
3	7.12	–0.64	0.27	0.38	–0.14	–0.13	0.68	0.01
4	–1.54	1.13	0.09	0.59	–0.45	–0.42	0.29	0.03
5	0.24	–1.90	–0.82	0.49	–1.07	–1.08	0.15	0.07
6	0.26	–0.06	0.03	0.53	–0.40	–0.37	0.12	0.01
7	–0.16	0.13	–0.22	0.54	–0.61	–0.59	0.14	0.02
8	0.43	0.80	0.13	0.54	–0.33	–0.31	0.15	0.01
9	–0.02	0.59	3.57	0.55	2.47	6.31	0.15	0.37
10	4.58	0.29	0.57	0.45	0.11	0.10	0.31	0.00

Yet more algebra shows that the deletion residual can be expressed as

$$r'_j = \left(\frac{n-p-1}{n-p-r_j^2} \right)^{1/2} r_j,$$

which is a monotonic function of r_j that exaggerates values for which $|r_j| > 1$. As their derivation suggests, deletion residuals for outlying observations are more prominent than are the corresponding r_j.

Example 8.27 (Cycling data) Table 8.3 gives standardized residuals, deletion residuals, and measures of leverage and influence for the model with an intercept and three main effects fitted to these data. The design is balanced, and since $(X^{\mathrm{T}}X)^{-1} = \frac{1}{16}I_4$, all the leverages equal $\frac{1}{4}$; consequently the standardized residuals are a simple multiple of the raw residuals. As remarked in Example 8.22, the only unusual residual is for setup 6, whose deletion residual is strikingly large: there is strong evidence that this is an outlier. The corresponding Cook statistic, 0.56, is by far the largest, but it is unremarkable relative to $8/(n-2p) = 1$. The belt-and-braces statistician might repeat the analysis without this datum, but it makes little difference. ■

Exercises 8.6

1 Show that the standardized residuals r_j have means zero and variances $(n-p)/(n-p-2)$. What can you say about their joint distribution?

2 Table 8.11 shows simulated data on the dependence of $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$ on covariates x_1 and x_2. The residual sum of squares was 12.43.
(a) Choose a case and check the relationships between $\widehat{y}$, r, r', h, and C.
(b) Discuss the fit. If it is not adequate, explain what further steps you would take in analyzing the data.

3 Provide the details for Example 8.23.

4 Compute and interpret the leverages for Examples 8.9 and 8.20.

5 Use Exercise 8.5.2(a) with $C = -1$ to show that

$$(X^{\mathrm{T}}X - x_j x_j^{\mathrm{T}})^{-1} = (X^{\mathrm{T}}X)^{-1} + (1 - h_{jj})^{-1}(X^{\mathrm{T}}X)^{-1}x_j x_j^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1};$$

it may help to note that $h_{jj} = x_j^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}x_j$. Hence show that

$$\widehat{\beta}_{-j} = (X^{\mathrm{T}}X - x_j x_j^{\mathrm{T}})^{-1}(X^{\mathrm{T}}y - x_j y_j) = \widehat{\beta} - (1 - h_{jj})^{-1}(X^{\mathrm{T}}X)^{-1}x_j(y_j - \widehat{y}_j),$$

deduce that $\widehat{y} - \widehat{y}_{-j} = (1 - h_{jj})^{-1}X(X^{\mathrm{T}}X)^{-1}x_j(y_j - \widehat{y}_j)$, and finally that

$$C_j = \frac{(\widehat{y} - \widehat{y}_{-j})^{\mathrm{T}}(\widehat{y} - \widehat{y}_{-j})}{ps^2} = \frac{r_j^2 h_{jj}}{p(1 - h_{jj})}.$$

6 Suppose that the straight-line regression model $y = \beta_0 + \beta_1 x + \varepsilon$ is fitted to data in which $x_1 = \cdots = x_{n-1} = -a$ and $x_n = (n-1)a$, for some positive a. Show that although y_n completely determines the estimate of β_1, $C_n = 0$. Is Cook's distance an effective measure of influence in this situation?

8.7 Model Building

8.7.1 General

Once the context for a regression problem is known and the data have been scrutinized for outliers, missing values, and so forth, a model must be built. Related investigations will often suggest a form for it, the main initial questions concerning the choice of response and explanatory variables.

The purpose of the analysis determines one or perhaps more responses, which may combine several of the original variables. Once it is chosen, questions arise about whether individual responses are correlated, and if their variance is constant. If not, it may be necessary to use weighted or generalized least squares (Section 8.2.4), or to consider transformations. These may also be suggested by constraints, for example that the response is positive, but it is then also good to consider more general classes of models discussed in Chapter 10.

Scatterplots of the response against potential explanatory variables and of these variables against each another are needed to screen out bad data, to suggest which covariates are likely to be important, and perhaps also to indicate suitable transformations. Dimensional considerations or subject-matter arguments, for example that certain regression coefficients should be positive, may suggest fruitful combinations of covariates or particular relations between them and the response.

It may be clear that the response depends on a few variables, and that possible models can be fitted and compared using F and related tests. Once some suitable models have been found, the techniques of model checking outlined in Section 8.6 can be applied. Often unexpected discrepancies between

a fitted model and data will lead to further thought, and then to more cycles of model-fitting, checking, and interpretation, iterated until a broadly satisfactory model has been found.

If p is much larger than n, then the design matrix must be cut down to size. One possibility is to use *principal components regression*. The basis of this is the *spectral decomposition*, which enables us to write $X^{\mathrm{T}}X = UDU^{\mathrm{T}}$, where D is the diagonal matrix $\mathrm{diag}(d_1, \ldots, d_p)$ containing the ordered eigenvalues $d_p \geq \cdots \geq d_1 \geq 0$ of $X^{\mathrm{T}}X$, and the columns of U are the corresponding eigenvectors. The matrix U can be chosen so that $UU^{\mathrm{T}} = U^{\mathrm{T}}U = I$. The idea is to form the design matrix from the columns of $Z = XU$, which are called *principal components*. The first principal component, z_1, is the linear combination $z = Xu$ of the columns of X for which $z^{\mathrm{T}}z$ is largest, the next, z_2, is the linear combination that maximizes $z_2^{\mathrm{T}}z_2$ subject to $z_1^{\mathrm{T}}z_2 = 0$, the third, z_3, maximizes $z_3^{\mathrm{T}}z_3$ subject to $z_1^{\mathrm{T}}z_2 = z_1^{\mathrm{T}}z_2 = 0$, and so forth. The hope is that much of the dependence of the response on the columns of X will be concentrated in these first few z_rs, in which case a good low-dimensional regression model may be obtainable. Sometimes it is useful to centre the columns of X by subtracting their averages, or to scale them by dividing centred columns by their standard deviations. The resulting principal components do not equal those for X.

Principal components and corresponding parameter estimates may be uninterpretable in terms of the original covariates, though this drawback is less critical when the goal of analysis is prediction.

8.7.2 Collinearity

If there is a nonzero vector c such that $Xc = 0$, the columns of the design matrix are said to be *collinear*. Then X has rank less than p and $X^{\mathrm{T}}X$ has no unique inverse. The simplest example of this arises in straight-line regression: if all the x_j are equal, it is impossible to find unique parameter estimates (Example 8.6). This difficulty arises more generally, because linear dependence among the columns of the design matrix means that some combinations of parameters cannot be estimated from the data; collinearity leads to indeterminable estimates with infinite variances. Related difficulties arise if the columns of X are almost collinear.

The matrix $X^{\mathrm{T}}X$ is invertible only if all its eigenvalues $d_p \geq \cdots \geq d_1 \geq 0$ are positive. Even if $X^{\mathrm{T}}X$ is invertible, however, the estimators can be very poor. The squared distance between $\widehat{\beta}$ and β is expressible as

$$(\widehat{\beta} - \beta)^{\mathrm{T}}(\widehat{\beta} - \beta) \stackrel{D}{=} \sigma^2 \sum_{r=1}^{p} Z_r^2/d_r, \quad \text{where} \quad Z_1, \ldots, Z_p \stackrel{\text{iid}}{\sim} N(0,1).$$

Thus $(\widehat{\beta}-\beta)^{\mathrm{T}}(\widehat{\beta}-\beta)$ has mean and variance

$$\sigma^2 \sum_{r=1}^{p} d_r^{-1}, \quad 2\sigma^4 \sum_{r=1}^{p} d_r^{-2},$$

bounded below respectively by σ^2/d_1 and $2\sigma^4/d_1^2$, and $\widehat{\beta}$ may be far distant from β for small d_1. The practical implication is that parameter estimates from different but related datasets may vary greatly, giving apparently contradictory interpretations of the same phenomenon.

Diagnostics to warn of collinearity can be based on functions of the d_r such as the *condition number* $(d_p/d_1)^{1/2}$, but its statistical interpretation is not clear-cut. The condition number is sometimes reduced by replacing X with the matrix obtained on dropping the column of ones if any and centering the remaining columns, or by using the corresponding correlation matrix.

The most straightforward solution to collinearity or near collinearity is to drop columns from the design matrix until the estimates are better behaved.

A more systematic approach to dealing with weak design matrices is *ridge regression*, which starts by rewriting the original model $y = 1\beta_0 + X_1\beta_1 + \varepsilon$ as $y = 1\beta_0 + Z\gamma + \varepsilon$, where $Z^{\mathrm{T}}1 = 0$ and the diagonal of $Z^{\mathrm{T}}Z$ consists of ns. This involves centring each column of X_1 by subtracting its average, then dividing by its standard deviation, and multiplying by $n^{1/2}$. This centring and rescaling ensures that the elements of γ and of β have the same interpretations apart from a change of scale, unlike with principal components regression. Then the least squares estimates are $\widehat{\beta}_0 = \overline{y}$ and $\widehat{\gamma} = (Z^{\mathrm{T}}Z)^{-1}Z^{\mathrm{T}}y$. The idea is to replace $Z^{\mathrm{T}}Z$ by $Z^{\mathrm{T}}Z + \lambda I_{p-1}$, where $\lambda \geq 0$ is called the *ridge parameter*. The corresponding estimates, $\widehat{\gamma}_\lambda = (Z^{\mathrm{T}}Z + \lambda I_{p-1})^{-1}Z^{\mathrm{T}}y$, are biased unless $\lambda = 0$, when they are the least squares estimates of γ. Large values of λ increase the bias by shrinking the estimates towards the origin, but this decreases their variance. The value of λ is chosen empirically by minimization of a criterion such as the *cross-validation sum of squares*

$$\mathrm{CV}(\lambda) = \sum_{j=1}^{n} (y_j - \widehat{y_j^{-}})^2,$$

where $\widehat{y_j^{-}}$ is the fitted value for y_j predicted from the ridge regression model obtained when the jth case is deleted. Cross-validation, introduced in Section 7.1.2, is here used to assess how well the ridge regression fit would predict a new set of independent data like the original observations. A variant approach chooses λ to minimize the *generalized cross-validation sum of squares*,

$$\mathrm{GCV}(\lambda) = \sum_{j=1}^{n} \frac{(y_j - \widehat{y}_j)^2}{\{1 - \mathrm{tr}(H_\lambda)/n\}^2},$$

Table 8.12 Parameter estimates and their standard errors for the full model and a reduced model fitted to the cement data.

Parameter	Full model		Reduced model	
	Estimate	Standard error	Estimate	Standard error
β_0	62.41	70.07	71.64	14.14
β_1	1.55	0.74	1.45	0.12
β_2	0.51	0.72	0.42	0.19
β_3	0.10	0.75		
β_4	–0.14	0.71	–0.24	0.17

where $H_\lambda = n^{-1}1_n 1_n^{\mathrm{T}} + Z(Z^{\mathrm{T}}Z + \lambda I_{p-1})^{-1}Z^{\mathrm{T}}$ is the hat matrix corresponding to the ridge regression, and the vector of fitted values $\widehat{y} = H_\lambda y$ depends on λ. We discuss these in more detail on page 585, though in another context.

Estimates such as $\widehat{\gamma}_\lambda$ that shrink towards a common value, here $\gamma = 0$, may also be derived by Bayesian arguments (Chapter 11).

Example 8.28 (Cement data) The astute reader will have realized that if the middle four columns of Table 8.1 are percentages, they may sum to 100. In fact they sum to $(99, 97, 95, 97, 98, 97, 97, 98, 96, 98, 98, 98, 98)$. As there is a column of ones in the design matrix for the full model, its columns are nearly dependent: estimation of five parameters is almost impossible. This is reflected by the standard errors in Table 8.12. The standard error for $\widehat{\beta}_0$ is vastly inflated by inclusion of x_3 because β_0 is almost impossible to estimate, whereas the other estimates are less badly affected.

The residual sum of squares for model without x_3 is 47.97, only slightly larger than that for the full model, 47.86. Thus inclusion of x_3 changes the fit of the model very little, but has a drastic effect on the precision of parameter estimation.

The eigenvalues of $X^{\mathrm{T}}X$ with all five columns of X are 44676, 5965.4, 810.0, 105.4 and 0.00012. The condition number of 6056 indicates strong ill-conditioning, and $\sum d_r^{-1} = 821$ seems very large.

The left panel of Figure 8.7 shows how the parameter estimates $\widehat{\gamma}_\lambda$ depend on the ridge parameter λ. All change fairly sharply as λ increases from zero, and are more stable for $\lambda > 0.2$. The right panel shows that GCV(λ) decreases sharply when λ increases from zero, and is minimized when $\lambda \doteq 0.3$. The dotted lines show that when x_3 is dropped both the $\widehat{\gamma}_\lambda$ and GCV(λ) depend much less on λ, consistent with the discussion above. ■

8.7.3 Automatic variable selection

The screening and selection of many explanatory variables may be onerous. With p covariates, each to be included or not, at least 2^p possible design ma-

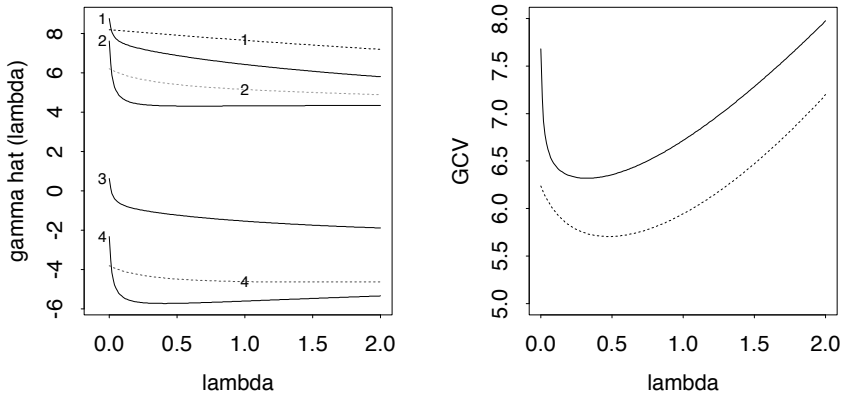


Figure 8.7 Ridge regression analysis of cement data. Left: variation of elements of $\widehat{\gamma}_\lambda$ as a function of λ, for models with all four covariates (solid) and with x_1, x_2, and x_4 only (dots). Right: generalized cross-validation criterion GCV(λ) for these models.

trices must be fitted even before accounting for transformations, combinations of covariates, and so forth. Consequently automatic procedures for variable selection are widely used if p is large. While valuable as screening procedures, they are no substitute for careful model-building incorporating knowledge of the system under study and should be treated as a backstop; their output should always be considered critically.

Stepwise methods

Forward selection takes as baseline the model with an intercept only. Each term is added separately to this, and the base model for the next stage is taken to be the model with the intercept and the term that most reduces the sum of squares. Each of the remaining terms is added to the new base model, and the process continued, stopping if at any stage the F statistic for the largest reduction in sum of squares is not significant or if the design matrix is rank deficient.

Backward elimination starts from the model containing all terms, and then successively drops the least significant term at each stage. It stops when no term can be deleted without increasing the sum of squares significantly.

Backward elimination is generally the preferable of the two because its initial estimate of σ^2 will usually be better than that for forward selection, though at the possible expense of an unstable initial model. They may yield different final models.

In *stepwise regression* four options are considered at each stage: add a term, delete a term, swap a term in the model for one not in the model, or stop. This algorithm is often used in practice.

These three procedures have been shown to fit complicated models to completely random data, and although widely used they have no theoretical basis.

Table 8.13 Data on light water reactors (LWR) constructed in the USA (Cox and Snell, 1981, p. 81). The covariates are `date` (date construction permit issued), `T1` (time between application for and issue of permit), `T2` (time between issue of operating license and construction permit), `capacity` (power plant capacity in MWe), `PR` (=1 if LWR already present on site), `NE` (=1 if constructed in north-east region of USA), `CT` (=1 if cooling tower used), `BW` (=1 if nuclear steam supply system manufactured by Babcock–Wilcox), `N` (cumulative number of power plants constructed by each architect-engineer), `PT` (=1 if partial turnkey plant).

	cost	date	T_1	T_2	capacity	PR	NE	CT	BW	N	PT
1	460.05	68.58	14	46	687	0	1	0	0	14	0
2	452.99	67.33	10	73	1065	0	0	1	0	1	0
3	443.22	67.33	10	85	1065	1	0	1	0	1	0
4	652.32	68.00	11	67	1065	0	1	1	0	12	0
5	642.23	68.00	11	78	1065	1	1	1	0	12	0
6	345.39	67.92	13	51	514	0	1	1	0	3	0
7	272.37	68.17	12	50	822	0	0	0	0	5	0
8	317.21	68.42	14	59	457	0	0	0	0	1	0
9	457.12	68.42	15	55	822	1	0	0	0	5	0
10	690.19	68.33	12	71	792	0	1	1	1	2	0
11	350.63	68.58	12	64	560	0	0	0	0	3	0
12	402.59	68.75	13	47	790	0	1	0	0	6	0
13	412.18	68.42	15	62	530	0	0	1	0	2	0
14	495.58	68.92	17	52	1050	0	0	0	0	7	0
15	394.36	68.92	13	65	850	0	0	0	1	16	0
16	423.32	68.42	11	67	778	0	0	0	0	3	0
17	712.27	69.50	18	60	845	0	1	0	0	17	0
18	289.66	68.42	15	76	530	1	0	1	0	2	0
19	881.24	69.17	15	67	1090	0	0	0	0	1	0
20	490.88	68.92	16	59	1050	1	0	0	0	8	0
21	567.79	68.75	11	70	913	0	0	1	1	15	0
22	665.99	70.92	22	57	828	1	1	0	0	20	0
23	621.45	69.67	16	59	786	0	0	1	0	18	0
24	608.80	70.08	19	58	821	1	0	0	0	3	0
25	473.64	70.42	19	44	538	0	0	1	0	19	0
26	697.14	71.08	20	57	1130	0	0	1	0	21	0
27	207.51	67.25	13	63	745	0	0	0	0	8	1
28	288.48	67.17	9	48	821	0	0	1	0	7	1
29	284.88	67.83	12	63	886	0	0	0	1	11	1
30	280.36	67.83	12	71	886	1	0	0	1	11	1
31	217.38	67.25	13	72	745	1	0	0	0	8	1
32	270.71	67.83	7	80	886	1	0	0	1	11	1

This arbitrariness is reflected in rules for deciding which terms to include, some of which use tables of the F or t distributions. Others simply drop a term from the model if its F statistic is less than a number such as 4, and otherwise include the term. Sometimes a theoretically-motivated criterion such as AIC is used.

Example 8.29 (Nuclear plant data) Table 8.13 contains data on the cost of 32 light water reactors. The cost (in dollars $\times 10^{-6}$ adjusted to a 1976 base) is the quantity of interest, and the others are explanatory variables.

Costs are typically relative. Moreover large costs are likely to vary more than small ones, so it seems sensible to take log(`cost`) as the response y. For consistency we also take logs of the other quantitative covariates, fitting linear models using `date`, log(`T1`), log(`T2`), log(`capacity`), `PR`, `NE`, `CT`, log(`N`), and

ble 8.14
ameter estimates
l standard errors
linear models
ed to nuclear
nts data; forward
l backward
icate models
ed by forward
ction and
kward
nination.

	Full model		Backward		Forward	
	Est (SE)	t	Est (SE)	t	Est (SE)	t
Constant	−14.24 (4.229)	−3.37	−13.26 (3.140)	−4.22	−7.627 (2.875)	−2.66
`date`	0.209 (0.065)	3.21	0.212 (0.043)	4.91	0.136 (0.040)	3.38
`log(T1)`	0.092 (0.244)	0.38				
`log(T2)`	0.290 (0.273)	1.05				
`log(cap)`	0.694 (0.136)	5.10	0.723 (0.119)	6.09	0.671 (0.141)	4.75
`PR`	−0.092 (0.077)	−1.20				
`NE`	0.258 (0.077)	3.35	0.249 (0.074)	3.36		
`CT`	0.120 (0.066)	1.82	0.140 (0.060)	2.32		
`BW`	0.033 (0.101)	0.33				
`log(N)`	−0.080 (0.046)	−1.74	−0.088 (0.042)	−2.11		
`PT`	−0.224 (0.123)	−1.83	−0.226 (0.114)	−1.99	−0.490 (0.103)	−4.77
Residual SE (df)	0.164 (21)		0.159 (25)		0.195 (28)	

`PT`. The last of these indicates six plants for which there were partial turnkey guarantees, and some subsidies may be hidden in their costs.

Estimates and standard errors for the full model and those found by backward elimination and forward selection are given in Table 8.14. Backward elimination starts by refitting the model without `BW` and then considering the t statistics for the remaining variables, dropping the next least significant, here `log(T1)`, and so forth. The effects for the variables retained are strengthened; most are highly significant. Forward selection chooses a smaller model with larger residual sum of squares, and this results in smaller t statistics. Stepwise selection starting from this model yields the model chosen by backward elimination. Examination of residuals for this suggests no difficulty, and we are left with a model in which cost increases with capacity, though not proportionally, with presence of a cooling tower, with date, and in the north-east region of the USA, but is decreased by a partial turnkey guarantee, and with architect's experience. ■

Likelihood criteria

A more satisfactory approach is to fit all reasonable models and adopt the one that minimizes some overall measure of discrepancy. One such measure is the residual sum of squares, but this continues to decrease as the number of parameters increases and always yields the model with all possible terms. This suggests that model complexity be penalized by balancing it against a measure of fit. We now discuss one approach to this.

Suppose that the data were generated by a true model g under which the responses Y_j are independent normal variables with means μ_j and variances σ^2 and let $\mathrm{E}_g(\cdot)$ denote expectation with respect to this model. Following the

discussion in Section 4.7, our ideal would be to choose the candidate model $f(y;\theta)$ to minimize the loss when predicting a new sample like the old one,

The scaling factor 2 is included for comparability with AIC.

$$\mathrm{E}_g\left(\mathrm{E}_g^+\left[2\sum_{j=1}^{n}\log\left\{\frac{g(Y_j^+)}{f(Y_j^+;\widehat{\theta})}\right\}\right]\right). \tag{8.31}$$

Here $Y_1^+,\ldots,Y_n^+$ is another sample independent of $Y_1,\ldots,Y_n$ but with the same distribution, E_g^+ denotes expectation over $Y_1^+,\ldots,Y_n^+$, and $\widehat{\theta}$ is the maximum likelihood estimator of θ based on $Y_1,\ldots,Y_n$.

If the candidate model is normal, then θ comprises the mean responses $\mu_1,\ldots,\mu_n$ and σ^2, with maximum likelihood estimators $\widehat{\mu}_1,\ldots,\widehat{\mu}_n$ and $\widehat{\sigma}^2$. Then the sum in (8.31) equals

$$\tfrac{1}{2}\sum_{j=1}^{n}\left\{\log\widehat{\sigma}^2+\frac{(Y_j^+-\widehat{\mu}_j)^2}{\widehat{\sigma}^2}-\log\sigma^2-\frac{(Y_j^+-\mu_j)^2}{\sigma^2}\right\},$$

and hence the inner expectation is

$$\sum_{j=1}^{n}\left\{\log\widehat{\sigma}^2+\frac{\sigma^2}{\widehat{\sigma}^2}+\frac{(\mu_j-\widehat{\mu}_j)^2}{\widehat{\sigma}^2}-\log\sigma^2-1\right\}.$$

Suppose that in our earlier terminology a candidate linear model with full-rank $n\times p$ design matrix X is correct, that is, the true model is nested within it. Then the vector $\mu=(\mu_1,\ldots,\mu_n)^{\mathrm{T}}$ of true means lies in the column space of X and there is a $p\times 1$ vector β such that $\mu=X\beta$. Hence $\widehat{\mu}=(\widehat{\mu}_1,\ldots,\widehat{\mu}_n)^{\mathrm{T}}$ is normal with mean μ, from which it follows that $\sum(\mu_j-\widehat{\mu}_j)^2=(\widehat{\mu}-\mu)^{\mathrm{T}}(\widehat{\mu}-\mu)\sim\sigma^2\chi^2_p$ independent of $n\widehat{\sigma}^2\sim\sigma^2\chi^2_{n-p}$. Now the expected values of a χ^2_ν variable and of its inverse are ν and $(\nu-2)^{-1}$, provided $\nu>2$, and so (8.31) equals

$$n\mathrm{E}_g\left(\log\widehat{\sigma}^2\right)+\frac{n^2}{n-p-2}+\frac{np}{n-p-2}-n\log\sigma^2-n,$$

or equivalently for our purposes,

$$n\mathrm{E}_g\left(\log\widehat{\sigma}^2\right)+\frac{n(n+p)}{n-p-2}.$$

This is estimated unbiasedly by the *corrected information criterion*

$$\mathrm{AIC_c}=n\log\widehat{\sigma}^2+n\frac{1+p/n}{1-(p+2)/n},$$

and the 'best' candidate model is taken to be that which minimizes this. Taylor expansion gives $\mathrm{AIC_c}\doteq n\log\widehat{\sigma}^2+n+2(p+1)+O(p^2/n)$, and for large n and fixed p this will select the same model as $\mathrm{AIC}=n\log\widehat{\sigma}^2+2p$. When p is comparable with n, $\mathrm{AIC_c}$ penalizes model dimension more severely.

Table 8.15
Number of times models were selected using various model selection criteria in 50 repetitions using simulated normal data for each of 20 design matrices. The true model has $p = 3$.

n		Number of covariates						
		1	2	3	4	5	6	7
10	C_p		131	504	91	63	83	128
	BIC		72	373	97	83	109	266
	AIC		52	329	97	91	125	306
	AIC_c	15	398	565	18	4		
20	C_p		4	673	121	88	61	53
	BIC		6	781	104	52	30	27
	AIC		2	577	144	104	76	97
	AIC_c		8	859	94	30	8	1
40	C_p			712	107	73	66	42
	BIC			904	56	20	15	5
	AIC			673	114	90	69	54
	AIC_c			786	105	52	41	16

A widely used related criterion is

$$C_p = \frac{SS_p}{s^2} + 2p - n,$$

where SS_p is the residual sum of squares for the fitted model and s^2 is an estimate of σ^2; C_p can be derived as an approximation to AIC (Problem 8.16), though its original motivation was different. In some cases σ^2 can be estimated from the full model, but care is needed because the choice of s^2 is critical to successful use of C_p.

Example 8.30 (Simulation study) Twenty different $n \times 7$ design matrices X were constructed using standard normal variables, centered and scaled so that each column of X had mean zero and unit variance. The parameter vector was $\beta = (3, 2, 1, 0, 0, 0, 0)^{\mathrm{T}}$, so the true model had three covariates, and the errors were taken to be independent standard normal variables. Then the models with the first p columns of X were fitted for $p = 1, \ldots, 7$, and the best of these was selected using AIC, AIC_c, the Bayesian criterion BIC, and C_p. This procedure was performed 50 times for each design matrix.

Table 8.15 shows the results of this experiment. For $n = 10$ and 20, AIC_c has the highest chance of selecting the true model, and moreover the models selected using it are the least dispersed because of the stronger penalty applied, at least for p comparable with n. For $n = 40$ the consistent criterion BIC is most likely to select the true model. In practice, however, the true model would rarely be among those fitted, and so AIC_c seems the best of the criteria considered, particularly when p is comparable with n. ■

Example 8.31 (Nuclear plant data) When $\mathrm{AIC_c}$ is computed for the 2^{10} possible models in Example 8.29, the model chosen by backward elimination is selected, with $\mathrm{AIC_c} = -71.24$. Two nearby models have $\mathrm{AIC_c}$ within 2 of the minimum, namely those without `log(N)` and without `PT`, but dropping these covariates together increases $\mathrm{AIC_c}$ sharply. The interpretation and overall fit are changed little by dropping them singly, so we retain them. ■

Plots of the contributions to these criteria from individual observations can be useful in diagnosing whether particular cases strongly influence model choice.

There may be several different models whose values of $\mathrm{AIC_c}$ are similarly low. If a single model is needed the choice among them should if possible be based on subject-matter considerations. If there are several equally plausible models with quite different interpretations, then it is important to say so.

Inference after model selection

One reason that automatic variable selection should if possible be avoided is its consequence for subsequent inference. To illustrate this, consider a straight-line regression model $y = \beta_0 + x\beta_1 + \varepsilon$, based on n pairs (x_j, y_j) with $\sum x_j = 0$ and independent normal errors with mean zero and known variance σ^2. Then the least squares estimate $\widehat{\beta}_1$ is normally distributed with mean β_1 and variance $v = \sigma^2 / \sum x_j$, and following the discussion in Section 8.3.2 we would base inference for β_1 on $Z = (\widehat{\beta}_1 - \beta_1)/v^{1/2}$, whose distribution is standard normal when model selection is not taken into account. Suppose, however, that before attempting to construct a confidence interval for β_1, we test for inclusion of the covariate x in the model, declaring that it should be included if $|\widehat{\beta}_1/v^{1/2}| > z_{1-\alpha}$. If not, we declare that $\beta_1 = 0$ and use the simpler model $y = \beta_0 + \varepsilon$. Now as $\widehat{\beta}_1 = \beta_1 + v^{1/2}Z$, post-model selection inference for β_1 given that x has been included will be based on the conditional density of Z given that $|Z + \beta_1/v^{1/2}| > z_{1-\alpha}$, which is

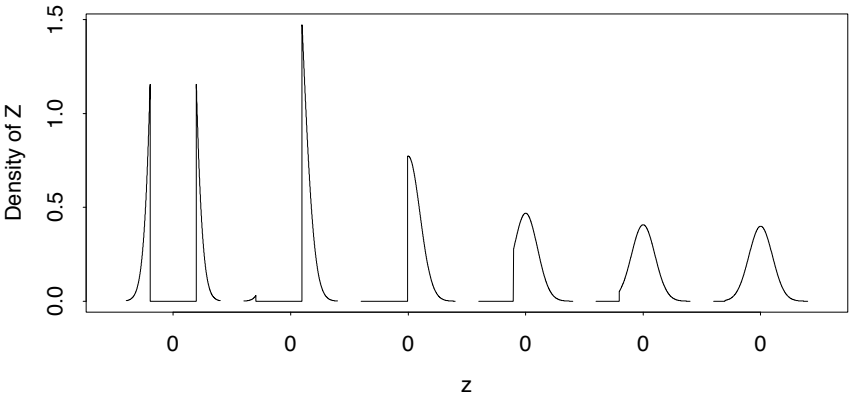
$z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution.

$$\phi_\delta(z) = \frac{\phi(z)\,\{H(z < z_\alpha - \delta) + 1 - H(z < -z_\alpha - \delta)\}}{\Phi(z_\alpha - \delta) + \Phi(z_\alpha + \delta)}, \quad -\infty < z < \infty,$$

$H(u)$ is the Heaviside function.

where $\delta = \beta_1/v^{1/2}$ is the standardized slope. Figure 8.8 displays $\phi_\delta(z)$ for $\delta = 0, 1, \ldots, 5$ and $\alpha = 0.025$, corresponding to two-sided testing at the 5% level. When $\beta_1 = 0$, for example, Z considered conditionally takes values in the tails of the standard normal distribution but not in its centre. Conditional on variable selection, Z is clearly far from pivotal unless $|\delta| \gg 0$. Hence it is only a sensible basis for inference on β_1 if the regression on x is very strong.

In practice there are three complications: the error variance σ^2 is unknown, there are typically many covariates, and the true model is not among those fitted. However the broad conclusion applies: if variables are selected automatically, the only covariates for which subsequent inference using the standard



gure 8.8
tribution of the posed pivot Z for erence on the pe parameter in a aight-line ression model, ditional on lusion of slope in model, for $0, 1, \ldots, 5$ (left ight) and testing inclusion at the level. Conditional inclusion, Z is r-pivotal only if $\gg 0$.

confidence intervals is reliable are those for which the evidence for inclusion is overwhelming, that is, for which it is clear that $|\delta| \gg 0$. Other covariates should be considered in the light of previous knowledge and the context of the model.

Model uncertainty

Inference is often performed after comparing different competing models, and the questions arise if, when, and how one should allow for this. Consider for example the quantity β_0 in the two models M_0 and M_1 in which $y = \beta_0 + \varepsilon$ and $y = \beta_0 + x\beta_1 + \varepsilon$, where $\mathrm{E}(\varepsilon) = 0$. It is sometimes suggested that one should somehow average the variances of the estimators $\widehat{\beta}_0$ across the models, but this is inappropriate because the interpretation of β_0 is model-dependent. Although the same symbol is used, β_0 represents the unconditional response mean $\mathrm{E}(Y)$ under M_0, while under M_1 it represents the conditional mean $\mathrm{E}(Y \mid x = 0)$. Hence the meaning of β_0 depends on the context and inference for it must be conditioned on the model in which it appears: averaging is meaningless unless the quantity of interest has the same interpretation for all models considered. In particular, the interpretation of regression coefficients typically depends on the model in which they appear. Having said this, one situation in which the quantity of interest has a model-free interpretation is prediction, and below we treat the simplest example of this.

Consider using the fits of M_0 and M_1 to estimate the mean $\mu_+ = \beta_0 + x_+\beta_1$ of a future variable Y_+ with covariate $x_+ \neq 0$, assuming the error ε to be normal with mean zero and known variance σ^2; note that μ_+ has the same interpretation under both models. Suppose that n independent pairs (x_j, y_j) are available and that $\sum x_j = 0$, so that $\widehat{\beta}_0 = \overline{y}$ with variance σ^2/n under either model, independent of the slope estimate $\widehat{\beta}_1$ with variance $v =$

$\sigma^2/\sum x_j^2$. The estimators of μ_+ and their biases, variances, and mean squared errors are

Model	Estimator	Bias	Variance	MSE
M_0 :	$\widehat{\mu}_+^0 = \widehat{\beta}_0$,	$x_+\beta_1$,	σ^2/n,	$\sigma^2/n + x_+^2\beta_1^2$,
M_1 :	$\widehat{\mu}_+^1 = \widehat{\beta}_0 + x_+\widehat{\beta}_1$,	0,	$\sigma^2/n + x_+^2 v$,	$\sigma^2/n + x_+^2 v$,

so $\widehat{\mu}_+^0$ improves on $\widehat{\mu}_+^1$ if $|\delta| < 1$, where $\delta = \beta_1/v^{1/2}$ is the standardized slope.

This suggests that it may be possible to construct a better estimator of μ_+ by choosing $\widehat{\mu}_+^0$ if an estimator of δ is close enough to zero, and otherwise taking $\widehat{\mu}_+^1$. If we decide between the models on the basis that M_1 is indicated when $|\widehat{\beta}_1|/v^{1/2} > z_{1-\alpha}$, corresponding to a two-sided test of the hypothesis that $\beta_1 = 0$ at level $(1-2\alpha)$, then the overall estimator is

$I(\cdot)$ is the indicator random variable of its event.

$$\begin{aligned}\widehat{\mu}_+ &= \widehat{\beta}_0 + x_+\widehat{\beta}_1\left\{I\left(\widehat{\beta}_1/v^{1/2} < -z_{1-\alpha}\right) + I\left(\widehat{\beta}_1/v^{1/2} > z_{1-\alpha}\right)\right\} \\ &= \widehat{\beta}_0 + x_+ v^{1/2}(\delta + Z)\left\{I\left(Z < z_\alpha - \delta\right) + I\left(Z > z_{1-\alpha} - \delta\right)\right\},\end{aligned}$$

where we have written $\widehat{\beta}_1 = v^{1/2}(\delta + Z)$, with $Z = (\widehat{\beta}_1 - \beta_1)/v^{1/2}$ a standard normal variable; note that $-z_{1-\alpha} = z_\alpha$. The bias and variance of $\widehat{\mu}_+$ are

$$\mathrm{E}\left(\widehat{\mu}_+ - \mu_+\right) = x_+ v^{1/2}\,\mathrm{E}(Q), \quad \mathrm{var}\left(\widehat{\mu}_+\right) = \frac{\sigma^2}{n} + x_+^2 v\,\mathrm{var}(Q),$$

where $Q = (\delta + Z)\left\{I(Z < z_\alpha - \delta) + I(Z > z_{1-\alpha} - \delta)\right\} - \delta$. As $v = \sigma^2/\sum x_j^2$, the bias is $O(n^{-1/2})$ and the variance is $O(n^{-1})$, while the mean squared error is $\sigma^2/n + x_+^2 v\left\{\mathrm{E}(Q)^2 + \mathrm{var}(Q)\right\}$. Elementary calculations give the functions $\mathrm{E}(Q)$, $\mathrm{var}(Q)$, and $\mathrm{E}(Q)^2 + \mathrm{var}(Q)$, which are shown in the upper right panel of Figure 8.9 for $\alpha = 0.025$, corresponding to choosing between the models at the two-sided 95% level. As we might have anticipated, $\widehat{\mu}_+$ is generally biased towards zero because of the possibility of using the simpler estimator $\widehat{\mu}_+^0$ even if $\beta_1 \neq 0$; its bias tends to zero when $|\delta| \gg 0$. The variance of $\widehat{\mu}_+$ is largest when $|\delta| \doteq 2$, and then decreases to the limit corresponding to use of $\widehat{\mu}_+^1$.

One difficulty with $\widehat{\mu}_+$ is that the indicator variables badly inflate its bias and variance. A simple way to avoid this is to use a weighted combination of $\widehat{\mu}_+^0$ and $\widehat{\mu}_+^1$. Take for example the estimator

$$\widehat{\mu}_+^w = (1-W)\widehat{\mu}_+^0 + W\widehat{\mu}_+^1 = (1-W)\widehat{\beta}_0 + W\left(\widehat{\beta}_0 + x_+\widehat{\beta}_1\right),$$

where the weight

$$W = \frac{\exp(-\mathrm{AIC}_1/2)}{\exp(-\mathrm{AIC}_1/2) + \exp(-\mathrm{AIC}_0/2)}$$

depends on the information criteria AIC_0 and AIC_1 for the two models. If $\mathrm{AIC}_1 \ll \mathrm{AIC}_0$, then $W \doteq 1$, the data give a strong preference for M_1, and $\widehat{\mu}_+^w \doteq \widehat{\mu}_+^1$. If on the other hand $\beta_1 = 0$, then W slightly favours M_0 but the estimators under both models are unbiased.

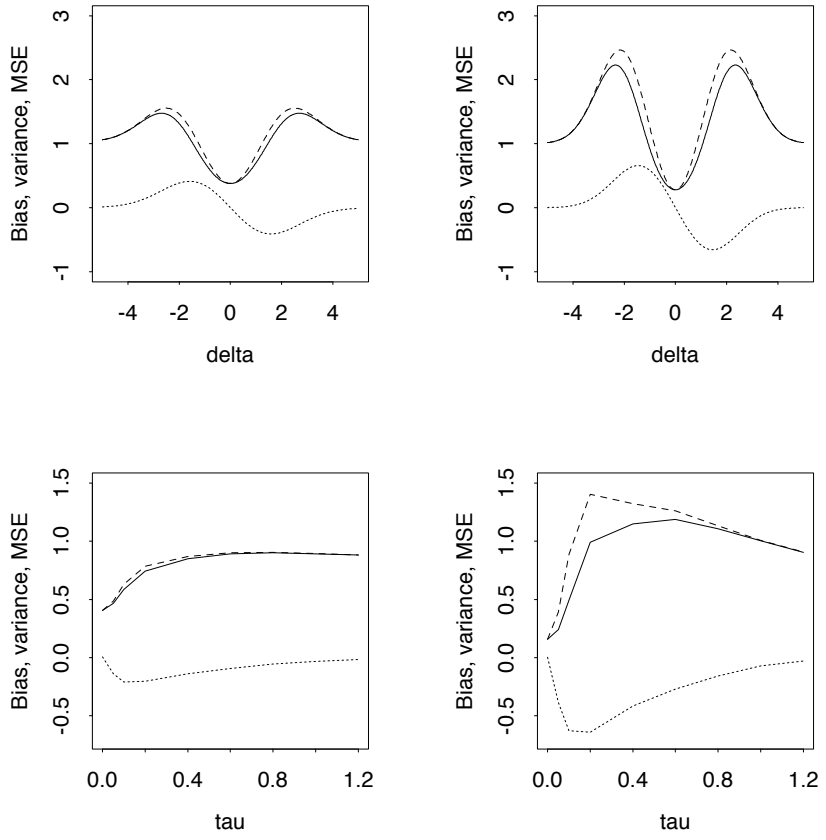


Figure 8.9
Properties of estimators of $\beta_0 + x_+\beta_1$ in the straight-line regression model. Left: bias (dots), variance (solid) and mean squared error (dashes) for weighted estimator $\widehat{\mu}_+^w$. Right: corresponding quantities for model-choice estimator $\widehat{\mu}_+$. The weighted estimator improves considerably on the model-choice estimator. The upper panels are for theoretical calculations, and the lower ones for the simulation experiment described in Example 8.32.

Under our simplifying assumptions, $\mathrm{AIC}_0 - \mathrm{AIC}_1 = \widehat{\beta}_1^2/v - 2 = (\delta + Z)^2 - 2$, and as $\widehat{\mu}_+^w = \widehat{\beta}_0 + x_+ W\widehat{\beta}_1$, the quantity that corresponds to Q above is $Q^w = (\delta + Z)G\left\{(\delta + Z)^2/2 - 1\right\} - \delta$, where $G(u) = \exp(u)/\{1 + \exp(u)\}$. The bias and variance of $\widehat{\mu}_+^w$ depend on those of Q^w, which are shown in the upper left panel of Figure 8.9. Both are smaller than the values for $\widehat{\mu}_+$, and the mean squared error is considerably reduced. Evidently $\widehat{\mu}_+^w$ improves on $\widehat{\mu}_+^1$ over a wide range of values of δ, while its mean squared error is smaller than that of $\widehat{\mu}_+$. The weighted estimator $\widehat{\mu}_+^w$ clearly improves on the model-choice estimator $\widehat{\mu}_+$.

Example 8.32 (Simulation study) To assess how this approach performs in a slightly more realistic setting, we performed a small simulation study with linear model data simulated in the same way as in Example 8.30, now with $n = 15$ and $\beta^{\mathrm{T}} = \tau(0, 4, 3, 2, 1, 1, 0, 0)$; thus $p = 8$ including a constant vector.

We then fitted the eight models with a constant only, constant plus the first covariate, constant plus first and second covariates, and so forth, and combined the corresponding estimators and AIC-based weights, to obtain a weighted estimator $\widehat{\theta}$ of $\theta = 1_8^{\mathrm{T}}\beta$. We compared this with the estimator $\widehat{\theta}_+$ obtained from the 'best' model, this being chosen as the model minimizing $-2\widehat{\ell}_q + 3.84q$, where $\widehat{\ell}_q$ is the log likelihood obtained when fitting the model with q parameters. This information criterion is constructed to give probability 0.05 of selecting the more complex of two nested models differing by one parameter, when in fact the simpler model is correct. This criterion is intended to mimic hypothesis testing procedures for model selection, such as backward elimination.

This experiment was repeated with 20 different response vectors for each of 250 design matrices: 5000 datasets, for $\tau = 0, 0.05, 0.1, 0.2, 0.4, \ldots, 1.2$. The lower panels of Figure 8.9 show the bias, variance, and mean squared error of $\widehat{\theta}$ and $\widehat{\theta}_+$. The results bear out the preceding toy analysis: the weighted estimator has lower mean squared error except when the regression effects are small. ■

Although we have only considered the simplest situation, our broad conclusion generalizes to more complex settings: sharp choices among estimators from different models tends to give worse predictions than do estimators interpolating smoothly among them.

Exercises 8.7

1 Consider the cement data of Example 8.3, where $n = 13$. The residual sums of squares for all models that include an intercept are given in Exercise 8.5.1.
(a) Use forward selection, backward elimination, and stepwise selection to select models for these data, including variables significant at the 5% level.
(b) Use C_p to select a model for these data.

2 Another criterion for model selection is to choose the covariates that minimize the cross-validated sum of squares $\sum(y_j - x_j^{\mathrm{T}}\widehat{\beta}_{-j})^2$, where $\widehat{\beta}_{-j}$ is the estimate of β obtained when the jth case is deleted. Show this is equivalent to minimizing $\sum(y_j - x_j^{\mathrm{T}}\widehat{\beta})^2/(1-h_{jj})^2$, and compare computational aspects of this approach with those based on AIC.

8.8 Bibliographic Notes

There are books on all aspects of the linear model. Seber (1977) and Searle (1971) give a thorough discussion of the theory, while Draper and Smith (1981), Weisberg (1985), Wetherill (1986) and Rawlings (1988) have somwhat more practical emphases; see also Sen and Srivastava (1990) and Jørgensen

(1997a). Most of these books cover the central topics of this chapter in more detail. Scheffé (1959) is a classic account of the analysis of variance.

Robust approaches to regression are described by Li (1985), and in more detail in Huber (1981), Hampel *et al.* (1986), and Rousseeuw and Leroy (1987).

Davison and Hinkley (1997) and Efron and Tibshirani (1993) give accounts of bootstrap methods, which are simulation approaches to finding standard errors, confidence limits and so forth, for use with awkward estimators.

The formal analysis of transformations was discussed by Box and Cox (1964) and further developed by many others; for book-length discussions see Atkinson (1985) and Carroll and Ruppert (1988). The test for non-additivity was suggested by Tukey (1949); see also Hinkley (1985). Books on general regression diagnostics include Cook and Weisberg (1982), Belsley *et al.* (1980) and Chatterjee and Hadi (1988). Belsley (1991) focuses on problems of collinearity. Shorter accounts of aspects of model-checking are Davison and Snell (1991) and Davison and Tsai (1992). Atkinson and Riani (2000) describe how diagnostic procedures may be used to give reliable strategies for data analysis.

Stone and Brooks (1990) and their discussants give numerous references and comparison of various approaches to regression situations with fewer observations than covariates, such as principal components regression and partial least squares. Perhaps the most widespread of these is ridge regression (Hoerl and Kennard, 1970a,b; Hoerl *et al.*, 1985). Brown (1993) is a book-length treatment of these and related methods.

Variable selection for the linear model has been intensively studied. Linhart and Zucchini (1986) and Miller (1990) give useful surveys, now somewhat dated owing to the considerable amount of work in the 1990s. Model selection based on AIC was suggested by Akaike (1973) in a much-cited paper, though related criteria such as C_p were already in use (Mallows, 1973). Schwartz (1978) proposed use of BIC, and Hurvich and Tsai (1989, 1991) derive the modified AIC with improved small-sample properties. McQuarrie and Tsai (1998) give a comprehensive discussion of these and related criteria. Pötscher (1991) and Hurvich and Tsai (1990) give theoretical and numerical results on inference after model selection in linear models. More general discussion and many further references may be found in Chatfield (1995) and Burnham and Anderson (2002).

8.9 Problems

1 Consider Table 8.16. Formulate the design matrix X for the model in which $\mathrm{E}(\text{Yield}) = \beta_i + \beta_3(z-2)$, estimate the parameters and test whether $\beta_1 = \beta_2$.

2 Suppose that random variables Y_{gj}, $j = 1, \ldots, n_g$, $g = 1, \ldots, G$, are independent and that they satisfy the normal linear model $Y_{gj} = x_g^{\mathrm{T}}\beta + \varepsilon_{gj}$. Write down the

Table 8.16
Rescaled yields (tonnes/Ha) when two varieties of corn were treated with five levels of fertiliser.

Variety	Level of fertilizer, z				
	0	1	2	3	4
1	0.2	0.6	0.5	0.8	0.9
2	0.1	0.2	0.4	0.6	0.7

covariate matrix for this model, and show that the least squares estimates can be written as $(X_1^{\mathrm{T}}WX_1)^{-1}X_1^{\mathrm{T}}WZ$, where $W = \mathrm{diag}\{n_1, \ldots, n_G\}$, and the gth element of Z is $n_g^{-1}\sum_j Y_{gj}$. Hence show that weighted least squares based on Z and unweighted least squares based on Y give the same parameter estimates and confidence intervals, when σ^2 is known. Why do they differ if σ^2 is unknown, unless $n_g \equiv 1$?
Discuss how the residuals for the two setups differ, and say which is preferable for model-checking.

3 Let $Y_1, \ldots, Y_n$ and $Z_1, \ldots, Z_m$ be two independent random samples from the $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ distributions respectively. Consider comparison of the model in which $\sigma_1^2 = \sigma_2^2$ and the model in which no restriction is placed on the variances, with no restriction on the means in either case. Show that the likelihood ratio statistic W_{p} to compare these models is large when the ratio $T = \sum(Y_j - \overline{Y})^2 / \sum(Z_j - \overline{Z})^2$ is large or small. Show that T is proportional to a random variable with the F distribution, and discuss whether the model of equal variances is plausible for the maize data of Example 1.1.

4 Find the expected information matrix for the parameters $(\beta_0, \beta_1, \sigma^2)$ of the normal straight-line regression model (5.2).

5 The usual linear model $y = X\beta + \varepsilon$ is thought to apply to a set of data, and it is assumed that the ε_j are independent with means zero and variances σ^2, so that the data are summarized in terms of the usual least squares estimates and estimate of σ^2, $\widehat{\beta}$ and S^2. Unknown to the unfortunate investigator, in fact $\mathrm{var}(\varepsilon_j) = v_j\sigma^2$, and $v_1, \ldots, v_n$ are unequal. Show that $\widehat{\beta}$ remains unbiased for β and find its actual covariance matrix.

6 Suppose that y satisfies a quadratic regression, that is,

$$y = \beta_0 + x\beta_1 + x^2\beta_2 + \varepsilon,$$

and that we can control the values of x. It is decided to choose $x = \pm a$ r times each and $x = 0$ $n - 2r$ times.
(a) Derive explicit expressions for the least squares estimates. Are they uncorrelated? If not, can they easily be made so?
(b) What value of r is best if we intend to test for the adequacy of a linear regression?
(c) What value of r is best if we intend to predict y at $x = a/2$?

7 By rewriting $y - X\beta$ as $e + X\widehat{\beta} - X\beta$ and that $e^{\mathrm{T}}X = 0$, show that

$$(y - X\beta)^{\mathrm{T}}(y - X\beta) = SS(\widehat{\beta}) + (\widehat{\beta} - \beta)^{\mathrm{T}}X^{\mathrm{T}}X(\widehat{\beta} - \beta).$$

Hence show that that the likelihood for the normal linear model equals

$$\frac{1}{(2\pi)^{n/2}\sigma^n}\exp\left\{-\frac{SS(\widehat{\beta})}{2\sigma^2}-\frac{1}{2\sigma^2}(\widehat{\beta}-\beta)^{\mathrm{T}}X^{\mathrm{T}}X(\widehat{\beta}-\beta)\right\},$$

and use the factorization criterion to establish that $(\widehat{\beta}, SS(\widehat{\beta}))$ is a minimal sufficient statistic for (β, σ^2). The sample size n and the covariate matrix X are also needed to calculate the likelihood, so why are they not regarded as part of the minimal sufficient statistic?

8 Consider a normal linear regression $y = \beta_0 + \beta_1 x + \varepsilon$ in which the parameter of interest is $\psi = \beta_0/\beta_1$, to be estimated by $\widehat{\psi} = \widehat{\beta}_0/\widehat{\beta}_1$; let $\mathrm{var}(\widehat{\beta}_0) = \sigma^2 v_{00}$, $\mathrm{cov}(\widehat{\beta}_0, \widehat{\beta}_1) = \sigma^2 v_{01}$ and $\mathrm{var}(\widehat{\beta}_1) = \sigma^2 v_{11}$.
(a) Show that

$$\frac{\widehat{\beta}_0 - \psi\widehat{\beta}_1}{\{s^2\,(v_{00} - 2\psi v_{01} + \psi^2 v_{11})\}^{1/2}} \;\sim\; t_{n-p},$$

and hence deduce that a $(1-2\alpha)$ confidence interval for ψ is the set of values of ψ satisfying the inequality

$$\widehat{\beta}_0^2 - s^2 t_{n-p}^2(\alpha)v_{00} + 2\psi\left\{s^2 t_{n-p}^2(\alpha)v_{01} - \beta_0\beta_1\right\} + \psi^2\left\{\widehat{\beta}_1^2 - s^2 t_{n-p}^2(\alpha)v_{11}\right\} \le 0.$$

How would this change if the value of σ was known?
(b) By considering the coefficients on the left-hand-side of the inequality in (a), show that the confidence set can be empty, a finite interval, semi-infinite intervals stretching to $\pm\infty$, the entire real line, two disjoint semi-infinite intervals — six possibilities in all. In each case illustrate how the set could arise by sketching a set of data that might have given rise to it.
(c) A government Department of Fisheries needed to estimate how many of a certain species of fish there were in the sea, in order to know whether to continue to license commercial fishing. Each year an extensive sampling exercise was based on the numbers of fish caught, and this resulted in three numbers, y, x, and a standard deviation for y, σ. A simple model of fish population dynamics suggested that $y = \beta_0 + \beta_1 x + \varepsilon$, where the errors ε are independent, and the original population size was $\psi = \beta_0/\beta_1$. To simplify the calculations, suppose that in each year σ equalled 25. If the values of y and x had been

y :	160	150	100	80	100
x :	140	170	200	230	260

after five years, give a 95% confidence interval for ψ. Do you find it plausible that $\sigma = 25$? If not, give an appropriate interval for ψ.

9 Over a period of $2m+1$ years the quarterly gas consumption of a particular household may be represented by the model

$$Y_{ij} = \beta_i + \gamma j + \varepsilon_{ij}, \quad i = 1, \ldots, 4, j = -m, -m+1, \ldots, m-1, m,$$

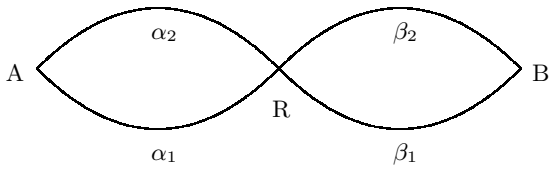
where the parameters β_i and γ are unknown, and $\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, \sigma^2)$. Find the least squares estimators and show that they are independent with variances $(2m+1)^{-1}\sigma^2$ and $\sigma^2/\left(8\sum_{i=1}^{m} i^2\right)$.

Show also that

$$(8m-1)^{-1}\left[\sum_{i=1}^{4}\sum_{j=-m}^{m} Y_{ij}^2 - (2m+1)\sum_{i=1}^{4}\overline{Y}_{i\cdot}^2 - \frac{2\sum_{j=-m}^{m} j\overline{Y}_{\cdot j}^2}{\sum_{i=1}^{m} i^2}\right]$$

is unbiased for σ^2, where $\overline{Y}_{i\cdot} = (2m+1)^{-1}\sum_{j=-m}^{m} Y_{ij}$ and $\overline{Y}_{\cdot j} = \frac{1}{4}\sum_{i=1}^{4} Y_{ij}$.

10 A statistician travels regularly from A to B by one of four possible routes, each route crossing a river bridge at R. The times taken for the possible segments of the journey are independent random variables with means as shown in the figure, each having variance $\sigma^2/2$.



He times the *complete* journey once by each route, obtaining observations y_{ij} distributed as random variables Y_{ij} having means $\mathrm{E}(Y_{ij}) = \alpha_i + \beta_j$, for $i, j = 1, 2$. Why it is not possible to estimate all the parameters from these observations? Now define $\mu = \alpha_1 + \beta_1$, $\gamma = \alpha_2 - \alpha_1$ and $\delta = \beta_2 - \beta_1$. Obtain expressions for the least squares estimates of μ, γ and δ and also for their variance matrix.
If the observed vector of times is $(y_{11}, y_{21}, y_{12}, y_{22}) = (124, 120, 128, 136)$ minutes, determine which route has the smallest estimated mean time. Obtain a 90% confidence interval for the mean on the assumption that the times are normally distributed.

11 Suppose that we wish to construct the likelihood ratio statistic for comparison of the two linear models $y = X_1\beta_1 + \varepsilon$ and $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$, where the components of ε are independent normal variables with mean zero and variance σ^2; call the corresponding residual sums of squares SS_1 and SS on ν_1 and ν degrees of freedom.
(a) Show that the maximum value of the log likelihood is $-\frac{1}{2}n(\log SS + 1 - \log n)$ for a model whose residual sum of squares is SS, and deduce that the likelihood ratio statistic for comparison of the models above is $W = n\log(SS_1/SS)$.
(b) By writing $SS_1 = SS + (SS_1 - SS)$, show that W is a monotonic function of the F statistic for comparison of the models.
(c) Show that $W \doteq (\nu_1 - \nu)F$ when n is large and ν is close to n, and say why F would usually be preferred to W.

12 Suppose that the denominator in the F statistic was replaced by $SS(\widehat{\beta}_1)/(n-q)$, giving F', say. Use the geometry of least squares to explain why F' does not have an F distribution, even if the simpler model is correct so that $SS(\widehat{\beta}_1) \sim \sigma^2\chi^2_{n-q}$. Show that F' is a monotone increasing function of F, that tends to be less than F if the simpler model is not adequate.

13 Table 8.17 gives results from $n = 10$ runs of a computer experiment to assess the accuracy of a hydrological model. The response y is the relative accuracy of predictions, and the covariates x_1, x_2, x_3, and x_4 represent parameters input

ble 8.17
idual sums of
ares for fits of
ear models to
put from $n = 10$
s of a
rological model.

Model	SS	Model	SS	Model	SS
- - - -	11.06	1 2 - -	5.56	1 2 3 -	4.75
1 - - -	5.96	1 - 3 -	4.78	1 2 - 4	0.74
- 2 - -	10.19	1 - - 4	1.34	1 - 3 4	0.83
- - 3 -	9.96	- 2 3 -	8.09	- 2 3 4	3.05
- - - 4	9.09	- 2 - 4	7.94		
		- - 3 4	6.51	1 2 3 4	0.69

to the model. The table gives the residual sums of squares for all normal linear models that include an intercept and the x_j.
Taking the level of significance to be 5%, select models for the data using (a) forward selection, (b) backward elimination, (c) stepwise model selection starting from the full model, and (d) C_p. Comment briefly.

14 In the normal straight-line regression model it is thought that a power transformation of the covariate may be needed, that is, the model

$$y = \beta_0 + \beta_1 x^{(\lambda)} + \varepsilon$$

may be suitable, where $x^{(\lambda)}$ is the power transformation

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log x, & \lambda = 0. \end{cases}$$

(a) Show by Taylor series expansion of $x^{(\lambda)}$ at $\lambda = 1$ that a test for power transformation can be based on the reduction in sum of squares when the constructed variable $x \log x$ is added to the model with linear predictor $\beta_0 + \beta_1 x$.
(b) Show that the profile log likelihood for λ is equivalent to $\ell_{\mathrm{p}}(\lambda) \equiv -\frac{n}{2} \log SS(\widehat{\beta}_\lambda)$, where $SS(\widehat{\beta}_\lambda)$ is the residual sum of squares for regression of y on the $n \times 2$ design matrix with a column of ones and the column consisting of the $x_j^{(\lambda)}$. Why is a Jacobian for the transformation not needed in this case, unlike in Example 8.23?
(Box and Tidwell, 1962)

15 Consider model $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$, which leads to least squares estimates

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \begin{pmatrix} X_1^{\mathrm{T}} X_1 & X_1^{\mathrm{T}} X_2 \\ X_2^{\mathrm{T}} X_1 & X_2^{\mathrm{T}} X_2 \end{pmatrix}^{-1} \begin{pmatrix} X_1^{\mathrm{T}} y \\ X_2^{\mathrm{T}} y \end{pmatrix}.$$

Let $H_1 = X_1(X_1^{\mathrm{T}} X_1)^{-1} X_1^{\mathrm{T}}$, $P_1 = I_n - H_1$, and define H_2 and P_2 similarly; notice that these projection matrices are symmetric and idempotent.
(a) Show that $\widehat{\beta}_2$ can be expressed as

$$(X_2^{\mathrm{T}} P_1 X_2)^{-1} X_2^{\mathrm{T}} y - (X_2^{\mathrm{T}} X_2)^{-1} X_2^{\mathrm{T}} X_1 (X_1^{\mathrm{T}} P_2 X_1)^{-1} X_1^{\mathrm{T}} y,$$

and use the result from Exercise 8.5.3 to deduce that $\widehat{\beta}_2 = (X_2^{\mathrm{T}} P_1 X_2)^{-1} X_2^{\mathrm{T}} P_1 y$, with variance matrix $\sigma^2 (X_2^{\mathrm{T}} P_1 X_2)^{-1}$. Note that $\widehat{\beta}_2$ is the parameter estimate from the regression of $P_1 y$ on the columns of $P_1 X_2$.
(b) Use the geometry of least squares to show that the residual sums of squares

for regression of y on X_1 and X_2 is the same as for the regression of $P_1 y$ on X_1 and X_2.
(c) Suppose that in a normal linear model, X_2 is a single column that depends on y only through the fitted values from regression of y on X_1, so that X_2 is itself random. Noting that the residuals $P_1 y$ are independent of the fitted values, $H_1 y$, and arguing conditionally on $H_1 y$, show that the t statistic for $\widehat{\beta}_2$ has a distribution that is independent of X_2. Hence give the unconditional distribution of (8.27).

Recall that a model is called correct if it contains all covariates with non-zero coefficients, and called true if it contains precisely these covariates.

16 (a) Show that AIC for a normal linear model with n responses, p covariates and unknown σ^2 may be written as $n \log \widehat{\sigma}^2 + 2p$, where $\widehat{\sigma}^2 = SS_p/n$ is the maximum likelihood estimate of σ^2. If $\widehat{\sigma}_0^2$ is the unbiased estimate under some fixed correct model with q covariates, show that use of AIC is equivalent to use of $n \log\{1+(\widehat{\sigma}^2-\widehat{\sigma}_0^2)/\widehat{\sigma}_0^2\}+2p$, and that this is roughly equal to $n(\widehat{\sigma}^2/\widehat{\sigma}_0^2-1)+2p$. Deduce that model selection using C_p approximates that using AIC.
(b) Show that $C_p = (q-p)(F-1)+p$, where F is the F statistic for comparison of the models with p and $q > p$ covariates, and deduce that if the model with p covariates is correct, then $\mathrm{E}(C_p) \doteq q$, but that otherwise $\mathrm{E}(C_p) > q$.

17 Consider the straight-line regression model $y_j = \alpha + \beta x_j + \sigma\varepsilon_j$, $j = 1, \ldots, n$. Suppose that $\sum x_j = 0$ and that the ε_j are independent with means zero, variances ε, and common density $f(\cdot)$.
(a) Write down the variance of the least squares estimate of β.
(b) Show that if σ is known, the log likelihood for the data is

$$\ell(\alpha, \beta) = -n \log \sigma + \sum_{j=1}^{n} \log f\left(\frac{y_j - \alpha - \beta x_j}{\sigma}\right),$$

derive the expected information matrix for α and β, and show that the asymptotic variance of the maximum likelihood estimate of β can be written as $\sigma^2/(i \sum x_j^2)$, where

$$i = \mathrm{E}\left\{-\frac{d^2 \log f(\varepsilon)}{d\varepsilon^2}\right\}.$$

Hence show that the the least squares estimate of β has asymptotic relative efficiency $i/v \times 100\%$.
(c) Show that the cumulant-generating function of the Gumbel distribution, $f(u) = \exp\{-u - \exp(-u)\}$, $-\infty < u < \infty$, is $\log \Gamma(1-t)$, and deduce that its variance is roughly 1.65. Find i for this distribution, and show that the asymptotic relative efficiency of least squares is about 61%.

With $\Gamma(t) = \int_0^\infty u^{t-1} e^{-u}\, du$, $\Gamma''(1) - \Gamma'(1)^2 \doteq 1.64493$.

18 Over a period of 90 days a study was carried out on 1500 women. Its purpose was to investigate the relation between obstetrical practices and the time spent in the delivery suite by women giving birth. One thing that greatly affects this time is whether or not a woman has previously given birth. Unfortunately this vital information was lost, giving the researchers three options: (a) abandon the study; (b) go back to the medical records and find which women had previously given birth (very time-consuming); or (c) for each day check how many women had previously given birth (relatively quick). The statistical question arising was whether (c) would recover enough information about the parameter of interest. Suppose that a linear model is appropriate for log time in delivery suite, and that the log time for a first delivery is normally distributed with mean $\mu + \alpha$ and variance σ^2, whereas for subsequent deliveries the mean time is μ. Suppose that the times for all the women are independent, and that for each there is

a probability π that the labour is her first, independent of the others. Further suppose that the women are divided into k groups corresponding to days and that each group has size m; the overall number is $n = mk$. Under (c), show that the average log time on day j, Z_j, is normally distributed with mean $\mu + R_j\alpha/m$ and variance σ^2/m, where R_j is binomial with probability π and denominator m. Hence show that the overall log likelihood is

$$\ell(\mu, \alpha) = -\tfrac{1}{2}k \log\left(2\pi\sigma^2/m\right) - \frac{m}{2\sigma^2}\sum_{j=1}^{k}(z_j - \mu - r_j\alpha/m)^2,$$

where z_j and r_j are the observed values of Z_j and R_j and we take π and σ^2 to be known. If R_j has mean $m\pi$ and variance $m\tau^2$, show that the inverse expected information matrix is

$$I(\mu, \alpha)^{-1} = \frac{\sigma^2}{n\tau^2}\begin{pmatrix} m\pi^2 + \tau^2 & -m\pi \\ -m\pi & m \end{pmatrix}.$$

(i) If $m = 1$, $\tau^2 = \pi(1 - \pi)$, and $\pi = n_1/n$, where $n = n_0 + n_1$, show that $I(\mu, \alpha)^{-1}$ equals the variance matrix for the two-sample regression model. Explain why.
(ii) If $\tau^2 = 0$, show that neither μ nor α is estimable; explain why.
(iii) If $\tau^2 = \pi(1 - \pi)$, show that μ is not estimable when $\pi = 1$, and that α is not estimable when $\pi = 0$ or $\pi = 1$. Explain why the conditions for these two parameters to be estimable differ in form.
(iv) Show that the effect of grouping, $(m > 1)$, is that $\mathrm{var}(\widehat{\alpha})$ is increased by a factor m regardless of π and σ^2.
(v) It was known that $\sigma^2 \doteq 0.2$, $m \doteq 1500/90$, $\pi \doteq 0.3$. Calculate the standard error for $\widehat{\alpha}$.
It was known from other studies that first deliveries are typically 20–25% longer than subsequent ones. Show that an effect of size $\alpha = \log(1.25)$ would be very likely to be detected based on the grouped data, but that an effect of size $\alpha = \log(1.20)$ would be less certain to be detected, and discuss the implications.

19 Suppose that model $y = X\beta + Z\gamma + \varepsilon$ holds, but that model $y = X\beta + \varepsilon$ is fitted, giving $\widehat{\beta} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$ with hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$ and residuals $e = y - X\widehat{\beta}$.
(a) Show that

$$e = (I - H)y = (I - H)Z\gamma + (I - H)\varepsilon,$$

and hence that $\mathrm{E}(e) = (I - H)Z\gamma$. What happens if Z lies in the space spanned by the columns of X?
(b) Now suppose that Z is a single column z. Explain how an *added variable plot* of the residuals from the regression of y on X against the residuals from the regression of z on X can help in deciding whether or not to add z to the design matrix.
(c) Discuss the interpretation of the added variable plots in Figure 8.10, bearing in mind the possibility of outliers and of a need to transform z before including it in the design matrix.

20 Figure 8.11 shows standardized residuals plotted against fitted values for linear models fitted to four different sets of data. In each case discuss the fit and explain briefly how you would try to remedy any deficiencies.

21 Data $(x_1, y_1), \ldots, (x_n, y_n)$ satisfy the straight-line regression model (5.3). In a

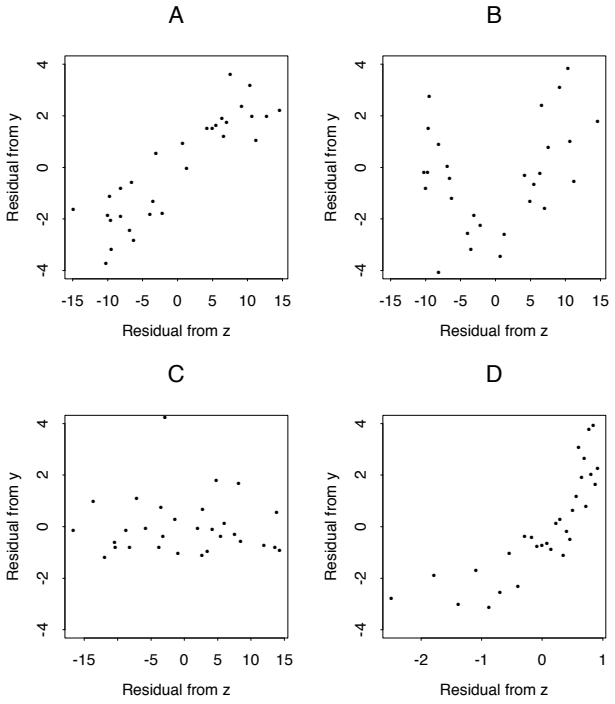


Figure 8.10 Added variable plots for four normal linear models.

calibration problem the value y_+ of a new response independent of the existing data has been observed, and inference is required for the unknown corresponding value x_+ of x.
(a) Let $s_x^2 = \sum(x_j - \overline{x})^2$ and let S^2 be the unbiased estimator of the error variance σ^2. Show that

$$T(x_+) = \frac{Y_+ - \widehat{\gamma}_0 - \widehat{\gamma}_1(x_+ - \overline{x})}{[S^2\{1 + n^{-1} + (x_+ - \overline{x})^2/s_x^2\}]^{1/2}}$$

is a pivot, and explain why the set

$$\mathcal{X}_{1-2\alpha} = \{x_+ : t_{n-2}(\alpha) \leq T(x_+) \leq t_{n-2}(1-\alpha)\}$$

contains x_+ with probability $1 - 2\alpha$.
(b) Show that the function $g(u) = (a + bu)/(c + u^2)^{1/2}$, $c > 0$, $a, b \neq 0$, has exactly one stationary point, at $\tilde{u} = -bc/a$, that sign $g(\tilde{u}) =$ sign a, that $g(\tilde{u})$ is a local maximum if $a > 0$ and a local minimum if $a < 0$, and that $\lim_{u\to\pm\infty} g(u) = \mp b$. Hence sketch $g(u)$ in the four possible cases $a, b < 0$, $a, b > 0$, $a < 0 < b$ and $b < 0 < a$.
(c) By setting $u = S(x_+ - \overline{x})/s_x$, show that $T(x_+)$ can be written in form $g(u)$. Deduce that $\mathcal{X}_{1-2\alpha}$ can be a finite interval, two semi-infinite intervals or the entire real line. Discuss.
(d) Show that if in fact $\gamma_1 = 0$, $\mathcal{X}_{1-2\alpha}$ has infinite length with probability $1-2\alpha$.
(e) A different approach considers x_+ to be an unknown parameter, and constructs the likelihood for β, σ^2 and x_+ based on the pairs (x_j, y_j) and y_+. Does the resulting profile log likelihood $\ell_{\mathrm{p}}(x_+)$ result in confidence sets such as those in (c)?

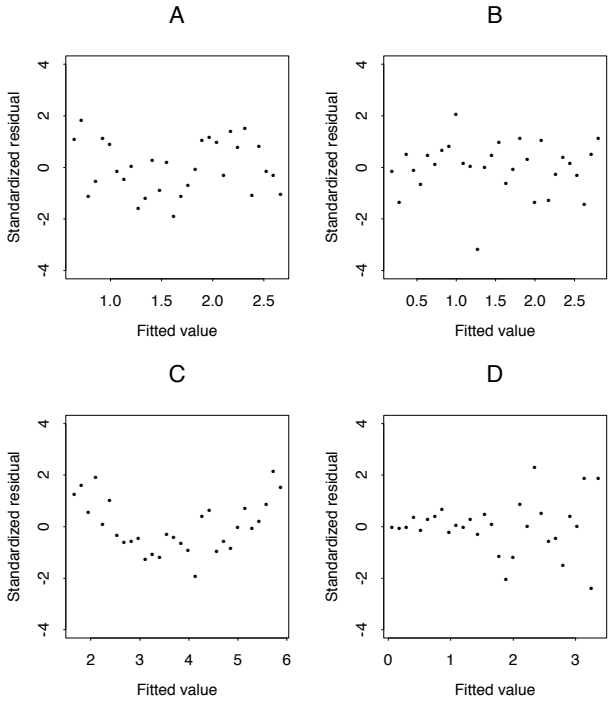


Figure 8.11
Standardized residuals plotted against fitted values for four normal linear models.