

Linear regression

Considerable effort in the environmental sciences is directed at predicting an environmental or ecological response from a collection of other variables. That is, an observed *response variable*, Y , is recorded alongside one or more *predictor variables*, and these latter quantities are used to describe the deterministic aspects of Y . If we denote the predictor variables as x_1, x_2, \dots, x_p , it is natural to model the deterministic aspects of the response via some function, say, $g(x_1, x_2, \dots, x_p; \boldsymbol{\beta})$, where $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \dots \ \beta_p]^T$ is a column vector of $p + 1$ unknown parameters. (A *vector* is an array of numbers arranged as a row or column. The superscript T indicates transposition of the vector, so that, for example, $[a_1 a_2]^T = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$. More generally, a *matrix* is an array of numbers arranged in a square or rectangular fashion; one can view a matrix as a collection of vectors, all of equal length. Background material on matrices and vectors appears in Appendix A. For a more general introduction to the use of matrix algebra in regression, see Neter *et al.*, 1996, Ch. 5.) We use the function $g(\cdot)$ to describe how Y changes as a function of the x_s .

As part of the model, we often include an additive error term to account for any random, or *stochastic*, aspects of the response. Formally, then, an observation Y_i is assumed to take the form

$$Y_i = g(x_{i1}, x_{i2}, \dots, x_{ip}; \boldsymbol{\beta}) + \varepsilon_i, \quad (1.1)$$

$i = 1, \dots, n$, where the additive error terms ε_i are assigned some form of probability distribution and the *sample size* n is the number of recorded observations. Unless otherwise specified, we assume the Y_s constitute a random sample of statistically independent observations. If Y represents a continuous measurement, it is common to take $\varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2)$, ‘i.i.d.’ being a shorthand notation for *independent and identically distributed* (see Appendix A). Coupled with the additivity assumption in (1.1), this is known as a *regression* of Y on the x_s .

2 Linear regression

Note also that we require the x_j predictor variables to be fixed values to which no stochastic variability may be ascribed (or, at least, that the analysis be conditioned on the observed pattern of the predictor variables).

We will devote a large portion of this text to environmetric analysis for a variety of regression problems. In this chapter, we give a short review of some elementary regression models, and then move on to a selection of more complex forms. We start with the most basic case: simple linear regression.

1.1 Simple linear regression

The simple linear case involves only one predictor variable ($p = 1$), and sets $g(x_{i1}; \boldsymbol{\beta})$ equal to a linear function of x_{i1} . For simplicity, when $p = 1$ we write x_{i1} as x_i . Equation (1.1) becomes

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

$i = 1, \dots, n$, and we call $\beta_0 + \beta_1 x_i$ the *linear predictor*. The linear predictor is the deterministic component of the regression model. Since this also models the population mean of Y_i , we often write $\mu(x_i) = \beta_0 + \beta_1 x_i$, and refer to $\mu(x)$ as the *mean response function*.

The simple linear regression model can also be expressed as the matrix equation $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\mathbf{Y} = [Y_1 \dots Y_n]^T$, $\boldsymbol{\varepsilon} = [\varepsilon_0 \dots \varepsilon_n]^T$ and \mathbf{X} is a matrix whose columns are the two vectors $\mathbf{J} = [1 \dots 1]^T$ – i.e., a column vector of ones – and $[x_1 \dots x_n]^T$.

As a first step in any regression analysis, we recommend that a graphical display of the data pairs (x_i, Y_i) be produced. Plotted, this is called a *scatterplot*; see Fig. 1.1 in Example 1.1, below. The scatterplot is used to visualize the data and begin the process of assessing the model fit: straight-line relationships suggest a simple linear model, while curvilinear relationships suggest a more complex model. We discuss nonlinear regression modeling in Chapter 2.

Under the common assumptions that $E[\varepsilon_i] = 0$ and $\text{Var}[\varepsilon_i] = \sigma^2$ for all $i = 1, \dots, n$, the model parameters in $\boldsymbol{\beta} = [\beta_0 \beta_1]^T$ have interpretations as the Y -intercept (β_0) and slope (β_1) of $\mu(x_i)$. In particular, for any unit increase in x_i , $\mu(x_i)$ increases by β_1 units. To estimate the unknown parameters we appeal to the least squares (LS) method, where the sum of squared errors $\sum_{i=1}^n \{Y_i - \mu(x_i)\}^2$ is minimized (LS estimation is reviewed in §A.4.1). The LS estimators of β_0 and β_1 here are

$$b_0 = \bar{Y} - b_1 \bar{x}$$

and

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i Y_i - \frac{1}{n} \sum_{i=1}^n x_i \sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n} (\sum_{i=1}^n x_i)^2}, \quad (1.2)$$

where $\bar{Y} = \sum_{i=1}^n Y_i/n$ and $\bar{x} = \sum_{i=1}^n x_i/n$. The algebra here can be simplified using matrix notation: if $\mathbf{b} = [b_0 \ b_1]^T$ is the vector of LS estimators, then $\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$, $(\mathbf{X}^T \mathbf{X})^{-1}$ being the *inverse* of the matrix $\mathbf{X}^T \mathbf{X}$ (see §A.4.3).

If we further assume that $\varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2)$, then the LS estimates will correspond to maximum likelihood (ML) estimates for β_0 and β_1 . (ML estimation is reviewed in §A.4.3.) The LS/ML estimate of the mean response, $\mu(x) = \beta_0 + \beta_1 x$, for any x , is simply $\hat{\mu}(x) = b_0 + b_1 x$.

We should warn that calculation of b_1 can be adversely affected by a number of factors. For example, if the x_i s are spaced unevenly, highly separated values of x_i can exert strong *leverage* on b_1 by pulling the estimated regression line too far up or down. (See the web applet at <http://www.stat.sc.edu/~west/javahtml/Regression.html> for a visual demonstration. Also see the discussion on regression diagnostics, below.) To avoid this, the predictor variables should be spaced as evenly as possible, or some transformation of the x_i s should be applied before performing the regression calculations. The natural logarithm is a typical choice here, since it tends to compress very disparate values. If when applying the logarithm, one of the x_i values is zero, say $x_1 = 0$, one can average the other log-transformed x_i s to approximate an equally spaced value associated with $x_1 = 0$. This is *consecutive-dose average spacing* (Margolin *et al.*, 1986): denote the transformed predictor by $u_i = \log(x_i)$, $i = 2, \dots, n$. Then at $x_1 = 0$, use

$$u_1 = u_2 - \frac{u_n - u_2}{n - 1}. \quad (1.3)$$

A useful tabular device for collecting important statistical information from a linear regression analysis is known as the *analysis of variance (ANOVA) table*. The table lays out *sums of squares* that measure variation in the data attributable to various components of the model. It also gives the *degrees of freedom* (df) for each component. The df represent the amount of information in the data available to estimate that particular source of variation. The ratio of a sum of squares to its corresponding df is called a *mean square*.

For example, to identify the amount of variability explained by the linear regression of Y on x , the sum of squares for regression is $\text{SSR} = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$, where $\hat{Y}_i = b_0 + b_1 x_i$ is the i th *predicted value* (also called a *fitted value*). SSR has degrees of freedom equal to the number of regression parameters estimated minus one; here, $\text{df}_r = 1$. Thus the mean square for regression when $p = 1$ is $\text{MSR} = \text{SSR}/1$.

We can also estimate the unknown variance parameter, σ^2 , via ANOVA computations. Find the sum of squared errors $\text{SSE} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ and divide this by the error df (the number of observations minus the number of regression parameters estimated), $\text{df}_e = n - 2$. The resulting *mean squared error* is

$$\text{MSE} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n - 2},$$

and this is an unbiased estimator of σ^2 . We often call $\sqrt{\text{MSE}}$ the *root mean squared error*.

4 Linear regression

We do not go into further detail here on the construction of sums of squares and ANOVA tables, although we will mention other aspects of linear modeling and ANOVA below. Readers unfamiliar with ANOVA computations can find useful expositions in texts on linear regression analysis, such as Neter *et al.* (1996) or Christensen (1996).

We use the MSE to calculate the *standard errors* of the LS/ML estimators. (A standard error is the square root or estimated square root of an estimator's variance; see §A.4.3.) Here, these are

$$se[b_0] = \sqrt{\text{MSE} \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}}$$

and

$$se[b_1] = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^n (x_i - \bar{x})^2}}. \quad (1.4)$$

Standard errors (and variances) quantify the variability of the point estimator, helping to gauge how meaningful the magnitude of a given estimate is. They also give insight into the impact of different experimental designs on estimating regression coefficients. For example, notice that $se[b_0]$ is smallest for x_i s chosen so that $\bar{x} = 0$, while $se[b_1]$ is minimized when $\sum_{i=1}^n (x_i - \bar{x})^2$ is taken to be as large as possible.

Similarly, the standard error of $\hat{\mu}(x)$ is

$$se[\hat{\mu}(x)] = \sqrt{\text{MSE} \left\{ \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}}.$$

Notice that, as with $\hat{\mu}(x)$, $se[\hat{\mu}(x)]$ varies with x . It attains its minimum at $x = \bar{x}$ and then increases as x departs from \bar{x} in either direction. One may say, therefore, that precision in $\hat{\mu}(x)$ is greatest near the center of the predictor range – i.e., at \bar{x} – and diminishes as x moves away from it. Indeed, if one drives x too far away from the predictor range, $se[\hat{\mu}(x)]$ can grow so large as to make $\hat{\mu}(x)$ essentially useless. This illustrates the oft-cited concern that *extrapolation* away from the range of the data leads to imprecise, inaccurate, and in some cases even senseless statistical predictions.

The standard errors are used in constructing statistical inferences on the β_j s or on $\mu(x)$. For example, notice that if $\beta_1 = 0$ then the predictor variable has no effect on the response and the simple linear model collapses to $Y_i = \beta_0 + \varepsilon_i$, a ‘constant + error’ model for Y . To assess this, assume that the $N(0, \sigma^2)$ assumption on the ε_i s is valid. Then, a $1 - \alpha$ *confidence interval* for β_1 is

$$b_1 \pm t_{\alpha/2}(n-2)se[b_1].$$

(The theory of confidence intervals is reviewed in §A.5.1.) An alternative inference is available by conducting a *hypothesis test* of the null hypothesis $H_0: \beta_1 = 0$ vs. the

alternative hypothesis $H_a: \beta_1 \neq 0$. (The theory of hypothesis tests is reviewed in §A.5.3.) Here, we find the test statistic

$$|t_{\text{calc}}| = \frac{|b_1|}{se[b_1]}$$

based on Student's t -distribution (§A.2.11), and reject H_0 when $|t_{\text{calc}}| \geq t_{\alpha/2}(n-2)$. (We use the subscript 'calc' to indicate a statistic that is wholly calculable from the data.) Equivalently, we can reject H_0 when the corresponding P -value, here

$$P = 2P \left[t(n-2) \geq \frac{|b_1|}{se(b_1)} \right],$$

drops below the preset *significance level* α (see §A.5.3).

For testing against a one-sided alternative such as $H_a: \beta_1 > 0$, we reject H_0 when $t_{\text{calc}} = b_1/se[b_1] \geq t_{\alpha}(n-2)$. The P -value is then $P[t(n-2) \geq b_1/se(b_1)]$. Similar constructions are available for β_0 ; for example, a $1 - \alpha$ confidence interval is $b_0 \pm t_{\alpha/2}(n-2)se[b_0]$.

All these operations can be conducted by computer, and indeed, many statistical computing packages perform simple linear regression. Herein, we highlight the SAS[®] system (SAS Institute Inc., 2000), which provides LS/ML estimates $\mathbf{b} = [b_0 \ b_1]^T$, their standard errors $se[b_j]$, an ANOVA table that includes an unbiased estimator of σ^2 via the MSE, and other summary statistics, via its PROC GLM or PROC REG procedures.

Example 1.1 (Motor vehicle CO₂) To illustrate use of the simple linear regression model, consider the following example. In the United Kingdom (and in most other industrialized nations) it has been noted that as motor vehicle use increases, so do emissions of various byproducts of hydrocarbon combustion. Public awareness of this potential polluting effect has bolstered industry aspirations to 'uncouple' detrimental emissions from vehicle use. In many cases, emission controls and other efforts have reduced the levels of hazardous pollutants such as small particulate matter (PM) and nitrogen oxides. One crucial counter-example to this trend, however, is the ongoing increases in the greenhouse gas carbon dioxide (CO₂). For example, Redfern *et al.* (2003) discuss data on $x = \text{UK motor vehicle use (in kilometers per year)}$ vs. $Y = \text{CO}_2 \text{ emissions (as a relative index; 1970 = 100)}$. Table 1.1 presents the data.

A plot of the data in Table 1.1 shows a clear, increasing, linear trend (Fig. 1.1). Assuming that the simple linear model with normal errors is appropriate for these data, we find the LS/ML estimates to be $b_0 = 28.3603$ and $b_1 = 0.7442$. The corresponding standard errors are $se[b_0] = 2.1349$ and $se[b_1] = 0.0127$. Since $n = 28$, a 95% confidence interval for β_1 is $0.7742 \pm t_{0.025}(26) \times 0.0127 = 0.7742 \pm 2.056 \times 0.0127 = 0.7742 \pm 0.0261$. (We find $t_{0.025}(26)$ from Table B.2 or via the SAS function `tinv`; see Fig. A.4.) Based on this 95% interval, the CO₂ index increases approximately 0.75 to 0.80 units (relative to 1970 levels) with each additional kilometer.

6 Linear regression

Table 1.1 Yearly CO₂ emissions (rel. index; 1970 = 100) vs. motor vehicle use (rel. km/yr; 1970 = 100) in the United Kingdom, 1971–1998

Year	1971	1972	1973	1974	1975	1976	1977
$x = \text{vehicle use}$	105.742	110.995	116.742	114.592	115.605	121.467	123.123
$Y = \text{CO}_2$	104.619	109.785	117.197	114.404	111.994	116.898	119.915
Year	1978	1979	1980	1981	1982	1983	1984
$x = \text{vehicle use}$	127.953	127.648	135.660	138.139	141.911	143.707	151.205
$Y = \text{CO}_2$	126.070	128.759	130.196	126.409	130.136	134.212	140.721
Year	1985	1986	1987	1988	1989	1990	1991
$x = \text{vehicle use}$	154.487	162.285	174.837	187.403	202.985	204.959	205.325
$Y = \text{CO}_2$	143.462	153.074	159.999	170.312	177.810	182.686	181.348
Year	1992	1993	1994	1995	1996	1997	1998
$x = \text{vehicle use}$	205.598	205.641	210.826	214.947	220.753	225.742	229.027
$Y = \text{CO}_2$	183.757	185.869	186.872	185.100	192.249	194.667	193.438

Source: Redfern *et al.* (2003).

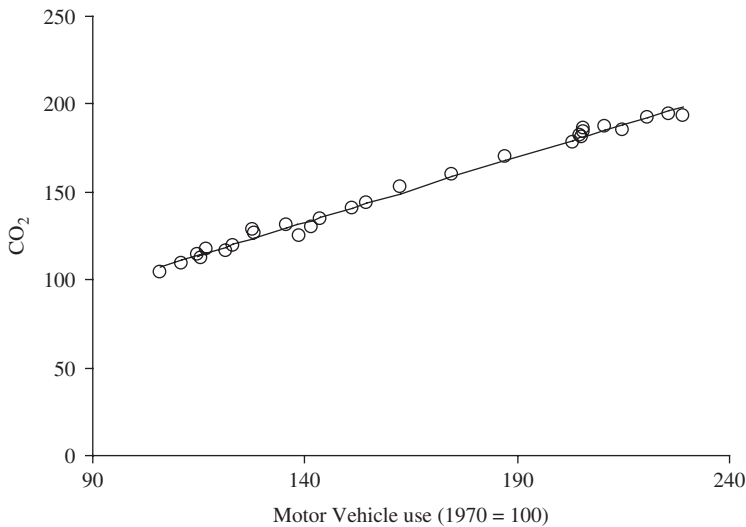



Figure 1.1 Scatterplot and estimated LS line for motor vehicle CO₂ data from Table 1.1

Alternatively, we can test the significance of the slope with these data. Specifically, since one would expect *a priori* that increased motor vehicle use would increase CO₂ emissions, the hypotheses $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 > 0$ are a natural choice. Suppose we set our significance level to $\alpha = 0.01$. For these data, the test statistic is $t_{\text{calc}} = b_1 / \text{se}[b_1] = 0.7742 / 0.0127 = 60.96$, with corresponding P -value $P[t(26) \geq 60.96] < 0.0001$. This is well below α , hence we conclude that a significant, increasing effect exists on CO₂ emissions associated with the observed pattern of motor vehicle use in the UK between 1971 and 1998. 

The sample size in Example 1.1, $n = 28$, is not atypical for a simple linear regression data set, but of course analysts can encounter much larger sample sizes in environmental practice. We will study selected examples of this in the chapters on nonlinear regression (Chapter 2), temporal data (Chapter 5), and spatially correlated data (Chapter 6), below.

Once a model has been fitted to data, it is important to assess the quality of the fit in order to gauge the validity of the consequent inferences and predictions. In practice, any statistical analysis of environmental data should include a critical examination of the assumptions made about the statistical model, in order to identify if any unsupported assumptions are being made and to alert the user to possible unanticipated or undesired consequences. At the simplest level, a numerical summary for the quality of a regression fit is the *coefficient of determination* $\{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})\}^2 / \{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2\}$, denoted as R^2 . This may also be computed from the ANOVA table as $R^2 = \text{SSR} / \{\text{SSR} + \text{SSE}\}$. Under a linear model, R^2 has interpretation as the proportion of variation in Y_i that can be attributed to the variation in x_i . If the predictor variable explains Y precisely (i.e., the x_i, Y_i pairs all coincide on a straight line), R^2 attains its maximum value of 1.0. Alternatively, if there is *no* linear relationship between x_i and Y_i (so $\beta_1 = 0$), $R^2 = 0.0$. As such, higher values of R^2 indicate higher-quality explanatory value in x_i .

More intricate *regression diagnostics* can include a broad variety of procedures for assessing model fit (Davison and Tsai, 1992; Neter *et al.*, 1996, Ch. 3). Most basic among these is study of the *residuals* $r_i = Y_i - \hat{Y}_i$. Almost every analysis of a regression relationship should include a graph of the residuals, r_i , against the predicted values, \hat{Y}_i (or, if $p = 1$, against x_i). Such a *residual plot* can provide information on a number of features. For instance, if there is an underlying curvilinear trend in the data that was not picked up by the original scatterplot, the residual plot may highlight the curvilinear aspects not explained by the simple linear terms. Or, if the assumption of variance homogeneity is inappropriate – i.e., if $\text{Var}[\varepsilon_i]$ is not constant over changing x_i – the residual plot may show a fan-shaped pattern of increasing or decreasing residuals (or both) as \hat{Y}_i increases. Figure 1.2 illustrates both these sorts of patterns. Notice in Fig. 1.2(b) that variability increases with increasing mean response; this sort of pattern is not uncommon with environmental data.

If the residual plot shows a generally uniform or random pattern, then evidence exists for a reasonable model fit.

Example 1.2 (Motor vehicle CO₂, cont'd) Returning to the data on motor vehicle use in the UK, we find $\text{SSR} = 26\,045.2953$ and $\text{SSE} = 196.0457$. This gives $R^2 = 0.9925$, from which it appears that variation in CO₂ emissions is strongly explained by variation in motor vehicle use.

Figure 1.3 shows the residual plot from the simple linear model fit. The residual points appear randomly dispersed, with no obvious structure or pattern. This suggests that the variability in CO₂ levels about the regression line is constant and so the homogeneous variance assumption is supported. One could also graph a histogram or normal probability plot of the residuals to assess the adequacy of the normality assumption. If the histogram appears roughly bell-shaped, or if the normal plot produces a roughly straight line, then the assumption of normal errors may be reasonable. For the residuals in Fig. 1.3, a normal probability plot constructed using PROC UNIVARIATE in SAS (via its `plot` option; output suppressed) does plot as

8 Linear regression

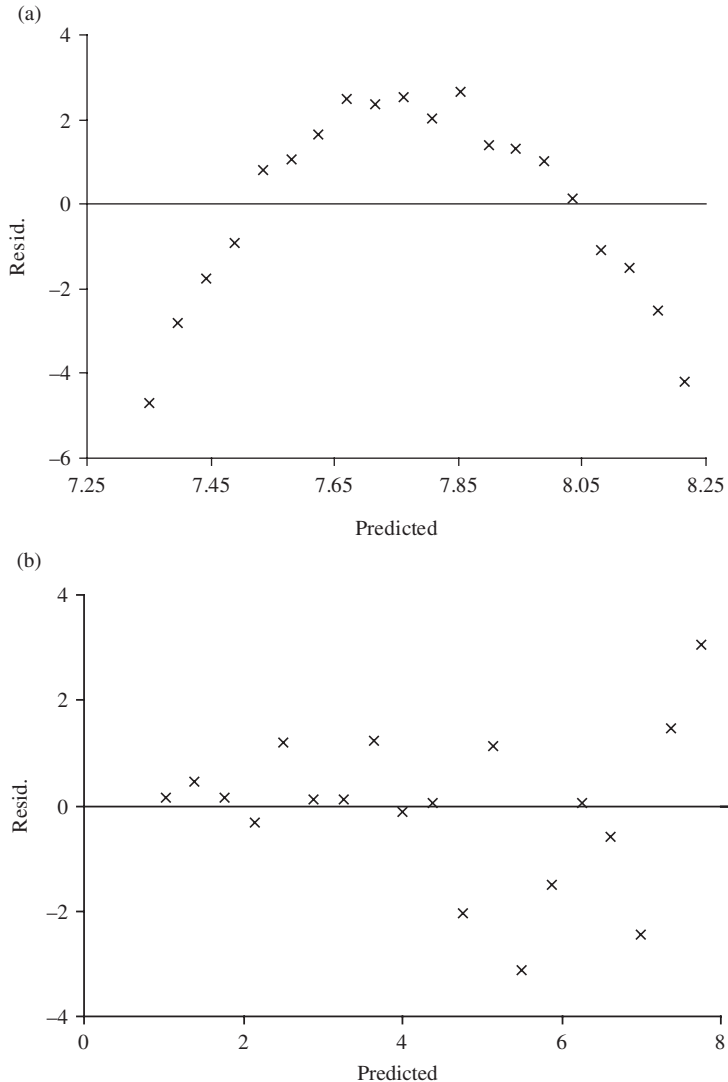


Figure 1.2 Typical residual plots in the presence of model misspecification. (a) Curvilinear residual trend indicates curvilinearity not fit by the model. (b) Widening residual spread indicates possible variance heterogeneity. Horizontal reference lines indicate residual = 0

roughly linear. Or one can call for normal probability plots directly in PROC REG, using the statement

```
plot nqq.*r. npp.*r.;
```

The plot statement in PROC REG can also be used to generate a residual plot, via

```
plot r.*p.;
```

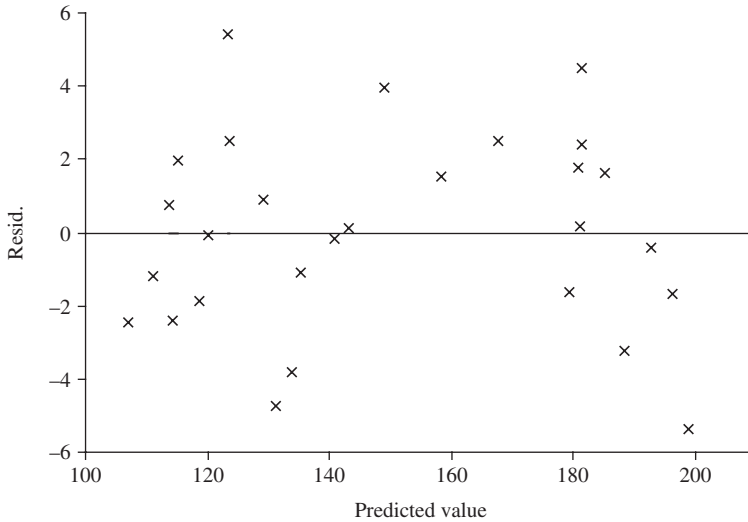


Figure 1.3 Residual plot for motor vehicle CO₂ data from Table 1.1. Horizontal bar indicates residual = 0

or an overlay of the data and the predicted regression line, via

```
plot Y*x p.*x/overlay;
```



When the residual plot identifies a departure from variance homogeneity, inferences on the unknown parameters based on the simple linear fit can be incorrect, and some adjustment is required. If the heterogeneous variation can be modeled or otherwise quantified, it is common to weight each observation in inverse proportion to its variance and apply weighted least squares (WLS; see §A.4.1). For example, suppose it is known or anticipated that the variance changes as a function of x_i , say $\text{Var}[Y_i] \propto h(x_i)$. Then, a common weighting scheme employs $w_i = 1/h(x_i)$.

For weights given as w_i , $i = 1, \dots, n$, the WLS estimators become

$$\tilde{b}_0 = \left(\sum_{i=1}^n w_i \right)^{-1} \left(\sum_{i=1}^n w_i Y_i - b_1 \sum_{i=1}^n w_i x_i \right) \quad (1.5)$$

and

$$\tilde{b}_1 = \frac{\sum_{i=1}^n w_i x_i Y_i - \left(\sum_{i=1}^n w_i \right)^{-1} \left(\sum_{i=1}^n w_i x_i \sum_{i=1}^n w_i Y_i \right)}{\sum_{i=1}^n w_i x_i^2 - \left(\sum_{i=1}^n w_i \right)^{-1} \left(\sum_{i=1}^n w_i x_i \right)^2}. \quad (1.6)$$

The standard errors require similar modification; for example,

$$se[\tilde{b}_1] = \frac{\sqrt{\tilde{\text{MSE}}}}{\sqrt{\sum_{i=1}^n w_i x_i^2 - (\sum_{i=1}^n w_i)^{-1} (\sum_{i=1}^n w_i x_i)^2}},$$

where $\tilde{\text{MSE}}$ is the weighted mean square $\sum_{i=1}^n w_i (Y_i - \tilde{b}_0 - \tilde{b}_1 x_i)^2 / (n - 2)$. Inferences on β_1 then mimic those described above for the simple linear case. In SAS, both PROC GLM and PROC REG can incorporate these (or any other) weighting schemes, using the `weight` statement. Neter *et al.* (1996, §10.1) give further details on the use of WLS methods.

If appropriate weights cannot be identified, it is often possible to stabilize the variances by transforming the original observations. A common transformation in many environmental applications is the (natural) logarithm: $V_i = \log(Y_i)$. This is part of a larger class of transformations, known as the Box–Cox power transformations (Box and Cox, 1964). The general form is $V_i = (Y_i^\lambda - 1)/\lambda$, for some specified transformation parameter λ . The natural logarithm is the limiting case at $\lambda = 0$. Other popular transformations include the square root ($\lambda = 1/2$), the quadratic ($\lambda = 2$), and the reciprocal ($\lambda = -1$). One can also estimate λ from the data, although this can lead to loss of independence among the V_i s. Users should proceed with caution when estimating a power transformation parameter; see Carroll and Ruppert (1988) for more on this and other issues regarding data transformation in regression. Another useful transformation, often employed with percentage data, is the *logit transform*: if Y_i is a percentage between 0 and 100, take $V_i = \log\{Y_i/(100 - Y_i)\}$. We employ this in Example 1.5, below.

Many other procedures are available for diagnosing and assessing model fit, correcting for various model perturbations and inadequacies, and analyzing linear relationships. A full description of all these methods for the simple linear model is beyond the scope of this chapter, however. Details can be found in the targeted textbook by Belsley *et al.* (1980), or in general texts on statistics such as Samuels and Witmer (2003, Ch. 12) and Neter *et al.* (1996, Chs. 1–5).

1.2 Multiple linear regression

The simplest statistical model for the case of $p > 1$ predictor variables in (1.1) employs a linear term for each predictor: set $g(x_{i1}, x_{i2}, \dots, x_{ip}; \boldsymbol{\beta}) = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip}$. This is a *multiple linear regression* model. The parameter β_j may be interpreted as the change in $E[Y_i]$ that occurs for a unit increase in x_{ij} – the ‘slope’ of the j th predictor – assuming all the other x -variables are held fixed. (When it is not possible to vary one predictor while holding all others constant, then of course this interpretation may not make sense. An example of such occurs with polynomial regression models; see §1.5.) We require $n > p + 1$.

Assuming, as above, that the errors satisfy $E[\varepsilon_i] = 0$ and $\text{Var}[\varepsilon_i] = \sigma^2$ for all $i = 1, \dots, n$, the LS estimators for $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \dots \ \beta_p]^T$ can be derived using multivariable

calculus. When the additional assumption is made that $\varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2)$, these LS estimates will correspond to ML estimates.

Unfortunately, the LS/ML estimators for β are not easily written in closed form. The effort can be accomplished using vector and matrix notation in similar fashion to that mentioned in §1.1, although actual calculation of the estimates is most efficiently performed by computer. Almost any statistical computing package can fit a multiple linear regression via LS or WLS methods; in Example 1.3, below, we illustrate use of SAS.

Similar to the simple linear case, we can test whether any particular predictor variable, x_{ij} , is important in modeling $E[Y_i]$ via appeal to a t -test: find $t_{\text{calc}} = b_j/se[b_j]$ and reject $H_0: \beta_j = 0$ in favor of $H_a: \beta_j \neq 0$ when $|t_{\text{calc}}| = |b_j|/se[b_j] \geq t_{\alpha/2}(n - p - 1)$. Note that this tests the significance of the j th predictor variable given that all the other predictor variables are present in the model. In this sense, we call it an *adjusted test* or a *partial test* of significance. Confidence intervals are similar; for example, a pointwise $1 - \alpha$ confidence interval for β_j is $b_j \pm t_{\alpha/2}(n - p - 1)se[b_j]$; $j = 1, \dots, p$. Notice the change in df_e from the simple linear case where $p = 1$: estimation of each additional β_j results in a loss of 1 additional df for error, so we have gone from $df_e = n - 2$ to $df_e = n - (p + 1)$.

We can also make statements on subsets or groupings of the β -parameters. For example, consider a test of the null hypothesis that a group of $k > 1$ of the β_j s is equal to zero, say, $H_0: \beta_{j+1} = \dots = \beta_{j+k} = 0$. Rejection of H_0 suggests that the corresponding group of k predictor variables has a significant impact on the regression relationship. A general approach for such a test involves construction of *discrepancy measures* that quantify the fit of the general (or *full*) model with all $p + 1$ of the β -parameters, and the *reduced model* with $p - k + 1$ (non-zero) β -parameters. For the multiple regression model with normally distributed errors, a useful discrepancy measure is the sum of squared errors $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$, where $\hat{Y}_i = b_0 + b_1x_{i1} + \dots + b_px_{ip}$ is the i th predicted value under the full model. For clarity, we augment the SSE notation by indicating if it is calculated under the full model (FM) or under the reduced model (RM): $SSE(\text{FM})$ or $SSE(\text{RM})$. The SSEs are used to quantify the relative quality of each model's fit to the data: if H_0 is false, we expect $SSE(\text{RM})$ to be larger than $SSE(\text{FM})$, since the model under which it is fitted fails to include important predictor variables. Corresponding to these terms, we also write the degrees of freedom associated with each error terms as $df_e(\text{FM})$ and $df_e(\text{RM})$, respectively. The difference between the two is $\Delta_e = df_e(\text{RM}) - df_e(\text{FM})$. Here, $df_e(\text{FM}) = n - p - 1$, while $df_e(\text{RM}) = n + k - p - 1$, so that $\Delta_e = k$ is the number of parameters constrained by the null hypothesis.

To use this discrepancy approach for testing H_0 , calculate the test statistic

$$F_{\text{calc}} = \frac{\{SSE(\text{RM}) - SSE(\text{FM})\}/\Delta_e}{SSE(\text{FM})/df_e(\text{FM})}, \quad (1.7)$$

which under H_0 is distributed as per an F -distribution with Δ_e and $df_e(\text{FM})$ degrees of freedom (§A.2.11). We denote this as $F_{\text{calc}} \sim F[\Delta_e, df_e(\text{FM})]$. Reject H_0 in favor of an alternative that allows at least one of the β_j s in H_0 to be non-zero when F_{calc} exceeds the appropriate upper- α F -critical point, $F_\alpha(\Delta_e, df_e[\text{FM}])$. For the multiple regression setting, this is $F_{\text{calc}} \geq F_\alpha(k, n - p - 1)$. The P -value is $P = \text{P}[F(k, n - p - 1) \geq F_{\text{calc}}]$. This testing strategy corresponds to a form of generalized likelihood ratio test (§A.5).

In many cases, the various measures in (1.7) can be read directly from an ANOVA table for the full model (Neter *et al.*, 1996, Ch. 16). For example, if $SSR(FM)$ is the full model's sum of squares for regression and the reduced model contains only the intercept β_0 (so $k = p$), $F_{\text{calc}} = \{SSR(FM)/p\}/MSE$. Also, an extension of the coefficient of determination from the simple linear setting is the *coefficient of multiple determination*: $R^2 = SSR(FM)/\{SSR(FM) + SSE(FM)\}$. As in the simple linear case, R^2 measures the proportion of variation in Y_i that can be accounted for by variation in the collection of x_{ij} s.

For this approach to be valid, the parameters represented under the RM must be a true subset of those under the FM. We say then that the models are *nested*. If the relationship between the RM and FM does not satisfy a nested hierarchy, F_{calc} under H_0 may not follow (or even approximate) an F -distribution. The family of models are then said to be *separate* (Cox, 1961, 1962); inferences for testing separate families are still an area of developing environmetric research (Hinde, 1992; Schork, 1993).

Example 1.3 (Soil pH) Edenharder *et al.* (2000) report on soil acidity in west-central Germany, as a function of various soil composition measures. For $Y = \text{soil pH}$, three predictor variables (all percentages) were employed: $x_{i1} = \text{soil texture (as clay)}$, $x_{i2} = \text{organic matter}$, and $x_{i3} = \text{carbonate composition (CaCO}_3 \text{ by weight)}$. The $n = 17$ data points are given in Table 1.2.

Table 1.2 Soil pH vs. soil composition variables in west-central Germany

$x_1 = \% \text{ Clay}$	$x_2 = \% \text{ Organics}$	$x_3 = \text{Carbonate}$	$Y = \text{pH}$
51.1	4.3	6.1	7.1
22.0	2.6	0.0	5.4
17.0	3.0	2.0	7.0
16.8	3.0	0.0	6.1
5.5	4.0	0.0	3.7
21.2	3.3	0.1	7.0
14.1	3.7	16.8	7.4
16.6	0.7	17.3	7.4
35.9	3.7	15.6	7.3
29.9	3.3	11.9	7.5
2.4	3.1	2.8	7.4
1.6	2.8	6.2	7.4
17.0	1.8	0.3	7.5
32.6	2.3	9.1	7.3
10.5	4.0	0.0	4.0
33.0	5.1	26.0	7.1
26.0	1.9	0.0	5.6

Source: Edenharder *et al.* (2000).

To fit the multiple linear regression model $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$ to the data in Table 1.2, we employ the SAS procedure PROC GLM, although one could also apply PROC REG profitably here, especially with its regression diagnostic plots as noted in Example 1.2. Sample SAS code is given in Fig. 1.4. (In some installations of SAS, the statement `run;` may be required at the end of the input code; however, we do not use it in the examples herein.)

The SAS output (edited for presentation purposes) is given in Fig. 1.5. (We will not go into particulars of the SAS code and output here, since we expect readers to be familiar with this level of detail for multiple linear regression. Where appropriate in future chapters, however, we will supply more background on selected SAS code and/or output.)

```
* SAS code to fit mult. lin. regr.;
data soil;
input Yph x1clay x2org x3carb @@;
datalines;
  7.1   51.1  4.3   6.1   5.4  22.0  2.6   0.0
  7.0   17.0  3.0   2.0   6.1  16.8  3.0   0.0
  3.7   5.5   4.0   0.0   7.0  21.2  3.3   0.1
  7.4   14.1  3.7  16.8   7.4  16.6  0.7  17.3
  7.3   35.9  3.7  15.6   7.5  29.9  3.3  11.9
  7.4    2.4  3.1   2.8   7.4   1.6  2.8   6.2
  7.5   17.0  1.8   0.3   7.3  32.6  2.3   9.1
  4.0   10.5  4.0   0.0   7.1  33.0  5.1  26.0
  5.6   26.0  1.9   0.0

proc glm;
  model Yph = x1clay x2org x3carb;
```

Figure 1.4 Sample SAS program to fit multiple linear regression model to soil pH data

```

                                The SAS System
                          General Linear Models Procedure

Dependent Variable: Yph

```

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	9.30703059	3.10234353	2.78	0.0834
Error	13	14.53296941	1.11792072		
Corrected Total	16	23.84000000			

	R-Square	Coeff Var	Root MSE	Yph Mean
	0.390396	16.01997	1.057318	6.600000

Parameter	Estimate	Standard Error	t Value	Pr > t
Intercept	7.030707780	0.85651288	8.21	<.0001
x1clay	0.016908806	0.02203381	0.77	0.4566
x2org	-0.424258440	0.26104819	-1.63	0.1281
x3carb	0.078999750	0.03537659	2.23	0.0437

Figure 1.5 SAS output (edited) from multiple linear regression fit for soil pH data

Environmental questions of interest with these data include: (i) do the soil texture and organic matter predictors (x_{i1} and x_{i2} , respectively) affect soil pH; and (ii) does carbonate composition (as quantified via x_{i3}) affect soil pH above and beyond any effects of soil texture and organic matter? To address issue (i), we can apply the discrepancy-measure approach from (1.7). Set the null hypothesis to $H_0: \beta_1 = \beta_2 = 0$. Figure 1.5 represents information on the full model, so we find $SSE(\text{FM}) = 14.533$ with $df_e(\text{FM}) = 13$. To fit the reduced model, we can use code similar to that in Fig. 1.4, replacing only the call to PROC GLM:

```
proc glm; model Yph=x3carb;
```

This produces $SSE(\text{RM}) = 17.775$ with $df_e(\text{RM}) = 15$ (output not shown; notice that this reduced model corresponds to a simple linear regression fit of Y on x_{i3}). Thus (1.7) becomes $F_{\text{calc}} = (3.242/2)/(14.533/13) = 1.621/1.118 = 1.450$. At $\alpha = 0.05$, we refer this to $F_{0.05}(2, 13) = 3.806$. (Critical points for F -distributions are best acquired by computer; see Fig. A.8.) Since 1.450 fails to exceed 3.806, we cannot reject H_0 , leading us to conclude that addition of x_{i1} and x_{i2} does not significantly improve the model that already contains x_{i3} as a predictor variable. Note that in SAS, it is possible to acquire the test statistic F_{calc} directly, using PROC REG and the command

```
test x1clay=0, x2org=0;
```

To address issue (ii) we can use the SAS output to construct a 95% confidence interval for β_3 . This is $b_3 \pm t_{0.025}(df_e)se[b_3]$, where for these data $df_e = 17 - 3 - 1 = 13$. From Table B.2 we find $t_{0.025}(13) = 2.1604$. Thus, using the information supplied in the last block of output in Fig. 1.5, we find the 95% interval to be $0.079 \pm 2.1604 \times 0.035 = 0.079 \pm 0.076$, or $0.003 < \beta_3 < 0.155$. That is, a unit (i.e. 1%) increase in CaCO_3 by weight appears to increase soil pH by between 0.003 and 0.155 units when both other predictors are held fixed. In terms of x_{i3} , we conclude that the carbonate predictor does have a significant effect on soil pH since this interval does not contain $\beta_3 = 0$. Notice that the corresponding P -value from Fig. 1.5 is $P = 0.0437$.

Exercise 1.4 studies the residuals under the full and reduced models in detail. 🌐

An important consideration when employing multiple linear regression models is the differing contributions of the various predictor variables. It can be the case that information in one of the predictors closely duplicates or overlaps with information in another predictor. The effect is known as *multicollinearity* among the predictor variables, and it can lead to computational instabilities, large statistical correlations, and inflated standard errors in the LS/ML estimators for β . A number of strategies exist to identify multicollinearity in a set of posited predictor variables. Most simply, examining pairwise scatterplots among all the predictor variables is an easy way to see if any two predictor variables are providing redundant information. More complicated diagnostics include the so-called *variance inflation factor (VIF)* for each of the p predictors. This is defined as $VIF_j = 1/(1 - R_j^2)$, where R_j^2 is the coefficient of multiple determination found from regressing the j th predictor, x_{ij} , on the other $p - 1$ predictor variables ($j = 1, \dots, p$). In SAS, PROC REG computes VIFs via the `vif`

option in the `model` statement. A set of predictors whose maximum VIF exceeds 10 is felt to be highly collinear. For the three predictor variables in Example 1.3, above, none of the VIFs exceeds 1.2, implying no serious concern regarding multicollinearity (Exercise 1.4).

Remedies for multicollinearity are primarily design-related: (i) to the best extent possible, identify predictors that provide separate information; (ii) avoid inclusion of predictors that are essentially the same variable (e.g., species abundance and species density in an ecological field study); (iii) always try to keep the number of predictor variables manageable (see below). One can also try transforming highly collinear predictor variables – via, say, a logarithmic or reciprocal transform – to decrease their VIFs. In some cases, as simple an operation as centering each variable about its mean by replacing x_{ij} by $u_{ij} = x_{ij} - \bar{x}_i$ can help alleviate multicollinearity. We employ this strategy with polynomial regression models in §1.5, below. One can also apply different forms of multivariate data reduction, such as principal components analysis, to the set of predictors to make them more amenable to regression modeling. We will not go into detail on such approaches here, however. For more on multivariate statistical methods see, for example, Manly (1994).

Notice that VIFs are determined from only the x_{ij} s, and so they may be computed prior to acquisition of the Y_i s. Thus, where possible, the VIFs can be calculated and a check for multicollinearity can precede sampling. Highly collinear predictors can be reassessed as to their anticipated value to the analysis, and removed if felt to be non-essential or marginal.

The construction and specification of predictor variables for a multiple linear regression can be a complex process. A natural approach, seen in many environmental studies, is to identify as many possible predictors as the investigator feels might affect the response and include all of these in the regression equation. Combinations of the original predictors are also possible, such as higher-order polynomial and cross-product terms; see §1.5. For example, a multiple linear regression with three predictor variables, x_{i1} , x_{i2} , and x_{i3} , may be expanded to second order by including the quadratic terms $x_{i4} = x_{i1}^2$, $x_{i5} = x_{i2}^2$, and $x_{i6} = x_{i3}^2$ and the second-order cross products $x_{i7} = x_{i1}x_{i2}$, $x_{i8} = x_{i1}x_{i3}$, and $x_{i9} = x_{i2}x_{i3}$. (The cross products are called ‘interaction’ terms. Note that it often does not make sense to include a higher-order term in a model without also including all its lower-order siblings. Thus, for example, if we were to include $x_{i9} = x_{i2}x_{i3}$ we would typically also include x_{i2} and x_{i3} . Similarly, if we include $x_{i6} = x_{i3}^2$ we would also include x_{i3} , etc.) Thus from three original x -variables we can construct a full second-order model with $p = 9$ predictors. Not all of these nine variables will necessarily be important in the model, but they can nonetheless be included and assessed for significance (as long as $n > p + 1 = 10$ observations are recorded). Exercise 1.5 explores this sort of modeling approach.

It is also possible to automate the variable selection effort. This involves studying in a stepwise, systematic fashion all possible models (or some prespecified subset, such as all possible second-order models) available from a given set of original x -variables. From these, an operating model is chosen that meets or exceeds some optimality criterion. For example, one might wish to minimize the MSE or maximize R^2 after adjusting for the total number of predictor variables employed in the model. From these exploratory efforts, conclusions can be drawn on the selected model, using the sorts of inferential and diagnostic methods described above. (The inferences

should be viewed as preliminary or tentative, however, since they are typically not corrected for the repeated *a priori* testing required to identify the selected model. Potscher and Novak, 1998, and Zhang, 1992, discuss some theoretical complexities associated with making formal inferences after variable selection; see also Olden and Jackson, 2000.)

The systematic search can be performed backwards from a maximal model (*backward*, or *backstep*, *elimination*), or forwards from some very simple progenitor model (*forward selection*); see Neter *et al.* (1996, Ch. 8). The effort is by necessity computationally intensive, and therefore is best performed by computer; for example, in SAS use PROC STEPWISE, or apply PROC REG with the `selection=` option in the `model` statement. (Indeed, all 2^p possible regression models for a set of p predictor variables can be constructed using PROC RSQUARE.) We should warn the user, however, that blindly or inattentively ceding to the computer the final decision(s) on which predictors to include in a regression model is often foolhardy, since the computer cannot incorporate important interpretations from the underlying subject-matter. No automated variable selection procedure should ever replace informed scientific judgment.

1.3 Qualitative predictors: ANOVA and ANCOVA models

1.3.1 ANOVA models

In many environmental regression problems, variables that are thought to affect the mean response may be qualitative rather than quantitative; for example, sex or ethnic status in an environmental health investigation, or habitat type in an ecological study. In such settings, it is still desirable to try and relate the mean response to some sort of linear predictor. A natural approach is to assign a set of codes or scores to the different levels of the qualitative variable(s) and build from these a series of quantitative predictor variables. For instance, suppose we remove the intercept term from the model to avoid complications in interpretation. Then, we could set $x_{i1} = 1$ if the i th observation corresponds to the first level of the (first) qualitative predictor (zero otherwise), $x_{i2} = 1$ if the i th observation corresponds to the second level of the (first) qualitative predictor (zero otherwise), etc. Other coding schemes are also possible (Neter *et al.*, 1996, §16.11).

In this manner, a multiple linear regression model can be constructed to account for qualitative as well as quantitative predictor variables. For the qualitative setting, however, interpretation of the regression coefficients as slopes, or as changes in effect, becomes questionable when using simple coding schemes such as those suggested above. As a result, we often write the model in a more traditional parameterization, based on a so-called ANOVA structure. (The name comes from use of an analysis of variance to assess the effects of the qualitative factor; see Neter *et al.*, 1996, Ch. 16.) To each qualitative factor the model assigns certain effect parameters, say, α_i for factor A, β_j for factor B, γ_k for factor C, etc. The factor indices vary over all levels of each factor. For instance, if a single factor, A, has a levels, then $i = 1, \dots, a$.

Assuming n_i observations are recorded at each combination of this single factor, we write

$$Y_{ij} = \theta + \alpha_i + \varepsilon_{ij}, \quad (1.8)$$

$i = 1, \dots, a, j = 1, \dots, n_i$. As above, we assume that the random error terms satisfy $\varepsilon_{ij} \sim \text{i.i.d. } N(0, \sigma^2)$. The standalone parameter θ may be interpreted as the *grand mean* of the model when the α_i parameters are viewed as deviations from θ due to the effects of factor A. This is a *one-factor ANOVA model*. The total sample size is $N = \sum_{i=1}^a n_i$. If n_i is constant at each level of i , say $n_i = n$, then we say the ANOVA is *balanced*, otherwise it is *unbalanced*.

Unfortunately, under this factor-effect parameterization there is not enough information in the data to estimate every parameter. As currently stated, the model in (1.8) has $a + 1$ regression parameters, $\theta, \alpha_1, \dots, \alpha_a$, but only a different groups from which to estimate these values. Thus we can estimate all a values of α_i (and σ^2 , using the usual MSE term), but not θ . To accommodate estimation of θ , we impose an *estimability constraint* on the α_i s. Many possible constraints exist; two of the more common are the *zero-sum constraint* $\sum_{i=1}^a \alpha_i = 0$ and the *corner-point constraint* $\alpha_a = 0$. If applied properly, the constraints do not affect tests or estimates of the differences between groups or estimates of other linear combinations of the group means; however, the interpretation of the parameter estimates is directly impacted by the constraint used. For example, under the zero-sum constraint, θ is the mean of all the a group means and α_i is the difference between the i th factor-level mean and the overall mean. By contrast, under the corner-point constraint θ is the mean of the last factor level while α_i is the difference between the i th factor-level mean and this last factor level's mean.

If a second factor, B, were considered in the study design such that n_{ij} observations are recorded at each combination of the two factors, we write $Y_{ijk} = \theta + \alpha_i + \beta_j + \varepsilon_{ijk}, i = 1, \dots, a, j = 1, \dots, b, k = 1, \dots, n_{ij}$. This is a *two-factor, main-effects ANOVA model*, so named since it contains effects due to only the two main factors. Allowing for the additional possibility that the two factors may interact calls for expansion of the model into

$$Y_{ijk} = \theta + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}, \quad (1.9)$$

$i = 1, \dots, a, j = 1, \dots, b, k = 1, \dots, n_{ij}$, with total sample size $N = \sum_{i=1}^a \sum_{j=1}^b n_{ij}$. The γ_{ij} s are the *interaction parameters*. As in (1.8), estimability constraints are required under this factor-effects parameterization. The zero-sum constraints are $\sum_{i=1}^a \alpha_i = 0, \sum_{j=1}^b \beta_j = 0, \sum_{i=1}^a \gamma_{ij} = 0$ for all j , and $\sum_{j=1}^b \gamma_{ij} = 0$ for all i . The corner-point constraints are $\alpha_a = 0, \beta_b = 0, \gamma_{aj} = 0$ for all j , and $\gamma_{ib} = 0$ for all i . Extensions to higher-factor models are also possible (Neter *et al.*, 1996, Ch. 23).

To analyze data under an ANOVA model, closed-form equations exist only in select cases. These provide point estimates of each mean level's effect and sums of squares for the ANOVA table; see Samuels and Witmer (2003, Ch. 11) or Neter *et al.* (1996, Ch. 16). Because of their correspondence to multiple linear regression models, however, ANOVA-type models may also be studied using regression methods, via computer. In SAS, for example, PROC GLM can analyze ANOVA-type models by

invoking the `class` statement; this is used to identify which predictors are qualitative ‘classification’ variables. PROC GLM can produce sequential sums of squares (SS) for testing the significance of model components in the sequential order in which they are fitted, partial SS for testing the significance of model components when they are fitted last in sequential order, the various corresponding mean squares, and from these the ANOVA table. Ratios of the factor-specific mean squares to the MSE produce F -statistics to assess the significance of any factor or any multi-factor interaction, using the general approach embodied in equation (1.7); see Neter *et al.* (1996, Chs. 16–22). For example, the null hypothesis of no $A \times B$ interaction under a full two-factor model is $H_0: \gamma_{11} = \dots = \gamma_{ab} = 0$. Assess this via the F -statistic

$$F_{\text{calc}} = \frac{\text{MS}[A \times B|A, B]}{\text{MSE}}$$

and reject H_0 at significance level δ if $F_{\text{calc}} \geq F_{\delta}([a-1][b-1], \text{df}_e)$, where df_e are the df associated with the error term. (The symbol ‘|’ is used to indicate sequencing or conditioning. Read it as ‘fitted after’ or ‘conditional on’.) For the full two-factor model, $\text{df}_e = N - ab + 1$.

As with our earlier comment on model construction with higher-order quantitative predictors, it does not make sense to include a higher-order qualitative term in a model without also including all associated lower-order terms. Thus, for example, if we were to include a two-factor interaction using the γ_{ij} s, we would typically also include both main effects terms via the α_i s and β_j s. This suggests a natural ordering for hypothesis testing: test interaction first via $H_0: \gamma_{ij} = 0$ for all i, j . If the interaction is insignificant, follow with separate tests of each main effect via $H_0: \alpha_i = 0$ for all i and $H_0: \beta_j = 0$ for all j . In some settings this strategy has a mathematical interpretation: if the interaction terms represent a departure from the additive effects of the two factors, testing interaction first is equivalent to evaluating whether a departure from additive effects is present before examining the separate additive effects of each main effect.

Example 1.4 (Macroinvertebrate ratios) Gonzalez and Manly (1998) give data on the change in abundance of two freshwater benthic macroinvertebrates (ephemeroptera and oligochaetes) in three New Zealand streams over the four different annual seasons. Ephemeroptera are pollution-sensitive, so by measuring the ratio Y_{ij} of ephemeroptera to oligochaetes we can assess potential stream pollution. We view the $a = 3$ streams as one factor, and the $b = 4$ seasons as a second factor; $n = 3$ independent random samples were taken at each of the 12 stream \times season combinations. (Following Gonzalez and Manly, we assume that the samples were separated enough in time so that no temporal correlation exists within streams across the different seasons for these data.) The data are given in Table 1.3. For them, we assume the full two-factor model (1.9). A SAS analysis via PROC GLM is given via the sample code in Fig. 1.6.

The output (edited) from PROC GLM appears in Fig. 1.7. In it, we begin by assessing the stream \times season interaction for these data. That is, compare the full model with linear predictor $\theta + \alpha_i + \beta_j + \gamma_{ij}$ to a reduced model whose linear predictor contains only the main-effect additive terms $\theta + \alpha_i + \beta_j$. Using either sequential

Table 1.3 Ratios of ephemeroptera to oligochaetes in three New Zealand streams

Stream	Season			
	Summer	Autumn	Winter	Spring
A	0.7, 8.5, 7.1	0.5, 1.1, 1.4	0.1, 0.1, 0.2	0.2, 0.1, 0.4
B	1.2, 0.7, 0.8	10.7, 19.9, 9.4	2.0, 6.3, 4.8	2.0, 1.6, 1.9
C	7.3, 10.4, 9.5	46.6, 20.3, 24.0	1.2, 0.8, 6.1	0.2, 0.1, 0.1

Source: Gonzalez and Manly (1998).


```

* SAS code to fit 2-factor ANOVA;
data macroinv;
input stream $ season $ Y @@;
datalines;
  A sum 0.7      A sum 8.5      A sum 7.1
  :           :           :
  C spr 0.2      C spr 0.1      C spr 0.1
proc glm;
class stream season;
model Y = stream season stream*season;
    
```

Figure 1.6 Sample SAS program to fit two-factor ANOVA model to macroinvertebrate ratio data

The SAS System						
General Linear Models Procedure						
		Class Level Information				
Class	Levels	Values				
stream	3	A	B	C		
season	4	fal	spr	sum	win	
Dependent Variable: Y						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	11	2512.389722	228.399066	10.18	<.0001	
Error	24	538.233333	22.426389			
Corrected Total	35	3050.623056				
R-Square		Coeff Var	Root MSE	Y Mean		
0.823566		81.84514	4.735651	5.786111		
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
stream	2	478.203889	239.101944	10.66	0.0005	
season	3	1080.727500	360.242500	16.06	<.0001	
stream*season	6	953.458333	158.909722	7.09	0.0002	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
stream	2	478.203889	239.101944	10.66	0.0005	
season	3	1080.727500	360.242500	16.06	<.0001	
stream*season	6	953.458333	158.909722	7.09	0.0002	

Figure 1.7 SAS output (edited) from two-factor ANOVA fit for macroinvertebrate ratio data

sums of squares (also called `Type I SS` in `PROC GLM`) or partial sums of squares (called `Type III SS` in `PROC GLM`), the F -statistic is $F_{\text{calc}} = (953.4583/6)/(538.2333/24) = 7.09$. (Since the design here is balanced, the Type I and Type III SS are identical.) Referring this to $F(6,24)$, `PROC GLM` gives the interaction P -value as $P = 0.0002$, implying that a significant stream \times season interaction exists. Standard practice requires therefore that further analysis of the stream or season main effects is inappropriate, since the main effects cannot be disentangled from each other (Neter *et al.*, 1996, §20.3). Further analysis of the stream and/or season effect must be stratified across levels of the other factor; for example, to assess if there is an effect due to season, perform an analysis of the season effect at each level of the stream factor. (That is, perform three separate one-factor ANOVAs.) A similar stratification is required to assess if there is a stream effect. In Exercise 1.8, such an analysis identifies selected significant effects for both seasons and streams. 

We should warn that complications can occur in two-factor (and higher) ANOVA if the design is unbalanced, that is, if $n_i \neq n$ for any i . Due to the lack of balance, point estimates of a factor's effects when averaged over the levels of the other factor(s) may not correspond to the marginal quantities of interest, and this can carry over into tests of hypotheses and other inferences on that factor. Estimation by what are called *least squares means* is often employed in these instances. (For example, in SAS one can apply the `lsmeans` command.) The details extend beyond the scope of our presentation here, however, and we note only that investigators must be careful when analyzing unbalanced multi-factor designs. For more on least squares means and analyses for unbalanced data, see Neter *et al.* (1996, Ch. 22) or Littell *et al.* (2002, §6.3).

1.3.2 ANCOVA models

It is also possible to work with a mix of both quantitative and qualitative predictor variables. The simplest example of this occurs when there is a single qualitative factor under study, but where it is known that the data vary according to some other, quantitative variable. Interest exists in assessing the effects of the qualitative factor, adjusted for any variation due to the quantitative factor. In this case, it is common to call the quantitative factor a 'covariate,' and apply an *analysis of covariance* (*ANCOVA*). The model may be written as

$$Y_{ij} = \theta + \alpha_i + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij}, \quad (1.10)$$

$i = 1, \dots, a, j = 1, \dots, n_i$. The total sample size is $N = \sum_{i=1}^a n_i$. In (1.10), α_i represents the differential factor-level effect and β is the slope parameter associated with the covariate. (The slope parameter is held constant under this model; this *equal-slopes assumption* requires that the covariate have the same effect across all levels of the qualitative factor.) An estimability constraint is again required on the α_i s, such

as the zero-sum constraint $\sum_{i=1}^a \alpha_i = 0$. The error terms take the usual assumption: $\varepsilon_{ij} \sim \text{i.i.d. } N(0, \sigma^2)$.

Notice the correction in (1.10) of the covariate by its mean $\bar{x} = \sum_{i=1}^a \sum_{j=1}^{n_i} x_{ij} / \sum_{i=1}^a n_i$. Many authors view this primarily as a computational aid, used to adjust the different estimates of the factor-level effects for any differential covariate effects. It yields a pertinent interpretation, however, as an evaluation of the differences among levels of the qualitative factor at a fixed level of $x = \bar{x}$ (Neter *et al.*, 1996, §25.2).

Inferences in an ANCOVA are most often directed at the qualitative variable, after ‘adjustment’ for the covariate effect. This is performed by constructing an ANOVA table with the covariate fitted before the qualitative factor, and assessing significance via the sequential sum of squares for the qualitative factor: $SS[\text{factor}|\text{covariate}]$. Applied in equation (1.7), this compares the full model in (1.10) to the reduced model $Y_{ij} = \theta + \beta(x_{ij} - \bar{x}) + \varepsilon_{ij}$. The resulting test statistic is $F_{\text{calc}} = MS[\text{factor}|\text{covariate}]/MSE$; this reflects the added contribution of the qualitative factor to the model. Reject the null hypothesis of no factor effect, $H_0: \alpha_1 = \dots = \alpha_a$, at significance level δ if $F_{\text{calc}} \geq F_{\delta}(a - 1, N - a - 1)$.

Example 1.5 (Faba bean genotoxicity) Menke *et al.* (2000) report on studies of the faba bean plant (*Vicia faba*) for use as an ecotoxicological indicator species. One specific endpoint investigated was DNA damage (percentage of DNA separating out during single-cell gel electrophoresis) in nuclei from *V. faba* root tip cells. Greater damage is indicated by higher percentages of separated DNA. Of interest was whether changing the treatment of the cells with different endonucleases produced differences in DNA damage. The study protocol also called for a variety of treatment times, since increasing treatment time was thought to increase genotoxic response and hence enhance the assay’s potential as a ecotoxicological screen.

Listed in Table 1.4 are the mean percentage responses of separated DNA as a function of treatment time ($x = 0, 10, 20$, or 30 min) and endonuclease treatment (*FokI*, *EcoRI*, *DNaseI* + MgCl_2 , or *DNaseI* + MnCl_2). Notice the highly unbalanced structure of the experimental design. We view exposure time as the quantitative covariate, and endonuclease treatment as the qualitative predictor to be tested.

One additional concern is the necessary assumption in (1.10) of variance homogeneity. Since percentage data are notorious for their variance heterogeneity, we apply a variance-stabilizing transform here: the logit of the observed percentages. That is, if P_{ij} is the mean percentage response for treatment i at time level j given in Table 1.4, we take $Y_{ij} = \log\{P_{ij}/(100 - P_{ij})\}$ as the response variable for analysis. (An alternative transformation is the arcsine square root: $V_{ij} = \sin^{-1} \sqrt{P_{ij}/100}$.) A plot of the data (Fig. 1.8) shows that the logits are roughly linear across the time span under study.

Figure 1.9 presents sample SAS code for fitting the equal-slopes ANCOVA model from (1.10) to the data in Table 1.4. Notice our continuing use of PROC GLM for the fit. (The `solution` option in the `model` statement calls for LS estimates of the α_i s and of β . The estimates of α_i depend on the choice of estimability constraint and so require careful interpretation. SAS indicates this in its output with a warning message that the estimators are ‘not unique’; see Fig. 1.10. The LS estimate of β is, however, a valid, unique, LS estimate of the common slope. The corresponding standard error can be used to construct confidence limits or test hypotheses about β ; see the comment below.)

22 Linear regression

Table 1.4 Mean percentage DNA separation in *V. faba* root tip cell nuclei

Endonuclease treatment	Treatment time (min)			
	0	10	20	30
<i>FokI</i>	20.5	65.4	68.4	78.1
	13.4	56.0	54.7	72.8
	17.2	57.3	62.8	81.9
	19.6	59.2	61.8	79.4
<i>EcoRI</i>	6.8	10.5	49.0	62.1
	13.0	12.6	40.6	50.2
DNaseI + MgCl ₂	18.2	24.5	28.0	48.2
	21.9	28.0	31.6	47.8
	2.6	–	56.2	–
	2.4	–	56.0	–
	9.9	–	52.7	–
	14.4	–	42.6	–
DNaseI + MnCl ₂	58.9	91.5	96.6	–
	65.2	89.9	96.6	–

Dashes indicate no observation at that time–treatment combination. Source: Menke *et al.* (2000).

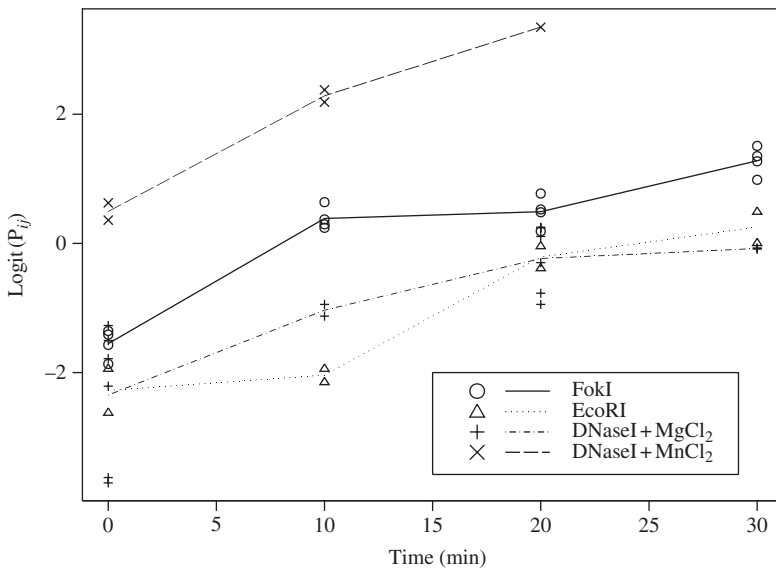


Figure 1.8 Scatterplot of logit-transformed proportions for faba bean genotoxicity data

```

* SAS code to fit equal-slopes ANCOVA;
data faba;
input etrt $ time percent @@;
  y = log(percent / (100 - percent));
  timebar = 13.47826;
  covar = time - timebar;
datalines;
FokI  0 20.5  FokI  0 13.4  FokI  0 17.2  FokI  0 19.6
      ⋮          ⋮          ⋮          ⋮
MnCl2 10 91.5 MnCl2 10 89.9 MnCl2 20 96.6 MnCl2 20 96.6

proc glm;
class etrt;
model Y = covar etrt / solution;

```

Figure 1.9 Sample SAS program to fit equal-slopes ANCOVA for faba bean genotoxicity data

The SAS System						
General Linear Models Procedure						
Dependent Variable: Y						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	4	96.5575940	24.1393985	70.53	<.0001	
Error	41	14.0331928	0.3422730			
Corrected Total	45	110.5907868				
	R-Square	Coeff Var	Root MSE	Y Mean		
	0.873107	-232.1263	0.585041	-0.252036		
Source	DF	Type I SS	Mean Square	F Value	Pr > F	
covar	1	38.19447717	38.19447717	111.59	<.0001	
etrt	3	58.36311682	19.45437227	56.84	<.0001	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
covar	1	45.39120872	45.39120872	132.62	<.0001	
etrt	3	58.36311682	19.45437227	56.84	<.0001	
Parameter		Estimate	Std. Error	t Value	Pr > t	
Intercept		2.362511350 B	0.24047205	9.82	<.0001	
covar		0.092543930	0.00803616	11.52	<.0001	
etrt EcoRI		-3.567986935 B	0.31850293	-11.20	<.0001	
etrt FokI		-2.352084490 B	0.28293472	-8.31	<.0001	
etrt MgCl2		-3.380745038 B	0.28078673	-12.04	<.0001	
etrt MnCl2		0.000000000 B	.	.	.	
NOTE: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.						

Figure 1.10 SAS output (edited) from equal-slopes ANCOVA fit for faba bean genotoxicity data

We find from Fig. 1.10 that the F -test of $H_0: \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4$, that is, whether endonuclease treatment makes a difference in the generation of DNA damage, produces a test statistic of $F_{\text{calc}} = 56.84$, on 3 and 41 df (using either the Type I or Type III SS term to build the F -statistic). The corresponding P -value is $P < 0.0001$, and is highly significant. We conclude that there is a clear difference among endonuclease treatments, after adjusting for possible differences in percent DNA damage due to treatment time. 🌱

Some environmental phenomena yield data that are best studied using an ANCOVA-type model where the quantitative predictor is itself of interest, making the null hypothesis $H_0: \beta = 0$ a primary concern. If so, then it is natural to reverse the order of the fit, and first test $H_0: \beta = 0$ via the F -statistic $F_{\text{calc}, \beta} = \text{MS}[\text{quantitative predictor} | \text{qualitative factor}] / \text{MSE}$. Here, we reject H_0 at significance level δ if $F_{\text{calc}, \beta} \geq F_{\delta}(1, N - a - 1)$. Failure to reject H_0 then allows for testing of the qualitative factor; that is, if $F_{\text{calc}, \beta} < F_{\delta}(1, N - a - 1)$, we can test $H_0: \alpha_1 = \dots = \alpha_a$ via $F_{\text{calc}, \alpha} = \text{MS}[\text{qualitative factor}] / \text{MSE}$. Reject the new H_0 at significance level δ if $F_{\text{calc}, \alpha} \geq F_{\delta}(a - 1, N - a - 1)$. If appropriate, (1.10) also can include polynomial terms for the covariate, such as the quadratic ANCOVA model

$$Y_{ij} = \theta + \alpha_i + \beta_1(x_{ij} - \bar{x}) + \beta_2(x_{ij} - \bar{x})^2 + \varepsilon_{ij}, \quad (1.11)$$

$i = 1, \dots, a, j = 1, \dots, n_i$.

As noted previously, the ANCOVA analysis as we have presented it makes the strong assumption that the covariate effect is the same across all levels of the qualitative factor. We can extend the model in (1.10) to assess this assumption: simply allow for differential slopes. That is, permit β to vary with the qualitative level index, producing

$$Y_{ij} = \theta + \alpha_i + \beta_i(x_{ij} - \bar{x}) + \varepsilon_{ij}, \quad (1.12)$$

$i = 1, \dots, a, j = 1, \dots, n_i$. Computationally, this is equivalent to including an interaction term; for example, in Fig. 1.9, use the alternative `model` statements

```
model y=covar etrt covar*etrt / solution;
```

or equivalently

```
model y=covar | etrt / solution;
```

Exercise 1.9 explores this with the faba bean genotoxicity data from Example 1.5.

1.4 Random-effects models

The ANOVA and ANCOVA models (1.8)–(1.12) employ a single, additive error term to describe any random variation (the *stochastic* components) in the data. All the other terms are assumed to be constant or fixed (the *deterministic* components of the model), and we often say that this combination represents a *fixed-effects model*. In some settings, however, random variation may be associated with one or more of the factors under study, and hence the fixed-effects assumption may be untenable. For instance, the factor levels under study may themselves be a random sample of all possible levels for that particular factor. This is called a *random effect*.

The consequences of a random-effect assumption on model equations such as (1.8) or (1.9) are relatively straightforward. Consider (1.8): if the treatment's effect is itself random, make the assumption that $\alpha_i \sim \text{i.i.d. } N(0, \tau^2)$. (We also assume that the α_i s are

statistically independent of the ε_{ijs} .) Notice that since the α_i s are now viewed as normal random variables, it makes no sense to place any estimability conditions on them.

The consequences of a random-effect assumption on the observations and on the analysis are somewhat more subtle. Equation (1.8) now leads to $E[Y_{ij}] = \theta$ and $\text{Var}[Y_{ij}] = \sigma^2 + \tau^2$. For simplicity, assume the design is balanced, so $n_i = n$. In this simple case with only a single qualitative factor, estimation of the factor effect parameters α_i is typically no longer relevant, since they are now random. Instead, we center attention on the contribution of each model component to the total variance $\text{Var}[Y_{ij}]$. These *variance components* (Searle *et al.*, 1992) may be estimated as $\hat{\sigma}^2 = \text{MSE}$ (again), and $\hat{\tau}^2 = (\text{MSA} - \text{MSE})/n$, where MSA is the mean square associated with the single (A) factor. Notice, however, that there is no guarantee that $\hat{\tau}^2$ will be non-negative, even though we assume $\tau^2 \geq 0$. A simple correction truncates $\hat{\tau}^2$ at zero if it falls below that value. More complex estimation schemes to accommodate this non-negativity, along with adaptations for designs when the experimental design is not balanced, are also possible; see Hocking (1996, Ch. 17), Kelly and Mathew (1994), or the dedicated text by Searle *et al.* (1992). To test for an effect due to factor A, we now write $H_0: \tau^2 = 0$. Rejection of H_0 occurs at significance level δ when $F_{\text{calc}} = \text{MSA}/\text{MSE} \geq F_{\delta}(a-1, a[n-1])$.

Another useful quantity associated with the variance components in a random-effects model is the *intraclass correlation coefficient*: $\tau^2/(\tau^2 + \sigma^2)$. This is the correlation between Y_{ij} and Y_{ih} ($j \neq h$) for any fixed level of i , and it measures the proportion of variation in Y accounted for by variation in the random effect.

Confidence intervals for the variance components or for certain functions of them, such as the intraclass correlation, are often straightforward to construct. For example, a $1 - \delta$ confidence interval for $\tau^2/(\tau^2 + \sigma^2)$ has the form $L/(1+L) < \tau^2/(\tau^2 + \sigma^2) < U/(1+U)$, for

$$L = \frac{1}{n} \left\{ \frac{F_{\text{calc}}}{F_{(1-\delta)/2}(a-1, a[n-1])} - 1 \right\}$$

$$U = \frac{1}{n} \left\{ \frac{F_{\text{calc}}}{F_{\delta/2}(a-1, a[n-1])} - 1 \right\},$$

and where $F_{\text{calc}} = \text{MSA}/\text{MSE}$ from above.

Extensions to the two-factor case in (1.9) follow in similar form. For example, if α_i , β_j , and γ_{ij} all represent random effects in (1.9), then we make the assumptions that $\alpha_i \sim \text{i.i.d. } N(0, \tau_\alpha^2)$, $\beta_j \sim \text{i.i.d. } N(0, \tau_\beta^2)$, and $\gamma_{ij} \sim \text{i.i.d. } N(0, \tau_\gamma^2)$, where the random variables are all assumed mutually independent of each other and of ε_{ijk} . Under such a multi-factor random-effects model, estimation and inference proceed in a manner similar to that for the single-factor case; see Neter *et al.* (1996, §24.2).

When some of the factors in a multi-factor model are random but others remain fixed we have a *mixed-effects model*, or simply a *mixed model*. In a mixed model, the inferences must change slightly to accommodate the mixed-effect features. We will not go into additional detail here on the subtleties of estimation and inference with mixed-effects models, although see §5.6.2 for a special case with temporally correlated data. Readers unfamiliar with mixed models may find Neter *et al.* (1996,

Ch. 24) a useful source; also see the review by Drum (2002). For fitting mixed models in SAS, the workbook by Littell *et al.* (1996) is particularly helpful.

1.5 Polynomial regression

In many environmental applications, the mean response is affected in a much more complex fashion than can be described by simple linear relationships. A natural extension of the simple linear relationship is the addition of higher-order polynomial terms. We hint at this above with the quadratic ANCOVA model of equation (1.11). Made formal, a p th-order polynomial regression model is

$$Y_i = \beta_0 + \beta_1(x_i - \bar{x}) + \beta_2(x_i - \bar{x})^2 + \cdots + \beta_p(x_i - \bar{x})^p + \varepsilon_i, \quad (1.13)$$

$i = 1, \dots, n$. We make the usual homogeneous-variance, normal-error assumption: $\varepsilon_i \sim \text{i.i.d. } N(0, \sigma^2)$. Also, we assume $p < n - 1$, and require that there be at least $p + 1$ distinguishable values among the x_i s.

This model is used to represent simple curvilinear relationships between $E[Y_i]$ and x_i . Notice that for any $p \geq 1$, (1.13) is in fact a special case of the multiple regression model from §1.2, where $x_{ij} = (x_i - \bar{x})^j$. (Centering the x_i s accomplishes two objectives: (i) it provides an interpretation for β_0 as the mean response at $x = \bar{x}$; and (ii) it helps avoid possible multicollinearity in the predictors (Bradley and Srivastava, 1979). Where possible, one can also *design* the study to avoid multicollinearity and still operate with a polynomial response model. That is, if the values of x_i can be chosen in advance, special forms called *orthogonal polynomials* exist that guarantee no multicollinearity among the polynomial regressors. See, for example, the review by Narula, 1979.) Thus all the statistical machinery available for fitting and analyzing a multiple linear regression is also available for fitting and analyzing polynomial regression models. This fact underlies much of the justification for employing these models: curvilinear regression relationships in the environmental sciences are often more complex than can be described by a simple polynomial function. Equation (1.13) becomes useful, however, when it can provide an approximation to the true nonlinear relationship over the range of x under study. In this case, the simplicity of use made possible by connection to multiple linear regression methodology becomes an attractive motivation. For example, to test the effect of the x -variable, assess $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ via an F -statistic as in (1.7). Or simple hypothesis tests on the individual regression coefficients follow from standard t -distribution theory. For instance, if $p = 2$ a test of $H_0: \beta_2 = 0$ using $t_{\text{calc}} = |b_2|/se[b_2]$ assesses whether the quadratic curvature is required in a model that already has a linear term present by comparing the full model with linear predictor $\beta_0 + \beta_1(x_i - \bar{x}) + \beta_2(x_i - \bar{x})^2$ to a reduced model with the simple linear predictor $\beta_0 + \beta_1(x_i - \bar{x})$. Note, however, that in the polynomial context β_j loses its interpretation as change in response to the j th predictor when holding the other predictors fixed. Clearly, one cannot vary $(x_i - \bar{x})^j$ while holding $(x_i - \bar{x})$, \dots , $(x_i - \bar{x})^{j-1}$, $(x_i - \bar{x})^{j+1}$, \dots , $(x_i - \bar{x})^p$ fixed.

The best-known and most useful polynomials are the quadratic (or parabolic) model at $p = 2$ and the cubic model at $p = 3$. The former is useful for curvilinear response that changes once and only once from strictly increasing (decreasing) to strictly decreasing (increasing). The latter is useful if the curvilinear response inflects such that its *rate of change* shifts from strictly decreasing to strictly increasing or vice versa; for example, the response might move from increasing to decreasing, and back to increasing. Of course, any order of polynomial may be fitted if the data are rich enough. We do not recommend use of (1.13) past $p = 3$ or perhaps $p = 4$, however, unless there is strong motivation from the underlying environmental subject-matter. Also, as noted in §1.2, it generally does not make sense to include a higher-order term without also including all the lower orders. Thus, whenever we include a certain order of polynomial, p , in a regression model we generally also include all the lower orders, $p - 1, p - 2, \dots, 2, 1$.

Example 1.6 (Rice yield) Seshu and Cady (1984) present data on yield of irrigated rice (*Oryza sativa* L.) as a function of minimum temperature, with a goal of understanding the environmental conditions that affect yield. (The temperatures were averages of daily minima over a 30-day period immediately after the rice plants flowered.) The data appear in Table 1.5. A plot (Fig. 1.11) shows a clear curvilinear shape to the temperature–yield curve, for which a quadratic polynomial may provide a good approximation. Towards this end, we consider the polynomial regression model (1.13) with $p = 2$.

Table 1.5 Rice yield (t/ha) vs. minimum temperature (°C)

$x = \text{min. temp.}$	$Y = \text{yield}$	$x = \text{min. temp.}$	$Y = \text{yield}$
29.2	2.3	23.4	3.0
28.1	3.1	23.4	4.4
27.2	2.8	23.2	3.1
26.4	2.4	23.1	2.6
26.3	3.6	23.1	4.8
26.2	2.3	23.0	3.2
26.2	3.8	22.9	3.3
26.0	3.1	22.5	3.1
25.9	2.4	22.5	3.4
25.7	2.1	22.4	3.2
24.5	3.5	21.7	4.2
24.4	3.8	21.2	4.5
24.0	3.1	20.0	4.7
23.9	2.9	19.2	5.0
23.9	3.2	19.0	6.2
23.7	3.0	19.0	6.0
23.7	4.5	18.8	6.1
23.7	3.5	18.0	7.3
23.6	3.3	18.0	6.6
23.5	3.7	17.4	6.2

Source: Seshu and Cady (1984).

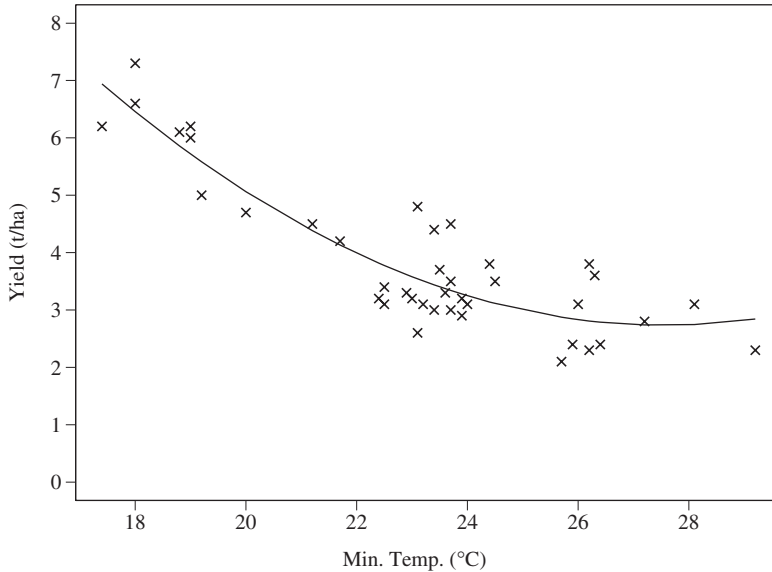


Figure 1.11 Scatterplot of observations (\times) and estimated LS parabola (—) for rice yield data from Table 1.5

A number of issues are of interest with these data. In particular, to assess if temperature significantly affects yield we set the null hypothesis to $H_0: \beta_1 = \beta_2 = 0$ and use an F -statistic based on the discrepancy measure in equation (1.7). Figure 1.12 gives sample SAS code to perform the LS fit, with associated output (edited) in Fig. 1.13. Therein, the F -statistic for testing whether all of the non intercept terms are zero is found under F Value in the main Analysis of Variance display. This is precisely the test of interest here, so for testing H_0 we take $F_{\text{calc}} = 78.09$. Referred to an $F(2,37)$ distribution, we see from the PROC REG output under $\text{Pr} > F$ that $P < 0.0001$. Clearly, there is a significant effect of temperature on yield.

```

* SAS code to fit quadratic regression;
data rice;
  input yield  mint @@;
  mintbar = 23.1975;
  x = mint - mintbar;
  x2 = x*x;
  datalines;
2.8  27.2   3.1  28.1   2.3  29.2   3.0  23.4
.:   .:   .:   .:   .:   .:   .:   .:
6.0  19.0   6.1  18.8   7.3  18.0   6.6  18.0
proc reg;
  model yield = x x2 ;

```

Figure 1.12 Sample SAS program to fit quadratic regression for rice yield data


The SAS System					
The REG Procedure					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	53.93171	26.96586	78.09	<.0001
Error	37	12.77604	0.34530		
Corrected Total	39	66.70775			
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	3.50874	0.11915	29.45	<.0001
x	1	-0.35563	0.03372	-10.55	<.0001
x2	1	0.04076	0.00939	4.34	0.0001

Figure 1.13 SAS output (edited) from quadratic regression fit for rice yield data

To study this effect further, we construct confidence intervals on the β_1, β_2 pair of regression coefficients. Pointwise intervals on each parameter are $b_j \pm t_{\alpha/2}(37)se[b_j]$, $j = 1, 2$, where the parameter estimates and standard errors can be read directly from Fig. 1.13 (under Parameter Estimate). Since the resulting intervals are based on the same set of data, however, some adjustment for multiplicity is required. The simplest is a Bonferroni adjustment (§A.5.4). Here, this affects only the critical point: the adjustment reduces α by a factor equal to the number of intervals being computed. Thus, the (two) Bonferroni-adjusted joint confidence intervals have the form $b_j \pm t_{\alpha/4}(37)se[b_j]$, $j = 1, 2$. At $\alpha = 0.05$, the necessary critical point is $t_{0.0125}(37)$. Using the SAS function `tinv` (see Fig. A.4), we find this as $t_{0.0125}(37) = 2.3363$. For β_1 , this produces the Bonferroni-adjusted interval $-0.3556 \pm 2.3363 \times 0.0337 = -0.3556 \pm 0.0788$. For β_2 , we find $0.0408 \pm 2.3363 \times 0.0094 = 0.0408 \pm 0.0219$. Both intervals fail to contain zero, suggesting that each term contributes significantly to the model (and, again, that the temperature–yield relationship is significant).

We can study the residuals from this LS/ML by plotting them against the predicted values; that is, plot $r_i = Y_i - \hat{Y}_i$ vs. $\hat{Y}_i = b_0 + b_1x_i + b_2x_i^2$. Sample SAS code to perform this is

```
proc reg;
  model yield = x x2 ;
  plot r. *p.;
```

The resulting residual plot (not shown) indicates no substantial diagnostic problems. 

Applications of polynomial models in the environmental sciences are far-reaching and diverse. The combination of curvilinear features and simple regression structure often allows for a number of interesting quantitative questions to be addressed. For example, the quadratic form ($p = 2$) of (1.13) using the mean-centered predictor $x - \bar{x}$ reaches a single optimum point (a minimum if the parabola is convex, a maximum if the parabola is concave) at $x_{\text{opt}} - \bar{x} = -\beta_1/2\beta_2$, or $x_{\text{opt}} = \bar{x} - \beta_1/2\beta_2$. At the core of this quantity is a ratio of parameters for which,

under a constant-variance, normal-error model, the ML estimator is simply the ratio of the point estimators. This leads to

$$\hat{x}_{\text{opt}} = \bar{x} - \frac{b_1}{2b_2}.$$

Unfortunately, although this point estimator is straightforward, statistical inferences on the ratio of parameters in x_{opt} can be difficult to construct. One possibility in this setting is to find confidence limits using a result due to Fieller (1940). *Fieller's theorem*, as it is known, gives a $1 - \alpha$ confidence interval on the ratio using the point estimates b_1 and b_2 , the standard errors $se[b_1]$ and $se[b_2]$, and an estimate of the covariance between b_1 and b_2 : $\hat{\sigma}_{12} = \text{Cov}[b_1, b_2]$. (Not to be confused with analysis of covariance from §1.3, the *covariance* here is a summary measure of the joint variability between b_1 and b_2 ; see §A.1.) These quantities are typically provided on standard computer outputs; for example, to display the estimated covariance in PROC REG use the `COVB` option in the `MODEL` statement. With these values, the Fieller confidence limits for x_{opt} can be calculated as

$$\begin{aligned} \hat{x}_{\text{opt}} + \frac{\gamma}{1 - \gamma} \left\{ (\hat{x}_{\text{opt}} - \bar{x}) + \frac{\hat{\sigma}_{12}}{2se^2[b_2]} \right\} \\ \pm \frac{t_{\alpha/2}(v)}{2(1 - \gamma)|b_2|} \left\{ se^2[b_1] + 4(\hat{x}_{\text{opt}} - \bar{x})[\hat{\sigma}_{12} + se^2[b_2](\hat{x}_{\text{opt}} - \bar{x})] \right. \\ \left. - \gamma \left(se^2[b_1] - \frac{\hat{\sigma}_{12}^2}{se^2[b_2]} \right) \right\}^{1/2}, \end{aligned} \quad (1.14)$$

where $\gamma = [t_{\alpha/2}(v)]^2 se^2(b_2)/b_2^2$ measures departure from symmetry in the distribution of \hat{x}_{opt} ($\gamma \rightarrow 0$ indicates greater symmetry), and $v = df_e$ from the ANOVA table for the quadratic regression fit. (See Exercise 1.16 for a derivation of Fieller's result.) In some applications of Fieller's theorem appeal is made to large-sample approximations, and in those cases we often take $v = \infty$, that is, we use the $N(0,1)$ limiting distribution for $t(v)$. In this case, set $t_{\alpha/2}(\infty) = z_{\alpha/2}$.

One might also consider use of the delta method (§A.6) to build approximate $1 - \alpha$ limits for x_{opt} . That is, from §A.5.1 we know that a $1 - \alpha$ Wald confidence interval is $\hat{x}_{\text{opt}} \pm z_{\alpha/2} se[\hat{x}_{\text{opt}}]$, and to complete the interval all we need is the standard error $se[\hat{x}_{\text{opt}}]$. This is the square root of the variance, $\text{Var}[\hat{x}_{\text{opt}}]$, which since \bar{x} is assumed fixed simplifies to $\text{Var}[\hat{x}_{\text{opt}}] = \text{Var}[\bar{x} - b_1/2b_2] = \text{Var}[-b_1/2b_2] = \frac{1}{4} \text{Var}[b_1/b_2]$. We use the delta method to approximate the variance of the ratio (see Example A.3) and use this in $\hat{x}_{\text{opt}} \pm z_{\alpha/2} \sqrt{\text{Var}[\hat{x}_{\text{opt}}]}$. This approximation will often give comparable results to that provided by Fieller's theorem, but can also suffer from selected instabilities in its true confidence level, relative to the Fieller result (Buonaccorsi, 2002). In general, we recommend use of Fieller's theorem (1.14) over the delta method in situations such as this.

Example 1.7 (Rice yield, cont'd) Returning to the rice yield data in Table 1.5, suppose there is interest in determining the temperature at which the minimum yield is attained, and, in particular, in providing a confidence interval for this value. (Once the daily minimum temperature enters this range, agricultural managers could be

alerted that poor yield conditions are imminent.) From the SAS analysis in Fig. 1.13, we find $\hat{x}_{\text{opt}} = \bar{x} - b_1/2b_2 = 23.197 + 4.362 = 27.559^\circ\text{C}$.

To calculate 95% Fieller limits from (1.14), we require $se[b_1] = 0.0337$, $se[b_2] = 0.0094$, and $t_{0.025}(37) = 2.0262$. For the estimated covariance, $\hat{\sigma}_{12}$, return to PROC REG and update the model statement to

```
model yield=x x2 / covb;
```

in Fig. 1.12. This produces the output

Covariance of Estimates			
COVB	INTERCEP	X	X2
INTERCEPT	0.0141959035	-0.000528327	-0.000700488
X	-0.000528327	0.0011370764	0.0000665211
X2	-0.000700488	0.0000665211	0.0000881977

from which we read $\hat{\sigma}_{12} = 6.65211 \times 10^{-5}$. Appeal to (1.14) then yields the 95% limits $25.903 < x_{\text{opt}} < 31.856^\circ\text{C}$. Over this range of minimum temperatures, we are 95% confident that the poorest conditions for average rice production will occur. 🌐

There are many other aspects to proper use and application of polynomial regression models, including proper regression diagnostics, variable selection, optimal design, higher-order polynomial models, and response-surface models. The exercises below study some of these issues to a limited extent. Interested readers can find more on these topics in regression textbooks, including Neter *et al.* (1996, §7.7) and Christensen (1996, §§7.11–12).

Exercises

- 1.1. Dalgård *et al.* (1994) reported on a study of mercury (Hg) toxicity in pregnant Faroe islanders, where potentially high mercury body burdens occur through the islanders' large consumption of pilot whale meat. Interest included the relation between x =daily Hg ingestion (calculated in μg) and Y =Hg concentration in the woman's umbilical cord blood (in $\mu\text{mol/l}$, recorded immediately after giving birth). A sample of $n=12$ women produced the following data:

x = Hg ingestion	1.4	49	90	96	108	125
Y = cord blood Hg	0.007	0.23	0.43	0.46	0.52	0.60
x = Hg ingestion	146	153	233	324	354	671
Y = cord blood Hg	0.70	0.73	1.12	1.56	1.70	3.22

- (a) Plot the data. Does a simple linear model seem reasonable for representing the relationship between x and Y ?

32 Linear regression

- (b) Assume the simple linear model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, and find the LS/ML estimators b_0 and b_1 . Also find the associated standard errors $se[b_0]$ and $se[b_1]$. Use these to find a 95% confidence interval for β_1 .
- (c) A natural question to ask is whether increasing Hg intake increases Hg cord blood concentrations. This translates into a test of $H_0: \beta_1 = 0$ vs. $H_a: \beta_1 > 0$. Test these hypotheses at significance level $\alpha = 0.01$. What do you conclude?
- (d) Calculate the predicted values and residuals from this LS/ML fit and construct a residual plot. What does the plot suggest?
- 1.2. Todd *et al.* (2001) report data on lead (Pb) concentrations in human bone tissue. For example, in human tibias they note that core Pb concentrations (in $\mu\text{g/g}$) can often be predicted from surface concentrations.
- (a) At the proximal end of the tibia (closest to the knee), data are available from $n = 9$ independent cadavers:

$x = \text{surface Pb } (\mu\text{g/g})$	22.5	5.2	11.8	16.5	12.5
$Y = \text{core Pb } (\mu\text{g/g})$	12.0	3.8	6.7	9.0	8.1
$x = \text{surface Pb } (\mu\text{g/g})$	21.5	22.4	13.6	5.8	
$Y = \text{core Pb } (\mu\text{g/g})$	13.9	14.2	7.8	3.1	

Analyze these data by fitting a simple linear regression model. Test if there is a significant linear relationship between surface and core Pb concentrations at significance level $\alpha = 0.05$. What do you conclude?

- (b) Calculate the predicted values and residuals from the fit in part (a) and construct a residual plot. What does the plot suggest?
- (c) Predict the mean value of core Pb concentration at the new surface Pb level $x = 25 \mu\text{g/g}$, that is, find $\hat{\mu}(25)$. Also find the standard error of your estimate and use it to calculate a 95% confidence interval on the true value of $\mu(25)$.
- (d) At the distal end of the tibia (farthest from the knee), $n = 9$ similar observations are taken:

$x = \text{surface Pb } (\mu\text{g/g})$	25.8	6.4	16.7	20.2	15.2
$Y = \text{core Pb } (\mu\text{g/g})$	14.4	5.4	8.8	8.2	11.0
$x = \text{surface Pb } (\mu\text{g/g})$	24.2	25.6	11.0	10.0	
$Y = \text{core Pb } (\mu\text{g/g})$	15.7	16.3	7.6	4.7	

Analyze these data by fitting a simple linear regression model. Test if there is a significant linear relationship between surface and core Pb concentrations at significance level $\alpha = 0.05$. What do you conclude?

- (e) Calculate the predicted values and residuals from the fit in part (d) and construct a residual plot. What does the plot suggest?
- 1.3. In a study of agricultural nutrient management and environmental quality, data were collected on the response of a feed crop to increasing nitrogen (N) fertilizer. (The goal was to use the minimum nitrogen possible to achieve target yields, consistent with good environmental practice.) The response variable was Y = yield (bu/acre), vs. x = applied nitrogen (lb/acre). The data are:

x = Nitrogen (lb/acre)	20	60	120	180
Y = yield (bu/acre)	22	17	39	61
	15	28	46	36
	43	34	42	71

- (a) Calculate the sample mean and sample variance at each level of x_i . Is the variance stable?
- (b) If your calculations in part (a) indicate that there is variance heterogeneity, fit a weighted LS regression, using x_i as the single predictor variable. (Use for your weights the reciprocals of the observed sample variances at each level of x_i . That is, calculate the sample variances, S_i^2 , for the three observations at each x_i , and use $w_i = 1/S_i^2$ for the weights. Notice that for all observations within a triplet, the weights will be equal.) Test the hypothesis that mean yield does not vary across nitrogen applications. Set $\alpha = 0.01$.
- 1.4. Return to the multiple linear regression analysis for the soil pH data from Example 1.3.
- (a) Calculate the VIFs for each of the predictor variables. Is multicollinearity a concern here?
- (b) Acquire the residuals and the predicted values from the full three-predictor fit, and construct a residual plot. Also construct individual plots of the residuals vs. each of the predictor variables. (In PROC REG, this can be accomplished by invoking the command

```
plot r.*(p. x1clay x2org x3carb);
```

after the model command.) Do you see anything unusual in the plots?

- 1.5. As part of a study on water resource management, Lu *et al.* (1999) present data on seven-day streamflow rates (in ft³/s) over a series of streams in west-central Florida. To model two-year recurrence flow, four possible hydrological predictor variables were collected: x_{i1} = drainage area (m²), x_{i2} = basin slope (ft/m), x_{i3} = soil-infiltration index (inches), and x_{i4} = rainfall index (inches). The response variable, Y_i , was taken to be the logarithm of the two-year recurrence flow rates. After removing any zero/censored flow rates, the $n = 45$ data values were:

Y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}	Y_i	x_{i1}	x_{i2}	x_{i3}	x_{i4}
13	180	0.93	2.66	5.27	1.7	24	3.66	2.05	10.06
170	640	1.09	4.14	3.15	1.2	65.3	5.83	2.05	10.06
3.1	41.2	14.96	12.2	6.19	4.0	80.0	4.93	2.05	8.56
13	469	2.00	4.12	8.54	9.1	149	5.03	2.55	9.56
1.1	83.6	2.04	3.08	5.34	40	135	4.96	2.70	8.98
4.8	89.2	1.78	2.28	5.34	12	107	3.60	2.05	8.98
4.7	75.0	3.57	5.38	5.34	43	335	3.45	2.14	4.94
0.9	23.0	4.17	5.38	5.34	20.0	28	2.07	5.38	4.94
16	58.9	1.66	4.88	5.34	0.1	9.5	6.70	2.05	8.98
8.9	60.9	3.12	4.21	8.04	4.8	110	3.52	2.72	8.98
9.1	38.8	4.9	4.38	8.04	70	220	3.87	2.77	4.94
36	379	1.4	2.88	8.04	0.8	375	2.41	2.73	4.94
8.1	109	3.81	5.14	8.04	1.4	35	4.12	2.23	4.94
7.7	160	1.26	5.20	8.98	2.4	72.5	3.54	2.42	7.23
26.0	390	1.25	4.05	8.98	0.9	182	1.97	4.11	7.23
4.8	121	3.88	2.05	8.98	30.0	570	1.35	2.70	11.42
100	826	1.38	4.05	8.04	2.3	145	1.75	2.71	11.42
3.2	330	1.68	2.08	6.93	73.0	810	1.24	5.37	11.42
110	367	1.3	3.04	6.93	13.0	40	2.10	5.38	9.27
1.7	132	4.06	2.03	6.93	330	1825	0.86	3.89	9.27
1.1	218	2.79	2.05	6.93	15.0	300	1.58	2.00	3.33
3	223	2.54	2.00	6.93	0.1	26	5.9	2.00	3.33
0.6	87.5	2.00	2.00	10.06					

- (a) Calculate the VIFs for each predictor variable. Is multicollinearity a concern?
- (b) Fit a multiple linear regression model to these data, using all four predictor variables. Perform a single (2 df) test to see if x_{i2} and x_{i4} are required in the model. Set $\alpha = 0.01$.
- (c) Lu *et al.* studied models with polynomial and cross-product terms to see if these could provide a better fit for the data. To explore this, define the new predictor variables $x_{i5} = (x_{i3} - \bar{x}_3)^2$ and $x_{i6} = (x_{i3} - \bar{x}_3)^3$, where $\bar{x}_3 = \sum_{i=1}^{45} x_{i3}/45$. Fit the four predictors x_{i1} , $x_{i3} - \bar{x}_3$, x_{i5} , and x_{i6} to the data, and calculate the new MSE. Is this smaller than the MSE from the fit in part (b)? (If so, it would indicate a more precise fit, since both models employ the same number of predictor variables.)
- (d) To your fit in part (c) also add the following cross-product terms: $x_{i7} = x_{i1}(x_{i3} - \bar{x}_3)$ and $x_{i8} = x_{i1}(x_{i3} - \bar{x}_3)^2$. Test to see if either or both these terms contribute significantly to the model fit. Operate at $\alpha = 0.01$ throughout.
- 1.6. Under the one-factor ANOVA model (1.8), the null hypothesis of no effect due to the qualitative factor is $H_0: \alpha_1 = \dots = \alpha_a$. Show that under either the zero-sum

constraint or the corner-point constraint, this is equivalent to $H_0: \alpha_1 = \dots = \alpha_a = 0$.

- 1.7. As part of a study on control of hazardous waste, Stelling and Sjerps (1999) report data on total polycyclic aromatic hydrocarbons (PAHs) observed in lots of fractionated demolition waste (in mg/kg). The original data are as follows:

Lot 1:	44.5, 86, 115, 120, 127, 136, 142, 147, 154, 240
Lot 2:	74.9, 85.4, 85.9, 97.5, 102, 130, 139, 151, 169, 245
Lot 3:	3.4, 11.2, 11.6, 21.9, 24.0, 29.8, 29.9, 29.9, 41.3, 51.3
Lot 4:	76.3, 97.7, 117, 120, 120, 121, 127, 132, 145, 249
Lot 5:	55, 74.2, 86.2, 114, 122, 125, 130, 137, 226, 258
Lot 6:	53.3, 56.3, 73.1, 85.2, 90.7, 91.8, 96.5, 99.4, 102, 126

Concentration data can often depart from normality. To adjust for this, apply a log-transformation and analyze the transformed data to assess if there is any difference in log-PAH concentrations across lots. Set your significance level to 10%. Include a residual analysis to check for any model irregularities.

- 1.8. Return to the macroinvertebrate ratio data in Table 1.3 and complete the two-factor analysis as follows.
- Since a significant interaction was observed, study the season effect by performing three separate one-factor ANOVAs, stratified over levels of stream. Is there a significant effect due to season at any of the three streams? Set your significance level to 1%.
 - Since, in effect, you performed three different tests in part (a), adjust your inferences via a Bonferroni correction (see §A.5.4). Do the results change at familywise significance level $\alpha = 0.01$?
 - Mimic the analysis in part (a) and study the stream effect by performing four separate one-factor ANOVAs, stratified over levels of season. Is there a significant effect due to streams in any of the four seasons? Set your significance level to 1%.
 - Since you performed four different tests in part (c), adjust your inferences via a Bonferroni correction (see §A.5.4). Do the results change at familywise significance level $\alpha = 0.01$?
- 1.9. Return to the faba bean genotoxicity data from Example 1.5.
- Test to see if the standard ANCOVA assumption of common slopes is valid for the logit-transformed data; that is, under (1.12) test $H_0: \beta_1 = \dots = \beta_a$. Operate at the 5% significance level.
 - Reanalyze these data using a one-factor ANOVA model, ignoring the covariate. Comment on whether it was beneficial to adjust for potential covariate differences in this analysis.

- 1.10. Malling and Delongchamp (2001) reported data on organ weight changes in laboratory mice exposed to the chemical mutagen ethylnitrosourea (ENU). Of interest was whether ENU exposure affects organ weights in these mice, compared to a group of control animals. The study protocol allowed for different-aged mice to be used, hence some adjustment for murine age should be included in any analysis of the ENU effect. This is essentially an ANCOVA problem: we view age as the quantitative covariate, and exposure condition (ENU or control) as the qualitative variable of interest. For spleen weights, the data are:

Controls: Age (d)	434	434	427	422	408
Spleen weight (mg)	71.4	85.9	72.9	75.3	77.2
ENU: Age (d)	392	435	434	427	427
Spleen weight (mg)	187.6	123.0	137.2	125.3	111.2

- (a) Plot the spleen weights as a function of age, overlaying the control and exposure groups on the same graph. Does the assumption of equal slopes – as used in (1.10) – appear valid for these data?
- (b) Based on your results from part (a), analyze these data using an appropriate ANCOVA model. Assess if there is an effect on spleen weights due to ENU exposure, after adjusting for any age effects. What do you conclude at the 5% significance level?
- (c) Perform a residual analysis to check for any model irregularities.
- 1.11. Return to the rice yield data in Example 1.6.

- (a) Seshu and Cady (1984) included a number of factors beyond minimum temperature in their study of environmental effects on rice yields. For example, they also measured x_{i3} = solar radiation (mWh/cm²) at each location. The full data are:

Y_i	x_{i1}	x_{i3}	Y_i	x_{i1}	x_{i3}
2.3	29.2	421	3.0	23.4	483
3.1	28.1	362	4.4	23.4	537
2.8	27.2	404	3.1	23.2	418
2.4	26.4	340	2.6	23.1	451
3.6	26.3	431	4.8	23.1	618
2.3	26.2	427	3.2	23.0	419
3.8	26.2	595	3.3	22.9	472
3.1	26.0	365	3.1	22.5	490
2.4	25.9	372	3.4	22.5	484
2.1	25.7	435	3.2	22.4	423
3.5	24.5	457	4.2	21.7	428
3.8	24.4	609	4.5	21.2	373

3.1	24.0	469	4.7	20.0	438
2.9	23.9	459	5.0	19.2	326
3.2	23.9	481	6.2	19.0	521
3.0	23.7	379	6.0	19.0	609
4.5	23.7	637	6.1	18.8	528
3.5	23.7	529	7.3	18.0	566
3.3	23.6	455	6.6	18.0	315
3.7	23.5	512	6.2	17.4	370

Study the impact of this additional predictor by fitting it along with x_{i1} = minimum temperature and $x_{i2} = x_{i1}^2$ (from Example 1.6). Also include a term for possible quadratic effects of x_{i3} = solar radiation; that is, include $x_{i4} = x_{i3}^2$ in the fit. (Remember to center each predictor variable about its mean.) Test to determine if the quadratic solar radiation term is significant at $\alpha = 0.01$. If the quadratic term is insignificant, test to determine if solar radiation contributes significantly to the model, after allowing for a quadratic effect due to minimum temperature. Here also, operate at $\alpha = 0.01$.

- (b) To your fit in part (a), add an additional term for interaction between temperature and radiation, $x_{i5} = x_{i1}x_{i3}$. Test to determine if the interaction is significant at $\alpha = 0.01$.
- (c) If either the solar radiation terms or the radiation \times temperature interaction (or all) were significant in your fit from parts (a) and (b), acquire the predicted values, \hat{Y}_i , and the residuals, $Y_i - \hat{Y}_i$, from the fit. Graph (i) a residual plot to assess the quality of the fit, and (ii) \hat{Y}_i vs. x_{i1} and x_{i3} in three dimensions. The latter plot will visualize the *response surface* (Khuri and Cornell, 1996) created by the joint effects of the two predictor variables. What do you conclude from the two plots?

- 1.12. Bates and Watts (1988, §A1.1) present data on the concentration of polychlorinated biphenyl (PCB, in parts per million) residues in trout from a northeastern US lake, as a function of age of the fish (in years). (PCBs are mixtures of organochlorines and other chemicals used in a wide variety of industrial and commercial applications prior to 1977. Their ubiquitous use led to ecosystem contamination in many developed nations. Study of their accumulation throughout the food web is an important concern.) The data are as follows:

Age	1	1	1	1	2	2	2
PCB conc.	0.6	1.6	0.5	1.2	2.0	1.3	2.5
Age	3	3	3	4	4	4	5
PCB conc.	2.2	2.4	1.2	3.5	4.1	5.1	5.7
Age	6	6	6	7	7	7	8
PCB conc.	3.4	9.7	8.6	4.0	5.5	10.5	17.5
Age	8	8	9	11	12	12	12
PCB conc.	13.4	4.5	30.4	12.4	13.4	26.2	7.4

- (a) Concentration data often depart from normality. To adjust for this, apply a log transformation to the PCB concentrations. Plot the data. What pattern do you see?
- (b) Bates and Watts (1988) suggest that a simple linear fit to these data can be improved by considering a transformation of the age variable. They suggest $x = (\text{age})^{1/3}$. Make this transformation and plot the transformed data. Comment on the result.
- (c) Fit a simple linear model to the transformed data; that is, fit $\log(\text{concentration}) = \beta_0 + \beta_1(\text{age})^{1/3} + \varepsilon$. Test if there is a significant effect due to (transformed) age at $\alpha = 0.01$.
- (d) Calculate residuals from your fit in part (c) and plot them against age. Comment on the pattern. Based on your results, suggest a revised model and fit it to the data. Identify whether this adds significantly to the model fit. Continue to operate at $\alpha = 0.01$.
- 1.13. In a study of the long-term effects of radiation exposure in humans, data were taken on the frequencies of mutations (as average chromosome aberrations in peripheral blood cells) seen in survivors of atomic radiation exposure in Hiroshima, Japan. Subjects were also assessed for their estimated radiation exposures (in rems). The data are as follows:

Radiation exposure	1.980	2.540	4.670	5.000	6.890	8.120
Mutation frequency	40.0057	41.2645	45.9816	45.9828	52.2148	54.7815
Radiation exposure	9.995	11.578	14.400	15.001	16.870	19.560
Mutation frequency	59.6586	63.0588	70.0128	71.3794	75.5131	80.6950

Since mutation frequencies often exhibit a right skew (to higher values), take as your response variable the natural logarithms: $Y_i = \log(\text{mutation frequency})$. Assume that these transformed variates are normally distributed with constant variance.

- (a) For $x_i =$ estimated radiation exposure, plot Y_i vs. x_i . What does the plot show?
- (b) Fit a simple linear model and assess if radiation exposure has a significant effect on log-mutant frequencies. Operate at $\alpha = 0.01$.
- (c) Find the predicted values and from these calculate residuals from your fit in part (b). Construct a residual plot and comment on the fit.
- (d) Based on your results in part (c), suggest an additional term for the model. Fit this term and identify whether it adds significantly to the model. Continue to operate at $\alpha = 0.01$.
- 1.14. Ribakov *et al.* (2001) report on engineering efforts to study building response to earthquake stress after reinforcement with active viscous damping. Observations were taken on $Y_i =$ peak displacements (cm) as a function of height

above ground level (represented here as $x_i =$ story number). Data from the 1989 Loma-Prieta, CA, earthquake were given as follows:

Story	1	2	3	4	5	6	7
Displacement	0.2	0.41	0.55	0.71	0.85	0.94	0.95

- Plot Y vs. x . What does the plot show?
 - Fit a quadratic regression model and assess if there is an effect of height on displacement. Use story number as a surrogate for height. (Remember to center the predictor variable about its mean.) Operate at $\alpha = 0.01$.
 - The effect of centering the predictor variable is especially striking here. What is the sample correlation between $x_i - \bar{x}$ and $(x_i - \bar{x})^2$? Examine the estimated covariance between b_1 and b_2 : fit the model with and without centering and study $\text{Cov}[b_1, b_2]$.
 - Estimate the story height at which peak displacement occurs. Include a 99% confidence interval (using Fieller's theorem) for this point. What cautions might you give regarding these estimates?
- 1.15. Similar to the study in Exercise 1.14, Ribakov *et al.* (2001) also report data on $Y_i =$ peak displacements (cm) as a function of height above ground level (as $x_i =$ story number) from the 1995 Eilat, Israel, earthquake, this time for buildings equipped with electrorheological dampers. The data are:

Story	1	2	3	4	5	6	7
Displacement	0.37	0.73	0.99	1.27	1.47	1.61	1.68

- Plot Y vs. x . What does the plot show?
 - Fit a quadratic regression model and assess if there is an effect of height (as story number) on displacement. Remember to center the predictor variable about its mean. Operate at $\alpha = 0.05$.
 - Estimate the story height at which peak displacement occurs. Include a 95% confidence interval (using Fieller's theorem) for this point. What cautions might you give regarding these estimates?
- 1.16. Derive the general version of Fieller's theorem (Buonaccorsi, 2002) via the following steps:
- Assume we have two parameters, θ_1 and θ_2 , for which we have unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$; that is, $E[\hat{\theta}_j] = \theta_j$, $j = 1, 2$. We desire confidence limits for the ratio $\varphi = \theta_1/\theta_2$. Start with $D = \hat{\theta}_1 - \varphi\hat{\theta}_2$. Show that $E[D] = 0$.
 - Suppose the standard error of $\hat{\theta}_j$ in part (a) is $se[\hat{\theta}_j]$, $j = 1, 2$. Also let $\hat{\sigma}_{12} = \text{Cov}[\hat{\theta}_1, \hat{\theta}_2]$. Show that the standard error of D , $se[D]$, is the square root of $se^2[\hat{\theta}_1] + \varphi^2 se^2[\hat{\theta}_2] - 2\varphi\hat{\sigma}_{12}$.

- (c) Assume that when standardized, D is approximately normal, $(D - E[D])/se[D] \sim N(0, 1)$, in large samples. (The dot notation above the \sim indicates that the distributional relationship is only approximate. The approximation improves as $n \rightarrow \infty$.) Then a set of confidence limits could be derived from the relationship $P\{|D - E[D]|/se[D] \leq z_{\alpha/2}\} = 1 - \alpha$. Show that this equivalent to

$$P\left\{(D - E[D])^2/se^2[D] \leq z_{\alpha/2}^2\right\} = 1 - \alpha.$$

- (d) Express the event from the probability statement in part (c) as $\{D - E[D]\}^2 \leq z_{\alpha/2}^2 se^2[D]$. (Why does the inequality's direction remain the same?) Substitute the appropriate quantities from parts (a) and (b) into this expression for each quantity involving D . Show that this leads to a quadratic inequality in ϕ .
- (e) Take the inequality in part (d) and operate with the points at equality. Solve for the two roots of this quadratic equation. For simplicity, let $\gamma = z_{\alpha/2}^2 se^2[\hat{\theta}_2]/\hat{\theta}_2^2$.
- (f) Manipulate the two roots to produce

$$\hat{\phi} + \frac{\gamma}{1 - \gamma} \left\{ \hat{\phi} - \frac{\hat{\sigma}_{12}}{se^2[\hat{\theta}_2]} \right\} \pm \frac{z_{\alpha/2}}{(1 - \gamma)|\hat{\theta}_2|} \left\{ se^2[\hat{\theta}_1] + \hat{\phi} \left(se^2[\hat{\theta}_2] \hat{\phi} - 2\hat{\sigma}_{12} \right) - \gamma \left(se^2[\hat{\theta}_1] - \frac{\hat{\sigma}_{12}^2}{se^2[\hat{\theta}_2]} \right) \right\}^{1/2},$$

where $\hat{\phi} = \hat{\theta}_1/\hat{\theta}_2$. Note that if the large-sample approximation in part (c) is in fact exact in small samples, then we would write $(D - E[D])/se[D] \sim t(v)$, where v are the df associated with estimating the standard error of D . If so, we would replace $z_{\alpha/2}^2$ with $t_{\alpha/2}^2(v)$.