A Handbook of Statistical Analyses Using R

SECOND EDITION

Brian S. Everitt and Torsten Hothorn

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Table 5.5: `students` data. Treatment and results of two tests in three groups of students.

treatment	low	high	treatment	low	high
AA	8	28	C	34	4
AA	18	28	C	34	4
AA	8	23	C	44	7
AA	12	20	C	39	5
AA	15	30	C	20	0
AA	12	32	C	43	11
AA	18	31	NC	50	5
AA	29	25	NC	57	51
AA	6	28	NC	62	52
AA	7	28	NC	56	52
AA	6	24	NC	59	40
AA	14	30	NC	61	68
AA	11	23	NC	66	49
AA	12	20	NC	57	49
C	46	13	NC	62	58
C	26	10	NC	47	58
C	47	22	NC	53	40
C	44	14			

Source: From Timm, N. H., *Applied Multivariate Analysis*, Springer, New York, 2002. With kind permission of Springer Science and Business Media.

CHAPTER 6

Simple and Multiple Linear Regression: How Old is the Universe and Cloud Seeding

6.1 Introduction

Freedman et al. (2001) give the relative velocity and the distance of 24 galaxies, according to measurements made using the Hubble Space Telescope – the data are contained in the **gamair** package accompanying Wood (2006), see Table 6.1. Velocities are assessed by measuring the Doppler red shift in the spectrum of light observed from the galaxies concerned, although some correction for 'local' velocity components is required. Distances are measured using the known relationship between the period of Cepheid variable stars and their luminosity. How can these data be used to estimate the age of the universe? Here we shall show how this can be done using simple linear regression.

Table 6.1: `hubble` data. Distance and velocity for 24 galaxies.

galaxy	velocity	distance	galaxy	velocity	distance
NGC0300	133	2.00	NGC3621	609	6.64
NGC0925	664	9.16	NGC4321	1433	15.21
NGC1326A	1794	16.14	NGC4414	619	17.70
NGC1365	1594	17.95	NGC4496A	1424	14.86
NGC1425	1473	21.88	NGC4548	1384	16.22
NGC2403	278	3.22	NGC4535	1444	15.78
NGC2541	714	11.22	NGC4536	1423	14.93
NGC2090	882	11.75	NGC4639	1403	21.98
NGC3031	80	3.63	NGC4725	1103	12.36
NGC3198	772	13.80	IC4182	318	4.49
NGC3351	642	10.00	NGC5253	232	3.15
NGC3368	768	10.52	NGC7331	999	14.72

Source: From Freedman W. L., et al., *The Astrophysical Journal*, 553, 47–72, 2001. With permission.

Table 6.2: `clouds` data. Cloud seeding experiments in Florida – see above for explanations of the variables.

seeding	time	sne	cloudcover	prewetness	echomotion	rainfall
no	0	1.75	13.4	0.274	stationary	12.85
yes	1	2.70	37.9	1.267	moving	5.52
yes	3	4.10	3.9	0.198	stationary	6.29
no	4	2.35	5.3	0.526	moving	6.11
yes	6	4.25	7.1	0.250	moving	2.45
no	9	1.60	6.9	0.018	stationary	3.61
no	18	1.30	4.6	0.307	moving	0.47
no	25	3.35	4.9	0.194	moving	4.56
no	27	2.85	12.1	0.751	moving	6.35
yes	28	2.20	5.2	0.084	moving	5.06
yes	29	4.40	4.1	0.236	moving	2.76
yes	32	3.10	2.8	0.214	moving	4.05
no	33	3.95	6.8	0.796	moving	5.74
yes	35	2.90	3.0	0.124	moving	4.84
yes	38	2.05	7.0	0.144	moving	11.86
no	39	4.00	11.3	0.398	moving	4.45
no	53	3.35	4.2	0.237	stationary	3.66
yes	55	3.70	3.3	0.960	moving	4.22
no	56	3.80	2.2	0.230	moving	1.16
yes	59	3.40	6.5	0.142	stationary	5.45
yes	65	3.15	3.1	0.073	moving	2.02
no	68	3.15	2.6	0.136	moving	0.82
yes	82	4.01	8.3	0.123	moving	1.09
no	83	4.65	7.4	0.168	moving	0.28

Weather modification, or cloud seeding, is the treatment of individual clouds or storm systems with various inorganic and organic materials in the hope of achieving an increase in rainfall. Introduction of such material into a cloud that contains supercooled water, that is, liquid water colder than zero degrees of Celsius, has the aim of inducing freezing, with the consequent ice particles growing at the expense of liquid droplets and becoming heavy enough to fall as rain from clouds that otherwise would produce none.

The data shown in Table 6.2 were collected in the summer of 1975 from an experiment to investigate the use of massive amounts of silver iodide (100 to 1000 grams per cloud) in cloud seeding to increase rainfall (Woodley et al., 1977). In the experiment, which was conducted in an area of Florida, 24 days were judged suitable for seeding on the basis that a measured suitability criterion, denoted *S-Ne*, was not less than 1.5. Here *S* is the 'seedability', the difference between the maximum height of a cloud if seeded and the same cloud if not seeded predicted by a suitable cloud model, and *Ne* is the number of

hours between 1300 and 1600 G.M.T. with 10 centimetre echoes in the target; this quantity biases the decision for experimentation against naturally rainy days. Consequently, optimal days for seeding are those on which seedability is large and the natural rainfall early in the day is small.

On suitable days, a decision was taken at random as to whether to seed or not. For each day the following variables were measured:

`seeding`: a factor indicating whether seeding action occurred (yes or no),

`time`: number of days after the first day of the experiment,

`cloudcover`: the percentage cloud cover in the experimental area, measured using radar,

`prewetness`: the total rainfall in the target area one hour before seeding (in cubic metres $\times 10^7$),

`echomotion`: a factor showing whether the radar echo was moving or stationary,

`rainfall`: the amount of rain in cubic metres $\times 10^7$,

`sne`: suitability criterion, see above.

The objective in analysing these data is to see how rainfall is related to the explanatory variables and, in particular, to determine the effectiveness of seeding. The method to be used is *multiple linear regression*.

6.2 Simple Linear Regression

Assume y_i represents the value of what is generally known as the *response variable* on the ith individual and that x_i represents the individual's values on what is most often called an *explanatory variable*. The simple linear regression model is

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where β_0 is the intercept and β_1 is the slope of the linear relationship assumed between the response and explanatory variables and ε_i is an error term. (The 'simple' here means that the model contains only a single explanatory variable; we shall deal with the situation where there are several explanatory variables in the next section.) The error terms are assumed to be independent random variables having a normal distribution with mean zero and constant variance σ^2.

The regression coefficients, β_0 and β_1, may be estimated as $\hat{\beta}_0$ and $\hat{\beta}_1$ using *least squares estimation*, in which the sum of squared differences between the observed values of the response variable y_i and the values 'predicted' by the regression equation $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is minimised, leading to the estimates;

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum\limits_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\end{aligned}$$

where $\bar{y}$ and $\bar{x}$ are the means of the response and explanatory variable, respectively.

The predicted values of the response variable y from the model are $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. The variance σ^2 of the error terms is estimated as

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

The estimated variance of the estimate of the slope parameter is

$$\mathsf{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2},$$

whereas the estimated variance of a predicted value y_{pred} at a given value of x, say x_0 is

$$\mathsf{Var}(y_{\text{pred}}) = \hat{\sigma}^2 \sqrt{\frac{1}{n} + 1 + \frac{(x_0 - \bar{x})^2}{\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}}.$$

In some applications of simple linear regression a model without an intercept is required (when the data is such that the line must go through the origin), i.e., a model of the form

$$y_i = \beta_1 x_i + \varepsilon_i.$$

In this case application of least squares gives the following estimator for β_1

$$\hat{\beta}_1 = \frac{\sum\limits_{i=1}^{n} x_i y_i}{\sum\limits_{i=1}^{n} x_i^2}. \tag{6.1}$$

6.3 Multiple Linear Regression

Assume y_i represents the value of the response variable on the ith individual, and that $x_{i1}, x_{i2}, \dots, x_{iq}$ represents the individual's values on q explanatory variables, with $i = 1, \dots, n$. The multiple linear regression model is given by

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_q x_{iq} + \varepsilon_i.$$

The error terms ε_i, $i = 1, \dots, n$, are assumed to be independent random variables having a normal distribution with mean zero and constant variance σ^2. Consequently, the distribution of the random response variable, y, is also normal with expected value given by the linear combination of the explanatory variables

$$\mathsf{E}(y|x_1, \dots, x_q) = \beta_0 + \beta_1 x_1 + \dots + \beta_q x_q$$

and with variance σ^2.

The parameters of the model β_k, $k = 1, \dots, q$, are known as regression coefficients with β_0 corresponding to the overall mean. The regression coefficients represent the expected change in the response variable associated with a unit change in the corresponding explanatory variable, when the remaining explanatory variables are held constant. The *linear* in multiple linear regression applies to the regression parameters, not to the response or explanatory variables. Consequently, models in which, for example, the logarithm of a response variable is modelled in terms of quadratic functions of some of the explanatory variables would be included in this class of models.

The multiple linear regression model can be written most conveniently for all n individuals by using matrices and vectors as $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ where $\mathbf{y}^\top = (y_1, \dots, y_n)$ is the vector of response variables, $\beta^\top = (\beta_0, \beta_1, \dots, \beta_q)$ is the vector of regression coefficients, and $\varepsilon^\top = (\varepsilon_1, \dots, \varepsilon_n)$ are the error terms. The *design* or *model matrix* $\mathbf{X}$ consists of the q continuously measured explanatory variables and a column of ones corresponding to the *intercept* term

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1q} \\ 1 & x_{21} & x_{22} & \dots & x_{2q} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nq} \end{pmatrix}.$$

In case one or more of the explanatory variables are nominal or ordinal variables, they are represented by a zero-one dummy coding. Assume that x_1 is a factor at m levels, the submatrix of $\mathbf{X}$ corresponding to x_1 is a $n \times m$ matrix of zeros and ones, where the jth element in the ith row is one when x_{i1} is at the jth level.

Assuming that the cross-product $\mathbf{X}^\top\mathbf{X}$ is non-singular, i.e., can be inverted, then the least squares estimator of the parameter vector β is unique and can be calculated by $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. The expectation and covariance of this estimator $\hat{\beta}$ are given by $\mathsf{E}(\hat{\beta}) = \beta$ and $\mathsf{Var}(\hat{\beta}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$. The diagonal elements of the covariance matrix $\mathsf{Var}(\hat{\beta})$ give the variances of $\hat{\beta}_j, j = 0, \dots, q$, whereas the off diagonal elements give the covariances between pairs of $\hat{\beta}_j$ and $\hat{\beta}_k$. The square roots of the diagonal elements of the covariance matrix are thus the standard errors of the estimates $\hat{\beta}_j$.

If the cross-product $\mathbf{X}^\top\mathbf{X}$ is singular we need to reformulate the model to $\mathbf{y} = \mathbf{X}\mathbf{C}\beta^\star + \varepsilon$ such that $\mathbf{X}^\star = \mathbf{X}\mathbf{C}$ has full rank. The matrix $\mathbf{C}$ is called the *contrast matrix* in S and R and the result of the model fit is an estimate $\hat{\beta}^\star$.

By default, a contrast matrix derived from *treatment contrasts* is used. For the theoretical details we refer to Searle (1971), the implementation of contrasts in S and R is discussed by Chambers and Hastie (1992) and Venables and Ripley (2002).

The regression analysis can be assessed using the following analysis of variance table (Table 6.3):

Table 6.3: Analysis of variance table for the multiple linear regression model.

Source of variation	Sum of squares	Degrees of freedom
Regression	$\sum\limits_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	q
Residual	$\sum\limits_{i=1}^{n} (\hat{y}_i - y_i)^2$	$n - q - 1$
Total	$\sum\limits_{i=1}^{n} (y_i - \bar{y})^2$	$n - 1$

where $\hat{y}_i$ is the predicted value of the response variable for the ith individual $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_q x_{q1}$ and $\bar{y} = \sum_{i=1}^{n} y_i/n$ is the mean of the response variable.

The mean square ratio

$$F = \frac{\sum\limits_{i=1}^{n} (\hat{y}_i - \bar{y})^2/q}{\sum\limits_{i=1}^{n} (\hat{y}_i - y_i)^2/(n - q - 1)}$$

provides an F-test of the general hypothesis

$$H_0 : \beta_1 = \cdots = \beta_q = 0.$$

Under H_0, the test statistic F has an F-distribution with q and $n - q - 1$ degrees of freedom. An estimate of the variance σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n - q - 1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

The correlation between the observed values y_i and the fitted values $\hat{y}_i$ is known as the *multiple correlation coefficient*. Individual regression coefficients can be assessed by using the ratio t-statistics $t_j = \hat{\beta}_j/\sqrt{\mathsf{Var}(\hat{\beta})_{jj}}$, although these ratios should be used only as rough guides to the 'significance' of the coefficients. The problem of selecting the 'best' subset of variables to be included in a model is one of the most delicate ones in statistics and we refer to Miller (2002) for the theoretical details and practical limitations (and see Exercise 6.4).

6.3.1 Regression Diagnostics

The possible influence of outliers and the checking of assumptions made in fitting the multiple regression model, i.e., constant variance and normality of error terms, can both be undertaken using a variety of diagnostic tools, of which the simplest and most well known are the estimated residuals, i.e., the differences between the observed values of the response and the fitted values of the response. In essence these residuals estimate the error terms in the simple and multiple linear regression model. So, after estimation, the next stage in the analysis should be an examination of such residuals from fitting the chosen model to check on the normality and constant variance assumptions and to identify outliers. The most useful plots of these residuals are:

- A plot of residuals against each explanatory variable in the model. The presence of a non-linear relationship, for example, may suggest that a higher-order term, in the explanatory variable should be considered.
- A plot of residuals against fitted values. If the variance of the residuals appears to increase with predicted value, a transformation of the response variable may be in order.
- A normal probability plot of the residuals. After all the systematic variation has been removed from the data, the residuals should look like a sample from a standard normal distribution. A plot of the ordered residuals against the expected order statistics from a normal distribution provides a graphical check of this assumption.

6.4 Analysis Using R

6.4.1 Estimating the Age of the Universe

Prior to applying a simple regression to the data it will be useful to look at a plot to assess their major features. The R code given in Figure 6.1 produces a scatterplot of velocity and distance. The diagram shows a clear, strong relationship between velocity and distance. The next step is to fit a simple linear regression model to the data, but in this case the nature of the data requires a model without intercept because if distance is zero so is relative speed. So the model to be fitted to these data is

$$\text{velocity} = \beta_1 \text{distance} + \varepsilon.$$

This is essentially what astronomers call Hubble's Law and β_1 is known as Hubble's constant; β_1^{-1} gives an approximate age of the universe.

To fit this model we are estimating β_1 using formula (6.1). Although this operation is rather easy

```
R> sum(hubble$distance * hubble$velocity) /
+      sum(hubble$distance^2)
[1] 76.58117
```

it is more convenient to apply R's linear modelling function

```
R> plot(velocity ~ distance, data = hubble)
```

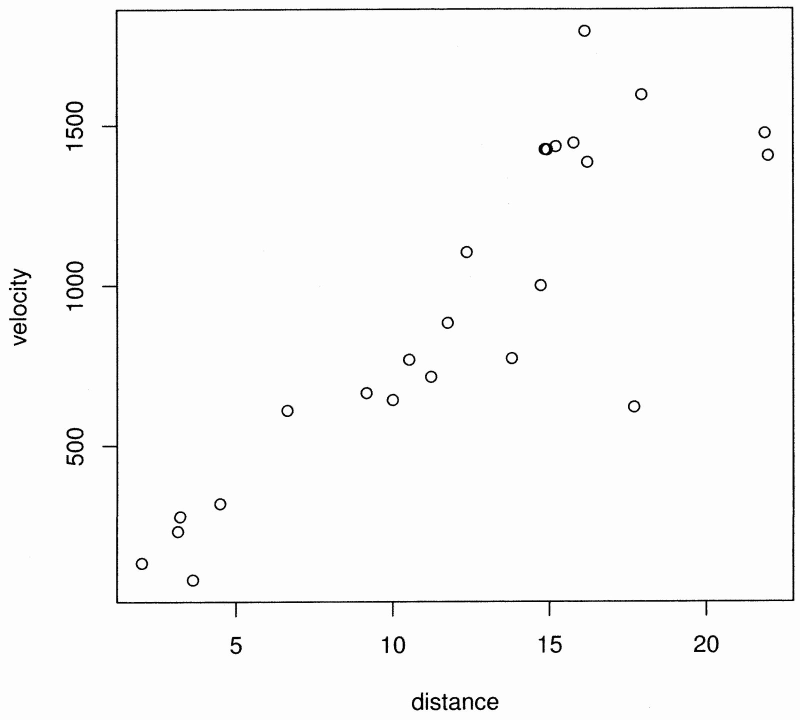

Figure 6.1 Scatterplot of velocity and distance.

```
R> hmod <- lm(velocity ~ distance - 1, data = hubble)
```

Note that the model formula specifies a model without intercept. We can now extract the estimated model coefficients via

```
R> coef(hmod)

distance
76.58117
```

and add this estimated regression line to the scatterplot; the result is shown in Figure 6.2. In addition, we produce a scatterplot of the residuals $y_i - \hat{y}_i$ against fitted values $\hat{y}_i$ to assess the quality of the model fit. It seems that for higher distance values the variance of velocity increases; however, we are interested in only the estimated parameter $\hat{\beta}_1$ which remains valid under variance heterogeneity (in contrast to t-tests and associated p-values).

Now we can use the estimated value of β_1 to find an approximate value

```
R> layout(matrix(1:2, ncol = 2))
R> plot(velocity ~ distance, data = hubble)
R> abline(hmod)
R> plot(hmod, which = 1)
```

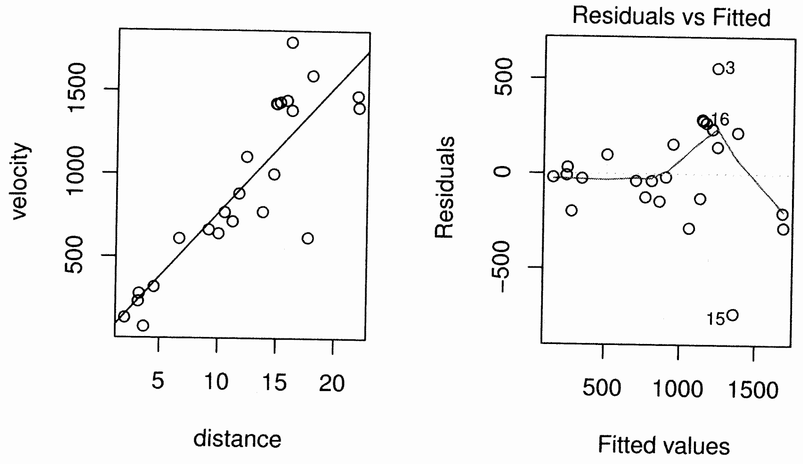

Figure 6.2 Scatterplot of velocity and distance with estimated regression line (left) and plot of residuals against fitted values (right).

for the age of the universe. The Hubble constant itself has units of km × $\text{sec}^{-1} \times \text{Mpc}^{-1}$. A mega-parsec (Mpc) is 3.09×10^{19}km, so we need to divide the estimated value of β_1 by this amount in order to obtain Hubble's constant with units of sec^{-1}. The approximate age of the universe in seconds will then be the inverse of this calculation. Carrying out the necessary computations

```
R> Mpc <- 3.09 * 10^19
R> ysec <- 60^2 * 24 * 365.25
R> Mpcyear <- Mpc / ysec
R> 1 / (coef(hmod) / Mpcyear)

   distance
12785935335
```

gives an estimated age of roughly 12.8 billion years.

6.4.2 Cloud Seeding

Again, a graphical display highlighting the most important aspects of the data will be helpful. Here we will construct boxplots of the rainfall in each category of the dichotomous explanatory variables and scatterplots of rainfall against each of the continuous explanatory variables.

Both the boxplots (Figure 6.3) and the scatterplots (Figure 6.4) show some evidence of outliers. The row names of the extreme observations in the clouds *data.frame* can be identified via

```
R> rownames(clouds)[clouds$rainfall %in% c(bxpseeding$out,
+                                            bxpecho$out)]

[1] "1"  "15"
```

where bxpseeding and bxpecho are variables created by boxplot in Figure 6.3. Now we shall not remove these observations but bear in mind during the modelling process that they may cause problems.

In this example it is sensible to assume that the effect that some of the other explanatory variables is modified by seeding and therefore consider a model that includes seeding as covariate and, furthermore, allows interaction terms for seeding with each of the covariates except time. This model can be described by the *formula*

```
R> clouds_formula <- rainfall ~ seeding +
+      seeding:(sne + cloudcover + prewetness + echomotion) +
+      time
```

and the design matrix $\mathbf{X}^\star$ can be computed via

```
R> Xstar <- model.matrix(clouds_formula, data = clouds)
```

By default, treatment contrasts have been applied to the dummy codings of the factors seeding and echomotion as can be seen from the inspection of the contrasts attribute of the model matrix

```
R> attr(Xstar, "contrasts")

$seeding
[1] "contr.treatment"

$echomotion
[1] "contr.treatment"
```

The default contrasts can be changed via the contrasts.arg argument to model.matrix or the contrasts argument to the fitting function, for example lm or aov as shown in Chapter 5.

However, such internals are hidden and performed by high-level model-fitting functions such as lm which will be used to fit the linear model defined by the *formula* clouds_formula:

```
R> clouds_lm <- lm(clouds_formula, data = clouds)
R> class(clouds_lm)

[1] "lm"
```

The results of the model fitting is an object of class *lm* for which a summary method showing the conventional regression analysis output is available. The

```
R> data("clouds", package = "HSAUR2")
R> layout(matrix(1:2, nrow = 2))
R> bxpseeding <- boxplot(rainfall ~ seeding, data = clouds,
+      ylab = "Rainfall", xlab = "Seeding")
R> bxpecho <- boxplot(rainfall ~ echomotion, data = clouds,
+      ylab = "Rainfall", xlab = "Echo Motion")
```

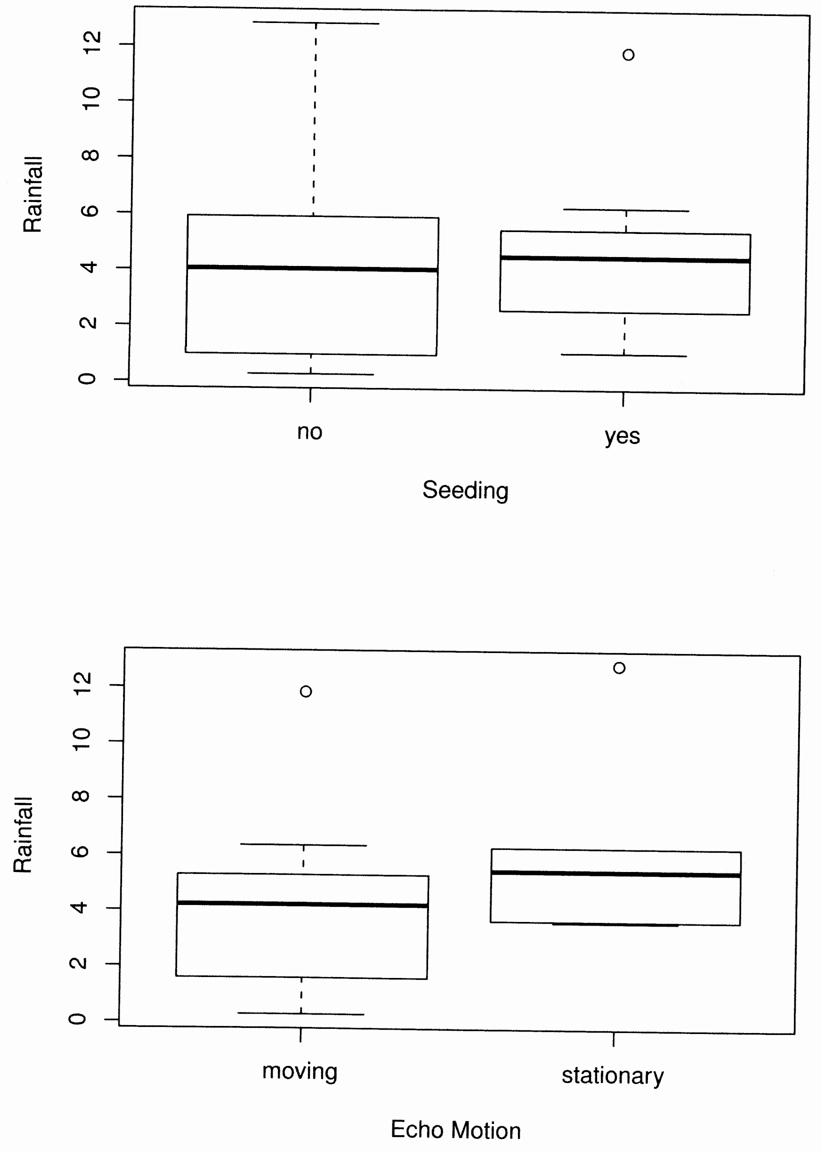

Figure 6.3 Boxplots of rainfall.

```
R> layout(matrix(1:4, nrow = 2))
R> plot(rainfall ~ time, data = clouds)
R> plot(rainfall ~ cloudcover, data = clouds)
R> plot(rainfall ~ sne, data = clouds, xlab="S-Ne criterion")
R> plot(rainfall ~ prewetness, data = clouds)
```

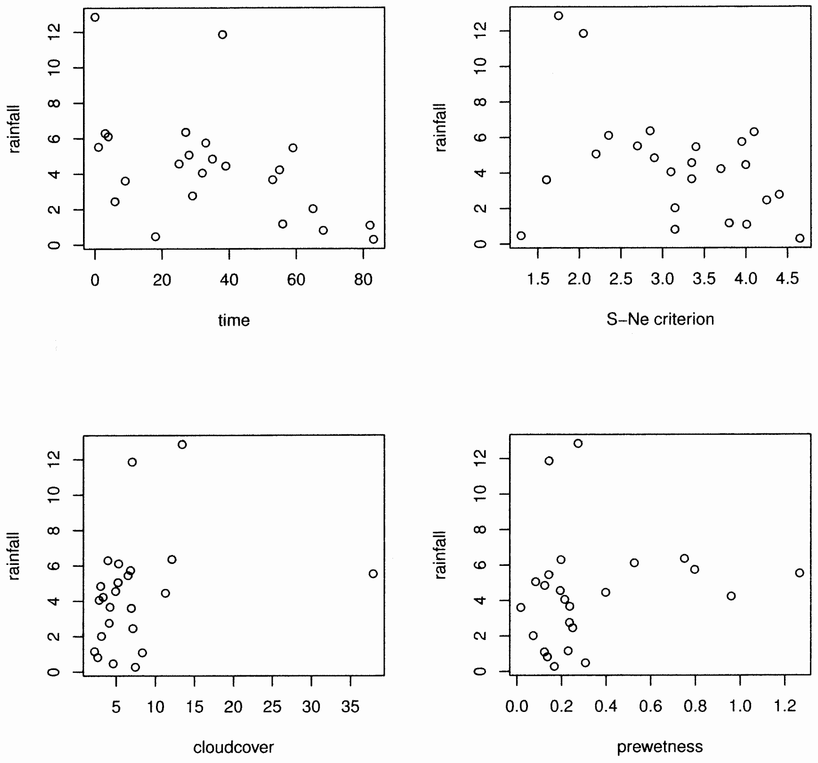

Figure 6.4 Scatterplots of `rainfall` against the continuous covariates.

output in Figure 6.5 shows the estimates $\hat{\beta}^{\star}$ with corresponding standard errors and t-statistics as well as the F-statistic with associated p-value.

Many methods are available for extracting components of the fitted model. The estimates $\hat{\beta}^{\star}$ can be assessed via

```
R> betastar <- coef(clouds_lm)
R> betastar
```

```
                      (Intercept)
                      -0.34624093
                       seedingyes
```

```
R> summary(clouds_lm)

Call:
lm(formula = clouds_formula, data = clouds)

Residuals:
    Min      1Q  Median      3Q     Max
-2.5259 -1.1486 -0.2704  1.0401  4.3913

Coefficients:
                                Estimate Std. Error t value
(Intercept)                     -0.34624    2.78773  -0.124
seedingyes                      15.68293    4.44627   3.527
time                            -0.04497    0.02505  -1.795
seedingno:sne                    0.41981    0.84453   0.497
seedingyes:sne                  -2.77738    0.92837  -2.992
seedingno:cloudcover             0.38786    0.21786   1.780
seedingyes:cloudcover           -0.09839    0.11029  -0.892
seedingno:prewetness             4.10834    3.60101   1.141
seedingyes:prewetness            1.55127    2.69287   0.576
seedingno:echomotionstationary   3.15281    1.93253   1.631
seedingyes:echomotionstationary  2.59060    1.81726   1.426
                                Pr(>|t|)
(Intercept)                      0.90306
seedingyes                       0.00372
time                             0.09590
seedingno:sne                    0.62742
seedingyes:sne                   0.01040
seedingno:cloudcover             0.09839
seedingyes:cloudcover            0.38854
seedingno:prewetness             0.27450
seedingyes:prewetness            0.57441
seedingno:echomotionstationary   0.12677
seedingyes:echomotionstationary  0.17757

Residual standard error: 2.205 on 13 degrees of freedom
Multiple R-squared: 0.7158,      Adjusted R-squared: 0.4972
F-statistic: 3.274 on 10 and 13 DF,  p-value: 0.02431
```

Figure 6.5 R output of the linear model fit for the `clouds` data.

```
       15.68293481
              time
       -0.04497427
     seedingno:sne
        0.41981393
    seedingyes:sne
       -2.77737613
```

```
           seedingno:cloudcover
                     0.38786207
          seedingyes:cloudcover
                    -0.09839285
           seedingno:prewetness
                     4.10834188
          seedingyes:prewetness
                     1.55127493
 seedingno:echomotionstationary
                     3.15281358
seedingyes:echomotionstationary
                     2.59059513
```

and the corresponding covariance matrix $\text{Cov}(\hat{\beta}^\star)$ is available from the `vcov` method

```
R> Vbetastar <- vcov(clouds_lm)
```

where the square roots of the diagonal elements are the standard errors as shown in Figure 6.5

```
R> sqrt(diag(Vbetastar))
```

```
                    (Intercept)
                     2.78773403
                     seedingyes
                     4.44626606
                           time
                     0.02505286
                  seedingno:sne
                     0.84452994
                 seedingyes:sne
                     0.92837010
           seedingno:cloudcover
                     0.21785501
          seedingyes:cloudcover
                     0.11028981
           seedingno:prewetness
                     3.60100694
          seedingyes:prewetness
                     2.69287308
 seedingno:echomotionstationary
                     1.93252592
seedingyes:echomotionstationary
                     1.81725973
```

The results of the linear model fit, as shown in Figure 6.5, suggests that rainfall can be increased by cloud seeding. Moreover, the model indicates that higher values of the S-Ne criterion lead to less rainfall, but only on days when cloud seeding happened, i.e., the interaction of seeding with S-Ne significantly affects rainfall. A suitable graph will help in the interpretation of this result. We can plot the relationship between rainfall and S-Ne for seeding and non-seeding days using the R code shown with Figure 6.6.

```
R> psymb <- as.numeric(clouds$seeding)
R> plot(rainfall ~ sne, data = clouds, pch = psymb,
+      xlab = "S-Ne criterion")
R> abline(lm(rainfall ~ sne, data = clouds,
+           subset = seeding == "no"))
R> abline(lm(rainfall ~ sne, data = clouds,
+           subset = seeding == "yes"), lty = 2)
R> legend("topright", legend = c("No seeding", "Seeding"),
+         pch = 1:2, lty = 1:2, bty = "n")
```

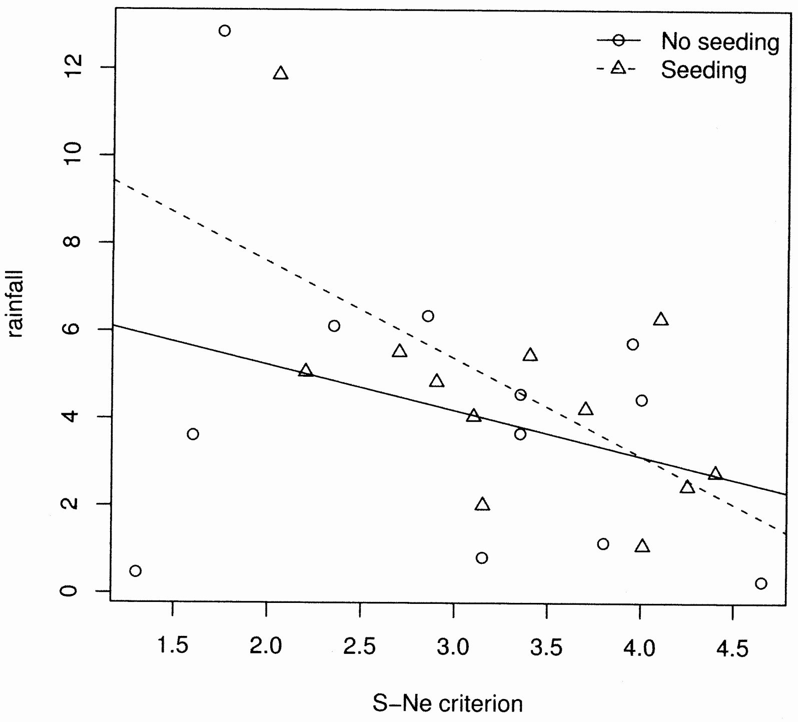

Figure 6.6 Regression relationship between S-Ne criterion and rainfall with and without seeding.

The plot suggests that for smaller S-Ne values, seeding produces greater rainfall than no seeding, whereas for larger values of S-Ne it tends to produce less. The cross-over occurs at an S-Ne value of approximately four which suggests that seeding is best carried out when S-Ne is less than four. But the number of observations is small and we should perhaps now consider the influence of any outlying observations on these results.

In order to investigate the quality of the model fit, we need access to the residuals and the fitted values. The residuals can be found by the `residuals` method and the fitted values of the response from the `fitted` (or `predict`) method

```
R> clouds_resid <- residuals(clouds_lm)
R> clouds_fitted <- fitted(clouds_lm)
```

Now the residuals and the fitted values can be used to construct diagnostic plots; for example the residual plot in Figure 6.7 where each observation is labelled by its number. Observations 1 and 15 give rather large residual values and the data should perhaps be reanalysed after these two observations are removed. The normal probability plot of the residuals shown in Figure 6.8 shows a reasonable agreement between theoretical and sample quantiles, however, observations 1 and 15 are extreme again.

A further diagnostic that is often very useful is an index plot of the Cook's distances for each observation. This statistic is defined as

$$D_k = \frac{1}{(q+1)\hat{\sigma}^2} \sum_{i=1}^{n} (\hat{y}_{i(k)} - y_i)^2$$

where $\hat{y}_{i(k)}$ is the fitted value of the ith observation when the kth observation is omitted from the model. The values of D_k assess the impact of the kth observation on the estimated regression coefficients. Values of D_k greater than one are suggestive that the corresponding observation has undue influence on the estimated regression coefficients (see Cook and Weisberg, 1982).

An index plot of the Cook's distances for each observation (and many other plots including those constructed above from using the basic functions) can be found from applying the `plot` method to the object that results from the application of the `lm` function. Figure 6.9 suggests that observations 2 and 18 have undue influence on the estimated regression coefficients, but the two outliers identified previously do not. Again it may be useful to look at the results after these two observations have been removed (see Exercise 6.2).

6.5 Summary

Multiple regression is used to assess the relationship between a set of explanatory variables and a response variable (with simple linear regression, there is a single exploratory variable). The response variable is assumed to be normally distributed with a mean that is a linear function of the explanatory variables and a variance that is independent of the explanatory variables. An important part of any regression analysis involves the graphical examination of residuals and other diagnostic statistics to help identify departures from assumptions.

```
R> plot(clouds_fitted, clouds_resid, xlab = "Fitted values",
+      ylab = "Residuals", type = "n",
+      ylim = max(abs(clouds_resid)) * c(-1, 1))
R> abline(h = 0, lty = 2)
R> text(clouds_fitted, clouds_resid, labels = rownames(clouds))
```

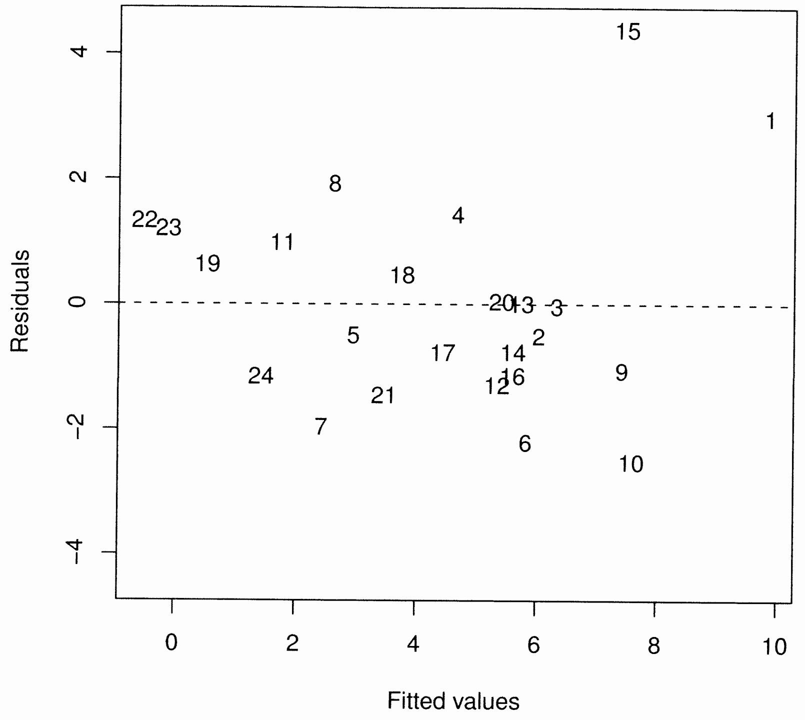

Figure 6.7 Plot of residuals against fitted values for `clouds` seeding data.

Exercises

Ex. 6.1 The simple residuals calculated as the difference between an observed and predicted value have a distribution that is scale dependent since the variance of each is a function of both σ^2 and the diagonal elements of the

```
R> qqnorm(clouds_resid, ylab = "Residuals")
R> qqline(clouds_resid)
```

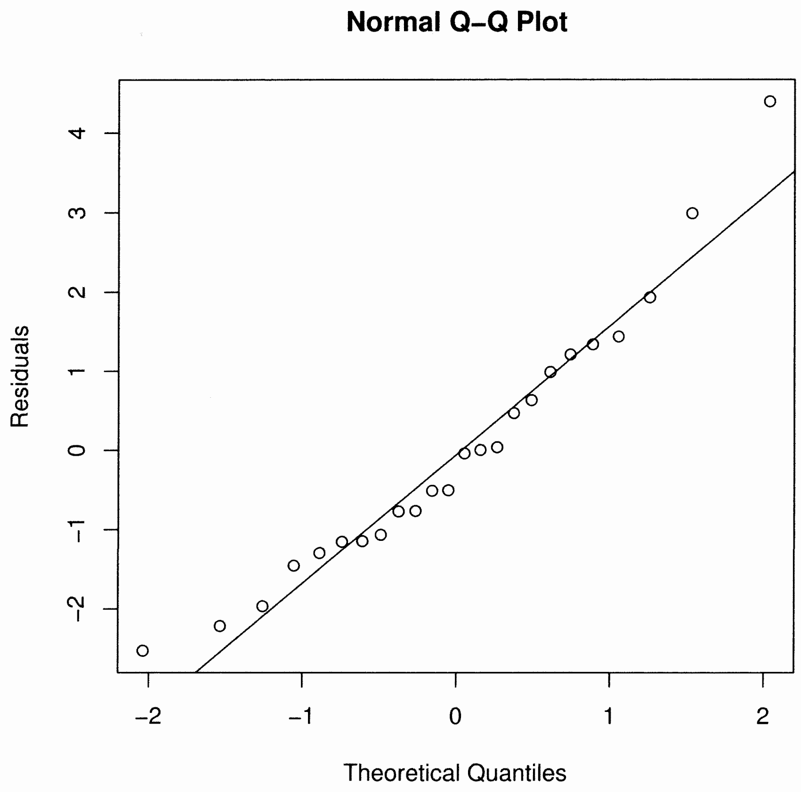

Figure 6.8 Normal probability plot of residuals from cloud seeding model `clouds_lm`.

hat matrix **H** given by

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top.$$

Consequently it is often more useful to work with the standardised version of the residuals that does not depend on either of these quantities. These standardised residuals are calculated as

$$r_i = \frac{y_i - \hat{y}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where $\hat{\sigma}^2$ is the estimator of σ^2 and h_{ii} is the ith diagonal element of **H**. Write an R function to calculate these residuals and use it to obtain some diagnostic plots similar to those mentioned in the text. (The elements of the hat matrix can be obtained from the `lm.influence` function.)

```
R> plot(clouds_lm)
```

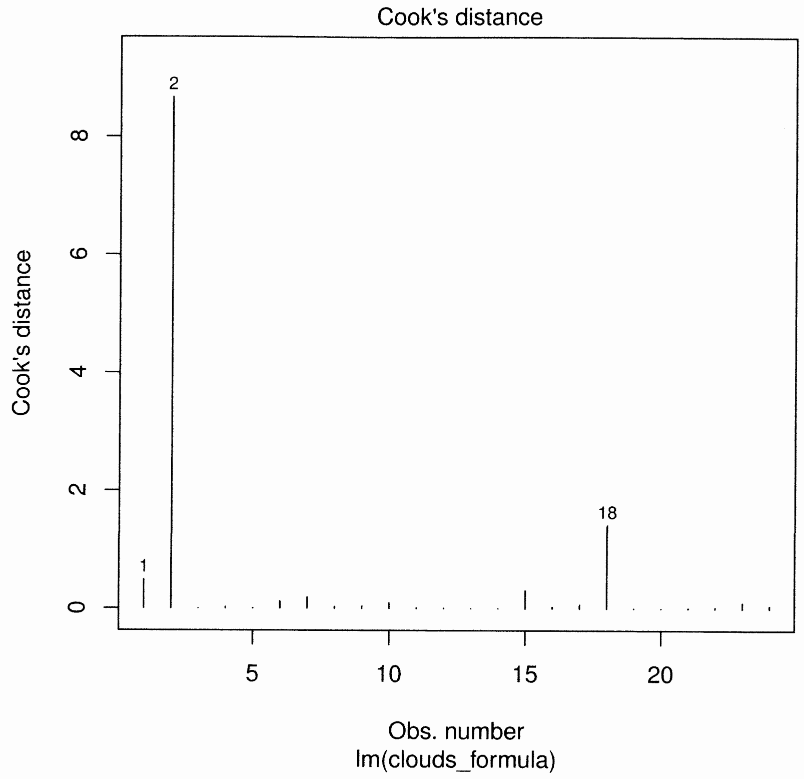

Figure 6.9 Index plot of Cook's distances for cloud seeding data.

Ex. 6.2 Investigate refitting the cloud seeding data after removing any observations which may give cause for concern.

Ex. 6.3 Show how the analysis of variance table for the data in Table 5.1 of the previous chapter can be constructed from the results of applying an appropriate multiple linear regression to the data.

Ex. 6.4 Investigate the use of the `leaps` function from package **leaps** (Lumley and Miller, 2009) for selecting the 'best' set of variables predicting rainfall in the cloud seeding data.

Ex. 6.5 Remove the observations for galaxies having leverage greater than 0.08 and refit the zero intercept model. What is the estimated age of the universe from this model?

Ex. 6.6 Fit a quadratic regression model, i.e, a model of the form

$$\text{velocity} = \beta_1 \times \text{distance} + \beta_2 \times \text{distance}^2 + \varepsilon,$$

to the `hubble` data and plot the fitted curve and the simple linear regression fit on a scatterplot of the data. Which model do you consider most sensible considering the nature of the data? (The 'quadratic model' here is still regarded as a linear regression model since the term *linear* relates to the parameters of the model not to the powers of the explanatory variable.)

CHAPTER 7

Logistic Regression and Generalised Linear Models: Blood Screening, Women's Role in Society, Colonic Polyps, and Driving and Back Pain

7.1 Introduction

The erythrocyte sedimentation rate (ESR) is the rate at which red blood cells (erythrocytes) settle out of suspension in blood plasma, when measured under standard conditions. If the ESR increases when the level of certain proteins in the blood plasma rise in association with conditions such as rheumatic diseases, chronic infections and malignant diseases, its determination might be useful in screening blood samples taken from people suspected of suffering from one of the conditions mentioned. The absolute value of the ESR is not of great importance; rather, less than 20mm/hr indicates a 'healthy' individual. To assess whether the ESR is a useful diagnostic tool, Collett and Jemain (1985) collected the data shown in Table 7.1. The question of interest is whether there is any association between the probability of an ESR reading greater than 20mm/hr and the levels of the two plasma proteins. If there is not then the determination of ESR would not be useful for diagnostic purposes.

Table 7.1: `plasma` data. Blood plasma data.

`fibrinogen`	`globulin`	`ESR`	`fibrinogen`	`globulin`	`ESR`
2.52	38	ESR < 20	2.88	30	ESR < 20
2.56	31	ESR < 20	2.65	46	ESR < 20
2.19	33	ESR < 20	2.28	36	ESR < 20
2.18	31	ESR < 20	2.67	39	ESR < 20
3.41	37	ESR < 20	2.29	31	ESR < 20
2.46	36	ESR < 20	2.15	31	ESR < 20
3.22	38	ESR < 20	2.54	28	ESR < 20
2.21	37	ESR < 20	3.34	30	ESR < 20
3.15	39	ESR < 20	2.99	36	ESR < 20
2.60	41	ESR < 20	3.32	35	ESR < 20
2.29	36	ESR < 20	5.06	37	ESR > 20
2.35	29	ESR < 20	3.34	32	ESR > 20
3.15	36	ESR < 20	2.38	37	ESR > 20
2.68	34	ESR < 20	3.53	46	ESR > 20