1.2.2. The data used to draw Figure 1.12 are in the file `Mitchell.txt`. Redraw the graph, but this time make the length of the horizontal axis at least four times the length of the vertical axis. Repeat Problem 1.2.1.

1.3. United Nations The data in the file `UN1.txt` contains *PPgdp*, the 2001 gross national product per person in US dollars, and *Fertility*, the birth rate per 1000 females in the population in the year 2000. The data are for 193 localities, mostly UN member countries, but also other areas such as Hong Kong that are not independent countries; the third variable on the file called *Locality* gives the name of the locality. The data were collected from http://unstats.un.org/unsd/demographic. In this problem, we will study the conditional distribution of *Fertility* given *PPgdp*.

1.3.1. Identify the predictor and the response.

1.3.2. Draw the scatterplot of *Fertility* on the vertical axis versus *PPgdp* on the horizontal axis and summarize the information in this graph. Does a straight-line mean function seem to be a plausible for a summary of this graph?

1.3.3. Draw the scatterplot of log(*Fertility*) versus log(*PPgdp*), using logs to the base two. Does the simple linear regression model seem plausible for a summary of this graph?

1.4. Old Faithful The data in the data file `oldfaith.txt` gives information about eruptions of Old Faithful Geyser during October 1980. Variables are the *Duration* in seconds of the current eruption, and the *Interval*, the time in minutes to the next eruption. The data were collected by volunteers and were provided by R. Hutchinson. Apart from missing data for the period from midnight to 6 AM, this is a complete record of eruptions for that month.

Old Faithful Geyser is an important tourist attraction, with up to several thousand people watching it erupt on pleasant summer days. The park service uses data like these to obtain a prediction equation for the time to the next eruption.

Draw the relevant summary graph for predicting interval from duration, and summarize your results.

1.5. Water run-off in the Sierras Can Southern California's water supply in future years be predicted from past data? One factor affecting water availability is stream run-off. If run-off could be predicted, engineers, planners and policy makers could do their jobs more efficiently. The data in the file `water.txt` contains 43 years' worth of precipitation measurements taken at six sites in the Sierra Nevada mountains (labelled *APMAM*, *APSAB*, *APSLAKE*, *OPBPC*, *OPRC*, and *OPSLAKE*), and stream run-off volume at a site near Bishop, California, labelled *BSAAM*. The data are from the UCLA Statistics WWW server.

Draw the scatterplot matrix for these data and summarize the information available from these plots.

CHAPTER 2

Simple Linear Regression

The *simple linear regression model* consists of the mean function and the variance function

$$\begin{aligned} \mathrm{E}(Y|X=x) &= \beta_0 + \beta_1 x \\ \mathrm{Var}(Y|X=x) &= \sigma^2 \end{aligned} \tag{2.1}$$

The parameters in the mean function are the intercept β_0, which is the value of $\mathrm{E}(Y|X=x)$ when x equals zero, and the slope β_1, which is the rate of change in $\mathrm{E}(Y|X=x)$ for a unit change in X; see Figure 2.1. By varying the parameters, we can get all possible straight lines. In most applications, parameters are unknown and must be estimated using data. The variance function in (2.1) is assumed to be constant, with a positive value σ^2 that is usually unknown.

Because the variance $\sigma^2 > 0$, the observed value of the ith response y_i will typically not equal its expected value $\mathrm{E}(Y|X=x_i)$. To account for this difference between the observed data and the expected value, statisticians have invented a quantity called a *statistical error*, or e_i, for case i defined implicitly by the equation $y_i = \mathrm{E}(Y|X=x_i) + e_i$ or explicitly by $e_i = y_i - \mathrm{E}(Y|X=x_i)$. The errors e_i depend on unknown parameters in the mean function and so are not observable quantities. They are random variables and correspond to the *vertical distance between the point y_i and the mean function* $\mathrm{E}(Y|X=x_i)$. In the heights data, page 2, the errors are the differences between the heights of particular daughters and the average height of all daughters with mothers of a given fixed height.

If the assumed mean function is incorrect, then the difference between the observed data and the incorrect mean function will have a non random component, as illustrated in Figure 2.2.

We make two important assumptions concerning the errors. First, we assume that $\mathrm{E}(e_i|x_i) = 0$, so if we could draw a scatterplot of the e_i versus the x_i, we would have a null scatterplot, with no patterns. The second assumption is that the errors are all *independent*, meaning that the value of the error for one case gives

Applied Linear Regression, Third Edition, by Sanford Weisberg
ISBN 0-471-66379-4

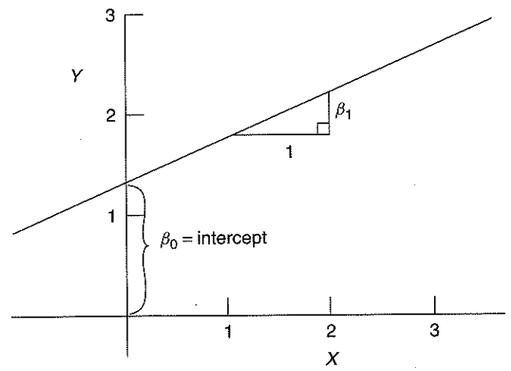

FIG. 2.1 Equation of a straight line $\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x$.

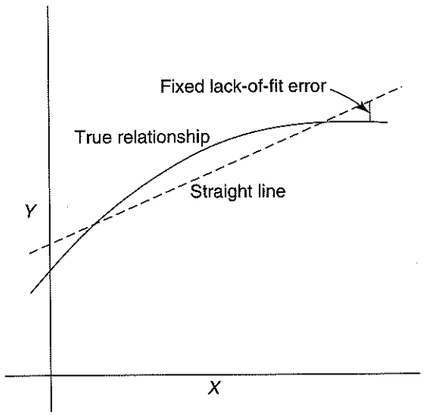

FIG. 2.2 Approximating a curved mean function by straight line cases adds a fixed component to the errors.

no information about the value of the error for another case. This is likely to be true in the examples in Chapter 1, although this assumption will not hold in all problems.

Errors are often assumed to be normally distributed, but normality is much stronger than we need. In this book, the normality assumption is used primarily to obtain tests and confidence statements with small samples. If the errors are thought to follow some different distribution, such as the Poisson or the Binomial,

other methods besides OLS may be more appropriate; we return to this topic in Chapter 12.

2.1 ORDINARY LEAST SQUARES ESTIMATION

Many methods have been suggested for obtaining estimates of parameters in a model. The method discussed here is called *ordinary least squares*, or OLS, in which parameter estimates are chosen to minimize a quantity called the *residual sum of squares*. A formal development of the least squares estimates is given in Appendix A.3.

Parameters are unknown quantities that characterize a model. *Estimates of parameters* are computable functions of data and are therefore *statistics*. To keep this distinction clear, parameters are denoted by Greek letters like α, β, γ and σ, and estimates of parameters are denoted by putting a "hat" over the corresponding Greek letter. For example, $\hat{\beta}_1$, read "beta one hat," is the estimator of β_1, and $\hat{\sigma}^2$ is the estimator of σ^2. The *fitted value* for case i is given by $\widehat{\mathrm{E}}(Y|X = x_i)$, for which we use the shorthand notation $\hat{y}_i$,

$$\hat{y}_i = \widehat{\mathrm{E}}(Y|X = x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{2.2}$$

Although the e_i are not parameters in the usual sense, we shall use the same hat notation to specify the residuals: the residual for the ith case, denoted $\hat{e}_i$, is given by the equation

$$\hat{e}_i = y_i - \widehat{\mathrm{E}}(Y|X = x_i) = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \quad i = 1, \ldots, n \tag{2.3}$$

which should be compared with the equation for the statistical errors,

$$e_i = y_i - (\beta_0 + \beta_1 x_i) \quad i = 1, \ldots, n$$

All least squares computations for simple regression depend only on averages, sums of squares and sums of cross-products. Definitions of the quantities used are given in Table 2.1. Sums of squares and cross-products have been centered by subtracting the average from each of the values before squaring or taking cross-products. Appropriate alternative formulas for computing the corrected sums of squares and cross products from uncorrected sums of squares and cross-products that are often given in elementary textbooks are useful for mathematical proofs, but they can be highly inaccurate when used on a computer and should be avoided.

Table 2.1 also lists definitions for the usual univariate and bivariate summary statistics, the sample averages $(\overline{x}, \overline{y})$, sample variances $(\mathrm{SD}_x^2, \mathrm{SD}_y^2)$, and estimated covariance and correlation (s_{xy}, r_{xy}). The "hat" rule described earlier would suggest that different symbols should be used for these quantities; for example, $\hat{\rho}_{xy}$ might be more appropriate for the sample correlation if the population correlation is ρ_{xy}.

The OLS estimators are those values β_0 and β_1 that minimize the function[1]

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{2.4}$$

When evaluated at $(\hat{\beta}_0, \hat{\beta}_1)$, we call the quantity $RSS(\hat{\beta}_0, \hat{\beta}_1)$ the *residual sum of squares*, or just *RSS*.

The least squares estimates can be derived in many ways, one of which is outlined in Appendix A.3. They are given by the expressions

$$\begin{aligned} \hat{\beta}_1 &= \frac{SXY}{SXX} = r_{xy}\frac{\mathrm{SD}_y}{\mathrm{SD}_x} = r_{xy}\left(\frac{SYY}{SXX}\right)^{1/2} \\ \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} \end{aligned} \tag{2.5}$$

The several forms for $\hat{\beta}_1$ are all equivalent.

We emphasize again that OLS produces *estimates* of parameters but not the actual values of the parameters. The data in Figure 2.3 were created by setting the x_i to be random sample of 20 numbers from a N(2, 1.5) distribution and then computing $y_i = 0.7 + 0.8x_i + e_i$, where the errors were N(0, 1) random numbers. For this graph, the true values of $\beta_0 = 0.7$ and $\beta_1 = 0.8$ are known. The graph of the true mean function is shown in Figure 2.3 as a dashed line, and it seems to match the data poorly compared to OLS, given by the solid line. Since OLS minimizes (2.4), it will always fit at least as well as, and generally better than, the true mean function.

Using Forbes' data, we will write $\overline{x}$ to be the sample mean of *Temp* and $\overline{y}$ to be the sample mean of *Lpres*. The quantities needed for computing the least squares estimators are

$$\begin{aligned} \overline{x} &= 202.95294 \quad SXX = 530.78235 \quad SXY = 475.31224 \\ \overline{y} &= 139.60529 \quad SYY = 427.79402 \end{aligned} \tag{2.6}$$

The quantity *SYY*, although not yet needed, is given for completeness. In the rare instances that regression calculations are not done using statistical software or a statistical calculator, intermediate calculations such as these should be done as accurately as possible, and rounding should be done only to final results. Using (2.6), we find

$$\begin{aligned} \hat{\beta}_1 &= \frac{SXY}{SXX} = 0.895 \\ \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} = -42.138 \end{aligned}$$

[1]We abuse notation by using the symbol for a fixed though unknown quantity like β_j as if it were a variable argument. Thus, for example, $RSS(\beta_0, \beta_1)$ is a function of two variables to be evaluated as its arguments β_0 and β_1 vary. The same abuse of notation is used in the discussion of confidence intervals.

The estimated line, given by either of the equations

$$\begin{aligned} \widehat{\mathrm{E}}(Lpres|Temp) &= -42.138 + 0.895Temp \\ &= 139.606 + 0.895(Temp - 202.953) \end{aligned}$$

was drawn in Figure 1.4a. The fit of this line to the data is excellent.

2.3 ESTIMATING σ^2

Since the variance σ^2 is essentially the average squared size of the e_i^2, we should expect that its estimator $\hat{\sigma}^2$ is obtained by averaging the squared residuals. Under the assumption that the errors are uncorrelated random variables with zero means and common variance σ^2, an unbiased estimate of σ^2 is obtained by dividing $RSS = \sum \hat{e}_i^2$ by its *degrees of freedom* (df), where residual df = number of cases minus the number of parameters in the mean function. For simple regression, residual df $= n - 2$, so the estimate of σ^2 is given by

$$\hat{\sigma}^2 = \frac{RSS}{n-2} \tag{2.7}$$

This quantity is called the *residual mean square*. In general, any sum of squares divided by its df is called a mean square. The residual sum of squares can be computed by squaring the residuals and adding them up. It can also be computed from the formula (Problem 2.9)

$$RSS = SYY - \frac{SXY^2}{SXX} = SYY - \hat{\beta}_1^2 SXX \tag{2.8}$$

Using the summaries for Forbes' data given at (2.6), we find

$$\begin{aligned} RSS &= 427.79402 - \frac{475.31224^2}{530.78235} \\ &= 2.15493 \end{aligned} \tag{2.9}$$

$$\sigma^2 = \frac{2.15493}{17-2} = 0.14366 \tag{2.10}$$

The square root of $\hat{\sigma}^2$, $\hat{\sigma} = \sqrt{0.14366} = 0.37903$ is often called the *standard error of regression*. It is in the same units as is the response variable.

If in addition to the assumptions made previously, the e_i are drawn from a normal distribution, then the residual mean square will be distributed as a multiple of a chi-squared random variable with df $= n - 2$, or in symbols,

$$(n-2)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

This is proved in more advanced books on linear models and is used to obtain the distribution of test statistics and also to make confidence statements concerning σ^2. In particular, this fact implies that $\mathrm{E}(\hat{\sigma}^2) = \sigma^2$, although normality is not required for unbiasedness.

2.4 PROPERTIES OF LEAST SQUARES ESTIMATES

The OLS estimates depend on data only through the statistics given in Table 2.1. This is both an advantage, making computing easy, and a disadvantage, since any two data sets for which these are identical give the same fitted regression, even if a straight-line model is appropriate for one but not the other, as we have seen in Anscombe's examples in Section 1.4. The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ can both be written as linear combinations of $y_1, \ldots, y_n$, for example, writing $c_i = (x_i - \overline{x})/SXX$ (see Appendix A.3)

$$\hat{\beta}_1 = \sum \left(\frac{x_i - \overline{x}}{SXX} \right) y_i = \sum c_i y_i$$

The fitted value at $x = \overline{x}$ is

$$\widehat{\mathrm{E}}(Y|X = \overline{x}) = \overline{y} - \hat{\beta}_1 \overline{x} + \hat{\beta}_1 \overline{x} = \overline{y}$$

so the fitted line must pass through the point $(\overline{x}, \overline{y})$, intuitively the center of the data. Finally, as long as the mean function includes an intercept, $\sum \hat{e}_i = 0$. Mean functions without an intercept will usually have $\sum \hat{e}_i \neq 0$.

Since the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the random e_is, the estimates are also random variables. If all the e_i have zero mean and the mean function is correct, then, as shown in Appendix A.4, the least squares estimates are unbiased,

$$E(\hat{\beta}_0) = \beta_0$$
$$E(\hat{\beta}_1) = \beta_1$$

The variance of the estimators, assuming $\mathrm{Var}(e_i) = \sigma^2, i = 1, \ldots, n$, and $\mathrm{Cov}(e_i, e_j) = 0, i \neq j$, are from Appendix A.4,

$$\mathrm{Var}(\hat{\beta}_1) = \sigma^2 \frac{1}{SXX}$$
$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{SXX} \right) \tag{2.11}$$

The two estimates are correlated, with covariance

$$\mathrm{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \frac{\overline{x}}{SXX} \tag{2.12}$$

The correlation between the estimates can be computed to be

$$\rho(\hat{\beta}_0, \hat{\beta}_1) = \frac{-\overline{x}}{\sqrt{SXX/n + \overline{x}^2}} = \frac{-\overline{x}}{\sqrt{(n-1)\mathrm{SD}_x^2/n + \overline{x}^2}}$$

This correlation can be close to plus or minus one if SD_x is small compared to $|\overline{x}|$ and can be made to equal zero if the predictor is centered to have sample mean zero.

The *Gauss–Markov theorem* provides an optimality result for OLS estimates. Among all estimates that are linear combinations of the ys and unbiased, the OLS estimates have the smallest variance. If one believes the assumptions and is interested in using linear unbiased estimates, the OLS estimates are the ones to use.

When the errors are normally distributed, the OLS estimates can be justified using a completely different argument, since they are then also maximum likelihood estimates, as discussed in any mathematical statistics text, for example, Casella and Berger (1990).

Under the assumption that errors are independent, normal with constant variance, which is written in symbols as

$$e_i \sim \mathrm{NID}(0, \sigma^2) \qquad i = 1, \ldots, n$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are also normally distributed, since they are linear functions of the y_is and hence of the e_i, with variances and covariances given by (2.11) and (2.12). These results are used to get confidence intervals and tests. Normality of estimates also holds without normality of errors if the sample size is large enough[2].

2.5 ESTIMATED VARIANCES

Estimates of $\mathrm{Var}(\hat{\beta}_0)$ and $\mathrm{Var}(\hat{\beta}_1)$ are obtained by substituting $\hat{\sigma}^2$ for σ^2 in (2.11). We use the symbol $\widehat{\mathrm{Var}}(\ )$ for an estimated variance. Thus

$$\widehat{\mathrm{Var}}(\hat{\beta}_1) = \hat{\sigma}^2 \frac{1}{SXX}$$
$$\widehat{\mathrm{Var}}(\hat{\beta}_0) = \hat{\sigma}^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{SXX} \right)$$

The square root of an estimated variance is called a *standard error*, for which we use the symbol se( ). The use of this notation is illustrated by

$$\mathrm{se}(\hat{\beta}_1) = \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}$$

[2]The main requirement for all estimates to be normally distributed in large samples is that $\max_i \left((x_i - \overline{x})^2 / SXX \right)$ must get close to zero as the sample size increases (Huber, 1981).

2.6 COMPARING MODELS: THE ANALYSIS OF VARIANCE

The analysis of variance provides a convenient method of comparing the fit of two or more mean functions for the same set of data. The methodology developed here is very useful in multiple regression and, with minor modification, in most regression problems.

An elementary alternative to the simple regression model suggests fitting the mean function

$$\mathrm{E}(Y|X = x) = \beta_0 \tag{2.13}$$

The mean function (2.13) is the same for all values of X. Fitting with this mean function is equivalent to finding the best line parallel to the horizontal or x-axis, as shown in Figure 2.4. The OLS estimate of the mean function is $\widehat{E(Y|X)} = \hat{\beta}_0$, where $\hat{\beta}_0$ is the value of β_0 that minimizes $\sum(y_i - \beta_0)^2$. The minimizer is given by

$$\hat{\beta}_0 = \overline{y} \tag{2.14}$$

The residual sum of squares is

$$\sum(y_i - \hat{\beta}_0)^2 = \sum(y_i - \overline{y})^2 = SYY \tag{2.15}$$

This residual sum of squares has $n - 1$ df, n cases minus one parameter in the mean function.

Next, consider the simple regression mean function obtained from (2.13) by adding a term that depends on X

$$\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x \tag{2.16}$$

Fitting this mean function is equivalent to finding the best line of arbitrary slope, as shown in Figure 2.4. The OLS estimates for this mean function are given by

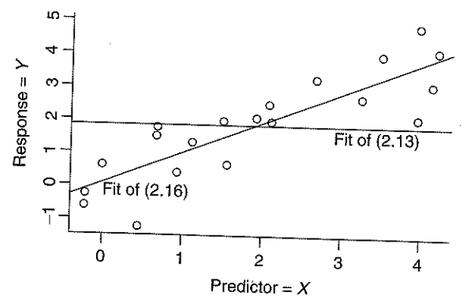

FIG. 2.4 Two mean functions compared by the analysis of variance.

(2.5). The estimates of β_0 under the two mean functions are different, just as the meaning of β_0 in the two mean functions is different. For (2.13), β_0 is the average of the y_is, but for (2.16), β_0 is the expected value of Y when $X = 0$.

For (2.16), the residual sum of squares, given in (2.8), is

$$RSS = SYY - \frac{(SXY)^2}{SXX} \tag{2.17}$$

As mentioned earlier, RSS has $n - 2$ df.

The difference between the sum of squares at (2.15) and that at (2.17) is the reduction in residual sum of squares due to enlarging the mean function from (2.13) to the simple regression mean function (2.16). This is the *sum of squares due to regression*, *SSreg*, defined by

$$\begin{aligned} SSreg &= SYY - RSS \\ &= SYY - \left(SYY - \frac{(SXY)^2}{SXX}\right) \\ &= \frac{(SXY)^2}{SXX} \end{aligned} \tag{2.18}$$

The df associated with *SSreg* is the difference in df for mean function (2.13), $n - 1$, and the df for mean function (2.16), $n - 2$, so the df for *SSreg* is $(n - 1) - (n - 2) = 1$ for simple regression. These results are often summarized in an analysis of variance table, abbreviated as *ANOVA*, given in Table 2.3. The column marked "Source" refers to descriptive labels given to the sums of squares; in more complicated tables, there may be many sources, and the labels given may be different in some computer programs. The df column gives the number of degrees of freedom associated with each named source. The next column gives the associated sum of squares. The mean square column is computed from the sum of squares column by dividing sums of squares by the corresponding df. The mean square on the residual line is just $\hat{\sigma}^2$, as already discussed.

The analysis of variance for Forbes' data is given in Table 2.4. Although this table will be produced by any linear regression software program, the entries in Table 2.4 can be constructed from the summary statistics given at (2.6).

The *ANOVA* is always computed relative to a specific larger mean function, here given by (2.16), and a smaller mean function obtained from the larger by setting

TABLE 2.3 The Analysis of Variance Table for Simple Regression

Source	df	SS	MS	F	p-value
Regression	1	*SSreg*	*SSreg*/1	$MSreg/\hat{\sigma}^2$	
Residual	$n-2$	*RSS*	$\hat{\sigma}^2 = RSS/(n-2)$		
Total	$n-1$	*SYY*			

TABLE 2.4 Analysis of Variance Table for Forbes' Data

Source	df	SS	MS	F	p-value
Regression on *Temp*	1	425.639	425.639	2962.79	≈ 0
Residual	15	2.155	0.144		

some parameters to zero, or occasionally setting them to some other known value. For example, equation (2.13) was obtained from (2.16) by setting $\beta_1 = 0$. The line in the *ANOVA* table for the total gives the residual sum of squares corresponding to the mean function with the fewest parameters. In the next chapter, the analysis of variance is applied to a sequence of mean functions, but the reference to a fixed large mean function remains intact.

2.6.1 The F-Test for Regression

If the sum of squares for regression *SSreg* is large, then the simple regression mean function $\mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x$ should be a significant improvement over the mean function given by (2.13), $\mathrm{E}(y|X = x) = \beta_0$. This is equivalent to saying that the additional parameter in the simple regression mean function β_1 is different from zero or that $\mathrm{E}(Y|X = x)$ is not constant as X varies. To formalize this notion, we need to be able to judge how large is "large." This is done by comparing the regression mean square, *SSreg* divided by its df, to the residual mean square $\hat{\sigma}^2$. We call this ratio F:

$$F = \frac{(SYY - RSS)/1}{\hat{\sigma}^2} = \frac{SSreg/1}{\hat{\sigma}^2} \tag{2.19}$$

F is just a rescaled version of $SSreg = SYY - RSS$, with larger values of *SSreg* resulting in larger values of F. Formally, we can consider testing the null hypothesis (NH) against the alternative hypothesis (AH)

$$\begin{aligned} &\text{NH:} \quad \mathrm{E}(Y|X = x) = \beta_0 \\ &\text{AH:} \quad \mathrm{E}(Y|X = x) = \beta_0 + \beta_1 x \end{aligned} \tag{2.20}$$

If the errors are $\mathrm{NID}(0, \sigma^2)$ or the sample size is large enough, then under NH (2.19) will follow an F-distribution with df associated with the numerator and denominator of (2.19), 1 and $n - 2$ for simple regression. This is written $F \sim F(1, n - 2)$. For Forbes' data, we compute

$$F = \frac{425.639}{0.144} = 2963$$

We obtain a significance level or p-value for this test by comparing F to the percentage points of the $F(1, n - 2)$-distribution. Most computer programs that fit regression models will include functions to computing percentage points of the F

and other standard distributions and will include the p-value along with the *ANOVA* table, as in Table 2.4. The p-value is shown as "approximately zero," meaning that, if the NH were true, the change of F exceeding its observed value is essentially zero. This is very strong evidence against NH and in favor of AH.

2.6.2 Interpreting p-values

Under the appropriate assumptions, the p-value is the conditional probability of observing a value of the computed statistic, here the value of F, as extreme or more extreme, here as large or larger, than the observed value, given that the NH is true. A small p-value provides evidence against the NH.

In some research areas, it has become traditional to adopt a *fixed significance level* when examining p-values. For example, if a fixed significance level of α is adopted, then we would say that an NH is rejected at level α if the p-value is less than α. The most common choice for α is 0.05, which would mean that, were the NH to be true, we would incorrectly find evidence against it about 5% of the time, or about 1 test in 20. Accept–reject rules like this are generally unnecessary for reasonable scientific inquiry. Simply reporting p-values and allowing readers to decide on significance seems a better approach.

There is an important distinction between statistical significance, the observation of a sufficiently small p-value, and scientific significance, observing an effect of sufficient magnitude to be meaningful. Judgment of the latter usually will require examination of more than just the p-value.

2.6.3 Power of Tests

When the NH is true, and all assumptions are met, the chance of incorrectly declaring an NH to be false at level α is just α. If $\alpha = 0.05$, then in 5% of tests where the NH is true we will get a p-value smaller than or equal to 0.05.

When the NH is false, we expect to see small p-values more often. The *power* of a test is defined to be the *probability of detecting a false NH*. For the hypothesis test (2.20), when the NH is false, it is shown in more advanced books on linear models (such as Seber, 1977) that the statistic F given by (2.19) has a *noncentral* F distribution, with 1 and $n - 2$ df, and with noncentrality parameter given by $SXX\beta_1^2/\sigma^2$. The larger the value of the non centrality parameter, the greater the power. The noncentrality is increased if β_1^2 is large, if SXX is large, either by spreading out the predictors or by increasing the sample size, or by decreasing σ^2.

2.7 THE COEFFICIENT OF DETERMINATION, R^2

If both sides of (2.18) are divided by SYY, we get

$$\frac{SSreg}{SYY} = 1 - \frac{RSS}{SYY} \tag{2.21}$$

The left-hand side of (2.21) is the proportion of variability of the response explained by regression on the predictor. The right-hand side consists of one minus the

remaining unexplained variability. This concept of dividing up the total variability according to whether or not it is explained is of sufficient importance that a special name is given to it. We define R^2, the coefficient of determination, to be

$$R^2 = \frac{SSreg}{SYY} = 1 - \frac{RSS}{SYY} \tag{2.22}$$

R^2 is computed from quantities that are available in the *ANOVA* table. It is a scale-free one-number summary of the strength of the relationship between the x_i and the y_i in the data. It generalizes nicely to multiple regression, depends only on the sums or squares and appears to be easy to interpret. For Forbes' data,

$$R^2 = \frac{SSreg}{SYY} = \frac{425.63910}{427.79402} = 0.995$$

and thus about 99.5% of the variability in the observed values or $100 \times \log(Pressure)$ is explained by boiling point. Since R^2 does not depend on units of measurement, we would get the same value if we had used logarithms with a different base, or if we did not multiply $\log(Pressure)$ by 100.

By appealing to (2.22) and to Table 2.1, we can write

$$R^2 = \frac{SSreg}{SYY} = \frac{(SXY)^2}{SXX \times SYY} = r_{xy}^2$$

and thus R^2 is the same as the square of the sample correlation between the predictor and the response.

2.8 CONFIDENCE INTERVALS AND TESTS

When the errors are NID$(0, \sigma^2)$, parameter estimates, fitted values, and predictions will be normally distributed because all of these are linear combinations of the y_i and hence of the e_i. Confidence intervals and tests can be based on the t-distribution, which is the appropriate distribution with normal estimates but using an estimate of variance $\hat{\sigma}^2$. Suppose we let $t(\alpha/2, d)$ be the value that cuts off $\alpha/2 \times 100\%$ in the *upper tail* of the t-distribution with d df. These values can be computed in most statistical packages or spreadsheet software[3].

2.8.1 The Intercept

The intercept is used to illustrate the general form of confidence intervals for normally distributed estimates. The standard error of the intercept is $\text{se}(\beta_0) = \hat{\sigma}(1/n + \bar{x}^2/SXX)^{1/2}$. Hence a $(1-\alpha) \times 100\%$ confidence interval for the intercept is the set of points β_0 in the interval

$$\hat{\beta}_0 - t(\alpha/2, n-2)\text{se}(\hat{\beta}_0) \le \beta_0 \le \hat{\beta}_0 + t(\alpha/2, n-2)\text{se}(\hat{\beta}_0)$$

[3]Such as the function `tinv` in Microsoft Excel, or the function `pt` in R or S-plus.

For Forbes' data, $\text{se}(\hat{\beta}_0) = 0.37903(1/17 + (202.95294)^2/530.78235)^{1/2} = 3.340$. For a 90% confidence interval, $t(0.05, 15) = 1.753$, and the interval is

$$-42.138 - 1.753(3.340) \le \beta_0 \le -42.136 + 1.753(3.340)$$
$$-47.993 \le \beta_0 \le -36.282$$

Ninety percent of such intervals will include the true value.

A hypothesis test of

$$\begin{array}{ll} \text{NH:} & \beta_0 = \beta_0^*, \quad \beta_1 \text{ arbitrary} \\ \text{AH:} & \beta_0 \ne \beta_0^*, \quad \beta_1 \text{ arbitrary} \end{array}$$

is obtained by computing the t-statistic

$$t = \frac{\hat{\beta}_0 - \beta_0^*}{\text{se}(\hat{\beta}_0)} \tag{2.23}$$

and referring this ratio to the t-distribution with $n-2$ df. For example, in Forbes' data, consider testing the NH $\beta_0 = -35$ against the alternative that $\beta_0 \ne -35$. The statistic is

$$t = \frac{-42.138 - (-35)}{3.340} = 2.137$$

which has a p-value near 0.05, providing some evidence against NH. This hypothesis test for these data is not one that would occur to most investigators and is used only as an illustration.

2.8.2 Slope

The standard error of $\hat{\beta}_1$ is $\text{se}(\hat{\beta}_1) = \hat{\sigma}/\sqrt{SXX} = 0.0164$. A 95% confidence interval for the slope is the set of β_1 such that

$$0.8955 - 2.131(0.0164) \le \beta_1 \le 0.8955 + 2.131(0.0164)$$
$$0.867 \le \beta_1 \le 0.930$$

As an example of a test for slope equal to zero, consider the Ft. Collins snowfall data presented on page 7. One can show, Problem 2.11, that the estimated slope is $\hat{\beta}_1 = 0.2035$, $\text{se}(\hat{\beta}_1) = 0.1310$. The test of interest is of

$$\begin{array}{ll} \text{NH:} & \beta_1 = 0 \\ \text{AH:} & \beta_1 \ne 0 \end{array} \tag{2.24}$$

For the Ft. Collins data, $t = (0.20335 - 0)/0.1310 = 1.553$. To get a significance level for this test, compare t with the $t(91)$ distribution; the two-sided p-value is

0.124, suggesting no evidence against the NH that *Early* and *Late* season snowfalls are independent.

Compare the hypothesis (2.24) with (2.20). Both appear to be identical. In fact,

$$t^2 = \left(\frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}\right)^2 = \frac{\hat{\beta}_1^2}{\hat{\sigma}^2/SXX} = \frac{\hat{\beta}_1^2 SXX}{\hat{\sigma}^2} = F$$

so the square of a t statistic with d df is equivalent to an F-statistic with $(1, d)$ df. In nonlinear and logistic regression models discussed later in the book, the analog of the t test will not be identical to the analog of the F test, and they can give conflicting conclusions. For linear regression models, no conflict occurs and the two tests are equivalent.

2.8.3 Prediction

The estimated mean function can be used to obtain values of the response for given values of the predictor. The two important variants of this problem are *prediction* and *estimation of fitted values*. Since prediction is more important, we discuss it first.

In prediction we have a new case, possibly a future value, not one used to estimate parameters, with observed value of the predictor x_*. We would like to know the value y_*, the corresponding response, but it has not yet been observed. We can use the estimated mean function to predict it. We assume that the data used to estimate the mean function are relevant to the new case, so the fitted model applies to it. In the heights example, we would probably be willing to apply the fitted mean function to mother–daughter pairs alive in England at the end of the nineteenth century. Whether the prediction would be reasonable for mother–daughter pairs in other countries or in other time periods is much less clear. In Forbes' problem, we would probably be willing to apply the results for altitudes in the range he studied. Given this additional assumption, a point prediction of y_*, say $\tilde{y}_*$, is just

$$\tilde{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$$

$\tilde{y}_*$ predicts the as yet unobserved y_*. The variability of this predictor has two sources: the variation in the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, and the variation due to the fact that y_* will not equal its expectation, since even if we knew the parameters exactly, the future value of the response will not generally equal its expectation. Using Appendix A.4,

$$\text{Var}(\tilde{y}_*|x_*) = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_* - \overline{x})^2}{SXX}\right) \tag{2.25}$$

Taking square roots and estimating σ^2 by $\hat{\sigma}^2$, we get the standard error of prediction (sepred) at x_*,

$$\text{sepred}(\tilde{y}_*|x_*) = \hat{\sigma}\left(1 + \frac{1}{n} + \frac{(x_* - \overline{x})^2}{SXX}\right)^{1/2} \tag{2.26}$$

A prediction interval uses multipliers from the t-distribution. For prediction of $100 \times \log(Pressure)$ for a location with $x_* = 200$, the point prediction is $\tilde{y}_* = -42.13778 + 0.89549(200) = 136.961$, with standard error of prediction

$$\begin{aligned}\text{sepred}(\tilde{y}_*|x_* = 200) &= 0.37903\left(1 + \frac{1}{17} + \frac{(200 - 202.95294)^2}{530.78235}\right)^{1/2} \\ &= 0.393\end{aligned}$$

Thus a 99% predictive interval is the set of all y_* such that

$$136.961 - 2.95(0.393) \le y_* \le 136.961 + 2.95(0.393)$$
$$135.803 \le y_* \le 138.119$$

More interesting would be a 99% prediction interval for *Pressure*, rather than for $100 \times \log(Pressure)$. A point prediction is just $10^{(136.961/100)} = 23.421$ inches of Mercury. The prediction interval is found by exponentiating the end points of the interval in log scale. Dividing by 100 and then exponentiating, we get

$$10^{135.803/100} \le Pressure \le 10^{138.119/100}$$
$$22.805 \le Pressure \le 24.054$$

In the original scale, the prediction interval is not symmetric about the point estimate.

For the heights data, Figure 2.5 is a plot of the estimated mean function given by the dashed line for the regression of *Dheight* on *Mheight* along with curves at

$$\hat{\beta}_0 + \hat{\beta}_1 x_* \pm t(.025, 1373)\text{se}(\widetilde{Dheight}_*|Mheight_*)$$

The vertical distance between the two solid curves for any value of *Mheight* corresponds to a 95% prediction interval for daughter's height given mother's height. Although not obvious from the graph because of the very large sample size, the interval is wider for mothers who were either relatively tall or short, as the curves bend outward from the narrowest point at $Mheight = \overline{Mheight}$.

2.8.4 Fitted Values

In rare problems, one may be interested in obtaining an estimate of $\text{E}(Y|X = x)$. In the heights data, this is like asking for the population mean height of all daughters of mothers with a particular height. This quantity is estimated by the fitted value $\hat{y} = \beta_0 + \beta_1 x$, and its standard error is

$$\text{sefit}(\tilde{y}_*|x_*) = \hat{\sigma}\left(\frac{1}{n} + \frac{(x_* - \overline{x})^2}{SXX}\right)^{1/2}$$

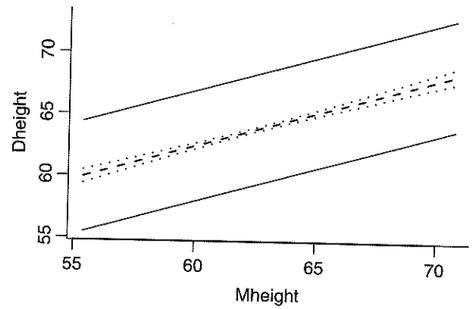

FIG. 2.5 Prediction intervals (solid lines) and intervals for fitted values (dashed lines) for the heights data.

To obtain confidence intervals, it is more usual to compute a simultaneous interval for all possible values of x. This is the same as first computing a joint confidence region for β_0 and β_1, and from these, computing the set of all possible mean functions with slope and intercept in the joint confidence set (Section 5.5). The confidence region for the mean function is the set of all y such that

$$(\hat{\beta}_0 + \hat{\beta}_1 x) - \text{sefit}(\hat{y}|x)[2F(\alpha; 2, n-2)]^{1/2} \le y$$
$$\le (\hat{\beta}_0 + \hat{\beta}_1 x) + \text{sefit}(\hat{y}|x)[2F(\alpha; 2, n-2)]^{1/2}$$

For multiple regression, replace $2F(\alpha; 2, n-2)$ by $p'F(\alpha; p', n-p')$, where p' is the number of parameters estimated in the mean function including the intercept. The simultaneous band for the fitted line for the heights data is shown in Figure 2.5 as the vertical distances between the two dotted lines. The prediction intervals are much wider than the confidence intervals. Why is this so (Problem 2.4)?

2.9 THE RESIDUALS

Plots of residuals versus other quantities are used to find failures of assumptions. The most common plot, especially useful in simple regression, is the plot of residuals versus the fitted values. A null plot would indicate no failure of assumptions. Curvature might indicate that the fitted mean function is inappropriate. Residuals that seem to increase or decrease in average magnitude with the fitted values might indicate nonconstant residual variance. A few relatively large residuals may be indicative of outliers, cases for which the model is somehow inappropriate.

The plot of residuals versus fitted values for the heights data is shown in Figure 2.6. This is a null plot, as it indicates no particular problems.

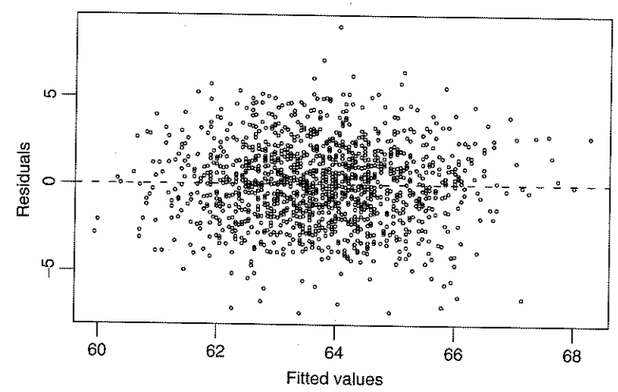

FIG. 2.6 Residuals versus fitted values for the heights data.

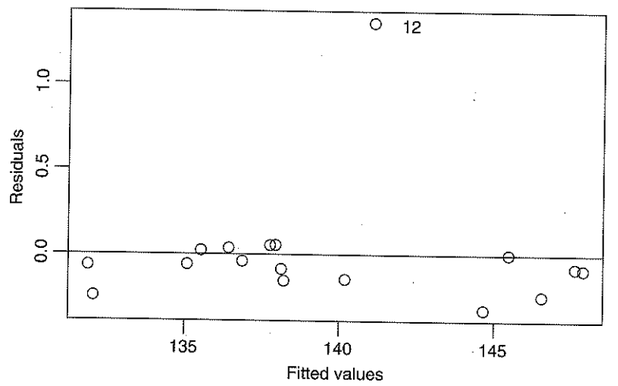

FIG. 2.7 Residual plot for Forbes' data.

The fitted values and residuals for Forbes' data are plotted in Figure 2.7. The residuals are generally small compared to the fitted values, and they do not follow any distinct pattern in Figure 2.7. The residual for case number 12 is about four times the size of the next largest residual in absolute value. This may suggest that the assumptions concerning the errors are not correct. Either $\text{Var}(100 \times \log(\textit{Pressure})|\textit{Temp})$ may not be constant or for case 12, the corresponding error may have a large fixed component. Forbes may have misread or miscopied the results of his calculations for this case, which would suggest that the numbers in

TABLE 2.5 Summary Statistics for Forbes' Data with All Data and with Case 12 Deleted

Quantity	All Data	Delete Case 12
$\hat{\beta}_0$	−42.138	−41.308
$\hat{\beta}_1$	0.895	0.891
$\text{se}(\hat{\beta}_0)$	3.340	1.001
$\text{se}(\hat{\beta}_1)$	0.016	0.005
$\hat{\sigma}$	0.379	0.113
R^2	0.995	1.000

the data do not correspond to the actual measurements. Forbes noted this possibility himself, by marking this pair of numbers in his paper as being "evidently a mistake", presumably because of the large observed residual.

Since we are concerned with the effects of case 12, we could refit the data, this time without case 12, and then examine the changes that occur in the estimates of parameters, fitted values, residual variance, and so on. This is summarized in Table 2.5, giving estimates of parameters, their standard errors, $\hat{\sigma}^2$, and the coefficient of determination R^2 with and without case 12. The estimates of parameters are essentially identical with and without case 12. In other regression problems, deletion of a single case can change everything. The effect of case 12 on standard errors is more marked: if case 12 is deleted, standard errors are decreased by a factor of about 3.1, and variances are decreased by a factor of about $3.1^2 \approx 10$. Inclusion of this case gives the appearance of less reliable results than would be suggested on the basis of the other 16 cases. In particular, prediction intervals of *Pressure* are much wider based on all the data than on the 16-case data, although the point predictions are nearly the same. The residual plot obtained when case 12 is deleted before computing indicates no obvious failures in the remaining 16 cases.

Two competing fits using the same mean function but somewhat different data are available, and they lead to slightly different conclusions, although the results of the two analyses agree more than they disagree. On the basis of the data, there is no real way to choose between the two, and we have no way of deciding which is the correct OLS analysis of the data. A good approach to this problem is to describe both or, in general, all plausible alternatives.

PROBLEMS

2.1. Height and weight data The table below and in the data file `htwt.txt` gives Ht = height in centimeters and Wt = weight in kilograms for a sample of $n = 10$ 18-year-old girls. The data are taken from a larger study described in Problem 3.1. Interest is in predicting weight from height.

Ht	*Wt*
169.6	71.2
166.8	58.2
157.1	56.0
181.1	64.5
158.4	53.0
165.6	52.4
166.7	56.8
156.5	49.2
168.1	55.6
165.3	77.8

2.1.1. Draw a scatterplot of Wt on the vertical axis versus Ht on the horizontal axis. On the basis of this plot, does a simple linear regression model make sense for these data? Why or why not?

2.1.2. Show that $\bar{x} = 165.52$, $\bar{y} = 59.47$, $SXX = 472.076$, $SYY = 731.961$, and $SXY = 274.786$. Compute estimates of the slope and the intercept for the regression of Y on X. Draw the fitted line on your scatterplot.

2.1.3. Obtain the estimate of σ^2 and find the estimated standard errors of $\hat{\beta}_0$ and $\hat{\beta}_1$. Also find the estimated covariance between $\hat{\beta}_0$ and $\hat{\beta}_1$. Compute the t-tests for the hypotheses that $\beta_0 = 0$ and that $\beta_1 = 0$ and find the appropriate p-values using two-sided tests.

2.1.4. Obtain the analysis of variance table and F-test for regression. Show numerically that $F = t^2$, where t was computed in Problem 2.1.3 for testing $\beta_1 = 0$.

2.2. More with Forbes' data An alternative approach to the analysis of Forbes' experiments comes from the Clausius–Clapeyron formula of classical thermodynamics, which dates to Clausius (1850). According to this theory, we should find that

$$\text{E}(Lpres|Temp) = \beta_0 + \beta_1 \frac{1}{Ktemp} \tag{2.27}$$

where $Ktemp$ is temperature in degrees Kelvin, which equals 255.37 plus $(5/9) \times Temp$. If we were to graph this mean function on a plot of $Lpres$ versus $Ktemp$, we would get a curve, not a straight line. However, we can estimate the parameters β_0 and β_1 using simple linear regression methods by defining u_1 to be the inverse of temperature in degrees Kelvin,

$$u_1 = \frac{1}{Ktemp} = \frac{1}{(5/9)Temp + 255.37}$$

Then the mean function (2.27) can be rewritten as

$$\mathrm{E}(Lpres|Temp) = \beta_0 + \beta_1 u_1 \tag{2.28}$$

for which simple linear regression is suitable. The notation we have used in (2.28) is a little different, as the left side of the equation says we are conditioning on *Temp*, but the variable *Temp* does not appear explicitly on the right side of the equation.

2.2.1. Draw the plot of *Lpres* versus u_1, and verify that apart from case 12 the 17 points in Forbes' data fall close to a straight line.

2.2.2. Compute the linear regression implied by (2.28), and summarize your results.

2.2.3. We now have two possible models for the same data based on the regression of *Lpres* on *Temp* used by Forbes, and (2.28) based on the Clausius–Clapeyron formula. To compare these two, draw the plot of the fitted values from Forbes' mean function fit versus the fitted values from (2.28). On the basis of these and any other computations you think might help, is it possible to prefer one approach over the other? Why?

2.2.4. In his original paper, Forbes provided additional data collected by the botanist Dr. Joseph Hooker on temperatures and boiling points measured often at higher altitudes in the Himalaya Mountains. The data for $n = 31$ locations is given in the file `hooker.txt`. Find the estimated mean function (2.28) for Hooker's data.

2.2.5. *This problem is not recommended unless you have access to a package with a programming language, like R, S-plus, Mathematica, or SAS IML.* For each of the cases in Hooker's data, compute the predicted values $\hat{y}$ and the standard error of prediction. Then compute $z = (Lpres - \hat{y})/\text{sepred}$. Each of the zs is a random variable, but if the model is correct, each has mean zero and standard deviation close to one. Compute the sample mean and standard deviation of the zs, and summarize results.

2.2.6. Repeat Problem 2.2.5, but this time predict and compute the z-scores for the 17 cases in Forbes data, again using the fitted mean function from Hooker's data. If the mean function for Hooker's data applies to Forbes' data, then each of the z-scores should have zero mean and standard deviation close to one. Compute the z-scores, compare them to those in the last problem and comment on the results.

2.3. Deviations from the mean Sometimes it is convenient to write the simple linear regression model in a different form that is a little easier to manipulate. Taking equation (2.1), and adding $\beta_1\overline{x} - \beta_1\overline{x}$, which equals zero, to the right-hand side, and combining terms, we can write

$$\begin{aligned} y_i &= \beta_0 + \beta_1\overline{x} + \beta_1 x_i - \beta_1\overline{x} + e_i \\ &= (\beta_0 + \beta_1\overline{x}) + \beta_1(x_i - \overline{x}) + e_i \\ &= \alpha + \beta_1(x_i - \overline{x}) + e_i \end{aligned} \tag{2.29}$$

where we have defined $\alpha = \beta_0 + \beta_1\overline{x}$. This is called the *deviations from the sample average form for simple regression.*

2.3.1. What is the meaning of the parameter α?

2.3.2. Show that the least squares estimates are

$$\hat{\alpha} = \overline{y}, \qquad \hat{\beta}_1 \text{ as given by (2.5)}$$

2.3.3. Find expressions for the variances of the estimates and the covariance between them.

2.4. Heights of mothers and daughters

2.4.1. For the heights data in the file `heights.txt`, compute the regression of *Dheight* on *Mheight*, and report the estimates, their standard errors, the value of the coefficient of determination, and the estimate of variance. Give the analysis of variance table that tests the hypothesis that $\mathrm{E}(Dheight|Mheight) = \beta_0$ versus the alternative that $\mathrm{E}(Dheight|Mheight) = \beta_0 + \beta_1 Mheight$, and write a sentence or two that summarizes the results of these computations.

2.4.2. Write the mean function in the deviations from the mean form as in Problem 2.3. For this particular problem, give an interpretation for the value of β_1. In particular, discuss the three cases of $\beta_1 = 1$, $\beta_1 < 1$ and $\beta_1 > 1$. Obtain a 99% confidence interval for β_1 from the data.

2.4.3. Obtain a prediction and 99% prediction interval for a daughter whose mother is 64 inches tall.

2.5. Smallmouth bass

2.5.1. Using the West Bearskin Lake smallmouth bass data in the file `wblake.txt`, obtain 95% intervals for the mean length at ages 2, 4 and 6 years.

2.5.2. Obtain a 95% interval for the mean length at age 9. Explain why this interval is likely to be untrustworthy.

2.5.3. The file `wblake2.txt` contains all the data for ages one to eight and, in addition, includes a few older fishes. Using the methods we have learned in this chapter, show that the simple linear regression model is not appropriate for this larger data set.

2.6. United Nations data Refer to the UN data in Problem 1.3, page 18.

2.6.1. Using base-ten logarithms, use a software package to compute the simple linear regression model corresponding to the graph in Problem 1.3.3, and get the analysis of variance table.

2.6.2. Draw the summary graph, and add the fitted line to the graph.

2.6.3. Test the hypothesis that the slope is zero versus the alternative that it is negative (a one-sided test). Give the significance level of the test and a sentence that summarizes the result.

2.6.4. Give the value of the coefficient of determination, and explain its meaning.

2.6.5. Increasing log(*PPgdp*) by one unit is the same as multiplying *PPgdp* by ten. If two localities differ in *PPgdp* by a factor of ten, give a 95% confidence interval on the difference in log(*Fertility*) for these two localities.

2.6.6. For a locality not in the data with $PPgdp = 1000$, obtain a point prediction and a 95% prediction interval for log(*Fertility*). If the interval (a, b) is a 95% prediction interval for log(*Fertility*), then a 95% prediction interval for *Fertility* is given by $(10^a, 10^b)$. Use this result to get a 95% prediction interval for *Fertility*.

2.6.7. Identify (1) the locality with the highest value of *Fertility*; (2) the locality with the lowest value of *Fertility*; and (3) the two localities with the largest positive residuals from the regression when both variables are in log scale, and the two countries with the largest negative residuals in log scales.

2.7. Regression through the origin Occasionally, a mean function in which the intercept is known *a priori* to be zero may be fit. This mean function is given by

$$\mathrm{E}(y|x) = \beta_1 x \tag{2.30}$$

The residual sum of squares for this model, assuming the errors are independent with common variance σ^2, is $RSS = \sum(y_i - \hat{\beta}_1 x_i)^2$.

2.7.1. Show that the least squares estimate of β_1 is $\hat{\beta}_1 = \sum x_i y_i / \sum x_i^2$. Show that $\hat{\beta}_1$ is unbiased and that $\mathrm{Var}(\hat{\beta}_1) = \sigma^2 / \sum x_i^2$. Find an expression for $\hat{\sigma}^2$. How many df does it have?

2.7.2. Derive the analysis of variance table with the larger model given by (2.16), but with the smaller model specified in (2.30). Show that the F-test derived from this table is numerically equivalent to the square of the t-test (2.23) with $\beta_0^* = 0$.

2.7.3. The data in Table 2.6 and in the file `snake.txt` give X = water content of snow on April 1 and Y = water yield from April to July in inches in the Snake River watershed in Wyoming for $n = 17$ years from 1919 to 1935 (from Wilm, 1950).

TABLE 2.6 Snake River Data for Problem 2.7

X	Y	X	Y
23.1	10.5	32.8	16.7
31.8	18.2	32.0	17.0
30.4	16.3	24.0	10.5
39.5	23.1	24.2	12.4
52.5	24.9	37.9	22.8
30.5	14.1	25.1	12.9
12.4	8.8	35.1	17.4
31.5	14.9	21.1	10.5
27.6	16.1		

Fit a regression through the origin and find $\hat{\beta}_1$ and σ^2. Obtain a 95% confidence interval for β_1. Test the hypothesis that the intercept is zero.

2.7.4. Plot the residuals versus the fitted values and comment on the adequacy of the mean function with zero intercept. In regression through the origin, $\sum \hat{e}_i \neq 0$.

2.8. Scale invariance

2.8.1. In the simple regression model (2.1), suppose the value of the predictor X is replaced by cX, where c is some non zero constant. How are $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$, R^2, and the t-test of NH: $\beta_1 = 0$ affected by this change?

2.8.2. Suppose each value of the response Y is replaced by dY, for some $d \neq 0$. Repeat 2.8.1.

2.9. Using Appendix A.3, verify equation (2.8).

2.10. Zipf's law Suppose we counted the number of times each word was used in the written works by Shakespeare, Alexander Hamilton, or some other author with a substantial written record (Table 2.7). Can we say anything about the frequencies of the most common words?

Suppose we let f_i be the rate per 1000 words of text for the ith most frequent word used. The linguist George Zipf (1902–1950) observed a law like relationship between rate and rank (Zipf, 1949),

$$\mathrm{E}(f_i|i) = a/i^b$$

and further observed that the exponent is close to $b = 1$. Taking logarithms of both sides, we get approximately

$$\mathrm{E}(\log(f_i)|\log(i)) = \log(a) - b\log(i) \tag{2.31}$$

TABLE 2.7 The Word Count Data

Word	The word
Hamilton	Rate per 1000 words of this word in the writings of Alexander Hamilton
HamiltonRank	Rank of this word in Hamilton's writings
Madison	Rate per 1000 words of this word in the writings of James Madison
MadisonRank	Rank of this word in Madison's writings
Jay	Rate per 1000 words of this word in the writings of John Jay
JayRank	Rank of this word in Jay's writings
Ulysses	Rate per 1000 words of this word in *Ulysses* by James Joyce
UlyssesRank	Rank of this word in *Ulysses*

Zipf's law has been applied to frequencies of many other classes of objects besides words, such as the frequency of visits to web pages on the internet and the frequencies of species of insects in an ecosystem.

The data in `MWwords.txt` give the frequencies of words in works from four different sources: the political writings of eighteenth-century American political figures Alexander Hamilton, James Madison, and John Jay, and the book *Ulysses* by twentieth-century Irish writer James Joyce. The data are from Mosteller and Wallace (1964, Table 8.1-1), and give the frequencies of 165 very common words. Several missing values occur in the data; these are really words that were used so infrequently that their count was not reported in Mosteller and Wallace's table.

2.10.1. Using only the 50 most frequent words in Hamilton's work (that is, using only rows in the data for which $HamiltonRank \le 50$), draw the appropriate summary graph, estimate the mean function (2.31), and summarize your results.

2.10.2. Test the hypothesis that $b = 1$ against the two-sided alternative and summarize.

2.10.3. Repeat Problem 2.10.1, but for words with rank of 75 or less, and with rank less than 100. For larger number of words, Zipf's law may break down. Does that seem to happen with these data?

2.11. For the Ft. Collins snow fall data discussed in Example 1.1, test the hypothesis that the slope is zero versus the alternative that it is not zero. Show that the t-test of this hypothesis is the same as the F-test; that is, $t^2 = F$.

2.12. Old Faithful Use the data from Problem 1.4, page 18.

2.12.1. Use simple linear regression methodology to obtain a prediction equation for *interval* from *duration*. Summarize your results in a way that might be useful for the nontechnical personnel who staff the Old Faithful Visitor's Center.

2.12.2. Construct a 95% confidence interval for

$$\mathrm{E}(interval|duration = 250)$$

2.12.3. An individual has just arrived at the end of an eruption that lasted 250 seconds. Give a 95% confidence interval for the time the individual will have to wait for the next eruption.

2.12.4. Estimate the 0.90 quantile of the conditional distribution of

$$interval|(duration = 250)$$

assuming that the population is normally distributed.

2.13. Windmills Energy can be produced from wind using windmills. Choosing a site for a *wind farm*, the location of the windmills, can be a multimillion dollar gamble. If wind is inadequate at the site, then the energy produced over the lifetime of the wind farm can be much less than the cost of building and operation. Prediction of long-term wind speed at a candidate site can be an important component in the decision to build or not to build. Since energy produced varies as the square of the wind speed, even small errors can have serious consequences.

The data in the file `wm1.txt` provides measurements that can be used to help in the prediction process. Data were collected every six hours for the year 2002, except that the month of May 2002 is missing. The values *Cspd* are the calculated wind speeds in meters per second at a candidate site for building a wind farm. These values were collected at tower erected on the site. The values *RSpd* are wind speeds at a *reference site*, which is a nearby location for which wind speeds have been recorded over a very long time period. Airports sometimes serve as reference sites, but in this case, the reference data comes from the National Center for Environmental Modeling; these data are described at http://dss.ucar.edu/datasets/ds090.0/. The reference is about 50 km south west of the candidate site. Both sites are in the northern part of South Dakota. The data were provided by Mark Ahlstrom and Rolf Miller of WindLogics.

2.13.1. Draw the scatterplot of the response *CSpd* versus the predictor *RSpd*. Is the simple linear regression model plausible for these data?

2.13.2. Fit the simple regression of the response on the predictor, and present the appropriate regression summaries.

2.13.3. Obtain a 95% prediction interval for *CSpd* at a time when $RSpd = 7.4285$.

2.13.4. For this problem, we revert to generic notation and let $x = RSpd$ and $y = CSpd$ and let n be the number of cases used in the regression ($n = 1116$ in the data we have used in this problem) and $\overline{x}$ and SXX defined from these n observations. Suppose we want to make predictions at m time points with values of wind speed $x_{*1}, .., x_{*m}$ that are different from the n cases used in constructing the prediction equation. Show that (1) the average of the m predictions is equal to the prediction taken at the average value $\overline{x}_*$ of the m values of the

predictor, and (2) using the first result, the standard error of the average of m predictions is

$$\text{se of average prediction} = \sqrt{\frac{\hat{\sigma}^2}{m} + \hat{\sigma}^2\left(\frac{1}{n} + \frac{(\overline{x}_* - \overline{x})^2}{SXX}\right)} \quad (2.32)$$

If m is very large, then the first term in the square root is negligible, and the standard error of average prediction is essentially the same as the standard error of a fitted value at $\overline{x}_*$.

2.13.5. For the period from January 1, 1948 to July 31, 2003, a total of $m = 62039$ wind speed measurements are available at the reference site, excluding the data from the year 2002. For these measurements, the average wind speed was $\overline{x}_* = 7.4285$. Give a 95% prediction interval on the long-term average wind speed at the candidate site. This long-term average of the past is then taken as an estimate of the long-term average of the future and can be used to help decide if the candidate is a suitable site for a wind farm.

CHAPTER 3

Multiple Regression

Multiple linear regression generalizes the simple linear regression model by allowing for many *terms* in a mean function rather than just one intercept and one slope.

3.1 ADDING A TERM TO A SIMPLE LINEAR REGRESSION MODEL

We start with a response Y and the simple linear regression mean function

$$\mathrm{E}(Y|X_1 = x_1) = \beta_0 + \beta_1 x_1$$

Now suppose we have a second variable X_2 with which to predict the response. By adding X_2 to the problem, we will get a mean function that depends on both the value of X_1 and the value of X_2,

$$\mathrm{E}(Y|X_1 = x_1, X_2 = x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \quad (3.1)$$

The main idea in adding X_2 is to explain the part of Y that has not already been explained by X_1.

United Nations Data

We will reconsider the United Nations data discussed in Problem 1.3. To the regression of log(*Fertility*), the base-two log fertility rate on log(*PPgdp*), the base-two log of the per person gross domestic product, we consider adding *Purban*, the percentage of the population that lives in an urban area. The data in the file `UN2.txt` give values for these three variables, as well as the name of the *Locality* for 193 localities, mostly countries, for which the United Nations provides data.

Figure 3.1 presents several graphical views of these data. Figure 3.1a can be viewed as a summary graph for the simple regression of log(*Fertility*) on log(*PPgdp*). The fitted mean function using OLS is

$$\widehat{\mathrm{E}}(\log(Fertility)|\log(PPgdp)) = 2.703 - 0.153\log(PPgdp)$$

Applied Linear Regression, Third Edition, by Sanford Weisberg
ISBN 0-471-66379-4