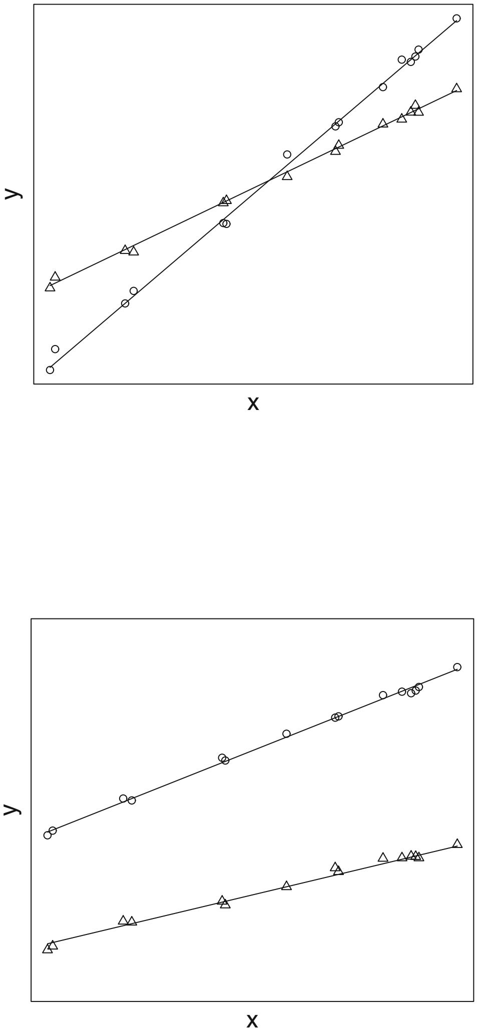

Figure 5.3 Top panel: Interaction term leads to convergence and then crossover for increasing x. Bottom panel: Interaction term leads to divergence of treatment effects.

We can then test for treatment effect, by testing

$$H_0 : \beta_2 = 0.$$

If the treatment ($\beta_2 z$) term is not significant, we can reduce again, to

$$\mu_i = \beta_0 + \beta_1 x_i \qquad \text{(common line)}.$$

We could, for completeness, then test for an age effect, by testing

$$H_0 : \beta_1 = 0$$

(though usually we would not do this – we know blood pressure *does* increase with age). The final, minimal model is

$$\mu_i = \beta_0.$$

These four models – with one, two, three and four parameters – are *nested models*. Each is successively a *sub-model* of the 'one above', with one more parameter. Equally, we have *nested hypotheses*

$$\begin{gathered} \beta_3 = 0, \\ \beta_2 (= \beta_3) = 0, \\ \beta_1 (= \beta_2 = \beta_3) = 0. \end{gathered}$$

Note 5.7

In the medical context above, we are interested in *treatments* (which is the better?). But we are only able to test for a treatment effect if there is *no interaction*. Otherwise, it is not a question of the *better treatment*, but of which treatment is better *for whom*.

5.2.1 Nested Models

Update. Using a full model, we may wish to simplify it by deleting non-significant terms. Some computer packages allow one to do this by using a special command. In S-Plus/R® the relevant command is `update`. F-tests for nested models may simply be performed as follows:

```
m1.lm<-lm(y~x variables)
m2.lm<-update(a.lm, ~. -x variables to be deleted)
anova(m1.lm, m2.lm, test="F")
```

Note the syntax: to delete a term, use `update` and

, ~ . - "comma tilde dot minus".

Akaike Information Criterion (AIC). If there are p parameters in the model,

$$AIC := -2\text{log-likelihood} + 2(p+1)$$

(p parameters, plus one for σ^2, the unknown variance). We then choose between competing models by trying to *minimise* AIC. The AIC is a *penalised log-likelihood*, penalised by the number of parameters (H. Akaike (1927–) in 1974).

The situation is like that of polynomial regression (§4.1). Adding more parameters gives a better fit. But, the Principle of Parsimony tells us to use as few parameters as possible. AIC gives a sensible compromise between

bad fit, over-simplification, too few parameters, and
good fit, over-interpretation, too many parameters.

Step. One can test the various sub-models nested within the full model automatically in S-Plus, by using the command `step`. This uses AIC to drop non-significant terms (Principle of Parsimony: the fewer terms, the better). The idea is to start with the *full* model, and end up with the *minimal adequate* model.

Unfortunately, it matters in what *order* the regressors or factors are specified in our current model. This is particularly true in *ill-conditioned* situations (Chapter 7), where the problem is numerically unstable. This is usually caused by *multicollinearity* (some regressors being nearly linear combinations of others). We will discuss multicollinearity and associated problems in more detail in Chapter 7. F-tests for nested models and stepwise methods for model selection are further discussed in Chapter 6.

5.3 Examples

Example 5.8 (Photoperiod example revisited)

Here we suppose that the data in Exercises 2.4 and 2.9 can be laid out as in Table 5.1 – we assume we have quantitative rather than purely qualitative information about the length of time that plants are exposed to light. We demonstrate that Analysis of Covariance can lead to a flexible class of models by combining methods from earlier chapters on regression and Analysis of Variance.

The simplest model that we consider is *Growth∼Genotype+Photoperiod*. This model has a different intercept for each different genotype. However, length of exposure to light is assumed to have the same effect on each plant irrespective of genotype. We can test for the significance of each term using an Analysis of Variance formulation analogous to the construction in Chapter 2. The sums-of-squares calculations are as follows. The total sum of squares and the genotype

Photoperiod	8h	12h	16h	24h
Genotype A	2	3	3	4
Genotype B	3	4	5	6
Genotype C	1	2	1	2
Genotype D	1	1	2	2
Genotype E	2	2	2	2
Genotype F	1	1	2	3

Table 5.1 Data for Example 5.8

sum of squares are calculated in exact accordance with the earlier analysis-of-variance calculations in Chapter 2:

$$\begin{aligned} SS &= 175 - (1/24)57^2 = 39.625, \\ SSG &= (1/4)(12^2 + 18^2 + 6^2 + 6^2 + 8^2 + 7^2) - (1/24)57^2 = 27.875. \end{aligned}$$

As before we have 23 total degrees of freedom and 5 degrees of freedom for genotype. In Chapter 1 we saw that the sum of squares explained by regression is given by

$$SSR := \sum\nolimits_i (\hat{y}_i - \overline{y})^2 = \frac{S_{xy}^2}{S_{xx}}.$$

Since photoperiod is now assumed to be a quantitative variable, we have only one degree of freedom in the ANOVA table. The sum-of-squares calculation for photoperiod becomes $77^2/840 = 7.058$. As before, the residual sum of squares is calculated by subtraction.

In the notation of Theorem 5.3 we find that

$$Z = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 8 \\ 1 & 12 \\ 1 & 16 \\ 1 & 24 \\ 1 & 8 \\ 1 & 12 \\ 1 & 16 \\ 1 & 24 \\ 1 & 8 \\ 1 & 12 \\ 1 & 16 \\ 1 & 24 \\ 1 & 8 \\ 1 & 12 \\ 1 & 16 \\ 1 & 24 \\ 1 & 8 \\ 1 & 12 \\ 1 & 16 \\ 1 & 24 \\ 1 & 8 \\ 1 & 12 \\ 1 & 16 \\ 1 & 24 \end{pmatrix}.$$

Using $\hat{\gamma}_A = (Z^T RZ)^{-1} Z^T RY$ gives

$$\hat{\gamma}_A = \begin{pmatrix} 1.5 \\ -1.5 \\ -1.5 \\ -1 \\ -1.25 \end{pmatrix}.$$

The regression sum of squares for genotype can then be calculated as $\hat{\gamma}_A Z^T RY = 27.875$ and we obtain, by subtraction, the resulting ANOVA table in Table 5.2. All terms for photoperiod and genotype are significant and we appear to need a different intercept term for each genotype.

A second model that we consider is *Photoperiod*∼*Genotype***Photoperiod*. This model is a more complicated extension of the first, allowing for the possibility of different intercepts *and* different slopes, dependent on genotype. As before, the degrees of freedom multiply to give five degrees of freedom for this

Source	df	Sum of Squares	Mean Square	F	p
Photoperiod	1	7.058	7.058	25.576	0.000
Genotype	5	27.875	5.575	20.201	0.000
Residual	17	4.692	0.276		
Total	23	39.625			

Table 5.2 ANOVA table for different intercepts model

interaction term. The sum-of-squares term of the Genotype:Photoperiod interaction term can be calculated as follows. In the notation of Theorem 5.3, we now have

$$Z = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 \\ 12 & 0 & 0 & 0 & 0 \\ 16 & 0 & 0 & 0 & 0 \\ 24 & 0 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 & 0 \\ 0 & 12 & 0 & 0 & 0 \\ 0 & 16 & 0 & 0 & 0 \\ 0 & 24 & 0 & 0 & 0 \\ 0 & 0 & 8 & 0 & 0 \\ 0 & 0 & 12 & 0 & 0 \\ 0 & 0 & 16 & 0 & 0 \\ 0 & 0 & 24 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 12 & 0 \\ 0 & 0 & 0 & 16 & 0 \\ 0 & 0 & 0 & 24 & 0 \\ 0 & 0 & 0 & 0 & 8 \\ 0 & 0 & 0 & 0 & 12 \\ 0 & 0 & 0 & 0 & 16 \\ 0 & 0 & 0 & 0 & 24 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 8 & 0 & 0 & 0 & 0 & 0 \\ 1 & 12 & 0 & 0 & 0 & 0 & 0 \\ 1 & 16 & 0 & 0 & 0 & 0 & 0 \\ 1 & 24 & 0 & 0 & 0 & 0 & 0 \\ 1 & 8 & 1 & 0 & 0 & 0 & 0 \\ 1 & 12 & 1 & 0 & 0 & 0 & 0 \\ 1 & 16 & 1 & 0 & 0 & 0 & 0 \\ 1 & 24 & 1 & 0 & 0 & 0 & 0 \\ 1 & 8 & 0 & 1 & 0 & 0 & 0 \\ 1 & 12 & 0 & 1 & 0 & 0 & 0 \\ 1 & 16 & 0 & 1 & 0 & 0 & 0 \\ 1 & 24 & 0 & 1 & 0 & 0 & 0 \\ 1 & 8 & 0 & 0 & 1 & 0 & 0 \\ 1 & 12 & 0 & 0 & 1 & 0 & 0 \\ 1 & 16 & 0 & 0 & 1 & 0 & 0 \\ 1 & 24 & 0 & 0 & 1 & 0 & 0 \\ 1 & 8 & 0 & 0 & 0 & 1 & 0 \\ 1 & 12 & 0 & 0 & 0 & 1 & 0 \\ 1 & 16 & 0 & 0 & 0 & 1 & 0 \\ 1 & 24 & 0 & 0 & 0 & 1 & 0 \\ 1 & 8 & 0 & 0 & 0 & 0 & 1 \\ 1 & 12 & 0 & 0 & 0 & 0 & 1 \\ 1 & 16 & 0 & 0 & 0 & 0 & 1 \\ 1 & 24 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

$\hat{\gamma}_A = (Z^T RZ)^{-1} Z^T RY$ gives

$$\hat{\gamma}_A = \begin{pmatrix} 0.071 \\ -0.071 \\ -0.043 \\ -0.114 \\ 0.021 \end{pmatrix}.$$

The sum of squares for the Genotype:Photoperiod term (Gen:Phot.) can then be calculated as $\hat{\gamma}_A Z^T RY = 3.149$ and we obtain the ANOVA table shown in Table 5.3. We see that the Genotype:Photoperiod interaction term is significant and the model with different slopes and different intercepts offers an improvement over the simpler model with just one slope but different intercepts.

Source	df	Sum of Squares	Mean Square	F	p
Photoperiod	1	7.058	7.058	54.898	0.000
Genotype		5	5.575	43.361	0.000
Gen:Phot. (Different slopes)	5	3.149	0.630	4.898	0.011
Residual	12	1.543	0.129		
Total	23	39.625			

Table 5.3 ANOVA table for model with different intercepts and different slopes

Example 5.9 (Exercise 1.6 revisited)

We saw a covert Analysis of Covariance example as early as the Exercises at the end of Chapter 1, in the half-marathon times in Table 1.2. The first model we consider is a model with different intercepts. The sum of squares for age is $114.795^2/747.5 = 17.629$. Fitting the model suggested in part (ii) of Exercise 1.6 gives a residual sum of squares of 43.679. The total sum of squares is $SS = 136.114$. Substituting gives a sum of squares of $136.114 - 43.679 - 17.629 = 74.805$ for club status. This result can alternatively be obtained as follows. We have that

$$\begin{aligned} Z &= (0,0,0,1,1,1,1,1,1,1,1,1,1,1)^T, \\ X &= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 42 & 43 & 44 & 46 & 48 & 49 & 50 & 51 & 57 & 59 & 60 & 61 & 62 & 63 \end{pmatrix}^T. \end{aligned}$$

We have that $\hat{\gamma}_A = (Z^T RZ)^{-1} Z^T RY = -7.673$ and the sum of squares for club status can be calculated as $\hat{\gamma}_A (Z^T RY) = (-7.673)(-9.749) = 74.805.$

The ANOVA table obtained is shown in Table 5.4. The term for club status is significant, but the age term is borderline insignificant. The calculations for the model with two different slopes according to club status is left as an exercise (see Exercise 5.1).

Source	df	Sum of squares	Mean Square	F	p
Age	1	17.629	17.629	4.440	0.059
Club membership	1	74.805	74.805	18.839	0.001
Residual	11	43.679	3.971		
Total	13	136.114			

Table 5.4 ANOVA table for different intercepts model

EXERCISES

5.1. Produce the ANOVA table for the model with different slopes for the data in Example 5.9.

5.2. In the notation of Theorem 5.3 show that

$$\mathrm{var}\left(\hat{\delta}_A\right) = \begin{pmatrix} (X^TX)^{-1} - LML^T & +LM \\ -ML^T & M \end{pmatrix},$$

where $M = (Z^TRZ)^{-1}$.

5.3. Suppose $Y_1, \ldots, Y_n$ are iid $N(\alpha, \sigma^2)$.
(i) Find the least-squares estimate of α.
(ii) Use Theorem 5.3 to estimate the augmented model

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

and verify the formulae for the estimates of the simple linear regression model in Chapter 1.

5.4. Repeat the analysis in Chapter 5.3 in S-Plus/R® using the commands `update` and `anova`.

5.5. The data in Table 5.5 come from an experiment measuring enzymatic reaction rates for treated (State=1) and untreated (State=0) cells exposed to different concentrations of substrate. Fit an Analysis of Covariance model to this data and interpret your findings.

State=0		State=1	
Concentration	Rate	Concentration	Rate
0.02	67	0.02	76
0.02	51	0.02	47
0.06	84	0.06	97
0.06	86	0.06	107
0.11	98	0.11	123
0.11	115	0.11	139
0.22	131	0.22	159
0.22	124	0.22	152
0.56	144	0.56	191
0.56	158	0.56	201
1.10	160	1.10	207
		1.10	200

Table 5.5 Data for Exercise 5.5

5.6. *ANCOVA on the log-scale.* Plot the data in Exercise 5.5. Does the assumption of a linear relationship appear reasonable? Log-transform both the independent variable and the response and try again. (This suggests a power-law relationship; these are extremely prevalent in the physical sciences.) Fit an Analysis of Covariance model and write out your final fitted model for the experimental rate of reaction.

5.7. The data in Table 5.6 is telephone usage (in 1000s) in various parts of the world. Fit an Analysis of Covariance model to the logged data, with time as an explanatory variable, using a different intercept term for each region. Test this model against the model with a different intercept *and* a different slope for each country.

	N. Am.	Europe	Asia	S. Am.	Oceania	Africa	Mid Am.
51	45939	21574	2876	1815	1646	89	555
56	60423	29990	4708	2568	2366	1411	733
57	64721	32510	5230	2695	2526	1546	773
58	68484	35218	6662	2845	2691	1663	836
59	71799	37598	6856	3000	2868	1769	911
60	76036	40341	8220	3145	3054	1905	1008
61	79831	43173	9053	3338	3224	2005	1076

Table 5.6 Data for Exercise 5.7

5.8. *Quadratic Analysis of Covariance model.* Suppose we have one explanatory variable X but that the data can also be split into two categories as denoted by a dummy variable Z. Write

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \gamma_0 Z + \gamma_1 ZX + \gamma_2 ZX^2 + \epsilon.$$

In addition to the possibility of different intercepts and different slopes this model allows for additional curvature, which can take different forms in each category. Suppose the first k observations are from the first category ($Z = 0$) and the remaining $n - k$ are from the second category ($Z = 1$).
(i) Write down the X matrix for this model.
Suggest appropriate F-tests to test:
(ii) The need for both quadratic terms,
(iii) The hypothesis $\gamma_2 = 0$ assuming $\beta_2 \neq 0$.

5.9. *Probability plots/normal probability plots.* Given an ordered sample x_i, an approximate test of normality can be defined by equating the theoretical and empirical cumulative distribution functions (CDFs):

$$\frac{i}{n} = \Phi\left(\frac{x_i - \mu}{\sigma}\right),$$

where $\Phi(\cdot)$ is the standard normal CDF. In practice, to avoid boundary effects, the approximate relation

$$\frac{i - \frac{1}{2}}{n} = \Phi\left(\frac{x_i - \mu}{\sigma}\right)$$

is often used (a 'continuity correction'; cf. Sheppard's correction, Kendall and Stuart (1977) §3.18–26).
(i) Use this approximate relation to derive a linear relationship and suggest a suitable graphical test of normality.
(ii) The following data represent a simulated sample of size 20 from $N(0, 1)$. Do these values seem reasonable using the above?

-2.501, -1.602, -1.178, -0.797, -0.698, -0.428, -0.156, -0.076, -0.032, 0.214, 0.290, 0.389, 0.469, 0.507, 0.644, 0.697, 0.820, 1.056, 1.145, 2.744

[Hint: In S-Plus/R® you may find the commands `ppoints` and `qqnorm` helpful.]
(iii) A random variable on $[0, L]$ has a power-law distribution if it has probability density $f(x) = ax^b$. Find the value of a and derive

an approximate goodness-of-fit test for this distribution by equating theoretical and empirical CDFs.

5.10. *Segmented/piecewise linear models.* Suppose we have the following data:

$$\begin{aligned} x &= (1, 2, 3, 4, 5, 6, 7, 8, 9), \\ y &= (1.8, 4.3, 5.6, 8.2, 9.1, 10.7, 11.5, 12.2, 14.0). \end{aligned}$$

Suppose it is known that a change-point occurs at $x = 5$, so that observations 1–4 lie on one straight line and observations 5–9 lie on another.

(i) Using dummy variables express this model as a linear model. Write down the X matrix. Fit this model and interpret the fitted parameters.

(ii) Assume that the location of the change-point is unknown and can occur at each of $x = \{4, 5, 6, 7\}$. Which choice of change-point offers the best fit to data?

(iii) Show that for a linear regression model the maximised likelihood function can be written as $\propto SSE$. Hence, show that AIC is equivalent to the penalty function

$$n \ln(SSE) + 2p.$$

Hence, compare the best fitting change-point model with linear and quadratic regression models with no change-point.

6
Linear Hypotheses

6.1 Minimisation Under Constraints

We have seen several examples of hypotheses on models encountered so far. For example, in dealing with polynomial regression §4.1 we met, when dealing with a polynomial model of degree k, the hypothesis that the degree was at most $k-1$ (that is, that the leading coefficient was zero). In Chapter 5, we encountered nested models, for example two general lines, including two parallel lines. We then met the hypothesis that the slopes were in fact equal (and so the lines were parallel). We can also conduct a statistical check of structural constraints (for instance, that the angles of a triangle sum to two right-angles – see Exercise 6.5).

We thus need to formulate a general framework for hypotheses of this kind, and for testing them. Since the whole thrust of the subject of regression is linearity, it is to be expected that our attention focuses on linear hypotheses.

The important quantities are the parameters β_i, $i = 1, \ldots, p$. Thus one expects to be testing hypotheses which impose linear constraints on these parameters. We shall be able to test k such constraints, where $k \leq p$. Assembling these into matrix form, we shall test a *linear hypothesis* (with respect to the parameters) of the matrix form

$$B\beta = c. \qquad (hyp)$$

Here B is a $k \times p$ matrix, β is the $p \times 1$ vector of parameters, and c is a $k \times 1$ vector of constants. We assume that matrix B has full rank: if not, there are linear

N.H. Bingham and J.M. Fry, *Regression: Linear Models in Statistics*,
Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-84882-969-5_6,

dependencies between rows of B; we then avoid redundancy by eliminating dependent rows, until remaining rows are linearly independent and B has full rank. Since $k \leq p$, we thus have that B has rank k.

We now seek to minimise the total sum of squares SS, with respect to variation of the parameters β, subject to the constraint (hyp). Now by (SSD) of §3.4,

$$SS = SSR + SSE.$$

Here SSE is a statistic, and can be calculated from the data y; it does not involve the unknown parameters β. Thus our task is actually to

$$\text{minimise} \qquad SSR = (\hat{\beta} - \beta)^T C (\hat{\beta} - \beta) \qquad \text{under} \qquad B\beta = c.$$

This *constrained minimisation* problem is solved by introducing *Lagrange multipliers*, $\lambda_1, \ldots, \lambda_k$, one for each component of the constraint equation (hyp). We solve instead the *unconstrained mimimisation problem*

$$\min \qquad \frac{1}{2} SSR + \lambda^T (B\beta - c),$$

where λ is the k-vector with ith component λ_i. Readers unfamiliar with Lagrange multipliers are advised to take the method on trust for the moment: we will soon produce our minimising value, and demonstrate that it does indeed achieve the minimum – or see e.g. Dineen (2001), Ch. 3 or Ostaszewski (1990), §15.6. (See also Exercises 6.4–6.6.) That is, we solve

$$\min \qquad \frac{1}{2} \sum \sum_{i,j=1}^{p} c_{ij} \left(\hat{\beta}_i - \beta_i\right) \left(\hat{\beta}_j - \beta_j\right) + \sum_{i=1}^{k} \lambda_j \left(\sum_{j=1}^{p} b_{ij} \beta_j - c_i \right).$$

For each $r = 1, \ldots, k$, we differentiate partially with respect to β_r and equate the result to zero. The double sum gives two terms, one with $i = r$ and one with $j = r$; as $C = (c_{ij})$ is symmetric, we obtain

$$-\sum_j c_{jr} \left(\hat{\beta}_j - \beta_j\right) + \sum_i \lambda_i b_{ir} = 0.$$

The terms above are the rth elements of the vectors $-C(\hat{\beta} - \beta)$ and $B^T \lambda$. So we may write this system of equations in matrix form as

$$B^T \lambda = C \left(\hat{\beta} - \beta\right). \qquad (a)$$

Now C is positive definite, so C^{-1} exists. Pre-multiply by BC^{-1} (B is $k \times p$, C^{-1} is $p \times p$):

$$BC^{-1} B^T \lambda = B \left(\hat{\beta} - \beta\right) = B\hat{\beta} - c,$$

by (hyp). Since C^{-1} is positive definite ($p \times p$) and B is full rank ($k \times p$), $BC^{-1}B^T$ is positive definite ($k \times k$). So we may solve for λ, obtaining

$$\lambda = \left(BC^{-1}B^T\right)^{-1} (B\hat{\beta} - c). \qquad (b)$$

We may now solve (a) and (b) for β, obtaining

$$\beta = \hat{\beta} - C^{-1}B^T\left(BC^{-1}B^T\right)^{-1}\left(B\hat{\beta} - c\right).$$

This is the required minimising value under (hyp), which we write as $\beta^\dagger$:

$$\beta^\dagger = \hat{\beta} - C^{-1}B^T\left(BC^{-1}B^T\right)^{-1}\left(B\hat{\beta} - c\right). \qquad (c)$$

In $SSR = (\hat{\beta} - \beta)^T C(\hat{\beta} - \beta)$, replace $\hat{\beta} - \beta$ by $(\hat{\beta} - \beta^\dagger) + (\beta^\dagger - \beta)$. This gives two squared terms, and a cross term,

$$2(\beta^\dagger - \beta)^T C(\hat{\beta} - \beta^\dagger),$$

which by (a) is

$$2(\beta^\dagger - \beta)^T B\lambda.$$

But $B\beta = c$ and $B\beta^\dagger = c$, by (hyp). So $B(\beta^\dagger - \beta) = 0$, $(\beta^\dagger - \beta)^T B = 0$, and the cross term is zero. So

$$SSR = (\hat{\beta} - \beta)^T C(\hat{\beta} - \beta) = (\hat{\beta} - \beta^\dagger)^T C(\hat{\beta} - \beta^\dagger) + (\beta^\dagger - \beta)^T C(\beta^\dagger - \beta). \quad (d)$$

The second term on the right is non-negative, and is zero only for $\beta = \beta^\dagger$, giving

Theorem 6.1

Under the linear constraint (hyp), the value

$$\beta^\dagger = \hat{\beta} - C^{-1}B^T(BC^{-1}B^T)^{-1}(B\hat{\beta} - c)$$

is the unique minimising value of the quadratic form SSR in β.
(i) The unique minimum of SS under (hyp) is

$$SS^* = SSR + (\hat{\beta} - \beta^\dagger)^T C(\hat{\beta} - \beta^\dagger).$$

Multiplying (c) by B confirms that $B\beta^\dagger = c$ – that is, that $\beta^\dagger$ does satisfy (hyp). Now (d) shows directly that $\beta^\dagger$ is indeed the minimising value of SSR and so of SS. Thus those unfamiliar with Lagrange multipliers may see directly from (d) that the result of the theorem is true.

Proposition 6.2

$E(SS^*) = (n - p + k)\sigma^2.$

Proof

The matrix B is $k \times p$ ($k \leq p$), and has full rank k. So some $k \times k$ sub–matrix of B is non-singular. We can if necessary relabel columns so that the first k columns form this non-singular $k \times k$ sub–matrix. We can then solve the linear system of equations

$$B\beta = c$$

to find $\beta_1, \ldots, \beta_k$ – in terms of the remaining parameters $\beta_{k+1}, \ldots, \beta_{k+p}$. We can then express SS as a function of these $p - k$ parameters, and solve by ordinary least squares. This is then *unconstrained* least squares with $p - k$ parameters. We can then proceed as in Chapter 3 but with $p - k$ in place of p, obtaining $E(SS^*) = (n - p + k)\sigma^2$. □

6.2 Sum-of-Squares Decomposition and F-Test

Definition 6.3

The *sum of squares for the linear hypothesis*, SSH, is the difference between the constrained minimum SS^* and the unconstrained minimum SSE of SS. Thus

$$SSH := SS^* - SSE = (\hat{\beta} - \beta^\dagger)^T C(\hat{\beta} - \beta^\dagger).$$

We proceed to find its distribution. As usual, we reduce the distribution theory to matrix algebra, using symmetric projections.

Now

$$\hat{\beta} - \beta^\dagger = C^{-1}B^T \left(BC^{-1}B^T\right)^{-1} \left(B\hat{\beta} - c\right),$$

by (i) of the Theorem above. So

$$B\hat{\beta} - c = B\left(\hat{\beta} - \beta\right) + (B\beta - c) = B\left(\hat{\beta} - \beta\right),$$

under the constraint (*hyp*). But

$$\begin{aligned} \hat{\beta} - \beta &= C^{-1}A^T y - \beta \\ &= C^{-1}A^T y - C^{-1}A^T A\beta \\ &= C^{-1}A^T(y - A\beta). \end{aligned}$$

Combining,

$$\hat{\beta} - \beta^\dagger = C^{-1}B^T \left(BC^{-1}B^T\right)^{-1} BC^{-1}A^T(y - A\beta),$$

so we see that

$$\begin{aligned}\left(\hat{\beta}-\beta^{\dagger}\right)^{T} C &= (y-A\beta)^{T} AC^{-1}B^{T}\left(BC^{-1}B^{T}\right)BC^{-1}C \\ &= (y-A\beta)^{T} AC^{-1}B^{T}\left(BC^{-1}B^{T}\right)B.\end{aligned}$$

Substituting these two expressions into the definition of SSH above, we see that SSH is

$$(y-A\beta)^{T} AC^{-1}B^{T}\left(BC^{-1}B^{T}\right)^{-1}B.C^{-1}B^{T}\left(BC^{-1}B^{T}\right)^{-1}BC^{-1}A^{T}(y-A\beta),$$

which simplifies, giving

$$SSH = (y-A\beta)^{T} D(y-A\beta),$$

say, where

$$D := AC^{-1}B^{T}\left(BC^{-1}B^{T}\right)^{-1}BC^{-1}A^{T}.$$

Now matrix D is *symmetric*, and

$$D^{2} = AC^{-1}B^{T}\left(BC^{-1}B^{T}\right)^{-1}BC^{-1}A^{T}.AC^{-1}B^{T}\left(BC^{-1}B^{T}\right)^{-1}BC^{-1}A^{T}$$

which simplifies to

$$\begin{aligned}D^{2} &= AC^{-1}B^{T}\left(BC^{-1}B^{T}\right)^{-1}BC^{-1}A^{T} \\ &= D,\end{aligned}$$

so D is also *idempotent*. So its rank is its trace, and D is a symmetric projection.

By the definition of SS^{*}, we have the sum-of-squares decomposition

$$SS^{*} := SSE + SSH.$$

Take expectations:

$$E(SS^{*}) = E(SSE) + E(SSH).$$

But

$$E(SSE) = (n-p)\sigma^{2},$$

by §3.4, and

$$E(SS^{*}) = (n-p+k)\sigma^{2},$$

by Proposition 6.2 above. Combining,

$$E(SSH) = k\sigma^{2}.$$

Since SSH is a quadratic form in normal variates with matrix D, a symmetric projection, this shows as in §3.5.1, that D has rank k:

$$rank(D) = \text{trace}(D) = k,$$

the number of (scalar) constraints imposed by the (matrix) constraint (*hyp*).

Theorem 6.4 (Sum of Squares for Hypothesis, SSH)

(i) In the sum-of-squares decomposition

$$SS^* := SSE + SSH,$$

the terms on the right are independent.

(ii) The three quadratic forms are chi-square distributed, with

$$SS^*/\sigma^2 \sim \chi^2(n-p+k), \qquad SSE/\sigma^2 \sim \chi^2(n-p), \qquad SSH/\sigma^2 \sim \chi^2(k).$$

Proof

Since the ranks $n-p$ and k of the matrices of the quadratic forms on the right sum to the rank $n-p+k$ of that on the left, and we already know that quadratic forms in normal variates are chi-square distributed, the independence follows from Chi-Square Decomposition, §3.5. □

We are now ready to formulate a test of our linear hypothesis (*hyp*). This use of Fisher's F distribution to test a general linear hypothesis is due to S. Kołodziejcyzk (d. 1939) in 1935.

Theorem 6.5 (Kołodziejcyzk's Theorem)

We can test our linear hypothesis (*hyp*) by using the F-statistic

$$F := \frac{SSH/k}{SSE/(n-p)},$$

with large values of F evidence against (*hyp*). Thus at significance level α, we use critical region

$$F > F_\alpha(k, n-p),$$

the upper α-point of the Fisher F-distribution $F(k, n-p)$.

Proof

By the result above and the definition of the Fisher F-distribution as the ratio of independent chi-square variates divided by their degrees of freedom, our F-statistic has distribution $F(k, n-p)$. It remains to show that *large* values of F are evidence *against* (*hyp*) – that is, that a one-tailed test is appropriate.

Write

$$w = B\beta - c.$$

Thus $w = 0$ iff the linear hypothesis (hyp) is true; w is non-random, so constant (though unknown, as it involves the unknown parameters β). Now

$$B\hat{\beta} - c = B\left(\hat{\beta} - \beta\right) + (B\beta - c) = B\left(\hat{\beta} - \beta\right) + w.$$

Here $\hat{\beta} - \beta = C^{-1}A^T(y - A\beta)$ has mean zero and covariance matrix $\sigma^2 C^{-1}$ (Proposition 4.4). So $B\hat{\beta} - c$ and $B(\hat{\beta} - \beta)$ have covariance matrix $\sigma^2 BC^{-1}B^T$; $B(\hat{\beta} - \beta)$ has mean zero (as $\hat{\beta}^*$ is unbiased), and $B\beta - c$ has mean w. Now by Theorem 6.1,

$$\begin{aligned} SSH &= (\hat{\beta} - \beta^\dagger)^T C(\hat{\beta} - \beta^\dagger) \\ &= [C^{-1}B^T\left(BC^{-1}B^T\right)^{-1}(B\hat{\beta} - c)]^T C[C^{-1}B^T(BC^{-1}B^T)^{-1}(B\hat{\beta} - c)]. \end{aligned}$$

This is a quadratic form in $B\hat{\beta} - c$ (mean w, covariance matrix $\sigma^2 BC^{-1}B^T$) with matrix

$$(BC^{-1}B^T)^{-1}.BC^{-1}.C.C^{-1}B^T(BC^{-1}B^T)^{-1} = (BC^{-1}B^T)^{-1}.$$

So by the Trace Formula (Prop. 3.22),

$$E(SSH) = \text{trace}[(BC^{-1}B^T)^{-1}.\sigma^2 BC^{-1}B^T] + w^T(BC^{-1}B^T)^{-1}w.$$

The trace term is $\sigma^2\text{trace}(I_k)$ (B is $k \times p$, C^{-1} is $p \times p$, B^T is $p \times k$), or $\sigma^2 k$, giving

$$E(SSH) = \sigma^2 k + w^T(BC^{-1}B^T)^{-1}w.$$

Since C is positive definite, so is C^{-1}, and as B has full rank, so is $(BC^{-1}B^T)^{-1}$. The second term on the right is thus non-negative, and positive unless $w = 0$; that is, unless the linear hypothesis (hyp) is true. Thus large values of $E(SSH)$, so of SSH, so of $F := (SSH/k)/(SSE/(n-p))$, are associated with violation of (hyp). That is, a one-tailed test, rejecting (hyp) if F is too big, is appropriate. □

Note 6.6

The argument above makes no mention of distribution theory. Thus it holds also in the more general situation where we do not assume *normally distributed* errors, only uncorrelated errors with the same variance. A one-tailed F-test is indicated there too. However, the difficulty comes when choosing the critical region – the cut-off level above which we will reject the null hypothesis – the linear hypothesis (hyp). With normal errors, we know that the F-statistic has the F-distribution $F(k, n-p)$, and we can find the cut-off level $F_\alpha(k, n-p)$ using the significance level α and tables of the F-distribution. Without the assumption of normal errors, we do not know the distribution of the F-statistic – so although we still know that large values are evidence against (hyp), we lack a yardstick to tell us 'how big is too big'. In practice, we would probably still use tables of the F-distribution, 'by default'. This raises questions of how close to normality our error distribution is, and how sensitive to departures from normality the distribution of the F-statistic is – that is, how *robust* our procedure is against departures from normality. We leave such robustness questions to the next chapter, but note in passing that Robust Statistics is an important subject in its own right, on which many books have been written; see e.g. Huber (1981).

Note 6.7

To implement this procedure, we need to proceed as follows.

(i) Perform the regression analysis in the 'big' model, Model 1 say, obtaining our SSE, SSE_1 say.

(ii) Perform the regression analysis in the 'little' model, Model 2 say, obtaining similarly SSE_2.

(iii) The big model gives a better fit than the little model; the difference in fit is $SSH := SSE_2 - SSE_1$.

(iv) We normalise the difference in fit SSH by the number k of degrees of freedom by which they differ, obtaining SSH/k.

(v) This is the numerator of our F-statistic. The denominator is SSE_1 divided by its df.

This procedure can easily be implemented by hand – it is after all little more than two regression analyses. Being both so important and so straightforward, it has been packaged, and is automated in most of the major statistical packages.

In S-Plus/R®, for example, this procedure is embedded in the software used whenever we compare two nested models, and in particular in the automated procedures `update` and `step` of §5.2. As we shall see in §6.3 the theory motivates a host of sequential methods to automatically select from the range of possible models.

Example 6.8 (Brownlee's stack loss data)

This data set is famous in statistics for the number of times it has been analysed. The data in Table 6.1 relate stack loss – a measure of inefficiency – to a series of observations. Exploratory data analysis suggests close relationships between Stack Loss and Air Flow and between Water Temperature and Stack Loss.

We wish to test whether or not Acid Concentration can be removed from the model. This becomes a test of the hypothesis $\alpha_3 = 0$ in the model

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \epsilon.$$

Air Flow X_1			Water Temp X_2			Acid Conc. X_3			Stack Loss Y		
80	62	50	27	24	18	89	93	89	42	20	8
80	58	50	27	23	18	88	87	86	37	15	7
75	58	50	25	18	19	90	80	72	37	14	8
62	58	50	24	18	19	87	89	79	28	14	8
62	58	50	22	17	20	87	88	80	18	13	9
62	58	56	23	18	20	87	82	82	18	11	15
62	58	70	24	19	20	93	93	91	19	12	15

Table 6.1 Data for Example 6.8

Fitting the model with all three explanatory variables gives a residual sum of squares of 178.83 on 17 df The model with acid concentration excluded has a residual sum of squares of 188.795 on 16 df Our F-statistic becomes

$$F = \left(\frac{188.795 - 178.83}{1}\right)\left(\frac{16}{188.795}\right) = 0.85.$$

Testing against $F_{1,16}$ gives a p-value of 0.372. Thus, we accept the null hypothesis and conclude that Acid Concentration can be excluded from the model.

6.3 Applications: Sequential Methods

6.3.1 Forward selection

We start with the model containing the constant term. We consider all the explanatory variables in turn, choosing the variable for which SSH is largest. The procedure is repeated for $p = 2, 3, \ldots$, selecting at each stage the variable not currently included in the model with largest F statistic. The procedure terminates when either all variables are included in the model or the maximum F value fails to exceed some threshold F_{IN}.

Example 6.9

We illustrate forward selection by returning to the data in Example 6.8.

Step 1
We compute SSE(Air Flow) = 319.116, SSE(Water Temperature) = 483.151, SSE(Acid concentration) = 1738.442. Air flow is the candidate for entry into the model. $F = 104.201$ against $F_{1,19}$ to give $p = 0.000$ so air flow enters the model.

Step 2
The computations give SSE(Air Flow+Water Temperature) = 188.795 and SSE(Air Flow+Acid Concentration) = 309.1376. Thus, water temperature becomes our candidate for entry into the model. We obtain that $F = 12.425$ and testing against $F_{1,18}$ gives $p = 0.002$ so water temperature enters the model.

Step 3
The F-test of Example 6.8 shows that acid concentration does not enter the model.

6.3.2 Backward selection

Backward selection is an alternative to forward selection. We start using the full model using all p variables (recall $p << n$) and compute the F-statistic with $k = 1$ for each of the p-variables in turn. We eliminate the variable having smallest F-statistic from the model, provided F is less than some threshold F_{OUT}. The procedure is continued until either all the variables are excluded from the model or the smallest F fails to become less than F_{OUT}. When performing forward or backward selection the thresholds F_{IN} and F_{OUT} may change as the algorithms proceed. The most obvious approach is to choose an appropriate formal significance level, e.g. $p = 0.05$, and set the thresholds according to the critical values of the corresponding F-test.

Example 6.10

We illustrate backward selection by returning to the example.

Step 1
The F-test of Example 6.8 excludes acid concentration from the model.

Step 2
The calculations show that SSE(Air Flow +Water Temperature) = 188.795, SSE(Air Flow) = 319.116, SSE(Water Temperature) = 483.151. Thus water temperature becomes our candidate for exclusion. The resulting F-test is the same as in Step 2 of Example 6.9, and we see that no further terms can be excluded from the model.

6.3.3 Stepwise regression

In forward selection, once a variable is included in the model it is not removed. Similarly, in backward selection once a variable is excluded it is never reintroduced. The two algorithms may also give very different results when applied to the same data set. *Stepwise regression* aims to resolve these issues by combining forward selection and backward selection.

The algorithm starts with the simple model consisting solely of a constant term. The first step is a forward selection stage, followed by a backward selection step. The algorithm then alternates between forward and backward selection steps until no further variables are introduced at the forward selection stage. It is shown in Seber and Lee (2003) Ch. 12 that if $F_{OUT} \leq F_{IN}$ then the algorithm must eventually terminate.

Example 6.11 (Example 6.8 re-visited)

The forward selection steps see first Air Flow and then Water Temperature enter the model. Example 6.10 then shows that neither of these variables can be excluded at the backward selection phase. Example 6.8 then shows that Acid Concentration cannot enter the model in the final forward selection phase.

Note 6.12

Some additional discussion of stepwise methods can be found in Seber and Lee (2003), Ch. 12. The S-Plus/R® command `step` uses a variant of the above method based on AIC (§5.2.1), which works both with Linear Models (Chapters 1–7) and Generalised Linear Models (Chapter 8). The command `step` can also be used to perform forward and backward selection by specifying `direction`.

EXERCISES

6.1. Fit regression models to predict fuel consumption for the data set shown in Table 6.2 using
(i) Forward selection
(ii) Backward selection
(iii) Stepwise regression.
T is a qualitative variable taking the value 1 specifying a manual rather than an automatic gearbox. G denotes the number of gears, C denotes the number of carburettors. RAR is the rear-axle ratio, 1/4M t is the time taken to complete a quarter of a mile circuit. Cyls. gives the number of cylinders and Disp. is the car's displacement. (This is a classical data set extracted from the 1974 Motor Trend US magazine, and available as part of the `mtcars` dataset in R®.)

6.2. Show that the first step in forward selection is equivalent to choosing the variable most highly correlated with the response.

6.3. *All-subsets regression.*
(i) Suppose that we have p non-trivial explanatory variables and we always include a constant term. Show that the number of possible models to consider in all–subsets regression is $2^p - 1$.
(ii) How many possible models are suggested in Exercise 6.1?
(iii) Suppose it is feasible to fit no more than 100 regression models. How large does p have to be in order for all-subsets regression to become infeasible?

6.4. *Lagrange multipliers method.* Using the Lagrange multipliers method maximise $f(x, y) := xy$ subject to the constraint $x^2 + 8y^2 = 4$. [Hint: Set $L = xy + \lambda(x^2 + 8y^2 - 4)$, where λ is the Lagrange multiplier, and differentiate with respect to x and y. The resulting solution for λ transforms the constrained problem into an unconstrained problem.]

6.5. *Angles in a triangle.* A surveyor measures three angles of a triangle, α, β, γ $(\alpha + \beta + \gamma = \pi)$. Given one measurement of each of these angles, find the constrained least–squares solution to this problem by using Lagrange multipliers.

6.6. *Angles in a cyclic quadrilateral.* A surveyor measures four angles α, β, γ, δ which are known to satisfy the constraint $\alpha + \beta + \gamma + \delta = 2\pi$. If there is one observation for each of these angles Y_1, Y_2, Y_3, Y_4 say, find the constrained least–squares solution to this problem using Lagrange multipliers.

Mpg	Cyls.	Disp.	Hp	RAR	Weight	1/4M t	v/s	T.	G.	C.
21.0	6	160.0	110	3.90	2.620	16.46	0	1	4	4
21.0	6	160.0	110	3.90	2.875	17.02	0	1	4	4
22.8	4	108.0	93	3.85	2.320	18.61	1	1	4	1
21.4	6	258.0	110	3.08	3.215	19.44	1	0	3	1
18.7	8	360.0	175	3.15	3.440	17.02	0	0	3	2
18.1	6	225.0	105	2.76	3.460	20.22	1	0	3	1
14.3	8	360.0	245	3.21	3.570	15.84	0	0	3	4
24.4	4	146.7	62	3.69	3.190	20.00	1	0	4	2
22.8	4	140.8	95	3.92	3.150	22.90	1	0	4	2
19.2	6	167.6	123	3.92	3.440	18.30	1	0	4	4
17.8	6	167.6	123	3.92	3.440	18.90	1	0	4	4
16.4	8	275.8	180	3.07	4.070	17.40	0	0	3	3
17.3	8	275.8	180	3.07	3.730	17.60	0	0	3	3
15.2	8	275.8	180	3.07	3.780	18.00	0	0	3	3
10.4	8	472.0	205	2.93	5.250	17.98	0	0	3	4
10.4	8	460.0	215	3.00	5.424	17.82	0	0	3	4
4.7	8	440.0	230	3.23	5.345	17.42	0	0	3	4
32.4	4	78.7	66	4.08	2.200	19.47	1	1	4	1
30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
33.9	4	71.1	65	4.22	1.835	19.90	1	1	4	1
21.5	4	120.1	97	3.70	2.465	20.01	1	0	3	1
15.5	8	318.0	150	2.76	3.520	16.87	0	0	3	2
15.2	8	304.0	150	3.15	3.435	17.30	0	0	3	2
13.3	8	350.0	245	3.73	3.840	15.41	0	0	3	4
19.2	8	400.0	175	3.08	3.845	17.05	0	0	3	2
27.3	4	79.0	66	4.08	1.935	18.90	1	1	4	1
26.0	4	120.3	91	4.43	2.140	16.70	0	1	5	2
30.4	4	95.1	113	3.77	1.513	16.90	1	1	5	2
15.8	8	351.0	264	4.22	3.170	14.50	0	1	5	4
19.7	6	145.0	175	3.62	2.770	15.50	0	1	5	6
15.0	8	301.0	335	3.54	3.570	14.60	0	1	5	8
21.4	4	121.0	109	4.11	2.780	18.60	1	1	4	2

Table 6.2 Data for Exercise 6.1

6.7. Show that the regression treatment of one-way ANOVA and the F-test for linear hypotheses returns the original F-test in Theorem 2.8.

6.8. Use a regression formulation and a suitable F-test to test the hypothesis of no differences between treatments in Example 2.9.

6.9. Repeat Exercise 6.1, this time treating the 1/4M time as the dependant variable.

6.10. *Mixtures.* Often chemical experiments involve mixtures of ingredients. This introduces a constraint into the problem, typically of the form

$$x_1 + x_2 + \ldots + x_p = 1.$$

Suppose $x_1, \ldots, x_p$ are from a mixture experiment and satisfy the above constraint.
(i) Reformulate the full main effects model

$$y_i = \beta_0 + \beta_1 x_{1,i} + \ldots + \beta_p x_{p,i} + \epsilon_i,$$

using this constraint.
(ii) Suppose $p = 3$. The usual full second-order model is

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{11} x_1^2 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 \\ &+ \beta_{22} x_2^2 + \beta_{23} x_2 x_3 + \beta_{33} x_3^2 + \epsilon. \end{aligned}$$

Using your answer to (i) suggest a possible way to estimate this model. What is the general solution to this problem for $p \neq 3$?

6.11. *Testing linear hypotheses.*
(i) Test for the need to use a quadratic model in order to describe the following mixture experiment. $x_1 = (1, 0, 0, 0.5, 0.5, 0, 0.2, 0.3)$, $x_2 = (0, 1, 0, 0.5, 0, 0.5, 0.6, 0.5)$, $x_3 = (0, 0, 1, 0, 0.5, 0.5, 0.2, 0.2)$, $y = (40.9, 25.5, 28.6, 31.1, 24.9, 29.1, 27.0, 28.4)$.
(ii) Suppose we have the following data $x_1 = (-1, -1, 0, 1, 1)$, $x_2 = (-1, 0, 0, 0, 1)$, $y = (7.2, 8.1, 9.8, 12.3, 12.9)$. Fit the model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. Test the hypothesis that $\beta_1 = 2\beta_2$. Explain how this constrained model may be fitted using simple linear regression.

7
Model Checking and Transformation of Data

7.1 Deviations from Standard Assumptions

In the above, we have assumed several things:

(i) the mean $\mu = Ey$ is a linear function of the regressors, or of the parameters;

(ii) the errors are additive;

(iii) the errors are independent;

(iv) the errors are normally distributed (Gaussian);

(v) the errors have equal variance.

Any or all of these assumptions may be inadequate. We turn now to a discussion of how to assess the adequacy of our assumptions, and to what we can do when they are inadequate.

Residual Plots. We saw in §3.6 that the residuals e_i and fitted values y_i^* are independent. So a residual plot of e_i against y_i^* should not show any particular pattern. If it *does*, then this suggests that the model is inadequate.

Scatter Plots. Always begin with EDA. With one regressor, we look at the scatter plot of y_i against x_i. With more than one regressor, one can look at all scatter plots of pairs of variables. In S-Plus, this can be done by using the command `pairs`. For details, see for example the S-Plus Help facility, or Crawley (2002), Ch. 24 (especially p. 432–3).

N.H. Bingham and J.M. Fry, *Regression: Linear Models in Statistics*,
Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-84882-969-5_7,

With two regressors, we have a data cloud in three dimensions. This is a highly typical situation: real life is lived in three spatial dimensions, but we represent it – on paper, or on computer screens – in two dimensions. The mathematics needed for this – the mathematics of computer graphics, or of virtual reality – is based on projective geometry. In S-Plus, the command `brush` allows one, in effect, to 'pick up the data cloud and rotate it' (see the S-Plus Help facility, or Venables and Ripley (2002), for details). This may well reveal important structural features of our data. For example, if the data appears round from one direction, but elliptical from another, this tells one something valuable about its distribution, and may suggest some appropriate transformation of the data.

In higher dimensions, we lose the spatial intuition that comes naturally to us in three dimensions. This is a pity, but is unavoidable: many practical situations involve more than two regressors, and so more than three dimensions. One can still use `pairs` to look at two-dimensional scatter plots, but there are many more of these to look at, and combining these different pieces of visual information is not easy.

In higher dimensions, the technique of Projection Pursuit gives a systematic way of searching for adequate low-dimensional descriptions of the data.

Non-constant Variance. In Figure 7.2 the points 'fan out' towards the right, suggesting that the variance increases with the mean. One possibility is to use *weighted regression* (§4.7). Another possibility is to *transform* the data (see below and Draper and Smith (1998) Ch. 13 for further details).

Unaccounted-for Structure. If there is visible structure present, e.g. curvature, in the residual plot, this suggests that the model is not correct. We should return to the original scatter plot of y against x and reinspect. One possibility is to consider adding an extra term or terms to the model – for example, to try a quadratic rather than a linear fit, etc.

Outliers. These are unusual observations that do not conform to the pattern of the rest of the data. They are always worth *checking* (e.g., has the value been entered correctly, has a digit been mis-transcribed, has a decimal point been slipped, etc.?)

Such outliers may be unreliable, and distort the reliable data. If so, we can *trim* the data to remove them. On the other hand, such points, if genuine, may be highly informative.

The subject of how to get protection against such data contamination by removing aberrant data points is called Robust Statistics (touched on in §5.3). In particular, we can use Robust Regression.

Example 7.1 (Median v Mean)

As a measure of location (or central tendency), using medians rather than means gives us some protection against aberrant data points. Indeed, medians can withstand gross data contamination – up to half the data wrong – without failing completely (up to half the data can go off to infinity without dragging the median off to infinity with them). We say that the median has *breakdown point* 1/2, while the mean has breakdown point zero.

Detecting outliers via residual analysis. Residual analysis can be useful in gauging the extent to which individual observations may be expected to deviate from the underlying fitted model. As above, large residuals may point to problems with the original data. Alternatively they may indicate that a better model is needed, and suggest ways in which this may be achieved. The raw residuals are given by

$$e_i = y_i - x_i\hat{\beta}.$$

Scaled residuals are defined as

$$e_i^* = \frac{e_i}{\sqrt{m_{ii}}},$$

where the m_{ii} are the diagonal elements of the matrix M, where $M = I - P = I - X(X^TX)^{-1}X^T$. Under this construction the scaled residuals should now have equal variances (see Theorem 3.30). Scaled residuals can be further modified to define *standardised* or *internally studentised residuals* defined as

$$s_i = \frac{e_i^*}{\hat{\sigma}}.$$

The distribution of the internally studentised residuals is approximately t_{n-p}. However, the result is not exact since the numerator and denominator are not independent. There is one further type of residual commonly used: the *standardised deletion* or *externally studentised residual.* Suppose we wish to test the influence that observation i has on a fitted regression equation. Deleting observation i and refitting we obtain a *deletion residual*

$$e_{-i} = y_i - x_i^T\hat{\beta}_{-i},$$

where $\hat{\beta}_{-i}$ is the estimate obtained *excluding* observation i. Working as above we can define a *standardised* deletion residual s_{-i}. It can be shown, see e.g. Seber and Lee (2003) Ch. 10, that

$$s_{-i} = \frac{s_i\sqrt{n-p-1}}{\sqrt{n-p-s_i^2}}.$$

Further, if the model is correctly defined, these externally studentised residuals have an *exact* t_{n-p-1} distribution. Residual plots can be generated automatically in S-Plus/R® using the command `plot`. In R® this produces a plot of residuals against fitted values, a normal probability plot of standardised residuals (the relevant command here is `qqnorm`), a plot of the square root of the absolute standardised residuals against fitted values, and a plot of standardised residuals versus leverage with control limits indicating critical values for Cook's distances. (See below for further details.)

Influential Data Points. A point has high *leverage* if omitting it causes a big change in the fit. For example, with one regressor x, an x_i far from $\bar{x}$ with an atypical y_i will have high leverage. The leverage of observation i is given by h_{ii} – the diagonal elements of the hat matrix H or projection matrix P. In R® the leverages can be retrieved using the command `hat`. As an illustration we consider an admittedly contrived example in Huber (1981) and also cited in Atkinson (1985). Data consist of $x = -4, -3, -2, -1, 0, 10$, $y = 2.48, 0.73, -0.04, -1.44, -1.32, 0.00$ and the effect of including or excluding the apparent outlier at $x = 10$ has a dramatic impact upon the line of best fit (see Figure 7.1).

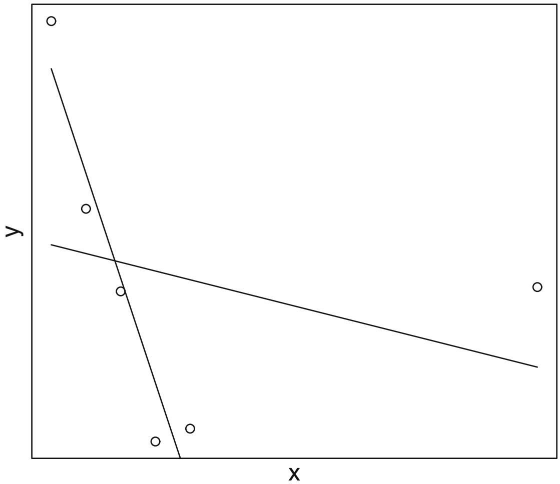

Figure 7.1 Effect of influential observation on line of best fit

Cook's distance. The Cook's distance D_i of observation i combines leverage and residuals – as can be seen from the definition (here $H = (h_{ij}) = P$)

$$D_i = \frac{s_i^2 h_{ii}}{p(1 - h_{ii})}.$$

Large values of Cook's distance occur if an observation is *both* outlying (large s_i) with high leverage (large h_{ii}). Plots of Cook's distance can be obtained as part of the output automatically generated in S-Plus/R® using the command `plot`. It can be shown that

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}_{-i}\right)^T X^T X \left(\hat{\beta} - \hat{\beta}_{-i}\right)}{p\hat{\sigma}^2},$$

where $\hat{\beta}_{-i}$ is the parameter estimate $\hat{\beta}$ obtained when the ith observation is *excluded.* Thus D_i does indeed serve as a measure of the *influence* of observation i. It provides an appropriate measure of the 'distance' from $\hat{\beta}$ to $\hat{\beta}_{-i}$.

Note 7.2

1. For further background on Cook's distance and related matters, we refer to Cook and Weisberg (1982).
2. This 'leave one out' idea is often useful in statistics. It leads to the method of *cross-validation* (CV).

Bias and Mallows's C_p statistic. Suppose we fit the model

$$\mathbf{y} = X_1\beta_1 + \epsilon.$$

This leads to the least-squares estimate $\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T \mathbf{y}$. If our postulated model is correct then this estimate is unbiased (§3.3). Suppose however that the true underlying relationship is

$$\mathbf{y} = X_1\beta_1 + X_2\beta_2 + \epsilon.$$

Our least-squares estimate $\hat{\beta}_1$ now has expected value $\beta_1 + (X_1^T X_1)^{-1} X_1^T X_2 \beta_2$. Omitting X_2 leads to a bias of $(X_1^T X_1)^{-1} X_1^T X_2 \beta_2$. Note that this is 0 if $X_1^T X_2 = 0$, the orthogonality relation we met in §5.1.1 on orthogonal parameters.

Mallows's C_p statistic is defined as

$$C_p = \frac{SSE}{s^2} - (n - 2p),$$

where p is the number of model parameters and s^2 is an estimate of σ^2 obtained from a subjective choice of full model. We consider sub-models of the full model. If a model is approximately correct

$$E(C_p) \approx \frac{(n-p)\sigma^2}{\sigma^2} - (n-2p) = p.$$

If the model is incorrectly specified it is assumed $E(SSE) > \sigma^2$ and $E(C_p) > p$. Models can be compared using this method by plotting C_p against p. Suitable candidate models should lie close to the line $C_p = p$. Note, however that by definition $C_p = p$ for the full model.

Non-additive or non-Gaussian errors. These may be handled using Generalised Linear Models (see Chapter 8). Generalised Linear Models can be fitted in S-Plus and R® using the command `glm`. For background and details, see McCullagh and Nelder (1989).

Correlated Errors. These are always very dangerous in Statistics! *Independent* errors tend to *cancel*. This is the substance of the Law of Large Numbers (LLN), that says

$$\bar{x} \to Ex \qquad (n \to \infty)$$

– sample means tend to population means as sample size increases. Similarly for sample variances and other sample quantities. This is basically why Statistics works. One does not even need to have *independent* errors: weakly dependent errors (which may be defined precisely, in a variety of ways) exhibit similar cancellation behaviour. By contrast, *strongly dependent* errors need *not* cancel. Here, increasing the sample size merely replicates existing readings, and if these are way off this does not help us (as in Note 1.3).

Correlated errors may have some special structure – e.g., in time or in space. Accordingly, one would then have to use special methods to reflect this – Time Series or Spatial Statistics; see Chapter 9. Correlated errors may be detected using the Durbin–Watson test or, more crudely, using a runs test (see Draper and Smith (1998), Ch. 7).

7.2 Transformation of Data

If the residual plot ‘funnels out’ one may try a transformation of data, such as $y \mapsto \log y$ or $y \mapsto \sqrt{y}$ (see Figure 7.2).

If on the other hand the residual plot ‘funnels in’ one may instead try $y \mapsto y^2$, etc (see Figure 7.3).

Is there a general procedure? One such approach was provided in a famous paper Box and Cox (1964). Box and Cox proposed a one-parameter family of

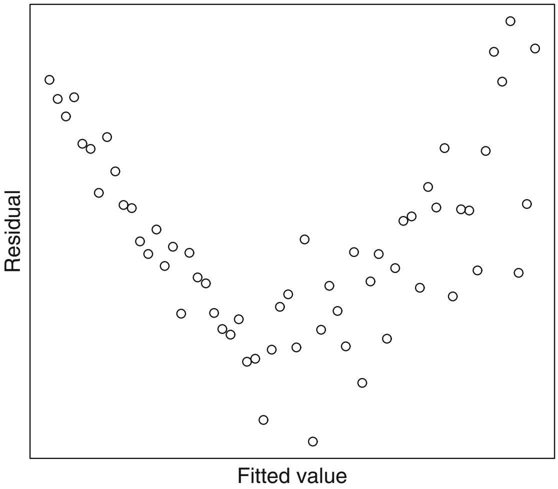

Figure 7.2 Plot showing 'funnelling out' of residuals

power transformations that included a *logarithmic* transformation as a special case. With λ as parameter, this is

$$y \mapsto \begin{cases} (y^\lambda - 1)/\lambda & \text{if} \quad \lambda \neq 0, \\ \log y & \text{if} \quad \lambda = 0. \end{cases}$$

Note that this is an indeterminate form at $\lambda = 0$, but since

$$\frac{y^\lambda - 1}{\lambda} = \frac{e^{\lambda \log y} - 1}{\lambda},$$

$$\frac{d}{d\lambda}\left(e^{\lambda \log y} - 1\right) = \log y . e^{\lambda \log y} = \log y \qquad \text{if } \lambda = 0,$$

L'Hospital's Rule gives

$$(y^\lambda - 1)/\lambda \to \log y \qquad (\lambda \to 0).$$

So we may *define* $(y^\lambda - 1)/\lambda$ as $\log y$ for $\lambda = 0$, to include $\lambda = 0$ with $\lambda \neq 0$ above.

One may – indeed, should – proceed *adaptively* by allowing the *data* to suggest which value of λ might be suitable. This is done in S-Plus by the command `boxcox`.

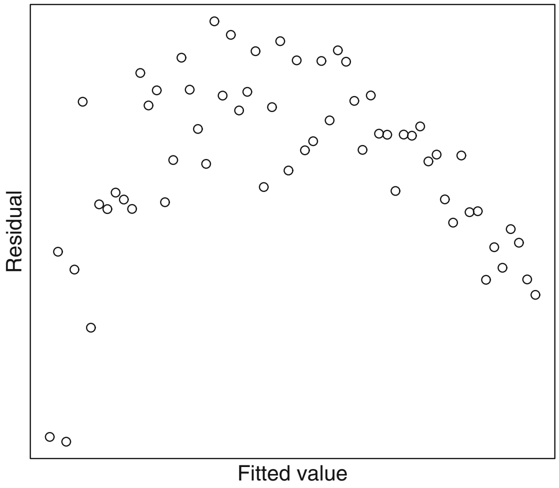

Figure 7.3 Plot showing ‘funnelling in’ of residuals

Example 7.3 (Timber Example)

The value of timber yielded by a tree is the response variable. This is measured only when the tree is cut down and sawn up. To help the forestry worker decide which trees to fell, the predictor variables used are girth (‘circumference’ – though the tree trunks are not perfect circles) and height. These can be easily measured without interfering with the tree – girth by use of a tape measure (at some fixed height above the ground), height by use of a surveying instrument and trigonometry.

Venables and Ripley (2002) contains a data library MASS, which includes a data set timber:

```
attach(timber)
names(timber)
[1] “volume” “girth” “height”
boxcox(volume) ~ (girth + height)
```

Dimensional Analysis. The data-driven choice of Box–Cox parameter λ seems to be close to 1/3. This is predictable on dimensional grounds: volume is in cubic metres, girth and height in metres (or centimetres). It thus always pays to be aware of *units*.

There is a whole subject of *Dimensional Analysis* devoted to such things (see e.g. Focken (1953)). A background in Physics is valuable here.

7.3 Variance-Stabilising Transformations

In the exploratory data analysis (EDA), the scatter plot may suggest that the variance is not constant throughout the range of values of the predictor variable(s). But, the theory of the Linear Model *assumes* constant variance. Where this standing assumption seems to be violated, we may seek a systematic way to *stabilise* the variance – to make it constant (or roughly so), as the theory requires.

If the response variable is y, we do this by seeking a suitable function g (sufficiently smooth – say, twice continuously differentiable), and then *transforming* our data by

$$y \mapsto g(y).$$

Suppose y has mean μ:

$$Ey = \mu.$$

Taylor expand $g(y)$ about $y = \mu$:

$$g(y) = g(\mu) + (y - \mu)g'(\mu) + \frac{1}{2}(y - \mu)^2 g''(\mu) + \ldots$$

Suppose the bulk of the response values y are fairly closely bunched around the mean μ. Then, approximately, we can treat $y - \mu$ as small; then $(y - \mu)^2$ is negligible (at least to a first approximation, which is all we are attempting here). Then

$$g(y) \sim g(\mu) + (y - \mu)g'(\mu).$$

Take expectations: as $Ey = \mu$, the linear term goes out, giving $Eg(y) \sim g(\mu)$. So

$$g(y) - g(\mu) \sim g(y) - Eg(y) \sim g'(\mu)(y - \mu).$$

Square both sides:

$$[g(y) - g(\mu)]^2 \sim [g'(\mu)]^2 (y - \mu)^2.$$

Take expectations: as $Ey = \mu$ and $Eg(y) \sim g(\mu)$, this says

$$\mathrm{var}(g(y)) \sim [g'(\mu)]^2 \mathrm{var}(y).$$

Regression. So if

$$E(y_i|x_i) = \mu_i, \qquad \mathrm{var}(y_i|x_i) = \sigma_i^2,$$

we use EDA to try to find some link between the means μ_i and the variances σ_i^2. Suppose we try $\sigma_i^2 = H(\mu_i)$, or

$$\sigma^2 = H(\mu).$$

Then by above,

$$\mathrm{var}(g(y)) \sim [g'(\mu)]^2\sigma^2 = [g'(\mu)]^2 H(\mu).$$

We want *constant variance*, c^2 say. So we want

$$[g'(\mu)]^2 H(\mu) = c^2, \qquad g'(\mu) = \frac{c}{\sqrt{H(\mu)}}, \qquad g(y) = c\int \frac{dy}{\sqrt{H(y)}}.$$

Note 7.4

The idea of variance-stabilising transformations (like so much else in Statistics!) goes back to Fisher. He found the density of the sample correlation coefficient r^2 in the bivariate normal distribution – a complicated function involving the population correlation coefficient ρ^2, simplifying somewhat in the case $\rho = 0$ (see e.g. Kendall and Stuart (1977), §16.27, 28). But Fisher's z transformation of 1921 (Kendall and Stuart (1977), §16.33)

$$r = \tanh z, \qquad z = \frac{1}{2}\log\left(\frac{1+r}{1-r}\right), \qquad \rho = \tanh\zeta, \qquad \zeta = \frac{1}{2}\log\left(\frac{1+\rho}{1-\rho}\right)$$

gives z approximately normal, with variance almost independent of ρ:

$$z \sim N(0, 1/(n-1)).$$

Taylor's Power Law. The following empirical law was proposed by R. L. Taylor in 1961 (Taylor (1961)):

log variance against log mean is roughly *linear* with slope γ between 1 and 2.

Both these extreme cases can occur. An example of slope 1 is the Poisson distribution, where the mean and the variance are the same. An example of slope 2 occurs with a Gamma-distributed error structure, important in Generalised Linear Models (Chapter 8).

With $H(\mu) = \mu^\gamma$ above, this gives variance

$$v = \sigma^2 = H(\mu) = \mu^\gamma.$$

Transform to

$$g(y) = c\int \frac{dy}{\sqrt{H(y)}} = c\int \frac{dy}{y^{\frac{1}{2}\gamma}} = c\left(y^{1-\frac{1}{2}\gamma} - y_0^{1-\frac{1}{2}\gamma}\right).$$

This is of Box–Cox type, with

$$\lambda = 1 - \frac{1}{2}\gamma.$$

Taylor's suggested range $1 \leq \gamma \leq 2$ gives

$$0 \leq 1 - \frac{1}{2}\gamma \leq \frac{1}{2}.$$

Note that this range includes the logarithmic transformation (Box–Cox, $\lambda = 0$), and the cube–root transformation ($\lambda = 1/3$) in the timber example. Partly for dimensional reasons as above, common choices for λ include $\lambda = -1/2, 0, 1/3, 1/2, (1), 3/2$ (if $\lambda = 1$ we do not need to transform). An empirical choice of λ (e.g. by Box–Cox as above) close to one of these may suggest choosing λ as this value, and/or a theoretical examination with dimensional considerations in mind.

Delta Method. A similar method applies to *reparametrisation.* Suppose we choose a parameter θ. If the true value is θ_0 and the maximum-likelihood estimator is $\hat{\theta}$, then under suitable regularity conditions a central limit theorem (CLT) will hold:

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right)/\sigma \to N(0,1) \qquad (n \to \infty).$$

Now suppose that one wishes to change parameter, and work instead with ϕ, where

$$\phi := g(\theta).$$

Then the same method (Taylor expansion about the mean) enables one to transfer this CLT for our estimate of θ to a CLT for our estimate of ϕ:

$$\sqrt{n}\left(\hat{\phi} - \phi_0\right)/\left(g'\left(\theta_0\right)\sigma\right) \to N(0,1) \qquad (n \to \infty).$$

Example 7.5 (Variance and standard deviation)

It is convenient to be able to change at will from using variance σ^2 as a parameter to using standard deviation σ. Mathematically the change is trivial, and it is also trivial computationally (given a calculator). Using the delta-method, it is statistically straightforward to transfer the results of a maximum-likelihood estimation from one to the other.

7.4 Multicollinearity

Recall the distribution theory of the bivariate normal distribution (§1.5). If we are regressing y on x, *but* y is (exactly) a *linear* function of x, then $\rho = \pm 1$, the bivariate normal density does not exist, and the *two*-dimensional setting is wrong – the situation is really *one*-dimensional. Similar remarks apply for the multivariate normal distribution (§4.3). When we assume the covariance matrix Σ is non-singular, the density exists and is given by Edgeworth's Theorem; when Σ is singular, the density does not exist. The situation is similar again in the context of Multiple Regression in Chapter 3. There, we assumed that the design matrix A ($n \times p$, with $n >> p$) has *full rank* p. A will have defective rank ($< p$) if there are linear relationships between regressors. In all these cases, we have a general situation which is non-degenerate, but which contains a special situation which is degenerate. The right way to handle this is to identify the degeneracy and its cause. By reformulating the problem in a suitably lower dimension, we can change the situation which is degenerate in the higher-dimensional setting into one which is non-degenerate if handled in its natural dimension. To summarise: to escape degeneracy, one needs to identify the linear dependence relationship which causes it. One can then eliminate dependent variables, begin again with only linearly independent variables, and avoid degeneracy.

The problem remains that in Statistics we are handling data, and data are uncertain. Not only do they contain sampling error, but having sampled our data we have to round them (to the number of decimal places or significant figures we – or the default option of our computer package – choose to work to). We may well be in the general situation, where things are non-degenerate, and there are no non-trivial linear dependence relations. *Nevertheless*, there may be *approximate* linear dependence relations. If so, then rounding error may lead us close to degeneracy (or even to it): our problem is then *numerically unstable.* This phenomenon is known as *multicollinearity.*

Multiple Regression is inherently prone to problems of this kind. One reason is that the more regressors we have, the more ways there are for some of them to be at least approximately linearly dependent on others. This will then cause the problems mentioned above. Our best defence against multicollinearity is to be alert to the danger, and in particular to watch for possible approximate linear dependence relations between regressors. If we can identify such, we have made two important gains:

(i) we can avoid the numerical instability associated with multicollinearity, and reduce the dimension and thus the computational complexity,

(ii) we have identified important structural information about the problem by identifying an approximate link between regressors.

The problem of multicollinearity in fact bedevils the whole subject of Multiple Regression, and is surprisingly common. It is one reason why the subject is 'an art as well as a science'. It is also a reason why automated computer procedures such as the S-Plus commands `step` and `update` produce different outcomes depending on the *order* in which variables are declared in the model.

Example 7.6 (Concrete example)

The following example is due to Woods et al. (1932). It is a very good illustration of multicollinearity and how to handle it.

In a study of the production of concrete, the response variable Y is the amount of heat (calories per gram) released while the concrete sets. There are four regressors $X_1, \ldots, X_4$ representing the percentages (by weight rounded to the nearest integer) of the chemically relevant constituents from which the concrete is made. The data are shown in Table 7.1 below.

n	Y	X_1	X_2	X_3	X_4
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.9	10	68	8	12

Table 7.1 Data for concrete example

Here the X_i are not exact percentages, due to rounding error and the presence of between 1% and 5% of other chemically relevant compounds. However, X_1, X_2, X_3, X_4 are rounded percentages and so sum to near 100 (cf. the mixture models of Exercise 6.10). So, strong (negative) correlations are anticipated, and we expect that we will not need all of $X_1, \ldots, X_4$ in our chosen model. In this simple example we can fit models using all possible combinations of variables and the results are shown in Table 7.2. Here we cycle through, using as an intuitive guide the proportion of the variability in the data explained by each model as defined by the R^2 statistic (see Chapter 3).

Model	$100R^2$	Model	$100R^2$	Model	$100R^2$
X_1	53.29	X_1 X_2	97.98	X_1 X_2 X_3	98.32
X_2	66.85	X_1 X_3	54.68	X_1 X_2 X_4	98.32
X_3	28.61	X_1 X_4	97.28	X_1 X_3 X_4	98.2
X_4	67.59	X_2 X_3	84.93	X_2 X_3 X_4	97.33
		X_2 X_4	68.18	X_1 X_2 X_3 X_4	98.32
		X_3 X_4	93.69		

Table 7.2 All-subsets regression for Example 7.6

The multicollinearity is well illustrated by the fact that omitting either X_3 or X_4 from the full model does not seem to have much of an effect. Further, the models with just one term do not appear sufficient. Here the t-tests generated as standard output in many computer software packages, in this case R®[1] using the `summary.lm` command, prove illuminating. When fitting the full model X_1 X_2 X_3 X_4 we obtain the output in Table 7.3 below:

Coefficient	Estimate	Standard Error	t-value	p-value
Intercept	58.683	68.501	0.857	0.417
X_1	1.584	0.728	2.176	0.061
X_2	0.552	0.708	0.780	0.458
X_3	0.134	0.738	0.182	0.860
X_4	-0.107	0.693	-0.154	0.882

Table 7.3 *R* output for Example 7.6

So despite the high value of R^2, tests for individual model components in the model are non-significant. This in itself suggests possible multicollinearity. Looking at Table 7.2, model selection appears to come down to a choice between the best two-term model X_1 X_2 and the best three-term models X_1 X_2 X_3 and X_1 X_2 X_4. When testing X_1 X_2 X_3 versus X_1 X_2 we get a t-statistic of 0.209 for X_3 suggesting that X_3 can be safely excluded from the model. A similar analysis for the X_1 X_2 X_4 gives a p-value of 0.211 suggesting that X_4 can also be safely omitted from the model. Thus, X_1 X_2 appears to be the best model and the multicollinearity inherent in the problem suggests that a model half the

[1] R®: A language and environment for statistical computing. © 2009 R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 http://www.R-project.org

size of the full model will suffice. In larger problems one might suggest using stepwise regression or backward selection starting with the full model, rather than the all-subsets regression approach we considered here.

Regression Diagnostics. A regression analysis is likely to involve an iterative process in which a range of plausible alternative models are examined and compared, before our final model is chosen. This process of *model checking* involves, in particular, looking at unusual or suspicious data points, deficiencies in model fit, etc. This whole process of model examination and criticism is known as *Regression Diagnostics.* For reasons of space, we must refer for background and detail to one of the specialist monographs on the subject, e.g. Atkinson (1985), Atkinson and Riani (2000).

EXERCISES

7.1. Revisit the concrete example using,
(i) stepwise selection starting with the full model,
(ii) backward selection starting with the full model,
(iii) forward selection from the null constant model.

7.2. *Square root transformation for count data.* Counts of rare events are often thought to be approximately Poisson distributed. The transformation $\sqrt{Y}$ or $\sqrt{Y+1}$, if some counts are small, is often thought to be effective in modelling count data. The data in Table 7.4 give a count of the number of poppy plants in oats.
(i) Fit an Analysis of Variance model using the raw data. Does a plot of residuals against fitted values suggest a transformation?
(ii) Interpret the model in (i).
(iii) Re-fit the model in (i-ii) using a square–root transformation. How do your findings change?

Treatment	A	B	C	D	E
Block 1	438	538	77	17	18
Block 2	442	422	61	31	26
Block 3	319	377	157	87	77
Block 4	380	315	52	16	20

Table 7.4 Data for Exercise 7.2

7.3. *Arc sine transformation for proportions.* If we denote the empirical proportions by $\hat{p}$, we replace $\hat{p}$ by introducing the transformation $y = \sin^{-1}(\sqrt{\hat{p}})$. In this angular scale proportions near zero or one are spread out to increase their variance and make the assumption of homogenous errors more realistic. (With small values of $n < 50$ the suggestion is to replace zero or one by $\frac{1}{4n}$ or $1 - \frac{1}{4n}$.) The data in Table 7.5 give the percentage of unusable ears of corn.
(i) Fit an Analysis of Variance model using the raw data. Does a plot of residuals against fitted values suggest a transformation?
(ii) Interpret the model in (i).
(iii) Re-fit the model in (i–ii) using the suggested transformation. How do your findings change?

Block	1	2	3	4	5	6
Treatment A	42.4	34.4	24.1	39.5	55.5	49.1
Treatment B	33.3	33.3	5.0	26.3	30.2	28.6
Treatment C	8.5	21.9	6.2	16.0	13.5	15.4
Treatment D	16.6	19.3	16.6	2.1	11.1	11.1

Table 7.5 Data for Exercise 7.3

7.4. The data in Table 7.6 give the numbers of four kinds of plankton caught in different hauls.
(i) Fit an Analysis of Variance model using the raw data. Does a plot of residuals against fitted values suggest a transformation of the response?
(ii) Calculate the mean and range $(\max(y) - \min(y))$ for each species and repeat using the logged response. Comment.
(iii) Fit an Analysis of Variance model using both raw and logged numbers, and interpret the results.

7.5. Repeat Exercise 7.4 using
(i) The square-root transformation of Exercise 7.2.
(ii) Taylor's power law.

7.6. *The delta method: Approximation formulae for moments of transformed random variables.* Suppose the random vector U satisfies $E(U) = \mu$, $\mathrm{var}(U) = \Sigma_U$, $V = f(U)$ for some smooth function f. Let F_{ij} be the matrix of derivatives defined by

$$F_{ij}(u) = \left(\frac{\partial u}{\partial v}\right)_{ij} = \left(\frac{\partial f}{\partial v}\right)_{ij} = \frac{\partial f_i}{\partial v_j}.$$

Haul	Type I	Type II	Type III	Type IV
1	895	1520	43300	11000
2	540	1610	32800	8600
3	1020	1900	28800	8260
4	470	1350	34600	9830
5	428	980	27800	7600
6	620	1710	32800	9650
7	760	1930	28100	8900
8	537	1960	18900	6060
9	845	1840	31400	10200
10	1050	2410	39500	15500
11	387	1520	29000	9250
12	497	1685	22300	7900

Table 7.6 Data for Exercise 7.4

We wish to construct simple estimates for the mean and variance of V. Set

$$V \approx f(\mu) + F(\mu)(u - \mu).$$

Taking expectations then gives

$$E(V) \approx f(\mu).$$

(i) Show that $\Sigma_V \approx F(\mu)\Sigma_U F(\mu)^T$.
(ii) Let $U{\sim}Po(\mu)$ and $V = \sqrt{U}$. Give approximate expressions for the mean and variance of V.
(iii) Repeat (ii) for $V = \log(U + 1)$. What happens if $\mu >> 1$?

7.7. Show, using the delta method, how you might obtain parameter estimates and estimated standard errors for the power-law model $y = \alpha x^{\beta}$.

7.8. *Analysis using graphics in S-Plus/R®*. Re-examine the plots shown in Figures 7.2 and 7.3. The R®-code which produced these plots is shown below. What is the effect of the commands `xaxt/yaxt="n"`? Use `?par` to see other options. Experiment and produce your own examples to show funnelling out and funnelling in of residuals.

Code for funnels out/in plot

```
y2<-(x2+rnorm(60, 0, 0.7))∧2/y2<-(1+x2+rnorm(60, 0,
     0.35))∧0.5
a.lm<-lm(y2∼x2)
plot(y2-a.lm$resid, a.lm$resid, xaxt‘"n", yaxt="n",
     ylab="Residual", xlab="Fitted value")
```

7.9. For the simple linear model in Exercise 1.6, calculate leverage, Cook's distances, residuals, externally studentised residuals and internally studentised residuals.

7.10. Revisit the simulated data example in Exercise 3.4 using techniques introduced in this chapter.

8
Generalised Linear Models

8.1 Introduction

In previous chapters, we have studied the model

$$y = A\beta + \epsilon,$$

where the mean $Ey = A\beta$ depends linearly on the parameters β, the errors are normal (Gaussian), and the errors are additive. We have also seen (Chapter 7) that in some situations, a transformation of the problem may help to correct some departure from our standard model assumptions. For example, in §7.3 on variance-stabilising transformations, we transformed our data from y to some function $g(y)$, to make the variance constant (at least approximately). We did not there address the effect on the *error structure* of so doing. Of course, $g(y) = g(A\beta + \epsilon)$ as above will *not* have an additive Gaussian error structure any more, even approximately, in general.

The function of this chapter is to generalise linear models beyond our earlier framework, so as to broaden our scope and address such questions. The material is too advanced to allow a full treatment here, and we refer for background and detail to the (numerous) references cited below, in particular to McCullagh and Nelder (1989) and to Venables and Ripley (2002), Ch. 7.

We recall that in earlier chapters the Method of Least Squares and the Method of Maximum Likelihood were equivalent. When we go beyond this framework, this convenient feature is no longer present. We use the Method of Maximum Likelihood (equivalent above to the Method of Least Squares, but no

N.H. Bingham and J.M. Fry, *Regression: Linear Models in Statistics*,
Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-84882-969-5_8,

longer so in general). This involves us in finding the maximum of the likelihood L, or equivalently the log-likelihood $\ell := \log L$, by solving the *likelihood equation*

$$\ell' = 0.$$

Unfortunately, this equation will no longer have a solution in closed form. Instead, we must proceed as we do when solving a transcendental (or even algebraic) equation

$$f(x) = 0,$$

and proceed numerically. The standard procedure is to use an *iterative* method: to begin with some starting value, x_0 say, and improve it by finding some better approximation x_1 to the required root. This procedure can be iterated: to go from a current approximation x_n to a better approximation x_{n+1}. The usual method here is *Newton–Raphson iteration* (or the *tangent method*):

$$x_{n+1} := x_n - f(x_n)/f'(x_n).$$

This effectively replaces the graph of the function f near the point $x = x_n$ by its tangent at x_n. In the context of statistics, the derivative ℓ' of the log-likelihood function is called the *score function*, s, and the use of iterative methods to solve the likelihood equation is called *Fisher's method of scoring* (see e.g. Kendall and Stuart (1979), §18.21).

Implementation of such an iterative solution by hand is highly laborious, and the standard cases have been programmed and implemented in statistical packages. One consequence is that (at least at the undergraduate level relevant here) in order to implement procedures involving Generalised Linear Models (GLMs), one really needs a statistical package which includes them. The package GLIM®[1] is designed with just this in mind (Aitkin *et al.* (1989), or Crawley (1993)), and also GenStat®[2] (McConway et al. (1999)). For S-Plus for GLMs, we refer to Venables and Ripley (2002), Ch. 7, Crawley (2002), Ch. 27. Unfortunately, the package Minitab® (admirably simple, and very useful for much of the material of this book) does not include GLMs.

Generalised Linear Models, or GLMs, arise principally from the work of the English statistician John A. Nelder (1924–2010); the term is due to Nelder and Wedderburn in 1972; the standard work on the subject is McCullagh and Nelder (1989). As noted above, GLMs may be implemented in GLIM® or GenStat®; the relevant command in S-Plus/R® is `glm`, with the family of error distributions specified, as well as the regressors; see below for examples.

[1] GLIM® is a registered trademark of The Royal Statistical Society.

[2] GenStat® is a registered trademark of VSN International Limited, 5 The Waterhouse, Waterhouse Street, Hemel Hempstead, HP1 1ES, UK.

8.2 Definitions and examples

Just as with a linear model, we have regressors, or stimulus variables, $x_1, \ldots, x_p$ say, and a response variable y, which depends on these via a *linear predictor*

$$\eta = \beta_1 x_1 + \ldots + \beta_p x_p,$$

where the β_i are parameters. The mean $\mu = Ey$ depends on this linear predictor η, but whereas in the linear case $\mu = \eta$, we now allow μ to be some smooth invertible function of η, and so also, η is a smooth invertible function of μ. We write

$$\mu = m(\eta), \qquad \eta = m^{-1}(\mu) = g(\mu),$$

where the function g is called the *link function* – it links the linear predictor to the mean. In the linear case, the link g is the identity; we shall see a range of other standard links below.

To complete the specification of the model, we need the distribution of the response variable y, not just its mean μ; that is, we need to specify the *error structure*. We assume that each observation y_i is *independent* and has a density f of the form

$$\exp\left\{\frac{\omega_i(y_i\theta_i - b(\theta_i))}{\phi} + c(y, \phi)\right\},$$

where the parameter θ_i depends on the linear predictor η, ϕ is a scale parameter (which may or may not be known), the ω_i are a sequence of known weights, and $b(.)$ and $c(.)$ are functions. It is further assumed that

$$\mathrm{var}(y_i) = \frac{\phi}{\omega_i} V(\mu_i),$$

where $V(\cdot)$ is a variance function relating the variance of the y_i to the mean μ_i. It can be shown that in the notation above

$$\begin{aligned} E(y_i) &= b'(\theta_i), \\ \mathrm{var}(y_i) &= \frac{\phi}{\omega_i} b''(\theta_i). \end{aligned}$$

This functional form derives from the theory of *exponential families*, which lies beyond the scope of this book. For a monograph treatment, see e.g. Brown (1986). Suffice it here to say that the parametric families which have a fully satisfactory inference theory are the exponential families. So the assumption above is not arbitrary, but is underpinned by this theory, and GLMs are tractable because of it.

The case when

$$\theta = \eta$$

is particularly important. When it occurs, the link function is called *canonical.* (See also Exercise 8.1).

Example 8.1 (Canonical forms)

1. *Normal.* Here $f(y;\theta,\phi)$ is given by

$$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2}(y-\mu)^2/\sigma^2\} = \exp\{(y\mu-\mu^2/2)/\sigma^2 - \frac{1}{2}(y^2/\sigma^2+\log(2\pi\sigma^2))\}.$$

So $\theta=\mu$, the scale parameter is simply the variance σ^2, and the link function g is the *identity* function:

$$g(\mu)=\mu.$$

This of course merely embeds the general linear model, with normal error structure, into the *generalised* linear model as a special case, and was to be expected. The normal distribution is the obvious choice – the 'default option' – for measurement data on the whole line.

2. *Poisson.* Here the mean μ is the Poisson parameter λ, and $f(k;\lambda) = e^{-\lambda}\lambda^k/k!$. Writing y for k to conform with the above,

$$f(y;\lambda)=\exp\{y\log\lambda-\lambda-\log y!\}.$$

So $\theta=\log\lambda$. So the canonical link, when $\theta=\eta=\log\lambda$, is the *logarithm*:

$$\eta=\log\lambda.$$

This explains the presence of the logarithm in §8.3 below on *log-linear models.* The Poisson distribution is the default option for count data (on the non-negative integers). Note also that in this case the scale parameter ϕ is simply $\phi=1$.

3. *Gamma.* The gamma density $\Gamma(\lambda,\alpha)$ is defined, for parameters α, $\lambda>0$, as

$$f(x)=\frac{\lambda^\alpha}{\Gamma(\alpha)}e^{-\lambda x}x^{\alpha-1}.$$

The mean is

$$\mu=\alpha/\lambda,$$

and as

$$\begin{aligned} f(x) &= \exp\{-\lambda x+(\alpha-1)\log x+\alpha\log\lambda-\log\Gamma(\alpha)\} \\ &= \exp\left\{(-\alpha)\frac{x}{\mu}+\dots\right\}, \end{aligned}$$

the canonical link is the inverse function:

$$\eta=1/\mu,$$

and we can also read off that the scale parameter is given by

$$\phi = 1/\alpha.$$

The gamma density is the default option for measurement data on the positive half-line. It is often used with the log-link

$$\eta = \log \mu$$

and we shall meet such examples below (see Exercises 8.7).

Other standard examples, included in S-Plus/R®, are the inverse Gaussian family (Exercise 8.9), the binomial (whose special case the Bernoulli, for binary data, we discuss below in §8.3), and the logit, probit and complementary log-log cases (see §8.3 also).

One other pleasant feature of the general linear (normal) case that does not carry over to GLMs is the distribution theory – independent sums of squares, chi-square distributed, leading to F-tests and Analysis of Variance. The distribution theory of GLMs is less simple and clear-cut. Instead of Analysis of Variance, one has *analysis of deviance.* This gives one a means of assessing model fit, and of comparing one model with another – and in particular, of choosing between two or more nested models. For further background and detail, we refer to McCullagh and Nelder (1989), Venables and Ripley (2002), Ch. 7, but we outline the basic procedures in the following two subsections.

8.2.1 Statistical testing and model comparisons

The *scaled deviance* metric is a measure of the distance between the observed y_i and the fitted $\hat{\mu}_i$ of a given model, and is defined as

$$\begin{aligned} S(y, \hat{\mu}) &= 2\left(l(y; \phi, y) - l(\hat{\mu}; \phi, y)\right), \\ &= \frac{2}{\phi}\sum\nolimits_i \omega_i \left[y_i\left(\theta(y_i) - \hat{\theta}_i\right) - \left(b(\theta(y_i)) - b\left(\hat{\theta}_i\right)\right)\right], \end{aligned}$$

where l denotes log-likelihood. We define the *residual deviance* or *deviance* which is the scaled deviance multiplied by the scale parameter ϕ:

$$D(y, \hat{\mu}) = \phi S(y, \hat{\mu}) = 2\sum\nolimits_i \omega_i \left[y_i\left(\theta(y_i) - \hat{\theta}_i\right) - \left(b(\theta(y_i)) - b\left(\hat{\theta}_i\right)\right)\right].$$

Both the scaled deviance and the residual deviance are important and enable both statistical testing of hypotheses and model comparisons. (Note that the scaled deviance retains the scale parameter ϕ, which is then eliminated from the residual deviance by the above.)

Example 8.2

In the case of the normal linear model, the residual deviance leads to the residual sum of squares:

$$D(y,\hat{\mu}) = SSE = \sum\nolimits_i (y_i - \hat{\mu})^2.$$

To see this we note that, written as a function of the μ_i, the log-likelihood function is

$$l(\mu|\phi, y) = \frac{1}{2\phi}\sum\nolimits_i (y_i - \mu_i)^2 + C,$$

where C is constant with respect to μ. We have that

$$D(\hat{\mu}|\phi, y) = 2\phi\left[\frac{-\sum(y_i - y_i)^2 + \sum(y_i - \hat{\mu}_i)^2}{2\phi}\right] = \sum(y_i - \hat{\mu}_i)^2.$$

The residual deviance can also be calculated for a range of common probability distributions (see Exercise 8.2).

Nested models. Nested models can be formally compared using generalised likelihood ratio tests. Suppose Model 1 is $\eta = X\beta$ and Model 2 is $\eta = X\beta + Z\gamma$ with $\text{rank}(Z) = r$. Model 1 has dimension p_1 and Model 2 has dimension $p_2 = p_1 + r$. The test statistic is

$$\begin{aligned} 2(l_2 - l_1) &= S(y;\hat{\mu}_1) - S(y;\hat{\mu}_2), \\ &= \frac{D(y;\hat{\mu}_1) - D(y;\hat{\mu}_2)}{\phi}. \end{aligned}$$

If the scale parameter ϕ is known, then the asymptotic distribution of this test statistic should be χ^2_r. This likelihood ratio test also suggests an admittedly rough measure of absolute fit by comparing the residual deviance to χ^2_{n-p}, with high values indicating lack of fit. If ϕ is unknown, one suggestion is to estimate ϕ using Model 2 and then treat ϕ as known. Alternatively, it is often customary to use the F-test

$$\frac{D(y;\hat{\mu}_1) - D(y;\hat{\mu}_2)}{\hat{\phi}r} \sim F_{r,n-p_2},$$

by analogy with the theory of Chapter 6. However, this must be used with caution in non-Gaussian cases. A skeleton analysis of deviance is outlined in Table 8.1, and should proceed as follows:
(i) Test $S(y;\mu_2)$ versus χ^2_{n-p-1} for an admittedly rough test of model accuracy for model 2.
(ii) Test $S(y;\mu_1) - S(y;\mu_2)$ versus χ^2_r to test the hypothesis $Z = 0$.

Source	Scaled Deviance	df
Model 2 after fitting Model 1	$S(y;\mu_1)$-$S(y;\mu_2)$	r
Model 2	$S(y;\mu_2)$	$n-p_1-r$
Model 1	$S(y;\mu_1)$	$n-p_1$

Table 8.1 Skeleton analysis of deviance

Usually more than two models would be compared in the same way. The reader should also note that methods of model selection similar to those discussed in Chapter 6 – namely forward and backward selection and sequential methods – also apply here.

t-tests. Approximate t-tests for individual parameters can be constructed by comparing

$$T = \frac{\hat{\beta}_j - \beta_j}{\text{e.s.e}(\hat{\beta}_j)}$$

to t_{n-p} where $\hat{\beta}_j$ is the estimate of β_j and e.s.e denotes the associated estimated standard error. This is partly by analogy with the theory of the Gaussian linear model but also as a way of treating a near-Gaussian situation more robustly. Approximate inference can also be conducted using the delta method of Exercise 7.6. Whilst useful in model simplification, tests based on analysis of deviance are usually preferred when testing between different models. *Non-nested* models may be compared using the following generalisation of AIC:

$$\text{AIC}(\hat{\mu}) = D(y;\hat{\mu}) + 2p\hat{\phi},$$

where μ denotes the fitted values and p the number of parameters of a given model.

8.2.2 Analysis of residuals

There are four types of residuals commonly encountered in Generalised Linear Models and roughly analogous to the various types of residuals defined for the general linear model in Chapter 7. The *response* or *raw* residuals are simply given by

$$e_i = y_i - \hat{\mu}_i.$$

The *Pearson* residuals are defined as

$$e_{P,i} = \sqrt{\omega_i}\frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}} = \sqrt{\phi}\frac{y_i - \hat{\mu}_i}{\sqrt{\hat{V}(y_i)}},$$

since $\mathrm{var}(y_i) = (\phi/\omega_i)V(\mu_i)$ by assumption. This is simply $(y_i - \hat{\mu}_i)/\sqrt{\hat{V}(y_i)}$ appropriately scaled so as to remove the dispersion parameter ϕ. A Pearson χ^2 statistic can be defined as

$$\chi^2 = \chi^2(y, \hat{\mu}) = \sum e_{P,i}^2,$$

and can be shown to be asymptotically equivalent to the deviance D. *Working residuals* are defined as

$$e_{W,i} = \frac{(y_i - \hat{\mu}_i)}{d\mu_i/d\eta_i},$$

and are derived as part of the iterative model fitting process. *Deviance residuals* are defined as

$$e_{D,i} = \mathrm{sgn}(y_i - \hat{\mu}_i)2\omega_i \left[y_i \left(\theta(y_i) - \hat{\theta}_i \right) - \left(b(\theta(y_i)) - b\left(\hat{\theta}_i\right) \right) \right],$$

where the sign function sgn (or signum) is defined by

$$sgn(x) = \begin{cases} -1 & x < 0 \\ 0 & x = 0 \\ 1 & x > 0. \end{cases}$$

This definition ensures that $\sum e_{D,i}^2 = D$. If ϕ is not equal to one, the residuals may be multiplied by $\sqrt{\phi}$ or its estimate to produce *scaled* versions of these residuals. Plots of residuals can be used in the usual way to check model adequacy – testing for nonlinearity, outliers, autocorrelation, etc – by plotting against individual covariates or against the $\hat{\mu}_i$ or the $\hat{\eta}_i$. However, in contrast to the general linear model, a Normal probability plot of residuals is unlikely to be helpful. Also, aspects of the data, e.g. Poisson data for small counts, may cause naturally occurring patterns in the residuals which should not then be interpreted as indicating model inadequacy.

8.2.3 Athletics times

Example 8.3

We give a further illustrative example of a gamma Generalised Linear Model by returning to our discussion of athletics times. For distance races, speed decreases with distance, and so the time t taken increases faster than the distance d. Because there are no natural units here, of distance or time, and the relationship between t and d is smooth, *Fechner's Law* applies (Gustav Fechner (1801–1887) in 1860), according to which the relationship should be a *power law*:

$$t = ad^b$$

(see e.g. Hand (2004), §5.6, where it is attributed to Stevens). Here a is *pace*, or time per unit distance (traditionally reckoned in minutes and seconds per mile, or per kilometre), and so is an indicator of the quality of the athlete, while b is dimensionless (and is thought to be much the same for all athletes – see Bingham and Rashid (2008) for background). This is an instance of *Buckingham's Pi Theorem* (Edgar Buckingham (1867–1940) in 1914), according to which a physically meaningful relationship between n physical variables, k of which are independent, can be expressed in terms of $p = n - k$ dimensionless quantities; here $n = 3$ (t, a, d), $k = 2$ (t, d), $p = 1$ (b).

Taking this relationship for the mean $t = ET$ for the actual running time T, one has

$$t = ET = ad^b, \qquad \log(ET) = \log a + b \log d = \alpha + b \log d,$$

say, giving a linear predictor (in $(1, \log d)$) with coefficients α, b. This gives the systematic part of the model; as $\eta = \log \mu$ (with $\mu = ET$ the mean), the link function is log. As time and distance are positive, we take the random part of the model (or error law) as Gamma distributed:

$$T \sim \Gamma(\lambda, \mu).$$

An alternative would be to use an ordinary linear model with Gaussian errors, as in Chapter 3:

$$\log T = \alpha + b \log d + \epsilon, \qquad \epsilon \sim N(0, \sigma^2).$$

With age also present, one needs an age-dependent version of the above: using c in place of a above,

$$ET = c(a)t^b,$$

where in view of our earlier studies one uses a linear model for $c(a)$:

$$Ec(a) = \alpha_1 + \alpha_2 a.$$

The resulting compound model is of *hierarchical* type, as in Nelder, Lee and Pawitan (2006). Here, an approximate solution is possible using the simpler gamma Generalised Linear Model if instead we assume

$$\log(ET) = \alpha_1 + \alpha_2 \log a + b \ \log d.$$

In this case we can use a Gamma Generalised Linear Model with log-link. In S-Plus/R® the relevant syntax required is

```
m1.glm<-glm(time∼log(age)+log(distance),family=Gamma(link="log"))
summary(m1.glm)
```

The results obtained for the marathon/half-marathon data (Table 1.1, Exercise 1.3) are shown in Table 8.2, and give similar results to those using a log-transformation and a normal linear model in Example 3.37. As there, the log(age) value of about 1/3 is consistent (for age$\sim$60, $ET\sim$180) with the Rule of Thumb: expect to lose a minute a year on the marathon through ageing alone.

	Value	Std. Error	t value
Intercept	0.542	0.214	2.538
log(age)	0.334	0.051	6.512
log(distance)	1.017	0.015	67.198

Table 8.2 Regression results for Example 8.3

8.3 Binary models

Logits.

Suppose that we are dealing with a situation where the response y is success or failure (or, life or death) or of zero-one, or Boolean, type. Then if

$$Ey = p,$$

$p \in [0,1]$, and in non-trivial situations, $p \in (0,1)$. Then the relevant distribution is *Bernoulli*, with *parameter* p, $B(p)$:

$$p = P(y=1), \qquad q := 1-p = P(Y=0),$$

$$\mathrm{var}(y) = pq = p(1-p).$$

Interpreting p as the probability of success and $q = 1-p$ as that of failure, the *odds* on success are $p/q = p/(1-p)$, and the *log-odds*, more natural from some points of view, are

$$\log\left(\frac{p}{1-p}\right).$$

Thinking of success or failure as survival or death in a medical context of treatment for some disease, the log-odds for survival may depend on covariates: age might well be relevant, so too might length of treatment, how early the disease was diagnosed, treatment type, gender, blood group etc. The simplest plausible model is to assume that the log-odds of survival depend on some *linear predictor* η – a linear combination $\eta = \sum_j a_j \beta_j$ of parameters β_j, just as before (cf. §9.5 below on survival analysis). With data $y_1, \ldots, y_n$ as before, and writing

$$Ey_i = p_i \qquad (i = 1, \ldots, n),$$

we need a double-suffix notation just as before, obtaining

$$\log\{p_i/(1-p_i)\} = \sum\nolimits_{j=1}^{p} a_{ij}\beta_j, \qquad (i = 1, \ldots, n).$$

There are three salient features here:
(i) The function

$$g(p) = \log\{p/(1-p)\},$$

the link function, which links mean response $p = Ey$ to the linear predictor.
(ii) The *distributions* ('error structure'), which belong to the *Bernoulli* family $B(p)$, a special case of the *binomial* family $B(n, p)$, under which

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad (k = 0, 1, \ldots, n).$$

(iii) The function V giving the variance in terms of the mean:

$$V(p) = p(1-p),$$

called the *variance function.*

The model above is called the *logit* model (from log-odds), or *logistic* model (as if $\eta = \log\{p(1-p)\}$, $p = e^\eta/(1+e^\eta)$, the logistic function). Binary data are very important, and have been studied at book length; see e.g. McCullagh and Nelder (1989) Ch. 13, Cox and Snell (1989), and Collett (2003). The relevant S-Plus/R® commands are of the form

```
glm(y ~ ..., family = binomial)
```

We draw an illustrative example (as usual) from athletics times. The 'time to beat' for a club runner of reasonable standard in the marathon is three hours; let us interpret 'success' as breaking three hours. The sample version of the expected frequency p of success is the observed frequency, the proportion of successful runners. For a mass event (such as the London Marathon), which we suppose for simplicity has reached a steady state in terms of visibility, prestige etc., the systematic component of the observed variability in frequency of success from year to year is governed principally by the weather conditions: environmental factors such as temperature, humidity, wind and the like. At too high a temperature, the body is prone to dehydration and heat-stroke; at too low a temperature, the muscles cannot operate at peak efficiency. Performance thus suffers on either side of the optimum temperature, and a quadratic in temperature is suggested. On the other hand, humidity is simply bad: the more humid the air is, the harder it is for sweat to evaporate – and so perform its function, of cooling the body (heat is lost through evaporation). In an endurance event in humid air, the body suffers doubly: from fluid loss, and rise in core temperature. Thus a linear term in humidity is suggested.

Probits.

A very different way of producing a mean response in the interval $(0, 1)$ from a linear predictor is to apply the (standard) normal probability distribution function Φ. The model

$$p = \Phi(\alpha + \beta x)$$

(or some more complicated linear predictor) arises in bioassay, and is called a *probit* model. Writing $\eta = \sum_j \beta_j x_j$ for the linear predictor, the link function is now

$$\eta = g(p) = \Phi^{-1}(p).$$

Complementary log-log link.

In dilution assay, the probability p of a tube containing bacteria is related to the number $x = 0, 1, 2, \ldots$ of dilutions by

$$p = 1 - e^{-\lambda x}$$

for some parameter λ (the number of bacteria present is modelled by a Poisson distribution with this parameter). The link function here is

$$\eta = g(p) = \log(-\log(1 - p)) = \log \lambda + \log x.$$

Example 8.4

The data in Table 8.3 show the number of insects killed when exposed to different doses of insecticide.

Dose	Number	Number killed	% killed
10.7	50	44	88
8.2	49	42	86
5.6	46	24	52
4.3	48	16	33
3.1	50	6	12
0.5	49	0	0

Table 8.3 Data for Example 8.4

We wish to model these data using a Generalised Linear Model. A sensible starting point is to plot the empirical logits defined here as $\eta_{e,i} =$

$\log(y_i + 1/2) - \log(1 - y_i + 1/2)$, where the 1/2 guards against singularities in the likelihood function if $y_i = 0$ or $y_i = 1$. Here, a plot of the $\eta_{e,i}$ against log(dose) appears roughly linear suggesting a logarithmic term in dose. The model can be fitted in R® as follows. First, the count data needs to be stored as two columns of successes and failures (the command `cbind` is helpful here). The model is fitted with the following commands:

```
a.glm<-glm(data~log(dose), family=binomial)
summary(a.glm)
```

This gives a residual deviance of 1.595 with 4 df The deviance of the null model with only a constant term is 163.745 on 5 df Testing 1.595 against χ^2_4 gives a p-value of 0.810, so no evidence of lack of fit. The log(dose) term is highly significant. The analysis of deviance test gives $163.745 - 1.595 = 162.149$ on 1 df with $p = 0.000$. Probit and complementary log-log models can be fitted in S-Plus/R® using the following syntax (see Exercise 8.4):

```
a.glm<-glm(data~log(dose), family=binomial(link=probit))
a.glm<-glm(data~log(dose), family=binomial(link=cloglog))
```

8.4 Count data, contingency tables and log-linear models

Suppose we have n observations from a population, and we wish to study a characteristic which occurs in r possible types. We classify our observations, and count the numbers $n_1, \ldots, n_r$ of each type (so $n_1 + \ldots + n_r = n$). We may wish to test the hypothesis H_0 that type k occurs with probability p_k, where

$$\sum\nolimits_{k=1}^{r} p_k = 1.$$

Under this hypothesis, the expected number of type k is $e_k = np_k$; the observed number is $o_k = n_k$. Pearson's *chi-square goodness-of-fit test* (Karl Pearson (1857–1936), in 1900) uses the *chi-square statistic*

$$X^2 := \sum\nolimits_{k=1}^{r} (n_k - np_k)^2/(np_k) = \sum (o_k - e_k)^2/e_k.$$

Then for large samples, X^2 has approximately the distribution $\chi^2(r-1)$, the chi-square distribution with r df; large values of X^2 are evidence against H_0. The proof proceeds by using the multidimensional Central Limit Theorem to show that the random vector $(x_1, \ldots, x_r)$, where

$$x_k := (n_k - np_k)/\sqrt{np_k},$$

is asymptotically multivariate normal, with mean zero and (symmetric) covariance matrix

$$A = I - pp^T,$$

where p is the column vector

$$(\sqrt{p_1}, \ldots, \sqrt{p_r})^T.$$

Since $\sum_k p_k = 1$, A is idempotent; its trace, and so its rank, is $r - 1$. This loss of one degree of freedom corresponds to the one linear constraint satisfied (the n_k sum to n; the p_k sum to 1). From this, the limiting distribution $\chi^2(r-1)$ follows by Theorem 3.16. For details, see e.g. Cramér (1946), §30.1.

Now the distribution of the vector of observations $(n_1, \ldots, n_r)$ (for which $\sum_i n_i = n$) is *multinomial*:

$$P(n_1 = k_1, \ldots, n_r = k_r) = \binom{n}{k_1, \ldots, k_r} p_1^{k_1} \ldots p_r^{k_r},$$

for any non-negative integers $k_1, \ldots, k_r$ with sum n (the multinomial coefficient counts the number of ways in which the k_1 observations of type 1, etc., can be chosen; then $p_1^{k_1} \ldots p_r^{k_r}$ is the probability of observing these types for each such choice.

According to the *conditioning property* of the Poisson process (see e.g. Grimmett and Stirzaker (2001), §6.12–6.13), we obtain multinomial distributions when we condition a Poisson process on the number of points (in some region).

These theoretical considerations lie behind the use of GLMs with *Poisson errors* for the analysis of count data. The basic observation here is due to Nelder in 1974. In the linear model of previous chapters we had additive normal errors, and – regarded as a GLM – the identity link. We now have *multiplicative Poisson errors*, the multiplicativity corresponding to the logarithmic link.

We assume that the logarithm of the ith data point, $\mu_i = Ey_i$, is given by a linear combination of covariates:

$$\log \mu_i = \eta_i = \beta^T \mathbf{x}_i \qquad (i = 1, \ldots, n).$$

We shall refer to such models as *log-linear models*. For them, the link function is the logarithm:

$$g(\mu) = \log \mu.$$

Example 8.5 (Poisson modelling of sequences of small counts)

Suppose that we have the following (artificial) data in Table 8.4 and we wish to model this count data using a Poisson Generalised Linear Model.

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y	1	0	2	5	6	9	12	12	25	25	22	30	52	54

Table 8.4 Data for Example 8.5

A plot of the guarded logs, $\log(y_i+0.5)$, against x_i seems close to a straight line although there is perhaps a slight suggestion of curvature. The model with x on its own gives a residual deviance of 24.672 on 12 df The χ^2 goodness-of-fit test gives a p-value of 0.016, suggesting that the fit of this model is poor. The model with a quadratic term has a residual deviance of 13.986 on 11 df This model seems to fit better; the χ^2 goodness of fit test gives a p-value of 0.234, and the AIC of this model is 75.934. A plot of the guarded logs against $\log(x_i)$ also appears close to linear and $\log(x)$ thus seems a suitable candidate model. Fitting this model gives a residual deviance of 14.526 on 12 df and appears reasonable (χ^2 test gives $p = 0.268$). The AIC for this model is 74.474 and thus $\log(x)$ appears to be the best model.

All of this continues to apply when our counts are cross-classified by more than one characteristic. We consider first the case of *two* characteristics, partly because it is the simplest case, partly because we may conveniently display count data classified by two characteristics in the form of a *contingency table*. We may then, for example, test the null hypothesis that the two characteristics are independent by forming an appropriate chi-square statistic. For large samples, this will (under the null hypothesis) have approximately a chi-square distribution with df $(r-1)(s-1)$, where r and s are the numbers of forms of the two characteristics. For proof, and examples, see e.g. Cramér (1946), Ch. 30.

We may very well have more than two characteristics. Similar remarks apply, but the analysis is more complicated. Such situations are common in the social sciences – sociology, for example. Special software has been developed: SPSS®[3] (statistical package for the social sciences). Such multivariate count data is so important that it has been treated at book length; see e.g. Bishop et al. (1995), Plackett (1974), Fienberg (1980).

Another application area is insurance. A motor insurer might consider, when assessing the risk on a policy, the driver's age, annual mileage, sex, etc; also the type of vehicle (sports cars are often charged higher premiums), whether used for work, whether kept off-road, etc. A house insurer might consider number of rooms (or bedrooms), indicators of population density, postal code (information about soil conditions, and so subsidence risk, for buildings; about the ambient population, and so risk of burglary, for contents, etc.). The simplest

[3] SPSS® is a registered trademark of SPSS Inc., 233 S. Wacker Drive, 11th Floor, Chicago, IL 60606, USA, http://www.spss.com

way to use such information is to use a linear regression function, or linear predictor, as above, whence the relevance of GLMs. The S-Plus commands are much as before:

```
glm(y ∼ ..., family = poisson).
```

We note in passing that the parameter λ in the Poisson distribution $P(\lambda)$, giving its mean and also its variance, is most naturally viewed as a *rate*, or *intensity*, of a stochastic process – the *Poisson point process* with rate λ (in time, or in space) – which corresponds to a *risk* in the insurance context. Thus this material is best studied in tandem with a study of stochastic processes, for which we refer to, e.g., Haigh (2002), Ch. 8, as well as Grimmett and Stirzaker (2001), Ch. 6 cited earlier.

Example 8.6 (Skeleton analysis of 2×2 contingency tables)

For technical reasons, it can be important to distinguish between two cases of interest.

Two response variables. Both variables are random, only the total sample size $\sum_{ij} y_{ij}$ is fixed. The data in Exercise 7.4 are an example with two response variables.

One response variable and one observed variable. The setting here is a controlled experiment rather than an observational study. The design of the experiment fixes row or column totals *before* the full results of the experiment are known. One example of this is medical trials where patients are assigned different treatment groups, e.g. placebo/vaccine, etc. The interested reader is referred to Dobson and Barnett (2003), Ch. 9.

A range of different possible hypotheses applies in each of these two cases. Apart from unrealistic or very uncommon examples, the main interest lies in testing the hypothesis of no association between the two characteristics A and B. It can be shown that this reduces to testing the adequacy of the log-linear model

$$\log(Y) = \text{const.} + A + B.$$

The data in Table 8.5 give hair and eye colours for a group of subjects. We use Poisson log-linear models to test for an association between hair and eye colour. Fitting the model we obtain a residual deviance of 146.44 on 9 df leading to a p-value of 0.000 and we reject the null hypothesis of no association.

	Brown	Blue	Hazel	Green
Black hair	68	20	15	5
Brown hair	119	84	54	29
Red hair	26	17	14	14
Blond hair	7	94	10	16

Table 8.5 Data for Example 8.6

8.5 Over-dispersion and the Negative Binomial Distribution

The fact that a Poisson mean and variance coincide gives a yardstick by which to judge variability, or dispersion, of count data. If the variance-to-mean ratio observed is > 1, the data are called *over-dispersed* (if < 1, they are called *under-dispersed*, though this is less common). Equivalently, one may also use the ratio of standard error to mean (coefficient of variation), often preferred to the variance-mean ratio as it is dimensionless.

One model used for over-dispersion is to take a *Gamma mixture of Poissons*: take a Poisson distribution with random mean, M say, where M is Gamma distributed. Thus

$$P(Y = n|M = \lambda) = e^{-\lambda}\lambda^n/n!,$$

but (it is convenient here to reparametrise, from λ, $\alpha > 0$ to ν, $\tau > 0$) $M \sim \Gamma(\nu/\tau, \nu)$: M has density

$$f(y) = \frac{1}{\Gamma(\nu)}\left(\frac{\nu y}{\tau}\right)^{\nu} e^{-\nu y/\tau}\frac{1}{y} \qquad (y > 0).$$

Then unconditionally

$$\begin{aligned} P(Y = n) &= \int_0^\infty \frac{e^{-y}y^n}{n!}\frac{1}{\Gamma(\nu)}\left(\frac{\nu y}{\tau}\right)^{\nu} e^{-\nu y/\tau} y^{n+\nu-1}\,dy \\ &= \frac{\nu^\nu}{\tau^\nu}\frac{1}{n!\Gamma(\nu)}\frac{1}{(1+\nu/\tau)^{n+\nu}}\int_0^\infty e^{-u}u^{n+\nu-1}\,du \quad (y(1+\nu/\tau) = u) \\ &= \frac{\nu^\nu}{\tau^\nu(1+\nu/\tau)^{n+\nu}}\frac{\Gamma(n+\nu)}{n!\Gamma(\nu)}. \end{aligned}$$

This is the *Negative Binomial* distribution, $NB(\nu, \tau)$, in one of several parametrisations (compare McCullagh and Nelder (1989), p237 and p373). The mean is

$$\mu = \tau.$$

The variance is

$$V(\mu) = \tau + \tau^2/\nu = \mu + \mu^2/\nu.$$

The model is thus over-dispersed.

Since $\Gamma(1+x) = x\Gamma(x)$,

$$\frac{\Gamma(n+\nu)}{n!\Gamma(\nu)} = \frac{(n+\nu-1)(n+\nu-2)\ldots\nu}{n!},$$

and when ν is a positive integer, r say, this has the form of a binomial coefficient

$$\binom{n+r-1}{n} = \binom{n+r-1}{r-1}.$$

In this case,

$$P(Y=n) = \binom{n+r-1}{n} p^r q^n \qquad (n = 0, 1, \ldots),$$

writing

$$p := r/(\tau + r), \qquad q := 1 - p = \tau/(\tau + r).$$

The case $r = 1$ gives the *geometric* distribution, $G(p)$:

$$P(Y = n) = q^n p \qquad (n = 0, 1, \ldots),$$

the distribution of the number of failures before the first success in Bernoulli trials with parameter p ('tossing a p-coin'). This has mean q/p and variance q/p^2 (over-dispersed, since $p \in (0,1)$, so $1/p > 1$). The number of failures before the rth success has the negative binomial distribution in the form just obtained (the binomial coefficient counts the number of ways of distributing the n failures over the first $n+r-1$ trials; for each such way, these n failures and $r-1$ successes happen with probability $q^n p^{r-1}$; the $(n+r)$th trial is a success with probability p). So the number of failures before the rth success
(i) has the negative binomial distribution (which it is customary and convenient to parametrise as $NB(r,p)$ in this case);
(ii) is the sum of r independent copies of geometric random variables with distribution $G(p)$;
(iii) so has mean rq/p and variance rq/p^2 (agreeing with the above with $r = \nu$, $p = r/(\tau + r)$, $q = \tau/(\tau + r)$).
The Federalist.

The Federalist Papers were a series of essays on constitutional matters, published in 1787–1788 by Alexander Hamilton, John Jay and James Madison to persuade the citizens of New York State to ratify the U.S. Constitution. Authorship of a number of these papers, published anonymously, was later disputed between Hamilton and Madison. Their authorship has since been settled by a classic statistical study, based on the use of the negative binomial distribution for over-dispersed count data (for usage of key indicator words – 'whilst' and 'while' proved decisive); see Mosteller and Wallace (1984).

8.5.1 Practical applications: Analysis of over-dispersed models in R®

For binomial and Poisson families, the theory of Generalised Linear Models specifies that the dispersion parameter $\phi = 1$. Over-dispersion can be very common in practical applications and is typically characterised by the residual deviance differing significantly from its asymptotic expected value given by the residual degrees of freedom (Venables and Ripley (2002)). Note, however, that this theory is only asymptotic. We may crudely interpret over-dispersion as saying that data varies more than if the underlying model really were from a Poisson or binomial sample. A solution is to multiply the variance functions by a dispersion parameter ϕ, which then has to be estimated rather than simply assumed to be fixed at 1. Here, we skip technical details except to say that this is possible using a *quasi-likelihood* approach and can be easily implemented in R® using the Generalised Linear Model families `quasipoisson` and `quasibinomial`. We illustrate the procedure with an application to over-dispersed Poisson data.

Example 8.7

We wish to fit an appropriate Generalised Linear Model to the count data of Exercise 7.2. Fitting the model with both blocks and treatments gives a residual deviance of 242.46 on 12 df giving a clear indication of over-dispersion. A quasi-poisson model can be fitted with the following commands:

```
m1.glm<-glm(data~blocks+treatments, family=quasipoisson)
summary(m1.glm)
```

Since we have to estimate the dispersion parameter ϕ we use an F-test to distinguish between the models with blocks and treatments and the model with blocks only. We have that

$$F = \frac{\Delta\text{Residual deviance}}{\Delta\text{df}(\hat{\phi})} = \frac{3468.5 - 242.46}{4(21.939)} = 36.762.$$

Testing against $F_{4,12}$ gives a p-value of 0.000. Similar procedures can be used to test the effectiveness of blocking (see Exercise 8.5).

EXERCISES

8.1. *Canonical forms.* Show that these common probability distributions can be written in the canonical form of a Generalised Linear Model as shown in Table 8.6:

	Normal $N(\theta,\phi)$	Poisson $\mathrm{Po}(e^{\theta})$	Binomial $ny{\sim}\mathrm{Bi}\left(n, \frac{e^{\theta}}{1+e^{\theta}}\right)$	Gamma $\mathit{\Gamma}\left(\frac{1}{\phi}, -\frac{\theta}{\phi}\right)$
$\frac{\phi}{\omega}$	ϕ	1	n^{-1}	ϕ
$b(\theta)$	$\frac{\theta^2}{2}$	e^{θ}	$\log\left(1+e^{\theta}\right)$	$-\log(-\theta)$
$c(y,\theta)$	$-\frac{y^2}{2\phi} - \frac{\phi\log(2\pi)}{2}$	$-\log(y!)$	$\log\binom{n}{ny}$	$\left(\frac{1}{\phi}-1\right)\log\, y$ $-\frac{\log\phi}{\phi} + \log\ \phi$
$\mu = b'(\theta)$	θ	e^{θ}	$\frac{e^{\theta}}{1+e^{\theta}}$	$-\frac{1}{\theta}$
$b''(\theta)$	1	μ	$\mu(1-\mu)$	μ^2

Table 8.6 Canonical forms for Exercise 8.1

8.2. *(Residual) deviance calculations.* Show that for the following common probability distributions the residual deviances can be calculated as follows:

Poisson

$$2\sum\nolimits_i \left(y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) - (y_i - \hat{\mu}_i)\right),$$

Binomial

$$2\sum\nolimits_i n_i \left\{y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)\right\},$$

Gamma

$$2\sum\nolimits_i \log\left(\frac{\hat{\mu}_i}{y_i}\right) + \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}.$$

8.3. Test the hypothesis of no association between haul and number for the data in Exercise 7.4 using
(i) a Poisson log-linear model,
(ii) the Pearson χ^2 test of no association,
and comment on your findings.

8.4. Re-fit the data in Example 8.4 using
(i) a probit model,
(ii) a complementary log-log model,
(iii) an approximate method using general linear models.

8.5. Re-fit the data in Exercise 7.2 using a Poisson Generalised Linear Model, before switching to an over-dispersed Poisson model if this seems appropriate. Test for the effectiveness of blocking by seeing if the model with just the blocks term offers an improvement over the null model.

8.6. Suppose that we have the following data for the number of unusable ears of corn shown in Table 8.7. (Assume totals are out of 36.) Analyse these data by fitting a binomial Generalised Linear Model, using a quasi-binomial model if it appears that we have over-dispersion. Compare your results with an approximation using General Linear Models on similar data in Exercise 7.3 and interpret the results.

Block	1	2	3	4	5	6
Treatment A	15	12	9	14	20	18
Treatment B	12	12	2	9	11	10
Treatment C	3	8	2	6	5	6
Treatment D	6	7	6	1	4	4

Table 8.7 Data for Exercise 8.6

8.7. *Generalised Linear Model with Gamma errors.* Using the data in Exercise 1.6 fit a Gamma Generalised Linear Model. Interpret your findings and compare both with Exercise 1.6 and the analyses in §5.3. Write down the equation of your fitted model.

8.8. *Inverse Gaussian distribution.* The inverse Gaussian distribution is the distribution on the positive half-axis with probability density

$$f(y)=\sqrt{\frac{\lambda}{2\pi y^3}}\exp\left(\frac{-\lambda(y-\mu)^2}{2\mu^2 y}\right).$$

Show that this density lies in the exponential family (see Exercise 8.1).

8.9. *Generalised Linear Model with inverse Gaussian errors.* Repeat Exercise 8.7 using an inverse Gaussian Generalised Linear Model.

8.10. *The effect of ageing on athletic performance.* Using the fitted equations obtained in Exercises 8.7 and 8.9 and using $x = 63$, comment on the effect of

(i) ageing,

(ii) club status.

9
Other topics

9.1 Mixed models

In §5.1 we considered extending our initial model (M_0), with p parameters, to an augmented model M_A with a further q parameters. Here, as in Chapter 2, we have $p + q << n$, there are many fewer parameters than data points. We now turn to a situation with some similarities but with important contrasts. Here our initial model has *fixed effects*, but our augmented model adds *random effects*, which may be comparable in number to the sample size n.

We mention some representative situations in which such mixed models occur.

1. Longitudinal studies (or *panel data*). Suppose we wish to monitor the effect of some educational initiative. One may choose some representative sample or cohort of school children or students, and track their progress over time. Typically, the resulting data set consists of a large number (the size of the cohort) of short time series (the longer the time the more informative the study, but the more expensive it is, and the longer the delay before any useful policy decisions can be made). For background on longitudinal data, see Diggle et al. (2002).

Here one takes for granted that the children in the cohort differ – in ability, and in every other aspect of their individuality. One needs information on between-children variation (that is, on cohort variance); this becomes a parameter in the mixed model. The child effects are the random effects: if one repeated the study with a different cohort, these would be different. The educational aspects one wishes to study are the fixed effects.

N.H. Bingham and J.M. Fry, *Regression: Linear Models in Statistics*,
Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-84882-969-5_9,

2. Livestock studies. One may wish to follow the effect of some treatments – a diet, or dietary supplements, say – over time, on a cohort of livestock (cattle, sheep or pigs, say). Again, individual animals differ, and these give the random effects. The fixed effects are the objects of study.

The field of mixed models was pioneered in the US dairy industry by C. R. Henderson (1911–1989) from 1950 on, together with his student S. R. Searle (1928–). Searle is the author of standard works on linear models (Searle (1991)), variance components (Searle, Casella and McCulloch (1992)), and matrix theory for statisticians (Searle (1982)). Henderson was particularly interested in selection of sires (breeding bulls) in the dairy industry. His work is credited with having produced great gains in yields, of great economic value.

3. Athletics times. One may wish to study the effect of ageing on athletes past their peak. One way to do this is to extract from the race results of a particular race over successive years the performances of athletes competing repeatedly. Again, individual athletes differ; these are the random effects. Fixed effects one might be interested in include age, sex and club status. For background, see Bingham and Rashid (2008).

We shall follow the notation of §5.1 fairly closely. Thus we write

$$W = (X, Z)$$

for the new design matrix ($n \times (p+q)$). It is convenient to take the random effects – which as is customary we denote by u – to have zero mean (any additive terms coming from the mean Eu can be absorbed into the fixed effects). Thus the *linear mixed model* is defined by

$$y = X\beta + Zu + \epsilon, \qquad (LMM)$$

where (both means are zero and) the covariance matrices are given by

$$E\epsilon = Eu = 0, \ \mathrm{cov}(\epsilon, u) = 0, \quad R := \mathrm{var}\ \epsilon, \ D := \mathrm{var}\ u,$$

('R for regresssion, D for dispersion'). One can write (LMM) as an ordinary linear model,

$$y = X\beta + \epsilon^*, \qquad \epsilon^* := Zu + \epsilon.$$

By Proposition 4.5, this has covariance matrix

$$V := \mathrm{cov}\ \epsilon^* = ZDZ^T + R$$

('V for variance'). So by Theorem 3.5, the generalised least-squares solution is

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y. \qquad (GLS)$$

We now specify the distributions in our model by assuming that u is multivariate normal (multinormal), and that the conditional distribution of y given u is also multinormal:

$$y|u \sim N(X\beta + Zu, R), \qquad u \sim N(0, D). \qquad (NMM)$$

Then the (unconditional) distribution of y is a *normal mean mixture*, whence the name (NMM). Now the joint density $f(y, u)$ is

$$f(y,u) = f(y|u)f(u),$$

the product of the conditional density of y given u and the density of u. So

$$f(y,u) = const. \exp\{-\frac{1}{2}(y-X\beta-Zu)^T R^{-1}(y-X\beta-Zu)\}. \exp\left\{-\frac{1}{2}u^T D^{-1}u\right\}.$$

Thus to maximise the likelihood (with respect to β and u), we maximise $f(y, u)$, that is, we minimise:

$$\min \qquad (y - X\beta - Zu)^T R^{-1}(y - X\beta - Zu) + u^T D^{-1}u. \qquad (pen)$$

Note the different roles of the two terms. The first, which contains the data, comes from the likelihoood; the second comes from the random effects. It serves as a penalty term (the penalty we pay for not knowing the random effects). So we have here a *penalised likelihood* (recall we encountered penalised likelihood in §5.2.1, in connection with nested models and AIC).

The least-squares solution of Chapters 3, 4 gives the best linear unbiased estimator or BLUE (see §3.3). It is conventional to speak of *predictors*, rather than estimators, with random effects. The solution is thus a *best linear unbiased predictor*, or BLUP.

Theorem 9.1

The BLUPs – the solutions $\hat{\beta}$, $\hat{u}$, of the minimisation problem (MME) – satisfy

$$\left.\begin{array}{rcl} XR^{-1}X\hat{\beta} + X^T R^{-1}Z\hat{u} & = & X^T R^{-1}y, \\ ZR^{-1}X\hat{\beta} + \left[Z^T R^{-1}Z + D^{-1}\right]\hat{u} & = & Z^T R^{-1}y \end{array}\right\} \qquad (MME)$$

(Henderson's mixed model equations of 1950).

Proof

We use the vector calculus results of Exercises 3.6–3.7. If we expand the first term in (pen) above, we obtain nine terms, but the quadratic form in y does not involve β or u, so we discard it; this with the second term above gives nine terms, all scalars, so all their own transposes. This allows us to combine three pairs of terms, reducing to six terms, two linear in β, two linear in u and two cross terms in β and u; there is also a quadratic term in β, and two quadratic terms in u, which we can combine. Setting the partial derivatives with respect to β and u equal to zero then gives

$$\begin{aligned} -2y^T R^{-1} X + 2u^T Z^T R^{-1} X + 2\beta^T X^T R^{-1} X &= 0, \\ -2y^T R^{-1} Z + 2\beta^T X^T R^{-1} Z + 2u^T \left[Z^T R^{-1} Z + D^{-1} \right] &= 0, \end{aligned}$$

or

$$\left.\begin{aligned} X^T R^{-1} X\beta + X^T R^{-1} Z u &= X^T R^{-1} y, \\ Z^T R^{-1} X\beta + [Z^T R^{-1} Z + D^{-1}] u &= Z^T R^{-1} y, \end{aligned}\right\} \qquad (MME)$$

as required. □

9.1.1 Mixed models and Generalised Least Squares

To proceed, we need some matrix algebra. The next result is known as the *Sherman–Morrison–Woodbury formula*, or *Woodbury's formula* (of 1950).

Lemma 9.2 (Woodbury's Formula)

$$(A + UBV)^{-1} = A^{-1} - A^{-1}U.(I + BVA^{-1}U)^{-1}.BVA^{-1},$$

if all the matrix products are conformable and all the matrix inverses exist.

Proof

We have to show that if we pre-multiply or post-multiply the right by $A+UBV$ we get the identity I.

Pre-multiplying, we get four terms. Taking the first two as those from $(A + UBV)A^{-1}$, these are

$$\begin{aligned} &I + UBVA^{-1} - U(I + BVA^{-1}U)^{-1}BVA^{-1} \\ &\qquad -UBVA^{-1}U(I + BVA^{-1}U)^{-1}BVA^{-1}. \end{aligned}$$

The third and fourth terms combine, to give

$$I + UBVA^{-1} - U.BVA^{-1} = I,$$

as required. The proof for post-multiplying is similar. □

Applied in the context of §9.1 (where now $V := ZDZ^T + R$, as above), this gives

Corollary 9.3

(i)

$$V^{-1} := (ZDZ^T + R)^{-1} = R^{-1} - R^{-1}Z(Z^TR^{-1}Z + D^{-1})^{-1}ZR^{-1}.$$

(ii)

$$DZ^TV^{-1} = (Z^TR^{-1}Z + D^{-1})^{-1}Z^TR^{-1}.$$

Proof

For (i), we use Woodbury's Formula with R, Z, D, Z^T for A, U, B, V:

$$\begin{aligned}(R + ZDZ^T)^{-1} &= R^{-1} - R^{-1}Z.(I + DZ^TR^{-1}Z)^{-1}.DZ^TR^{-1}\\ &= R^{-1} - R^{-1}Z.[D(D^{-1} + Z^TR^{-1}Z)]^{-1}.DZ^TR^{-1}\\ &= R^{-1} - R^{-1}Z.(D^{-1} + Z^TR^{-1}Z)^{-1}.Z^TR^{-1}.\end{aligned}$$

For (ii), use Woodbury's Formula with D^{-1}, Z^T, R^{-1}, Z for A, U, B, V:

$$(D^{-1} + Z^TR^{-1}Z)^{-1} = D - DZ^T.(I + R^{-1}ZDZ^T)^{-1}.R^{-1}ZD,$$

so

$$(D^{-1}+Z^TR^{-1}Z)^{-1}Z^TR^{-1} = DZ^TR^{-1} - DZ^T(I+R^{-1}ZDZ^T)^{-1}R^{-1}ZDZ^TR^{-1}.$$

The right is equal to $DZ^T[I - (I + R^{-1}ZDZ^T)^{-1}R^{-1}ZDZ^T]R^{-1}$, or equivalently, to $DZ^T[I - (I + R^{-1}ZDZ^T)^{-1}\{(I + R^{-1}ZDZ^T) - I\}]R^{-1}$. Combining, we see that

$$\begin{aligned}(D^{-1} + Z^TR^{-1}Z)^{-1}Z^TR^{-1} &= DZ^T[I - I + (I + R^{-1}ZDZ^T)^{-1}]R^{-1}\\ &= DZ^T(R + ZDZ^T)^{-1}\\ &= DZ^TV^{-1},\end{aligned}$$

as required. □

Theorem 9.4

The BLUP $\hat{\beta}$ in Theorem 9.1 is the same as the generalised least-squares estimator:

$$\hat{\beta} = \left(X^TV^{-1}X\right)^{-1}X^TV^{-1}y. \qquad (GLS)$$

The BLUP $\hat{u}$ is given by either of

$$\hat{u} = \left(Z^T R^{-1} Z + D^{-1}\right)^{-1} Z^T R^{-1} \left(y - X\hat{\beta}\right)$$

or

$$\hat{u} = D Z^T V^{-1} \left(y - X\hat{\beta}\right).$$

Proof

We eliminate $\hat{u}$ between the two equations (MME). To do this, pre-multiply the second by $X^T R^{-1} Z (Z^T R^{-1} Z + D^{-1})^{-1}$ and subtract. We obtain that

$$X^T R^{-1} X\hat{\beta} - X^T R^{-1} Z \left(Z^T R^{-1} Z + D^{-1}\right)^{-1} Z^T R^{-1} X\hat{\beta} \qquad = \\ X^T R^{-1} y - X^T R^{-1} Z \left(Z^T R^{-1} Z + D^{-1}\right)^{-1} Z^T R^{-1} y. \qquad (a)$$

Substitute the matrix product on the right of Corollary 9.3(i) into both sides of (a):

$$X^T R^{-1} X\hat{\beta} - X^T \left\{R^{-1} - V^{-1}\right\} X\hat{\beta} = X^T R^{-1} y - X^T \left\{R^{-1} - V^{-1}\right\} y,$$

or

$$X^T V^{-1} X\hat{\beta} = X^T V^{-1} y,$$

which is

$$\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y,$$

as in (GLS).

The first form for $\hat{u}$ follows from the second equation in (MME). The second follows from this by Corollary 9.3(ii). □

The conditional density of u given y is

$$f(u|y) = f(y,u)/f(y) = f(y|u)f(u)/f(y)$$

(an instance of Bayes's Theorem: see e.g. Haigh (2002), §2.2). We obtain $f(y)$ from $f(y,u)$ by integrating out u (as in §1.5 on the bivariate normal). By above (below (NMM)), $f(y,u)$ is equal to a constant multiplied by

$$\exp\{-\frac{1}{2}[u^T (Z^T R^{-1} Z + D^{-1}) u - 2u^T Z^T R^{-1} (y - X\beta) + (y - X\beta)^T R^{-1} (y - X\beta)]\}.$$

This has the form of a multivariate normal. So by Theorem 4.25, $u|y$ is also multivariate normal. We can pick out *which* multivariate normal by identifying the mean and covariance from Edgeworth's Theorem, Theorem 4.16 (see also Note 4.30). Looking at the quadratic term in u above identifies the covariance matrix as $(Z^T R^{-1} Z + D^{-1})^{-1}$. Then looking at the linear term in u identifies the mean as

$$\left(Z^T R^{-1} Z + D^{-1}\right)^{-1} Z^T R^{-1} (y - X\beta).$$

Here β on the right is unknown; replacing it by its BLUP $\hat{\beta}$ gives the first form for $\hat{u}$ (recall from §4.5 that a regression is a conditional mean; this replacement of β by $\hat{\beta}$ is called a *plug-in estimator*). The interpretation of the second form of $\hat{u}$, in terms of the regression of u on y with $\hat{\beta}$ plugged in for β, is similar (as in (GLS), with $(X^T V^{-1} X)^{-1}$ replaced by $(I^T D^{-1} I)^{-1} = D$, X^T by Z^T and y by $y - X\hat{\beta}$.

Note 9.5

1. The use of Bayes's Theorem above is very natural in this context. In Bayesian Statistics, parameters are no longer unknown constants as here. Our initial uncertainty about them is expressed in terms of a distribution, given here by a density, the *prior density.* After sampling and obtaining our data, one uses Bayes's Theorem to update this prior density to a *posterior density.* From this Bayesian point of view, the distinction between fixed and random effects in the mixed model above evaporates. So one can expect simplification, and unification, in a Bayesian treatment of the Linear Model. However, one should first meet a treatment of Bayesian Statistics in general, and for this we must refer the reader elsewhere. For a Bayesian treatment of the Linear Model (fixed effects), see Williams (2001), §8.3.

Bayes's Theorem stems from the work of Thomas Bayes (1702–1761, posthumously in 1764). One of the founders of modern Bayesian Statistics was I. J. Good (1916–2009, from 1950 on). Good also pioneered penalised likelihood, which we met above and will meet again in §9.2 below.

2. In Henderson's mixed model equations (MME), one may combine β and u into one vector, v say, and express (MME) as one matrix equation, $Mv = c$ say. This may be solved as $v = M^{-1}c$. Here, one needs the inverse of the partitioned matrix M. We have encountered this in Exercise 4.10. The relevant Linear Algebra involves the *Schur complement*, and gives an alternative to the approach used above via Woodbury's Formula.

Example 9.6 (Mixed model analysis of ageing athletes)

We give a brief illustration of mixed models with an application to the athletics data in Table 9.1.

In S-Plus/R® the basic command is `lme`, although in R® this requires loading the package `nlme`. We fit a model using Restricted Maximum Likelihood (REML) with fixed effects for the intercept, age and club status, and a random intercept depending on each athlete.

Athlete	Age	Club	Time	Athlete	Age	Club	Time
1	38	0	91.500	4	41	0	91.167
1	39	0	89.383	4	42	0	90.917
1	40	0	93.633	4	43	0	90.883
1	41	0	93.200	4	44	0	92.217
1	42	0	93.533	4	45	1	94.283
1	43	1	92.717	4	46	0	99.100
2	53	1	96.017	5	54	1	105.400
2	54	1	98.733	5	55	1	104.700
2	55	1	98.117	5	56	1	106.383
2	56	1	91.383	5	57	1	106.600
2	58	1	93.167	5	58	1	107.267
2	57	1	88.950	5	59	1	111.133
3	37	1	83.183	6	57	1	90.250
3	38	1	83.500	6	59	1	88.400
3	39	1	83.283	6	60	1	89.450
3	40	1	81.500	6	61	1	96.380
3	41	1	85.233	6	62	1	94.620
3	42	0	82.017				

Table 9.1 Data for Example 9.6. The times are taken from athletes regularly competing in the Berkhamsted Half–Marathon 2002–2007.

```
m1.nlme<-lme(log(time)~club+log(age), random=~1|athlete)
summary(m1.nlme)
```

From the output, t-statistics show that the fixed effects term for age is significant ($p = 0.045$) but suggest that a fixed effects term for club status is not needed ($p = 0.708$). We repeat the analysis, excluding the fixed effects term for club status:

```
m2.nlme<-lme(log(time)~log(age), random=~1|athlete)
```

Next we fit a model with a fixed effect term for age, but allow for the possibility that this coefficient can vary randomly between athletes:

```
m3.nlme<-lme(log(time)~log(age), random=~1+log(age)|athlete)
```

The AIC for these latter two models are -114.883 and -112.378 respectively, so the most appropriate model appears to be the model with a random intercept term and a fixed age-effect term. Log(age) is significant in the chosen model – a t-test gives a p-value of 0.033. A 95% confidence interval for the coefficient of log(age) is 0.229 ± 0.209, consistent with earlier estimates in Examples 3.37 and 8.3, although this time this estimate has a higher level of uncertainty attached to it.

One reason why the ageing effect appears to be weaker here is that the Berkhamsted Half-Marathon (in March) is often used as a 'sharpener' for the London Marathon in April. One could allow for this by using a Boolean variable for London Marathon status (though full data here would be hard to obtain for any data set big enough for the effort to be worthwhile).

9.2 Non-parametric regression

In §4.1 on polynomial regression, we addressed the question of fitting a function $f(x)$ more general than a straight line through the data points in the least-squares sense. Because polynomials of high degree are badly behaved numerically, we restricted attention there to polynomials of low degree. This is a typical parametric setting.

However, we may need to go beyond this rather restricted setting, and if we do the number of parameters we use can increase. This provides more flexibility in fitting. We shall see below how spline functions are useful in this context. But the point here is that we can now move to a function-space setting, where the dimensionality of the function space is infinite. We will use only finitely many parameters. Nevertheless, because the number of parameters available is infinite, and because one usually uses the term *non-parametric* to describe situations with infinitely many parameters, this area is referred to as *non-parametric regression.*

The idea is to choose some suitable set of basic, or simple, functions, and then represent functions as finite linear combinations of these. We have met this before in §4.1, where the basic functions are powers, and §4.1.2, where they are orthogonal polynomials. The student will also have met such ideas in Fourier analysis, where we represent functions as series of sines and cosines (infinite series in theory, finite series in practice). Many other sets of basic functions are in common use – splines, to which we now turn, radial basis functions, wavelets, etc. The relevant area here is Approximation Theory, and we must refer to a text in that area for details and background; see e.g. Ruppert, Wand and Carroll (2003).

The above deals with functions of one variable, or problems with one covariate, but in Chapter 3 we already have extensive experience of problems with several covariates. A similar extension of the treatment to higher dimensions is possible here too. For brevity, we will confine such extensions to two dimensions. Non-parametric regression in two dimensions is important in Spatial Statistics, to which we return in the next subsection.

Recall that in §4.1 on polynomial regression we found that polynomials of high degree are numerically unstable. So if a polynomial of low degree does not suffice, one needs functions of some other kind, and a suitable function class is provided by *splines*. A spline of *degree* p is a continuous function f that is piecewise polynomial of degree p, that is, polynomial of degree p on subintervals $[x_i, x_{i+1}]$, where f and its derivatives f', ..., $f^{(p-1)}$ are continuous at the points x_i, called the *knots* of the spline. Typical splines are of the form

$$(x-a)_+^k, \qquad x_+^k := \begin{cases} x^k, & x \geq 0, \\ 0, & x < 0. \end{cases}$$

We shall restrict ourselves here to *cubic splines*, with $p = 3$; here f, f' and f'' are continuous across the knots x_i. These may be formed by linear combinations of functions of the above type, with $k \leq 3$ and a the knots x_i. It is possible and convenient, to restrict to *basic splines*, or *B-splines*. These are of local character, which is convenient numerically, and one can represent any spline as a linear combination of B-splines. For background and details, see e.g. de Boor (1978).

Suppose now we wish to approximate data y_i at points x_i. As with polynomial regression, we can approximate arbitrarily closely in the least-squares sense, but this is no use to us as the approximating functions are unsuitable. This is because they oscillate too wildly, or are insufficiently smooth. To control this, we need to *penalise* functions that are too rough. It turns out that a suitable measure of roughness for cubic splines is provided by the integral $\int (f'')^2$ of the squared second derivative. We are led to the minimisation problem

$$\min \quad \sum\nolimits_{i=1}^{n} (y_i - f(x_i))^2 + \lambda^2 \int (f''(x))^2 \, dx.$$

Here the first term is the sum of squares as before, the integral term is a *roughness penalty*, and λ^2 is called a *smoothing parameter*. (As the sum is of the same form as in the likelihood theory of earlier chapters, and the integral is a penalty term, the method here is called *penalised likelihood* or *penalised log-likelihood*.) With λ small, the roughness penalty is small and the minimiser is close to the least-squares solution as before; with λ large, the roughness penalty is large, and the minimiser will be smooth, at the expense of giving a worse least-squares fit. Since λ is under our control, we have a choice as to how much smoothness we wish, and at what cost in goodness of fit.

It turns out that the minimising function f above is necessarily a cubic spline with knots at the points x_i. This will be a linear combination of B-splines $B_j(x)$, with coefficients β_j say. Forming the β_j into a vector β also, the approximating f is then

$$f(x) = \beta^T B(x),$$

and the mimimisation problem is of the form

$$\min \quad \sum\nolimits_{i=1}^{n} \left(y_i - \beta^T B(x_i)\right)^2 + \lambda^2 \beta^T D \beta,$$

for some symmetric positive semi-definite matrix D whose entries are integrals of products of derivatives of the basic splines.

This minimisation problem is of the same form as that in §9.1 for BLUPS, and may be solved in the same way: smoothing splines are BLUPs. Let X be the matrix with ith row $B(x_i)^T$. One obtains the minimising β and fitted values $\hat{y}$ as

$$\hat{\beta} = (X^T X + \lambda^2 D)^{-1} X^T y, \qquad \hat{y} = X(X^T X + \lambda^2 D)^{-1} X^T y = S_\lambda y,$$

say, where S_λ is called the *smoother matrix.* Use of smoothing splines can be implemented in S-Plus/R® by the command `smooth.spline`; see Venables and Ripley (2002), §8.7. For background and details, see Green and Silverman (1994), Ruppert, Wand and Carroll (2003).

Splines were studied by I. J. Schoenberg (1903–1990) from 1946 on, and were used in Statistics by Grace Wahba (1934–) from 1970 on. The term spline derives from the flexible metal strips used by draughtsmen to construct smooth curves interpolating fixed points, in the days before computer-aided design (CAD). Penalised likelihood and roughness penalties go back to I. J. Good (with his student R. A. Gaskins) in 1971 (preceding the AIC in 1974).

9.2.1 Kriging

Kriging describes a technique for non-parametric regression in spatial problems in multiple (commonly three) dimensions. The original motivation was to model ore deposits in mining, though applications extend beyond geology and also typically include remote sensing and black-box modelling of computer experiments. The name kriging derives from the South African mining engineer D. G. Krige (1919–), and was further developed in the 1960s by the French mathematician G. Matheron (1930–2000) at the Paris School of Mines. The basic idea behind kriging is as follows. We observe data

$$(\mathbf{x_1}, y_1), \ldots, (\mathbf{x_n}, y_n),$$

where the $\mathbf{x_i} \in \mathbb{R}^d$ and the $y_i \in \mathbb{R}$. We might imagine the $\mathbf{x_i}$ as a sequence of co-ordinates and the y_i as corresponding to observed levels of mineral deposits. If $d = 2$, this picture corresponds to a three-dimensional plot in which y is the height. Given the observed sequence of $(\mathbf{x_i}, y_i)$ we wish to estimate the y values corresponding to a new set of data $\mathbf{x_0}$. We might, for example, envisage this set-up corresponding to predicting the levels of oil or mineral deposits, or some environmental pollutant etc., at a set of new locations given a set of historical measurements. The set-up for our basic kriging model is

$$y_i = \mu + S(\mathbf{x_i}) + \epsilon_i,$$

where $S(\mathbf{x})$ is a zero-mean stationary stochastic process in $\mathbb{R}^d$ with covariance matrix $\mathbf{C}$ independent of the ϵ_i, which are assumed iid $N(0, \sigma^2)$. However, this formulation can be made more general by choosing $\mu = \mu(x)$ (Venables and Ripley (2002), Ch. 15). It is usually assumed that

$$C_{ij} = \operatorname{cov}\left(S(\mathbf{x_i}, \mathbf{x_j})\right) = C(||\mathbf{x_i} - \mathbf{x_j}||), \qquad \text{(Isotropy)}$$

although more general models which do not make this assumption are possible. Suppose that the ϵ_i and $S(\cdot)$ are multivariate normal. By §4.6 the mean square error is minimised by the Conditional Mean Formula given by Theorem 4.25. We have that

$$\begin{pmatrix} \mathbf{y}(\mathbf{x_0}) \\ y(\mathbf{x_0}) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu\mathbf{1} \\ \mu \end{pmatrix}, \begin{pmatrix} (\mathbf{C} + \sigma^2 I) & c_0 \\ c_0^T & \sigma^2 \end{pmatrix} \right),$$

where $\mathbf{1}$ denotes a column vector of 1s. It follows that the optimal prediction (best linear predictor) for the unobserved $y(\mathbf{x_0})$ given the observed $\mathbf{y}(\mathbf{x_0})$ is given by

$$\hat{y}(\mathbf{x_0}) = \mu + c_0^T \left(\mathbf{C} + \sigma^2 I\right)^{-1} (\mathbf{y}(\mathbf{x_0}) - \mu\mathbf{1}). \qquad (BLP)$$

From first principles, it can be shown that this still gives the *best linear predictor (BLP)* when we no longer assume that $S(\mathbf{x})$ and ϵ_i are Gaussian. In practice $\mathbf{C}$ can be estimated using either maximum likelihood or variogram methods (some details can be found in Ruppert, Wand and Carroll (2003), Ch. 13 or Venables and Ripley (2002), Ch. 15). As presented in Ruppert, Wand and Carroll (2003) the full kriging algorithm is as follows:

1. Estimate the covariance function C, σ^2 and set $\mu = \overline{y}$.
2. Construct the estimated covariance matrix $\hat{\mathbf{C}} = C(||x_i - x_j||)$.
3. Set up a mesh of $\mathbf{x_0}$ values in the region of interest.
4. Using (BLP) construct a set of predicted values $\hat{y}(\mathbf{x_0})$.
5. Plot $\hat{y}(\mathbf{x_0})$ against x_0 to estimate the relevant spatial surface.

As briefly discussed in Ruppert, Wand and Carroll (2003), Ch. 13.3–4. it is possible to relate kriging to the non-parametric regression models with a non-parametric regression model using cubic splines. In particular, two-dimensional kriging can be shown to be equivalent to minimising

$$\sum\nolimits_{i=1}^{n} (y_i - f(x_1, x_2))^2 + \lambda \int\int \left(f^2_{x_1x_1} + 2f^2_{x_1x_2} + f^2_{x_2x_2}\right)\, dx_1\, dx_2.$$

This gives an integral of the sum of squares of second derivatives to generalise cubic splines; see e.g. Cressie (1993) §3.4.5 for further details.

The end product of a kriging study may well be some computer graphic, perhaps in (a two-dimensional representation of) three dimensions, perhaps in colour, etc. This would be used to assist policy makers in decision taking – e.g. whether or not to drill a new oil well or mine shaft in some location, whether or not to divert traffic, or deny residential planning permission, for environmental reasons, etc. Specialist software is needed for such purposes.

9.3 Experimental Design

9.3.1 Optimality criteria

We have already seen in §7.1 how to identify unusual data points, in terms of their *leverage* and *influence*. For example, Cook's distance D_i is defined by a quadratic form in the information matrix $C = A^T A$ formed from the design matrix A. Before conducting the statistical experiment that leads to our data y, the design matrix A is still at our disposal, and it is worth considering whether we can choose A in some good way, or better still, in some optimal way. This is indeed so, but there are a number of different possible optimality criteria. One criterion in common use is to maximise the determinant of the information matrix C, the determinant $|C|$ serving as a measure of quantity of information (recall from vector algebra that the volume of a parallelepiped with sides three 3-vectors is the determinant of their co-ordinates).

The situation is similar to that in our first course in Statistics, when we discussed estimation of parameters. Here two important measures of quality of an estimator $\hat{\theta}$ of a parameter θ are *bias*, $E\hat{\theta} - \theta$, and *precision*, measured by the inverse of the variance var θ; we can think of this variance as a measure of sampling error, or noise. We want to keep both noise and bias low, but it is pointless to diminish one at the expense of increasing the other. One thus has a *noise–bias tradeoff*, typical in Statistics. To choose how to make this trade–off, one needs some optimality criterion. This is usually done by choosing some *loss function* (or alternatively, some *utility function*). One then *minimises expected loss* (or *maximises expected utility*). This area of Statistics is called Decision Theory.

The situation here is similar. One needs some optimality criterion for the experimental design (there are a number in common use) – maximising the determinant as above corresponds to D-optimality – and seeks to optimise the design with respect to this criterion. For further detail, we must refer to a book on Optimal Experimental Design, for example Atkinson and Donev (1992).

9.3.2 Incomplete designs

In addition to the profoundly mathematical criteria above, there are also more tangible ways in which experimental design can bring benefits to experimenters by reducing the sample size requirements needed in order to perform a full analysis. It is frequently impractical, say in an agricultural experiment, to grow or include every combination of treatment and block. (Recall that in §2.7 every combination of treatment and block occurred *once*, with multiple replications possible in §2.8.)

Rather than admitting defeat and returning to one-way ANOVA (hence confounding treatment effects with block effects) we need some *incomplete design* which nonetheless enables all treatment and block effects to be estimated. The factors of treatment and block need to be *balanced*, meaning that any two treatments occur together in the same block an equal number of times. This leads to a set of designs known as *balanced incomplete block designs (BIBD).* These designs are usually tabulated, and can even be used in situations where the blocks are of insufficient size to accommodate one whole treatment allocation (provided that the allocation of experimental units is appropriately randomised). For full details and further reference we refer to Montgomery (1991), Ch. 6. Analysis of large experiments using fractions of the permissible factor combinations is also possible in so-called *factorial* experiments using *fractional factorial designs* (see Montgomery (1991) Ch. 9–12).

Example 9.7 (Latin Squares)

We consider briefly the simplest type of incomplete block design. Suppose we have (e.g.) five types of treatment (fertiliser) to apply to five different varieties of wheat on five different types of soil. This simple experiment leads to 125 different factor combinations in total. It is economically important to be able to test

$$H_0 : \text{The treatment (fertiliser) means are all equal,}$$

in such two-factor experiments (variety and soil type) with fewer than 125 readings. We can make do with 25 readings by means of a 5×5 *Latin square* (see Table 9.2). Each cell contains each fertiliser type once, showing that the design is indeed balanced. Given experimental observations, an ANOVA table with *three* factors (Soil type, Variety and Fertiliser) can be constructed by using the general methods of Chapter 2.

	Variety				
Soil Type	1	2	3	4	5
1	1	2	3	4	5
2	5	1	2	3	4
3	4	5	1	2	3
4	3	4	5	1	2
5	2	3	4	5	1

Table 9.2 5×5 Latin square design. Fertiliser allocations by Soil Type and Variety.

Analysis of $n \times n$ Latin squares. We show how to perform a skeleton ANOVA for a $n \times n$ Latin square design. The approach follows the same general outline laid out in Chapter 2, but generalises §2.6–2.7 by including three factors. In effect, we isolate treatment effects by 'blocking' over rows and columns. The model equation can be written as

$$X_{ijk} = \mu + r_i + c_j + t_k + \epsilon_{ijk}, \quad \epsilon_{ijk} \quad \text{iid} \quad N(0, \sigma^2),$$

for $i, j = 1 \ldots, n$, where $k = k(i, j)$ is the entry in the Latin square in position (i, j) in the matrix. Note $k = 1, \ldots, n$ also. The r_i, c_j, t_k denote row, column and treatment effects respectively and satisfy the usual constraints:

$$\sum\nolimits_i r_i = \sum\nolimits_j c_j = \sum\nolimits_k t_k = 0.$$

Write

$$\begin{aligned} R_i &= i\text{th} \quad \text{row total}, \quad X_{i\bullet} = R_i/n = i\text{th} \quad \text{row mean}, \\ C_j &= j\text{th} \quad \text{column total}, \quad X_{\bullet j} = C_j/n = j\text{th} \quad \text{column mean}, \end{aligned}$$

$$\begin{array}{rcl} T_k & = & k\text{th} \quad \text{treatment total}, \quad X_{(k)} = T_k/n = k\text{th} \quad \text{treatment mean}, \\ G & = & \text{grand total} = \sum_i \sum_j \sum_k X_{ijk}, \quad \overline{X} = G/n \quad \text{grand mean}. \end{array}$$

The following algebraic identity can be verified:

$$SS := SSR + SSC + SST + SSE,$$

where

$$\begin{array}{rcl} SS & := & \sum_i \sum_j \sum_k \left(X_{ijk} - \overline{X} \right)^2 = \sum_i \sum_j \sum_k X_{ijk}^2 - \dfrac{G^2}{n^2}, \\ SSR & := & n \sum_i \left(X_{i\bullet} - \overline{X} \right)^2 = \dfrac{1}{n} \sum_i R_i^2 - \dfrac{G^2}{n^2}, \\ SSC & := & n \sum_j \left(X_{\bullet j} - \overline{X} \right)^2 = \dfrac{1}{n} \sum_j C_j^2 - \dfrac{G^2}{n^2}, \\ SST & := & n \sum_k \left(X_{(k)} - \overline{X} \right)^2 = \dfrac{1}{n} \sum_k T_k^2 - \dfrac{G^2}{n^2}, \\ SSE & := & \sum_i \sum_j \sum_k \left(X_{ijk} - X_{i\bullet} - X_{\bullet j} - X_{(k)} + 2\overline{X} \right)^2, \end{array}$$

with $SSE = SS - SSR - SSC - SST$ as before. An Analysis of Variance of this model can be performed as laid out in Table 9.3.

Source	df	SS	Mean Square	F
Rows	$n-1$	SSR	$MSR = \frac{SSR}{n-1}$	MSR/MSE
Columns	$n-1$	SSC	$MSC = \frac{SSC}{n-1}$	MSC/MSE
Treatments	$n-1$	SST	$MST = \frac{SST}{n-1}$	MST/MSE
Residual	$(n-1)(n-2)$	SSE	$MSE = \frac{SSE}{(n-1)(n-2)}$	
Total	n^2-1	SS		

Table 9.3 ANOVA table for $n \times n$ Latin square

Note 9.8

While Experimental Design is a very useful and practical subject, it also uses a lot of interesting pure mathematics. One area important here is projective geometry over finite fields; see e.g. Hirschfeld (1998). Whereas the mathematics here is discrete, as one would expect since matrix theory is involved, important insights can be gained by using a continuous framework, and so analysis rather than algebra; see e.g. Wynn (1994).

Experimental Design is one of a number of areas pioneered by Fisher in his time at Rothamsted in the 1920s, and by his Rothamsted colleague Frank

Yates (1902–1994). Fisher published his book *The Design of Experiments* in 1935.

9.4 Time series

It often happens that data arrive sequentially in time. This may result in measurements being taken at regular intervals – for example, daily temperatures at noon at a certain meteorological recording station, or closing price of a particular stock, as well as such things as monthly trade figures and the like. We may suppose here that time is measured in discrete units, and that the nth reading is X_n. Then the data set $X = (X_n)$ is called a *time series* (TS).

One often finds in time series that high values tend to be followed by high values, or low values by low values. Typically this is the case when the underlying system has some dynamics (probably complicated and unknown) that tends to fluctuate about some mean value, but intermittently undergoes some perturbation away from the mean in some direction, this perturbation showing a marked tendency to persist for some time, rather than quickly die away.

In such cases one has a *persistence of memory* phenomenon; the question is how long does memory persist? Sometimes memory persists indefinitely, and the infinitely *remote past* continues to exert an influence (rather as the magnetism in a rock reflects the conditions when the rock solidified, in a former geological era, or tempered steel locks in its thermal history as a result of the tempering process). But more commonly only the recent past really influences the present. Using p for the number of parameters as usual, we may represent this by a model in which the present value X_t is influenced by the last p values $X_{t-1}, \ldots, X_{t-p}$. The simplest such model is a linear regression model, with these as covariates and X_t as dependent variable. This gives the model equation

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \epsilon_t. \qquad (AR(p))$$

Here the ϕ_i are the parameters, forming a vector ϕ, and the ϵ_t are independent errors, normally distributed with mean 0 and common variance σ^2. This gives an *autoregressive model* of *order* p, $AR(p)$, so called because the process X is *regressed on itself.*

For simplicity, we centre at means (that is, assume all $EX_t = 0$) and restrict to the case when $X = (X_n)$ is *stationary* (that is, its distribution is invariant under shifts in time). Then the covariance depends only on the time difference – or rather, its modulus, as the covariance is the same for two variables, either way round; similarly for the correlation, on dividing by the variance σ^2. Write this as $\rho(k)$ at lag k:

$$\rho(k) = \rho(-k) = E[X_t X_{t-k}].$$

Multiplying $(AR(p))$ by X_k and taking expectations gives

$$\rho(k) = \phi_1 \rho(k-1) + \ldots + \phi_p \rho(k-p) \qquad (k > 0). \qquad (YW)$$

These are the *Yule–Walker equations* (G. Udny Yule in 1926, Sir Gilbert Walker in 1931). One has a *difference equation* of *order* p, with *characteristic polynomial*

$$\lambda^p - \phi_1 \lambda^{p-1} - \ldots - \phi_p = 0.$$

If λ_1, ..., λ_p are the roots of this polynomial, then the general solution is

$$\rho(k) = c_1 \lambda_1^k + \ldots + c_p \lambda_p^k$$

(if the roots are distinct, with appropriate modification for repeated roots). Since $\rho(.)$ is a correlation, one has $|\rho(k)| \leq 1$ for all k, which forces

$$|\lambda_i| \leq 1 \qquad (i = 1, \ldots, p).$$

One may instead deal with *moving average* processes of *order* q,

$$X_t = \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t, \qquad (MA(q))$$

or with a combination,

$$X_t = \phi_1 X_{t-1} + \ldots + \phi_p X_{t-p} + \theta_1 \epsilon_{t-1} + \ldots + \theta_q \epsilon_{t-q} + \epsilon_t. \quad (ARMA(p,q))$$

The class of autoregressive moving average models, or $ARMA(p,q)$ processes, is quite rich and flexible, and is widely used. We refer to e.g. Box and Jenkins (1970), Brockwell and Davis (2002) for details and background.

9.4.1 Cointegration and spurious regression

Integrated processes. One standard technique used to reduce non-stationary time series to the stationary case is to *difference* them repeatedly (one differencing operation replaces X_t by $X_t - X_{t-1}$). If the series of dth differences is stationary but that of $(d-1)$th differences is not, the original series is said to be *integrated* of *order* d; one writes

$$(X_t) \sim I(d).$$

Cointegration. If $(X_t) \sim I(d)$, we say that (X_t) is *cointegrated* with *cointegration vector* α if $(\alpha^T X_t)$ is (integrated of) order less than d.

A simple example of cointegration arises in random walks. Suppose $X_n = \sum_{i=1}^{n} \xi_i$ with the ξ_n iid random variables, and $Y_n = X_n + \epsilon_n$, with the ϵ_n iid errors as above, is a noisy observation of X_n. Then the bivariate process $(X, Y) = (X_n, Y_n)$ is integrated of order 1, with cointegration vector $(1, -1)^T$.

Cointegrated series are series that tend to move together, and commonly occur in economics. These concepts arose in econometrics, in the work of R. F. Engle (1942–) and C. W. J. (Sir Clive) Granger (1934–2009) in 1987. Engle and Granger gave (in 1991) an illustrative example – the prices of tomatoes in North Carolina and South Carolina. These states are close enough for a significant price differential between the two to encourage sellers to transfer tomatoes to the state with currently higher prices to cash in; this movement would increase supply there and reduce it in the other state, so supply and demand would move the prices towards each other.

Engle and Granger received the Nobel Prize in Economics in 2003. The citation included the following:

> Most macroeconomic time series follow a stochastic trend, so that a temporary disturbance in, say, GDP has a long-lasting effect. These time series are called nonstationary; they differ from stationary series which do not grow over time, but fluctuate around a given value. Clive Granger demonstrated that the statistical methods used for stationary time series could yield wholly misleading results when applied to the analysis of nonstationary data. His significant discovery was that specific combinations of nonstationary time series may exhibit stationarity, thereby allowing for correct statistical inference. Granger called this phenomenon cointegration. He developed methods that have become invaluable in systems where short-run dynamics are affected by large random disturbances and long-run dynamics are restricted by economic equilibrium relationships. Examples include the relations between wealth and consumption, exchange rates and price levels, and short and long-term interest rates.

Spurious regression. Standard least-squares models work perfectly well if they are applied to *stationary* time series. But if they are applied to *non-stationary* time series, they can lead to spurious or nonsensical results. One can give examples of two time series that clearly have nothing to do with one another, because they come from quite unrelated contexts, but nevertheless have quite a high value of R^2. This would normally suggest that a correspondingly high proportion of the variability in one is accounted for by variability in the other – while in fact *none* of the variability is accounted for. This is the phenomenon of *spurious regression*, first identified by Yule in 1927, and later studied by

Granger and Newbold in 1974. We can largely avoid such pitfalls by restricting attention to stationary time series, as above.
ARCH and GARCH.
The terms homoscedastic and heteroscedastic are used to describe processes where the variance is constant or is variable. With Z_i independent and normal $N(0,1)$, the *autoregressive conditionally heteroscedastic (ARCH)* model of *order* p, or $ARCH(p)$, is defined by the model equations

$$X_t = \sigma_t Z_t, \qquad \sigma_t^2 = \alpha_0 + \sum\nolimits_{i=1}^{p} \alpha_i X_{t-i}^2, \tag{$ARCH(p)$}$$

for $\alpha_0 > 0$ and $\alpha_i \geq 0$. The $AR(p)$ character is seen on the right of the second equation; the *conditional variance* of X_t *given* the information available at time $t-1$ is σ_t^2, a function of $X_{t-1}, \ldots, X_{t-p}$, and so varies, hence the conditional heteroscedasticity. In the *generalised* ARCH model $GARCH(p,q)$, the variance becomes

$$\sigma_t^2 = \alpha_0 + \sum\nolimits_{i=1}^{p} \alpha_i X_{t-i}^2 + \sum\nolimits_{j=1}^{q} \beta_j X \sigma_{t-j}^2. \tag{$GARCH(p,q)$}$$

Both ARCH and GARCH models are widely used in econometrics; see e.g. Engle's Nobel Prize citation. We must refer to a specialist time series or econometrics textbook for more details; the point to note here is that regression methods are widely used in economics and econometrics.

Note 9.9

We observed in §1.2.2 and §7.1 that, while independent errors tend to cancel as in the Law of Large Numbers, strongly dependent errors need not do so and are very dangerous in Statistics. The time series models above, which can model the tendency of high or low values to follow each other, reflect this – though there we separate out the terms giving rise to this and put them in the main part of the model, rather than the error.

9.5 Survival analysis

We return to the Poisson point process, *Ppp*(λ) say, first discussed in §8.4. In the sequel the parameter λ has the interpretation of an *intensity* or *rate* as follows. For an interval I of length $|I|$, the number of points of the process (the number of *Poisson points*) is Poisson distributed with parameter $\lambda|I|$; the counts in disjoint intervals are independent. This use of an intensity parameter to measure exposure to risk (of mortality) is generalised below.

Suppose now we have a population of individuals, whose lifetimes are independent, each with distribution function F on $(0, \infty)$, which we will suppose to have density f. If T is the lifetime of a given individual, the conditional probability of death in a short interval $(t, t+h)$ given survival to time t is, writing $\overline{F}(t) := 1 - F(t) = P(T > t)$ for the tail of F,

$$P(T \in (t, t+h) | T > t) = P(T \in (t+h))/P(T > t) = hf(t)/\overline{F}(t),$$

to first order in h. We call the coefficient of h on the right the *hazard function*, $h(t)$. Thus

$$h(t) = f(t) \Big/ \int_t^\infty f(u)\, du = -D\left(\int_t^\infty f\right) \Big/ \int_t^\infty f,$$

and integrating one has

$$\log\left(\int_t^\infty f\right) = -\int_0^t h : \qquad \int_t^\infty f(u)\, du = \exp\left\{-\int_0^t h(u)\, du\right\}$$

(since f is a density, $\int_0^\infty f = 1$, giving the constant of integration).

Example 9.10

1. *The exponential distribution.* If F is the exponential distribution with parameter λ, $E(\lambda)$ say, $f(t) = \lambda e^{-\lambda t}$, $\overline{F}(t) = e^{-\lambda t}$, and $h(t) = \lambda$ is constant. This property of constant hazard rate captures the *lack of memory property* of the exponential distributions (for which see e.g. the sources cited in §8.4), or the lack of ageing property: given that an individual has survived to date, its further survival time has the same distribution as that of a new individual. This is suitable for modelling the lifetimes of certain components (lightbulbs, etc.) that fail without warning, but of course not suitable for modelling lifetimes of biological populations, which show ageing.
2. *The Weibull distribution.*

Here

$$f(t) = \lambda \nu^{-\lambda} t^{\lambda - 1} \exp\{-(t/\lambda)^\nu\},$$

with λ, ν positive parameters; this reduces to the exponential $E(\lambda)$ for $\nu = 1$.
3. *The Gompertz-Makeham distribution.*

This is a three-parameter family, with hazard function

$$h(t) = \lambda + ae^{bt}.$$

This includes the exponential case with $a = b = 0$, and allows one to model a baseline hazard (the constant term λ), with in addition a hazard growing exponentially with time (which can be used to model the winnowing effect of ageing in biological populations).

In medical statistics, one may be studying survival times in patients with a particular illness. One's data is then subject to *censoring*, in which patients may die from other causes, discontinue treatment, leave the area covered by the study, etc.

9.5.1 Proportional hazards

One is often interested in the effect of covariates on survival probabilities. For example, many cancers are age-related, so the patient's age is an obvious covariate. Many forms of cancer are affected by diet, or lifestyle factors. Thus the link between smoking and lung cancer is now well known, and similarly for exposure to asbestos. One's chances of contracting certain cancers (of the mouth, throat, oesophagus etc.) are affected by alcohol consumption. Breast cancer rates are linked to diet (western women, whose diets are rich in dairy products, are more prone to the disease than oriental women, whose diets are rich in rice and fish). Consumption of red meat is linked to cancer of the bowel, etc., and so is lack of fibre. Thus in studying survival rates for a particular cancer, one may identify a suitable set of covariates z relevant to this cancer. One may seek to use a linear combination $\beta^T z$ of such covariates with coefficients β, as in the multiple regression of Chapters 3 and 4. One might also superimpose this effect on some baseline hazard, modelled non-parametrically. One is led to model the hazard function by

$$h(t; z) = g(\beta^T z) h_0(t),$$

where the function g contains the parametric part $\beta^T z$ and the baseline hazard h_0 the non-parametric part. This is the *Cox proportional hazards model* (D. R. Cox in 1972). The name arises because if one compares the hazards for two individuals with covariates z_1, z_2, one obtains

$$h(t; z_1)/h(t; z_2) = g(\beta^T z_1)/g(\beta^T z_2),$$

as the baseline hazard term cancels.

The most common choices of g are:
(i) *Log-linear*: $g(x) = e^x$ (if $g(x) = e^{ax}$, one can absorb the constant a into β);
(ii) *Linear*: $g(x) = 1 + x$;
(iii) *Logistic*: $g(x) = \log(1 + x)$.

We confine ourselves here to the log-linear case, the commonest and most important. Here the hazard ratio is

$$h(t; z_1)/h(t; z_2) = \exp\left\{\beta^T (z_1 - z_2)\right\}.$$

Estimation of β by maximum likelihood must be done numerically (we omit the non-parametric estimation of h_0). For a sample of n individuals, with covariate vectors $z_1, \ldots, z_n$, the data consist of the point events occurring – the identities (or covariate values) and times of death or censoring of non-surviving individuals; see e.g. Venables and Ripley (2002), §13.3 for use of S-Plus here, and for theoretical background see e.g. Cox and Oakes (1984).

9.6 $p >> n$

We have constantly emphasised that the number p of parameters is to be kept small, to give an economical description of the data in accordance with the Principle of Parsimony, while the sample size n is much larger – the larger the better, as there is then more information. However, practical problems in areas such as *bioinformatics* have given rise to a new situation, in which this is reversed, and one now has p much larger than n. This happens with, for example, data arising from *microarrays*. Here p is the number of entries in a large array or matrix, and p being large enables many biomolecular probes to be carried out at the same time, so speeding up the experiment. But now new and efficient variable-selection algorithms are needed. Recent developments include that of LASSO (least absolute shrinkage and selection operator) and LARS (least angle regression). One seeks to use such techniques to eliminate most of the parameters, and reduce to a case with $p << n$ that can be handled by traditional methods. That is, one seeks systematic ways to take a large and complex problem, in which most of the parameters are unimportant, and focus in on the small subset of important parameters.

Solutions

Chapter 1

1.1

$$\begin{aligned} Q(\lambda) &= \lambda^2 \frac{1}{n}\sum\nolimits_1^n (x_i - \overline{x})^2 + 2\lambda \frac{1}{n}\sum\nolimits_1^n (x_i - \overline{x})(y_i - \overline{y}) + \frac{1}{n}\sum\nolimits_1^n (y_i - \overline{y})^2 \\ &= \lambda^2 \overline{(x-\overline{x})^2} + 2\lambda \overline{(x-\overline{x})(y-\overline{y})} + \overline{(y-\overline{y})^2} \\ &= \lambda^2 S_{xx} + 2\lambda S_{xy} + S_{yy}. \end{aligned}$$

Now $Q(\lambda) \geq 0$ for all λ, so $Q(\cdot)$ is a quadratic which does not change sign. So its discriminant is ≤ 0 (if it were > 0, there would be distinct real roots and a sign change in between). So ('$b^2 - 4ac \leq 0$'):

$$s_{xy}^2 \leq s_{xx} s_{yy} = s_x^2 s_y^2, \quad r^2 := (s_{xy}/s_x s_y)^2 \leq 1.$$

So

$$-1 \leq r \leq +1,$$

as required.

The extremal cases $r = \pm 1$, or $r^2 = 1$, have discriminant 0, that is $Q(\lambda)$ has a repeated real root, λ_0 say. But then $Q(\lambda_0)$ is the sum of squares of $\lambda_0(x_i - \overline{x}) + (y_i - \overline{y})$, which is zero. So each term is 0:

$$\lambda_0(x_i - \overline{x}) + (y_i - \overline{y}) = 0 \quad (i = 1, \ldots, n).$$

That is, all the points (x_i, y_i) $(i = 1, \ldots, n)$, lie on a straight line through the centroid $(\overline{x}, \overline{y})$ with slope $-\lambda_0$.

1.2
Similarly

$$\begin{aligned} Q(\lambda) &= E\left[\lambda^2(x-Ex)^2 + 2\lambda(x-Ex)(y-Ey) + (y-Ey)^2\right] \\ &= \lambda^2 E[(x-Ex)^2] + 2\lambda E[(x-Ex)(y-Ey)] + E\left[(y-Ey)^2\right] \\ &= \lambda^2\sigma_x^2 + 2\lambda\sigma_{xy} + \sigma_y^2. \end{aligned}$$

(i) As before $Q(\lambda)\geq 0$ for all λ, as the discriminant is ≤ 0, i.e.

$$\sigma_{xy}^2 \leq \sigma_x^2\sigma_y^2, \quad \rho := (\sigma_{xy}/\sigma_x\sigma_y)^2 \leq 1, \quad -1 \leq \rho \leq +1.$$

The extreme cases $\rho = \pm 1$ occur iff $Q(\lambda)$ has a repeated real root λ_0. Then

$$Q(\lambda_0) = E[(\lambda_0(x-Ex) + (y-Ey))^2] = 0.$$

So the random variable $\lambda_0(x-Ex) + (y-Ey)$ is zero (a.s. – except possibly on some set of probability 0). So all values of (x,y) lie on a straight line through the centroid (Ex, Ey) of slope $-\lambda_0$, a.s.

1.3
(i) Half-marathon: $a = 3.310$ (2.656, 3.964). $b = 0.296$ (0.132, 0.460).
Marathon: $a = 3.690$ (2.990, 4.396). $b = 0.378$ (0.202, 0.554).
(ii) Compare rule with model $y = e^a t^b$ and consider, for example, $\frac{dy}{dt}(\bar{t})$. Should obtain a reasonable level of agreement.

1.4
A plot gives little evidence of curvature and there does not seem to be much added benefit in fitting the quadratic term. Testing the hypothesis $c = 0$ gives a p-value of 0.675. The predicted values are 134.44 and 163.89 for the linear model and 131.15 and 161.42 for the quadratic model.

1.5
The condition in the text becomes

$$\begin{pmatrix} S_{uu} & S_{uv} \\ S_{uv} & S_{vv} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} S_{yu} \\ S_{yv} \end{pmatrix}.$$

We can write down the solution for $(a\ b)^T$ as

$$\begin{pmatrix} S_{uu} & S_{uv} \\ S_{uv} & S_{vv} \end{pmatrix}^{-1} \begin{pmatrix} S_{yu} \\ S_{yv} \end{pmatrix} = \frac{1}{S_{uu}S_{vv} - S_{uv}^2} \begin{pmatrix} S_{vv} & -S_{uv} \\ -S_{uv} & S_{uu} \end{pmatrix} \begin{pmatrix} S_{yu} \\ S_{yv} \end{pmatrix},$$

giving

$$a = \frac{S_{vv}S_{yu} - S_{uv}S_{yv}}{S_{uu}S_{vv} - S_{uv}^2}, \ b = \frac{S_{uu}S_{yv} - S_{uv}S_{yu}}{S_{uu}S_{vv} - S_{uv}^2}.$$

1.6

(i) A simple plot suggests that a quadratic model might fit the data well (leaving aside, for the moment, the question of interpretation). An increase in R^2, equivalently a large reduction in the residual sum of squares, suggests the quadratic model offers a meaningful improvement over the simple model $y = a + bx$. A t-test for $c = 0$ gives a p-value of 0.007.

(ii) t-tests give p-values of 0.001 (in both cases) that b and c are equal to zero. The model has an R^2 of 0.68, suggesting that this simple model explains a reasonable amount, around 70%, of the variability in the data. The estimate gives $c = -7.673$, suggesting that club membership has improved the half-marathon times by around seven and a half minutes.

1.7

(i) The residual sums of squares are 0.463 and 0.852, suggesting that the linear regression model is more appropriate.

(ii) A t-test gives a p-value of 0.647, suggesting that the quadratic term is not needed. (Note also the very small number of observations.)

1.8

A simple plot suggests a faster-than-linear growth in population. Sensible suggestions are fitting an exponential model using $\log(y) = a + bt$, or a quadratic model $y = a + bt + ct^2$. A simple plot of the resulting fits suggests the quadratic model is better, with all the terms in this model highly significant.

1.9

(i) Without loss of generality assume $g(\cdot)$ is a monotone increasing function. We have that $F_Y(x) = \mathrm{P}(g(X) \leq x) = \mathrm{P}(X \leq g^{-1}(x))$. It follows that

$$\begin{aligned} f_Y(x) &= \frac{d}{dx}\int_{-\infty}^{g^{-1}(x)} f_X(u)\,du, \\ &= f_X\left(g^{-1}(x)\right)\left(\frac{dg^{-1}(x)}{dx}\right). \end{aligned}$$

(ii)

$$\begin{aligned} \mathrm{P}(Y \leq x) &= \mathrm{P}(e^X \leq x) = \mathrm{P}(X \leq \log x), \\ f_Y(x) &= \frac{d}{dx}\int_{\infty}^{\log x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y-\mu)^2}{2\sigma^2}}\,dy, \\ &= \frac{1}{\sqrt{2\pi}\sigma} x^{-1} \exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}. \end{aligned}$$

1.10
(i) $\mathrm{P}\left(Y \leq x\right) = \mathrm{P}\left(r/U \leq x\right) = \mathrm{P}\left(U \geq r/x\right)$. We have that

$$\begin{aligned} f_Y(x) &= \frac{d}{dx} \int_{r/x}^{\infty} \frac{\left(\frac{1}{2}\right)^{\frac{r}{2}} u^{\frac{r}{2}-1} e^{-\frac{u}{2}}\ du}{\Gamma(\frac{r}{2})}, \\ &= \frac{\left(\frac{r}{x^2}\right)\left(\frac{1}{2}\right)^{\frac{r}{2}}\left(\frac{r}{x}\right)^{\frac{r}{2}-1} e^{-\frac{r}{2x}}}{\Gamma\left(\frac{r}{2}\right)}, \\ &= \frac{r^{\frac{r}{2}} x^{-1-\frac{r}{2}} e^{-\frac{r}{2x}}}{2^{\frac{r}{2}}\Gamma\left(\frac{r}{2}\right)}. \end{aligned}$$

(ii) $\mathrm{P}(Y \leq x) = \mathrm{P}(X \geq 1/x)$, and this gives

$$\begin{aligned} f_Y(x) &= \frac{d}{dx} \int_{\frac{1}{x}}^{\infty} \frac{u^{a-1} b^a e^{-bu}\ du}{\Gamma(a)}, \\ &= \frac{\left(\frac{1}{x^2}\right) b^a \left(\frac{1}{x}\right)^{a-1} e^{-b/x}}{\Gamma(a)}, \\ &= \frac{b^a x^{-1-a} e^{-b/x}}{\Gamma(a)}. \end{aligned}$$

Since the above expression is a probability density, and therefore integrates to one, this gives

$$\int_0^{\infty} x^{-1-a} e^{-b/x}\ dx = \frac{\Gamma(a)}{b^a}.$$

1.11
We have that $f(x,u) = f_Y(u)\phi(x|0,u)$ and $f_{t(r)}(x) = \int_0^{\infty} f(x,u)du$, where $\phi(\cdot)$ denotes the probability density of $N(0,u)$. Writing this out explicitly gives

$$\begin{aligned} f_{t_r}(x) &= \int_0^{\infty} \frac{r^{\frac{r}{2}} u^{-1-\frac{r}{2}} e^{-\frac{r}{2u}}}{2^{\frac{r}{2}}\Gamma\left(\frac{r}{2}\right)} . \frac{e^{-\frac{x^2}{2u}}}{\sqrt{2\pi} u^{\frac{1}{2}}}\ du, \\ &= \frac{r^{\frac{r}{2}}}{2^{\frac{r}{2}}\Gamma(\frac{r}{2})\sqrt{2\pi}} \int_0^{\infty} u^{-\frac{3}{2}-\frac{r}{2}} e^{-\left[\frac{r}{2}+\frac{x^2}{2}\right]\frac{1}{u}}\ du, \\ &= \frac{r^{\frac{r}{2}}}{2^{\frac{r}{2}}\Gamma(\frac{r}{2})\sqrt{2\pi}} \frac{\Gamma\left(\frac{r}{2}+\frac{1}{2}\right)}{\left[\frac{r}{2}+\frac{x^2}{2}\right]^{\left(\frac{1}{2}+\frac{r}{2}\right)}}, \\ &= \frac{\Gamma\left(\frac{r}{2}+\frac{1}{2}\right)}{\sqrt{\pi r}\Gamma(\frac{r}{2})} \left(1+\frac{x^2}{r}\right)^{-\frac{1}{2}(r+1)}. \end{aligned}$$