Chapter 12
Linear Regression

Regression is the central method in the analysis of neural data. This is partly because, in all its guises, it is the most widely applied technique. But it also played a crucial historical role[1] in the development of statistical thinking, and continues to form a core conceptual foundation for a great deal of statistical analysis. We introduced linear regression in Section 1.2.1 (on p. 10) by placing it in the context of curve-fitting, reviewing the method of least squares, and providing an explicit statement of the linear regression model. This enabled us to use linear regression as a concrete example of a statistical model, so that we could emphasize a few general points, including the role of models in expressing knowledge and uncertainty via inductive reasoning. The linear regression model is important not only because many noisy relationships are adequately described as linear, but also—as we tried to explain in Section 1.2.1—because the framework gives us a way of thinking about relationships between measured variables. For this reason, we began with the more general regression model in Eq. (1.2), i.e.,

$$Y_i = f(x_i) + \epsilon_i, \tag{12.1}$$

and only later, in Eq. (1.4), specified that $f(x)$ is taken to be linear, i.e.,

$$f(x) = \beta_0 + \beta_1 x. \tag{12.2}$$

Equation (1.2), repeated here as (12.1), gave substance to the diagram in Eq. (1.1), i.e.,

$$Y \longleftarrow X. \tag{12.3}$$

To incorporate multiple explanatory variables we replace $f(x)$ in (12.1) with $f(x_1, \ldots, x_p)$, and to extend beyond the additive form of noise in (12.1) we replace the diagram in (12.3) with

[1] See the appendix of Brown and Kass (2009).

R. E. Kass et al., *Analysis of Neural Data*,
Springer Series in Statistics, DOI: 10.1007/978-1-4614-9602-1_12,

$$Y \longleftarrow \begin{cases} \text{noise} \\ f(x_1, \ldots, x_p). \end{cases} \tag{12.4}$$

This diagram is supposed to indicate a variety of generalizations of linear regression which, together, form the class of methods known as *modern regression*.

In this chapter we provide a concise introduction to linear regression. In Sections 12.1–12.4 we treat the *simple linear regression model* given by

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{12.5}$$

for $i = 1, \ldots, n$, where ϵ_i is a random variable. The adjective "simple" refers to the single x variable on the right-hand side of (12.5). When there are two or more x variables on the right-hand side the terminology *multiple regression* is used instead. We go over some of the most fundamental aspects of multiple regression in Section 12.5. That section also lays the groundwork for modern regression. Generalizations are described in Chapters 14 and 15.

12.1 The Linear Regression Model

To help fix ideas, as we proceed we will refer to several examples.

Example 12.1 Neural correlates of reward in parietal cortex Platt and Glimcher (1999) suggested that cortical areas involved in sensory-motor processing may encode not only features of sensation and action but also key inputs to decision making. To support their claim they recorded neurons from the lateral intraparietal (LIP) region of monkeys during an eye movement task, and used linear regression to summarize the increasing trend in firing rate of intraparietal neurons with increasing expected gain in reward (volume of juice received) for successful completion of a task. Figure 12.1 shows plots of firing rate versus reward volume for a particular LIP neuron following onset of a visual cue. □

Example 2.1 (continued from p. 24) In their analysis of saccadic reaction time in hemispatial neglect, Behrmann et al. (2002) used linear regression in examining the modulation of saccadic reaction time as a function of angle to target by eye, head, or trunk orientation. We refer to this study in Section 12.5. □

In Chapter 1 we used Example 1.5 on neural conduction velocity to illustrate linear regression. Another plot of the neural conduction velocity data is provided again in Fig. 12.2.

Before we begin our discussion of statistical inference in linear regression, let us recall some of the things we said in Chapter 1 and provide a few basic formulas.

Given n data pairs (x_i, y_i), least squares finds $\hat{\beta}_0$ and $\hat{\beta}_1$ that satisfy

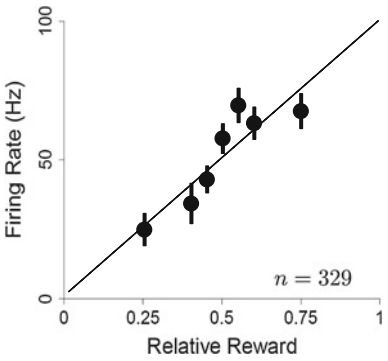

Fig. 12.1 Plots of firing rate (in spikes per second) versus reward volume (as fraction of the maximal possible reward volume). The plot represents firing rates during 200 ms following onset of a visual cue across 329 trials recorded from an LIP neuron. The 329 pairs of values have been reduced to 7 pairs, corresponding to seven distinct levels of the reward volume. Each of the 7 y_i values in the figure is a mean (among the trials with x_i as the reward volume), and error bars representing standard errors of each mean are also visible. A least-squares regression line is overlaid on the plot. Adapted from Platt and Glimcher (1999).

$$\sum_{i=1}^{n} \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2 = \min_{\beta_0^*, \beta_1^*} \sum_{i=1}^{n} \left(y_i - (\beta_0^* + \beta_1^* x_i)\right)^2 \tag{12.6}$$

where we use β_0^* and β_1^* as generic possible estimates of β_0 and β_1. The least-squares estimates (obtained by calculus) are

$$\hat{\beta}_1 = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{12.7}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{12.8}$$

The resulting fitted line

$$y = \hat{\beta}_0 + \hat{\beta}_1 x \tag{12.9}$$

is the *linear regression* line (and often "linear" is dropped).

Details: To be clear what we mean when we say that the least-squares estimates may be found by calculus, let us write

$$g(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 .$$

The formulas (12.8) and (12.7) may be obtained by computing the partial derivatives of $g(\beta_0, \beta_1)$ and then solving the equations

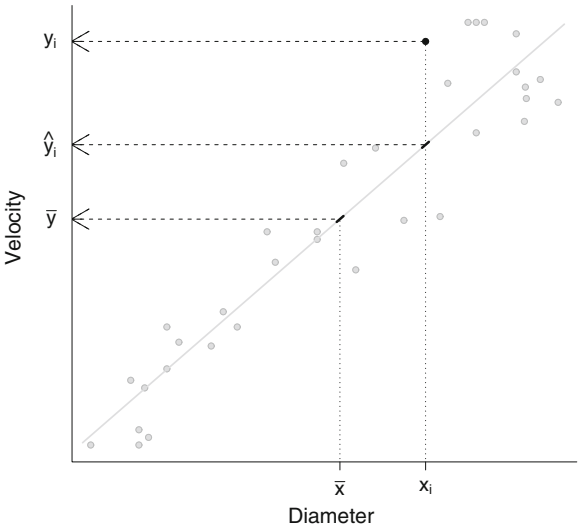

Fig. 12.2 Plot of the Hursh conduction velocity data set, for $5 < x < 15$, with data points in *gray* except for a particular point (x_i, y_i) which is shown in *black* to identify the corresponding fitted value $\hat{y}_i$. The ith residual is $y_i - \hat{y}_i$. The regression line also passes through the point $(\bar{x}, \bar{y})$, as indicated on the plot.

$$0 = \frac{\partial g}{\partial \beta_0}$$
$$0 = \frac{\partial g}{\partial \beta_1}.$$

□

The least-squares fitted values at each x_i are

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{12.10}$$

and the least-squares residuals are

$$e_i = y_i - \hat{y}_i. \tag{12.11}$$

See Fig. 12.2. If we plug (12.8) into (12.9) we get

$$y - \bar{y} = \hat{\beta}_1 (x - \bar{x}) \tag{12.12}$$

which shows that the regression line passes through the point $(\bar{x}, \bar{y})$, as may be seen in Fig. 12.2. Also, when we plug into (12.12) the (x, y) value (x_i, y_i) we get

$$y_i - \bar{y} = \hat{\beta}_1(x_i - \bar{x})$$

or

$$y_i = \bar{y} - \hat{\beta}_1(x_i - \bar{x}). \tag{12.13}$$

A few more lines of algebra show that using (12.13) in (12.11) gives

$$\sum_{i=1}^{n} e_i = 0, \tag{12.14}$$

which is useful as a math fact, and also can be important to keep in mind in data analysis: linear least squares residuals fail to satisfy (12.14) only when a numerical error has occurred.

Details: We have

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (y_i - \bar{y} + \bar{y} - \hat{y}_i). \tag{12.15}$$

Because $\sum y_i = n\bar{y}$ we have

$$\sum_{i=1}^{n} (y_i - \bar{y}) = 0 \tag{12.16}$$

and, similarly,

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0. \tag{12.17}$$

Combining (12.13) with (12.17) gives

$$\sum_{i=1}^{n} (\hat{y}_i - \bar{y}) = 0. \tag{12.18}$$

Finally, using (12.16) with (12.18) in (12.15) gives (12.14). □

It is worth drawing attention to one other interesting feature of the linear regression model. While (12.1) and (12.4) emphasize potential nonlinearity in the way a variable x, or multiple variables $x_1, \ldots, x_p$ may influence y, it turns out that linear regression may be used to fit some nonlinear relationships. This is discussed in Section 12.5.4. Here is a particularly simple, yet important additional example.

Example 12.2 BOLD hemodynamic response in fMRI In Fig. 1.3 of Example 1.3 we displayed fMRI images from a single subject during a simple finger-tapping task in response to a visual stimulus. As we said there, fMRI detects changes in

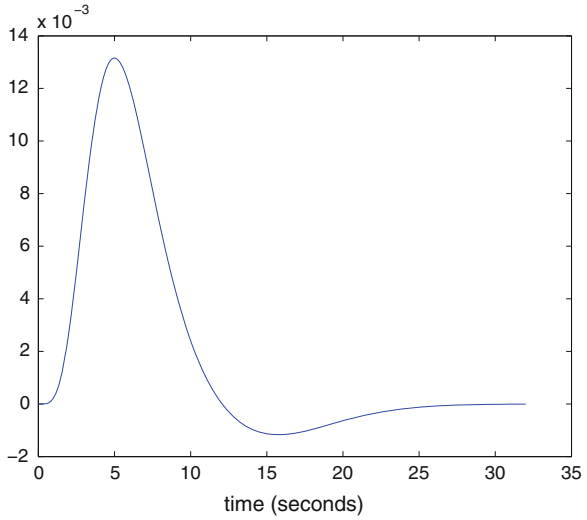

Fig. 12.3 The hemodynamic response function defined by Eq. (12.19).

blood oxygenation and the measurement is known as the BOLD signal, for Blood Oxygen-Level Dependent signal. The typical hemodynamic response that produces the signal is relatively slow, lasting roughly 20 s (seconds). Many experiments have shown, however, that it has a reasonably stable form (see Glover 1999). Software for analyzing fMRI data, such as BrainVoyager (see Goebel et al. 2006; Formisano et al. 2006), often uses a particular hemodynamic function. Figure 12.3 displays a plot of such a theoretical hemodynamic response function $h(t)$ defined by

$$h(t) = \left(\frac{t}{d_1}\right)^{a_1} \exp(-\frac{t-d_1}{b_1}) - c\left(\frac{t}{d_2}\right)^{a_2} \exp(-\frac{t-d_2}{b_2}) \tag{12.19}$$

where $a_1, b_1, d_1, a_2, b_2, d_2$ and c are parameters that have default values in the software. Using this function the fMRI data at a particular voxel (a particular small rectangular box in the brain) may be analyzed using linear regression. Let us suppose we have an on/off stimulus, as is often the case, and let $u_j = 1$ when the stimulus is on and 0 otherwise, $j = 1, \ldots, T$. The effect at time i of the stimulus being on at time j is assumed to follow the hemodynamic response function, i.e., the effect is determined by $h(t)$ where $t = i - j$ is the delay between the stimulus and the response time i. It is also assumed that the effects of multiple "on" stimuli at different times j produce additive effects at different time lags $i - j$. Therefore, the total stimulus effect at time i is[2]

$$x_i = \sum_{j<i} h(i-j)u_j. \tag{12.20}$$

[2] This expression is known as the *convolution* of the hemodynamic response function $h(t)$ with the stimulus function u_j.

The linear regression model (12.5) may then be fitted, and the coefficient β_1 represents the overall magnitude of the increased BOLD response due to the activity associated with the stimulus. □

12.1.1 *Linear regression assumes linearity of $f(x)$ and independence of the noise contributions at the various observed x values.*

The model (12.1) is *additive* in the sense that it assumes the noise, represented by ϵ_i is added to the function value $f(x_i)$ to get Y_i. This entails a *theoretical* relationship between x and y that holds except for the "errors" ϵ_i. Linear regression further specializes by taking $f(x)$ to be linear as in (12.2) so that we get the model (12.5). The ϵ_i's are assumed to satisfy

$$E(\epsilon_i) = 0$$

for all i, so that $E(Y_i) = \beta_0 + \beta_1 x_i$. In words, the linear relationship $y = \beta_0 + \beta_1 x$ is assumed to hold "on average," that is, apart from errors that are on average zero. Additivity of the errors and linearity of $E(Y_i)$ are the most fundamental assumptions of linear regression. In addition, the errors ϵ_i are assumed to be independent of each other. In Section 12.2.3 we show how lack of independence can distort statistical inferences about the regression model. The independence assumption may be violated when observations are recorded sequentially across time, in which case more elaborate *time series* methods are needed. These are discussed in Chapter 18.

Important, though less potentially problematic, additional assumptions are that the variances of the ϵ_i's are all equal, so that the variability of the errors does not change with the value of x, and that the errors are normally distributed. These latter two assumptions guarantee that the 95 % confidence intervals discussed in Section 12.3.1 have the correct probability .95 of covering the coefficients and the significance tests in Section 12.3.2 have the correct p-values. In sufficiently large samples the normality assumption becomes unnecessary, as the confidence intervals and significance tests will be valid, approximately (see (12.37)).

To summarize, the assumptions of linear regression may be enumerated, in order of importance, as follows:

(i) the linear regression model (12.5) holds;
(ii) the errors satisfy $E(\epsilon_i) = 0$ for all i;
(iii) the errors ϵ_i are independent of each other;
(iv) $V(\epsilon_i) = \sigma^2$ for all i (homogeneity of error variances), and
(v) $\epsilon_i \sim N(0, \sigma^2)$ (normality of the errors).

To repeat, the crucial assumptions are the first three: there is, on average, a linear relationship between Y and x, and the deviations from it are represented by independent errors.

12.1.2 The relative contribution of the linear signal to the total response variation is summarized by R^2.

As shown in Fig. 12.2, in Example 1.5 linear regression provides a very good representation of the relationship between x and y, with the points clustering tightly around the line. In other cases there is much more "noise" relative to "signal," meaning that the (x_i, y_i) values scatter more widely, so that the residuals tend to be much larger. In this section we describe two measures of residual deviation.

The error standard deviation σ (see item (iv) in the assumptions in Section 12.1.1) represents the average amount of deviation of each ϵ_i from zero. Thus, σ tells us how far off, on average, we would expect the line to be in predicting a value of y at any given x_i. It is estimated by $s = \sqrt{s^2}$ where

$$s^2 = \frac{1}{n-2} SSE \tag{12.21}$$

and

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{12.22}$$

is the *sum of squares for error* or the *residual sum of squares*. (Here $\hat{y}_i$ is defined by (12.10).) The variance estimate s^2 is then also called the *residual mean squared error* and we often write

$$MSE = s^2. \tag{12.23}$$

This definition of s makes it essentially the standard deviation of the residuals, except that $n-2$ is used in the denominator instead of $n-1$; here there are two parameters β_0 and β_1 being estimated so that two degrees of freedom are lost from n, rather than only one.

The other quantity, R^2, is interpreted as the fraction of the variability in Y that is attributable to the regression, as opposed to error. We begin by defining the *total sum of squares*

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2. \tag{12.24}$$

This represents the overall variability among the y_i values. We then define

$$R^2 = 1 - \frac{SSE}{SST}. \tag{12.25}$$

The fraction SSE/SST is the proportion of the variability in Y that is attributable to error, and R^2 is what's left over, which is attributable to the regression line. The value of R^2 is between 0 and 1. It is 0 when there is no linear relationship and 1 when there is a perfect linear relationship. If we define the *sum of squares due to regression* as the difference

$$SSR = SST - SSE \tag{12.26}$$

then we can re-write R^2 in the form

$$R^2 = \frac{SSR}{SST}. \tag{12.27}$$

From this version we get the interpretation of R^2 as "the proportion of variability of Y that is explained by X." In different terminology, we may think of SSR as the *signal* variability (often called "the variability due to regression") and SSE as the *noise* variability. Then $R^2 = SSR/(SSR + SSE)$ becomes the relative proportion of signal-to-noise variability. (The ratio of signal-to-noise variabilities[3] would be SSR/SSE.)

In (12.26) we defined the sum of squares due to regression by subtraction. There is a different way to define it, so that we may see how total variability (SST) is decomposed into regression (SSR) and error components (SSE). The derivation begins with the values $y_i, \hat{y}_i$, and $\bar{y}$, as shown in Fig. 12.2, where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Writing $y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$, we have

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

but after plugging in the definition of $\hat{y}_i$ from (12.10) some algebra shows that the cross-product term vanishes and, defining

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2, \tag{12.28}$$

we have

$$SST = SSR + SSE. \tag{12.29}$$

As we mention again in Section 12.5.3, the vanishing of the cross-product may be considered, geometrically, to be a consequence of the Pythagorean theorem. Equation (12.29) is important in understanding linear regression and analysis of variance: we think of the total variation as coming from different additive components, whose magnitudes we compare.

The estimated standard deviation s has the units of Y and is therefore interpretable —at least to the extent that the Y measurements themselves are interpretable. But R^2 is dimensionless. Unfortunately, there are no universal rules of thumb as to what constitutes a large value: in some applications one expects an R^2 of at least .99 while

[3] The signal-to-noise ratio is a term borrowed from engineering, where it refers to a ratio of the power for signal to the power for noise, and is usually reported in the log scale; under certain stochastic models it translates into a ratio of signal variance to noise variance.

in other applications an R^2 of .40 or less would be considered substantial. One gets a feeling for the size of R^2 mainly by examining, and thinking about, many specific examples.

12.1.3 *Theory shows that if the model were correct then the least-squares estimate would be likely to be accurate for large samples.*

In presenting the assumptions on p. 315 we noted that they were listed in order of importance and, in particular, normality of the errors is not essential. The following theoretical result substantiates the validity of least-squares for non-normal errors in large samples.

Theorem: Consistency of least squares estimators For the linear regression model (12.5) suppose conditions (i)–(iv) hold and let $x_1, x_2, \ldots, x_n, \ldots$ be a sequence of x values such that

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 \to \infty \tag{12.30}$$

as $n \to \infty$. Then the least-squares estimator defined by (12.7) satisfies

$$\begin{aligned} \hat{\beta}_1 &\xrightarrow{P} \beta_1 \\ \hat{\beta}_0 &\xrightarrow{P} \beta_0. \end{aligned} \tag{12.31}$$

In other words, under these conditions $\hat{\beta}_1$ and $\hat{\beta}_0$ are consistent estimators of β_1 and β_0.

Proof: This is essentially a consequence of the law of large numbers in a non-i.i.d. setting, where linear combinations of the Y_i values are being used according to (12.7) and (12.8). We omit the proof and refer the interested reader to Wu (1981), which examines a more general problem but provides extensive references and discussion. □

Note that to fit a line we must have at least 2 distinct values, so that not every observation can be made at the same x value. The condition (12.30) fails when, for all sufficiently large i and j, $x_i = x_j$. In other words, it rules out degenerate cases where essentially all the observations (i.e., all but finitely many of them) are made at a single x value.[4] We may interpret this asymptotic statement as saying that for all situations in which there is any hope of fitting a line to the data, as the sample size increases the least-squares estimator of the slope will converge to the true value.

[4] In fact, the results cited in Wu (1981) show that (12.30) is necessary and sufficient for (12.31).

12.2 Checking Assumptions

12.2.1 Residuals should represent unstructured noise.

In examining single batches of data, in Chapter 2, we have seen how the data may be used not only to estimate unknown quantities (there, an unknown mean μ) but also to check assumptions (in particular, the assumption of normality). This is even more important in regression analysis and is accomplished by analyzing the residuals defined in (12.11). Sometimes the residuals are replaced by *standardized residuals*. The ith standardized residual is $e_i/SD(e_i)$, where $SD(e_i)$ is the standard deviation of e_i. Dividing by the standard deviation puts the residuals on a familiar scale: since they are supposed to be normal, about 5 % of the standardized residuals should be either larger than 2 or smaller than -2. Standardized residuals that are a lot larger than 2 in magnitude might be considered outliers.

> *A detail:* There are two different ways to standardize the residuals. We have here taken $SD(e_i)$ to be the estimated standard deviation of e_i. The formula for $SD(e_i)$ involves the x_i values. An alternative would be to compute the sample variance of the residuals
>
> $$s_e^2 = \frac{1}{n-1}\sum(e_i - \bar{e})^2$$
>
> and take its square root. The standardization using $SD(e_i)$, which allows the n residual standard deviations to be different, is often called *studentization* (by analogy with the ratio that defines Student's t distribution, see p. 129). The statistical software packages we are most familiar with use $SD(e_i)$ to standardize the residuals. □

Two kinds of plots are used. Residual versus fit plots are supposed to reveal (i) nonlinearity, (ii) inhomogeneity variances, or (iii) outliers. Plots having structure of the kind that would indicate these problems are shown in Fig. 12.4. The first plot is typical of data with no systematic variation remaining after linear regression: the pattern is "random," specifically, it is consistent with errors that are independent and normally distributed, all having the same distribution. The second plot shows departure from linearity; the third indicates more variability for large fitted values than for smaller ones. The last plot has an outlier, indicating a point that is way off the fitted line.

Histograms and Q-Q plots of the residuals are also used to assess assumptions. These are supposed to (i) reveal outliers and (ii) check whether the errors may be described, at least approximately, by a normal distribution.

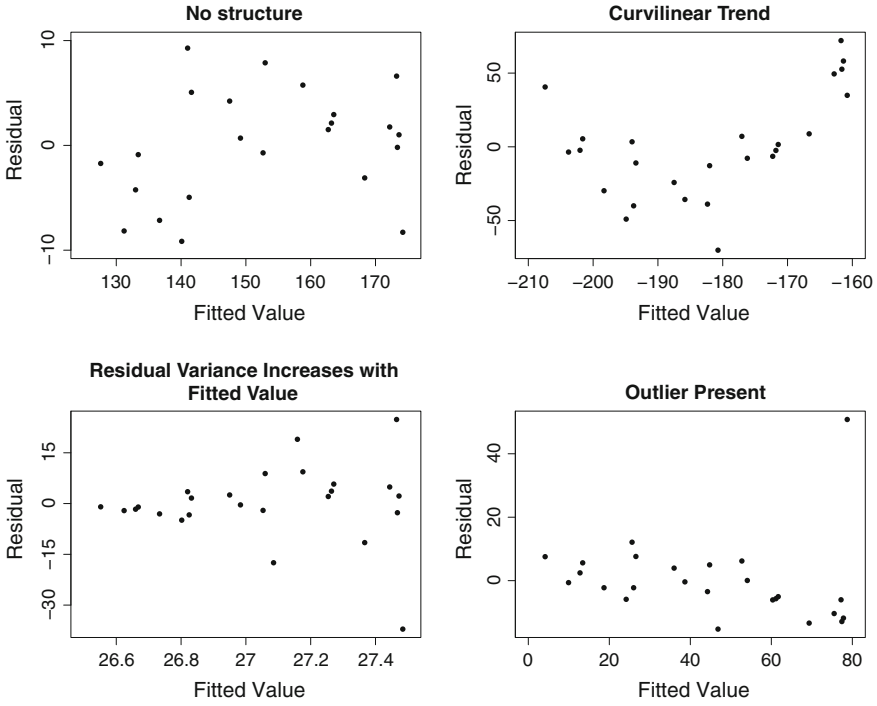

Fig. 12.4 Residual plots: the *Top Left* plot depicts unstructured noise while the latter three reveal structure, and thus deviations from the assumptions.

12.2.2 Graphical examination of (x, y) data can yield crucial information.

As we tried to emphasize in Chapters 1 and 2, it is important to examine data with exploratory methods, using visual summaries where possible. The following illustration gives a nice demonstration of how things can go wrong if one relies solely on the simplest numerical summaries of least-squares regression.

Illustration Figure 12.5 shows a striking example in which four sets of data all have the same regression equation and R^2, but only in the first case (data set 1) would the regression line appropriately summarize the relationship. In the second case (data set 2) the relationship is clearly nonlinear, in the third case there is a big outlier and removing it dramatically changes the regression. In the fourth case the slope of the line is determined entirely by the height of the point to the right of the graph; therefore, since each point is subject to some random fluctuation, one would have to be very cautious in drawing conclusions. □

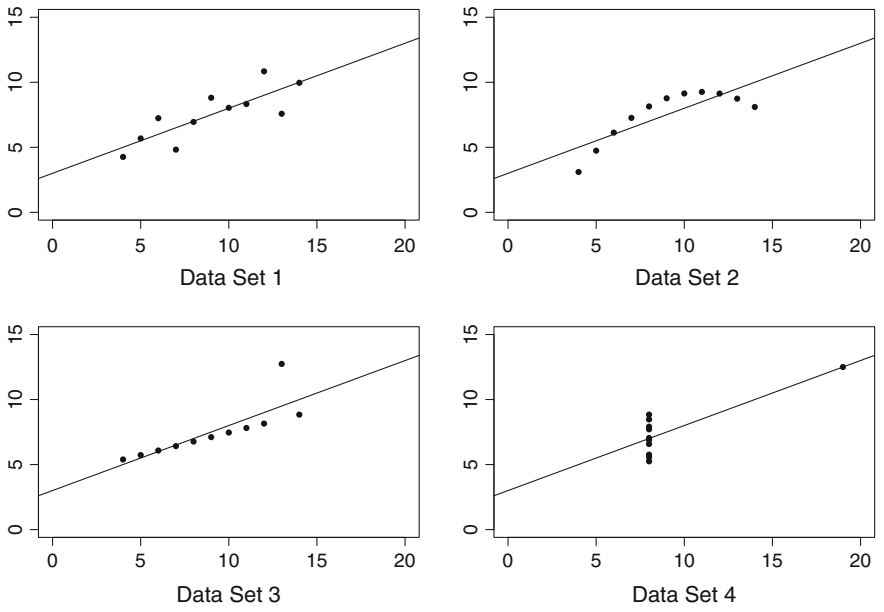

Fig. 12.5 Plots of four very different data sets all having the same fitted regression equation $Y = 3 + .5x$ and $R^2 = .667$. These were discussed in Anscombe (1973).

This illustration underscores the value of plotting the data when examining linear or curvilinear relationships.

12.2.3 Failure of independence among the errors can have substantial consequences.

In stating the assumptions of linear regression on p. 315 we stressed the importance of independence among the errors ϵ_i. To be concrete, we now consider how inference about the strength of the linear relationship between $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$, as measured by R^2, can be badly misled when the data are correlated. To do this we use a simple model of serial dependence: we put

$$U_t = \rho U_{t-1} + \delta_t \tag{12.32}$$

$$W_t = \rho W_{t-1} + \eta_t \tag{12.33}$$

for $t = 2, 3, \ldots, n$ where

$$\delta_t \sim N(0, 1)$$
$$\eta_t \sim N(0, 1)$$
$$U_1 \sim N(0, 1)$$
$$W_1 \sim N(0, 1)$$

all independently of each other. Models (12.32) and (12.33) are both examples of first-order *autoregressive models*, which we discuss further in Chapter 18, with *autocorrelation coefficient* ρ. According to these models the observations U_t and W_t are likely to be close to the respective values U_{t-1} and W_{t-1}, but with noise added. The variation in experimental data observed across time may often be described well using autoregressive models. Note that U_t and W_t are independent for all t. We simulate values $u_1, \ldots, u_n$ and $w_1, \ldots, w_n$ from (12.32) and (12.33), using $n = 100$, and we then define

$$x_i = u_i$$
$$y_i = w_i$$

for $i = 1, \ldots, n$ and compute R^2 from the regression of y on x. We could say that the correct linear model in this case is

$$Y_i = \epsilon_i$$

where ϵ_i follows the autoregressive model (12.33), so that in principle we should find $R^2 = 0$. Figure 12.6 gives the results from 100 simulations (each using $n = 100$). When the autocorrelation coefficient is zero, we get values of R^2 that deviate from 0 according to the null distribution so that about 5 % of the values are above the threshold corresponding to $p < .05$ and about 1 % of the values are above the threshold corresponding to $p < .01$. However, as the magnitude of the autocorrelation coefficient increases we find many values of R^2 that are substantial, many more than would be predicted by the null distribution—thus, the p-values are no longer accurate. In fact, for magnitudes of the autocorrelation that are close to 1 it becomes highly probable to get what would look like a "significant" correlation in the data, even though the x and y data were computer-generated to be independent.

This phenomenon may be appreciated further by contrasting the variation in independent normal data with data generated from model (12.32) with $\rho = .9$. As seen in the right-hand side of Fig. 12.7, data following this autoregressive model tend to have patches of values that are all either above 0 or below 0. If we imagine two such series, there are likely be patches of time where both series are very different from 0 and this will often lead to a substantial magnitude of the correlation coefficient computed across time.

The point is that one must be very careful about the assumption of independence in linear regression. When regression or correlation analysis is to be performed on data recorded across time, where dependence among errors is likely, the standard advice is to first *pre-whiten* the data by removing temporal structure (for instance, by fitting auto-regressive models and then analyzing the residuals) as discussed in Section 18.5.2.

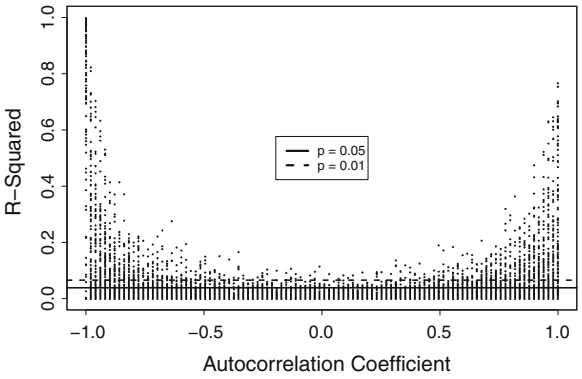

Fig. 12.6 Values of R^2 based on truly independent Y and X data that were simulated using (12.32) and (12.33), with $n = 100$. The x-axis of the plot gives the value of the autocorrelation coefficient ρ. The usual p-values, obtained from applying the t-distribution to (12.38), accurately represent the probability of deviation as large as the observed R^2 only when $\rho = 0$.

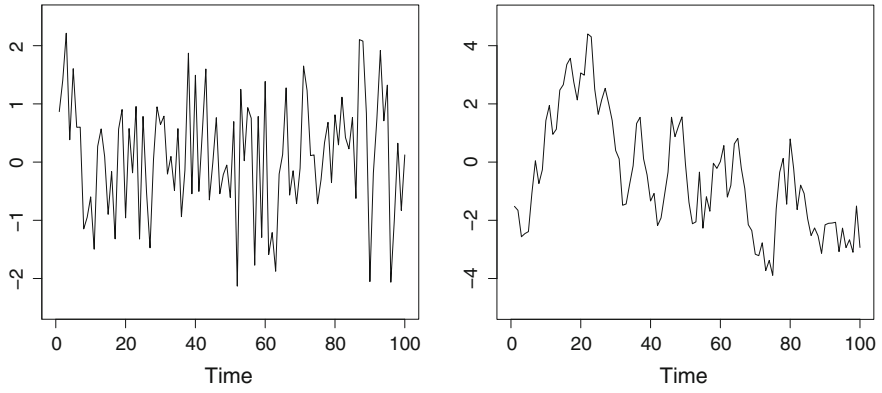

Fig. 12.7 Plots of artificial data against a variable representing time, which takes on values $1, 2, \ldots, 100$. The data values have been connected with lines. *Left* plot of 100 independent $N(0, 1)$ random values. *Right* plot of 100 values from an autoregressive model, as in (12.32) with $\rho = .9$. The independent values fluctuate without trends, while the autoregressive values show excursions of several successive values that are consistently positive or negative.

12.3 Evidence of a Linear Trend

12.3.1 Confidence intervals for slopes are based on SE, according to the general formula.

When reporting least-squares estimates, standard errors should also be supplied. That is, one reports either $\hat{\beta}_1 \pm SE(\hat{\beta}_1)$ or a confidence interval. Standard errors are given as standard output from regression software. The general formula for standard errors

in linear regression appears in Eq. (12.61). To get an approximate 95 % confidence interval for β_1 based on $\hat{\beta}_1$ and $SE(\hat{\beta}_1)$, we again use the general form given by (7.8), i.e.,

$$\text{approx. } 95\,\%\,\text{CI} = (\hat{\beta}_1 - 2 \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot SE(\hat{\beta}_1)). \tag{12.34}$$

An alternative, in small samples, is analogous to the small sample procedure in (7.31) used to estimate a population mean: we substitute for 2 the value $t_{.975,\nu}$, where now $\nu = n - 2$ because we have estimated two parameters (intercept and slope) and thus have lost two degrees of freedom. Thus, we would use the formula

$$95\,\%\,\text{CI} = (\hat{\beta}_1 - t_{.025,n-2} \cdot SE(\hat{\beta}_1), \hat{\beta}_1 + t_{.025,n-2} \cdot SE(\hat{\beta}_1)). \tag{12.35}$$

Example 1.5 (continued, see p. 11) Using least squares regression we found $\hat{\beta}_1 = 6.07$ and $SE(\hat{\beta}_1) = .14$. We would report this by saying that, on average, action potential velocity increases by $6.07 \pm .14$ m/s for every micron increase in diameter of a neuron. Applying (12.34), an approximate 95 % CI for the slope of the regression line is $6.07 \pm 2(.14)$ or (5.79, 6.35). For these data there were $n = 67$ observations, so we have $\nu = 65$ and $t_{.975,n-1} = 2.0$. Thus, the CI based on (12.35) is the same as that based on (12.34). □

Formula (12.34) may be justified by an extension of the theorem on the consistency of $\hat{\beta}_1$ in (12.31), which we present next.

Theorem: Asymptotic normality of least squares estimators For the linear regression model (12.5) suppose conditions (i)–(iv) hold and let $x_1, x_2, \ldots, x_n, \ldots$ be a sequence of x values such that

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \to c \tag{12.36}$$

for some positive constant c, as $n \to \infty$. Then the least-squares estimator defined by (12.7) satisfies

$$\begin{aligned} \frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_1)} &\xrightarrow{D} N(0, 1) \\ \frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} &\xrightarrow{D} N(0, 1) \end{aligned} \tag{12.37}$$

where $SE(\hat{\beta}_1)$ and $SE(\hat{\beta}_0)$ are the standard errors given by (12.61).

Proof: This is a consequence of the CLT, but requires some algebraic manipulation. We omit the proof and again refer the interested reader to Wu (1981) for references. □

The condition (12.36) implies (12.30). It would be satisfied if we were drawing x_i values from a fixed probability distribution.[5] In the context of a particular set of data, the x_i values, even when selected by an experimenter, are somehow spread out and thus could be conceived as coming from some probability distribution (one that is not concentrated on a single value). On the other hand, the Anscombe example in Section 12.2.2 is a reminder that sensible interpretations require the fitted line to represent well the relationship between the x_i and y_i values. In the theoretical world this is expressed by saying that the model assumptions (i)–(iv) are satisfied. In practice we would interpret the theorems guaranteeing consistency and asymptotic normality of least-squares estimators, according to (12.31) and (12.37), as saying that if the regression model does a good job in describing the variation in the data, and the sample size is not too small, then the approximate confidence interval in (12.34) will produce appropriate inferences. We typically do not need normality of the errors, as specified in assumption (v). What we need is normality of the estimator, as in (12.37).

12.3.2 Evidence in favor of a linear trend can be obtained from a t-test concerning the slope.

In Examples 1.5 and 12.1 it is obvious that there are linear trends in the data. This kind of increasing or decreasing tendency is sometimes a central issue in an analysis. Indeed, in Example 12.1 the quantitative relationship, meaning the number of additional spikes per second per additional drop of juice, is not essential. Rather, the main conclusion involved the qualitative finding of increasing firing rate with increasing reward. In problems such as this, it makes sense to assume that y is roughly linear in x but to consider the possibility that in fact the slope of the line is zero—meaning that y is actually constant, on average, as x changes; that is, that y is really not related to x at all. We formalize this possibility as the null hypothesis H_0: $\beta_1 = 0$ and we test it by applying the z-test discussed in Section 10.3.2. In the one-sample problem of testing H_0: $\mu = \mu_0$, considered in Section 10.3.3, the z-test is customarily replaced by a t-test, which inflates the p-value somewhat for small samples and is justified under the assumption of normality of the data. Similarly, in linear regression, the z-test may be replaced by a t-test under the assumption of normality of errors (assumption (v) on p. 315). The test statistic becomes the t-ratio,

$$t\text{-ratio} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}. \tag{12.38}$$

For large samples, under H_0, this statistic has a $N(0, 1)$ distribution, but for small samples, if assumption (v) is satisfied, under H_0 the t-ratio has a t distribution on $\nu = n - 2$ degrees of freedom. This is the basis for the p-value reported by most

[5] Beyond (12.30), condition (12.36) says that the x_i values do not diverge extremely quickly, which would make $\hat{\beta}_1$ converge faster than $1/\sqrt{n}$.

statistical software. Here, the degrees of freedom are $n - 2$ because two parameters β_1 and β_0 from n freely ranging data values y_i. Generally speaking, when the magnitude of the t-ratio is much larger than 2 the p-value will be small (much less than .05, perhaps less than .01) and there will be clear evidence against H_0: $\beta_1 = 0$ and in favor of the existence of a linear trend.

Example 1.5 (continued, see page 11) For the conduction velocity data, testing H_0: $\beta_1 = 0$ with (12.38) we obtained $p < 10^{-15}$. Keeping in mind that very extreme tail probabilities are not very meaningful (they are sensitive to small departures from normality of the estimator) we would report this result as very highly statistically significant with $p << .0001$, where the notation $<<$ is used to signify "much less than." □

Example 12.1 (continued from p. 310) For the data shown in Fig. 12.1 the authors reported $p < .0001$. □

In the data reported in Fig. 12.1 there are only 7 distinct values of x_i, with many firing rates (across many trials) corresponding to each reward level. Thus, the 329 data pairs have been aggregated to 7 pairs with the mean value of y_i reported for each x_i. It turns out that the fitted line based on means is the same as the fitted line based on all 329 values considered separately. However, depending on the details of the way the computation based on the means is carried out, the standard error may or may not agree with the standard error obtained by analyzing all 329 values. To capture the full regularity and variation in the data, the hypothesis test should be based on all 329 values.

12.3.3 The fitted relationship may not be accurate outside the range of the observed data.

We have so far ignored an interesting issue that arises in Example 1.5. There, the fitted line does not go through the origin $(0, 0)$. In fact, according to the fitted line, when the diameter of the nerve is 0, the conduction velocity becomes negative! Should we try to fix this?

It is possible to force the line through $(0, 0)$ by omitting the intercept in the fitting process. Regression software typically provides an option for leaving out the intercept. However, for this data set, and for many others, omission of the intercept may be unwise. The reason is that the relationship may well be nonlinear near the origin, and there are no data to determine the fitted relationship in that region. Instead, we would view the fitted relationship as accurate only for diameters that are within the range of values examined in the data. Put differently, when the linear regression model does a good job of representing the regularity and variability in the data it allows us to interpolate (predict values within the range of the data) but may not be trustworthy if we try to extrapolate (predict values outside the range of the data).

12.4 Correlation and Regression

Sometimes the "explanatory variable" x is observed, rather than fixed by the experimenter. In this case the pair (x, y) is observed and we may model this by considering a pair of random variables X and Y and their *joint* distribution. Recall (from Section 4.2.1) that the *correlation coefficient* ρ is a measure of linear association between X and Y. As we discussed in Section 4.2.1, the best linear predictor $\beta_0 + \beta_1 X$ of Y satisfies

$$\beta_1 = \frac{\sigma_Y}{\sigma_X} \cdot \rho \tag{12.39}$$

as in Eq. (4.9). Also, the theoretical regression of Y on X is defined (see Section 4.2.4) to be $E(Y|X = x)$, which is a function of x, and it may happen that this function is linear:

$$E(Y|X = x) = \beta_0 + \beta_1 x.$$

In Chapter 4 we noted that the regression is, in fact, linear when (X, Y) has a bivariate normal distribution and then (12.39) holds. This linearity, and its interpretation, was illustrated in Fig. 4.3. However, the right-hand plot in Fig. 4.3 concerns data, rather than a theoretical distribution, and there is an analogous formula and interpretation using the sample correlation r, which was defined in (4.7). Under the assumption of bivariate normality, it may be shown that the sample correlation r is the MLE of ρ.

The sample correlation is related to the relative proportion of signal-to-noise variability R^2 by $R^2 = r^2$. Important properties are the following:

- $-1 \le r \le 1$ with $r = 1$ when the points fall exactly on a line with positive slope and $r = -1$ when the points fall exactly on a line with negative slope;
- the value of r is unitless and does not change when either or both of the two variables are linearly rescaled (e.g., when x is replaced by $ax + b$);
- just as ρ measures linear association between random variables X and Y, so too may r be considered a measure of *linear* association.

As we said in discussing R^2, there are no general guidelines as to what constitutes a "large" value of the correlation coefficient. Interpretation depends on the application.

12.4.1 The correlation coefficient is determined by the regression coefficient and the standard deviations of x and y.

Equation (12.39) gives the relationship of the theoretical slope β_1 to the theoretical correlation coefficient ρ. For data pairs (x_i, y_i) we have the analogous formula

$$\hat{\beta}_1 = \frac{s_Y}{s_X} \cdot r.$$

As a consequence, if x and y have about the same variability, the fitted regression slope becomes approximately equal to the sample correlation. In some contexts it is useful to standardize x and y by dividing each variable by its standard deviation. When that is done, the regression slope will equal the sample correlation.

12.4.2 Association is not causation.

There are numerous examples of two variables having a high correlation while no one would seriously suggest that high values of one causes high values of the other. For instance, one author (Brownlee 1965) looked at data from many different countries and pointed out that the number of telephones per capita had a strong correlation with the death rate due to heart disease. In such situations there are confounding factors that, presumably, have an effect on both variables and thus create a "spurious" correlation. Only in well-performed experiments, often using randomization,[6] can one be confident there are no confounding factors. Indeed, discussion sections of articles typically include arguments as to why possible confounding factors are unlikely to explain reported results.

12.4.3 Confidence intervals for ρ may be based on a transformation of r.

The sample correlation coefficient r may be considered an estimate of the theoretical correlation ρ and, as we mentioned on p. 327, under the assumption of bivariate normality r is the MLE of ρ. To get approximate confidence intervals the large-sample theory of Section 8.4.3 may be applied.[7] If we have a random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ we may compute its sample correlation R_n, which is itself a random variable (so that when $X_1 = x_1, Y_1 = y_1, \ldots, X_n = x_n, Y_n = y_n$ we compute the sample correlation $R_n = r$ based on $(x_1, y_1), \ldots, (x_n, y_n)$). Now, if we consider a sequence of such samples from a bivariate normal distribution with correlation ρ it may be shown that

$$\frac{\sqrt{n}(R_n - \rho)}{(1 - \rho^2)} \xrightarrow{D} N(0, 1)$$

[6] Randomization refers to the random assignment of treatments to subjects, and to the process of randomly ordering treatment conditions; we discuss this further in Section 13.4.

[7] The usual derivation of the limiting normal distribution of r begins with an analytic calculation of the covariance matrix of (V_x, V_y, C) where $V_x = V(X)$, $V_y = V(Y)$, and $C = Cov(X, Y)$, in which (X, Y) is bivariate normal. That calculation provides an explicit formula for the covariance matrix in the limiting joint normal distribution of (V_x, V_y, C), and then propagation of uncertainty is applied as in Section 9.1.2.

as $n \to \infty$. This limiting normal distribution could be used to find confidence intervals. However, Fisher (1924) showed that a transformation of the correlation $R_n = r$ improves the limiting normal approximation. This is known as *Fisher's z transformation* (z because it creates a nearly $N(0, 1)$ distribution) defined by

$$z_r = \frac{1}{2} \log \left(\frac{1+r}{1-r} \right). \tag{12.40}$$

For the theoretical statement we again consider a sequence of bivariate normal random samples with sample correlations R_n and define

$$Z_R = \frac{1}{2} \log \left(\frac{1+R_n}{1-R_n} \right)$$

and

$$\zeta = \frac{1}{2} \log \left(\frac{1+\rho}{1-\rho} \right)$$

to get

$$\sqrt{n-3}(Z_R - \zeta) \xrightarrow{D} N(0, 1) \tag{12.41}$$

as $n \to \infty$ (see[8] p. 43 in DasGupta 2008). Consequently, we can define the lower and the upper bounds of an approximate 95 % confidence interval for the theoretical quantity ζ by

$$\begin{aligned} L_z &= z_r - 2\sqrt{\frac{1}{n-3}} \\ U_z &= z_r + 2\sqrt{\frac{1}{n-3}}. \end{aligned} \tag{12.42}$$

To get an approximate 95 % confidence interval for ρ we apply the inverse transformation

$$\rho = \frac{\exp(2\zeta) - 1}{\exp(2\zeta) + 1}$$

to L and U in (12.42) to get

$$\begin{aligned} L &= \frac{\exp(2L_z) - 1}{\exp(2L_z) + 1} \\ U &= \frac{\exp(2U_z) - 1}{\exp(2U_z) + 1}. \end{aligned} \tag{12.43}$$

[8] The z-transformation may be derived as a variance-stabilizing transformation, as on p. 232, beginning with the limiting result mentioned in footnote 7. More general results are given by Hawkins (1989).

Confidence interval for ρ
Suppose we have a random sample from a bivariate normal distribution with correlation ρ and $R_n = r$ is the sample correlation. Then an approximate 95 % confidence interval for ρ is given by (L, U) where L and U are defined by (12.43), (12.42), and (12.40).

The result (12.41) may also be used to test H_0:$\rho = 0$, which holds if and only if H_0: $\beta_1 = 0$. The procedure is to apply the z-test in Section 10.3.2 using

$$z_{obs} = \sqrt{n-3}z_r,$$

which is z_r divided by its large-sample standard deviation $1/\sqrt{n-3}$, and is thus a z-ratio.

12.4.4 When noise is added to two variables, their correlation diminishes.

When measurements are corrupted by noise, the magnitude of their correlation decreases. The precise statement is given in the theorem below, where we begin with two random variables U and W and then add noise to each, in the form of variables ϵ and δ. The noise-corrupted variables are then $X = U + \epsilon$ and $Y = W + \delta$.

Theorem: Attenuation of Correlation Suppose U and W are random variables having correlation ρ_{UW} and ϵ and δ are independent random variables that are also independent of U and V. Define $X = U + \epsilon$ and $Y = W + \delta$, and let ρ_{XY} be the correlation between X and Y. If $\rho_{UW} > 0$ then

$$0 < \rho_{XY} < \rho_{UW}.$$

If $\rho_{UW} < 0$ then

$$\rho_{UW} < \rho_{XY} < 0.$$

Proof details: We assume that $V(\epsilon) > 0$ and $V(\delta) > 0$ and we begin by writing

$$\begin{aligned} Cov(X, Y) &= Cov(U + \epsilon, W + \delta) \\ &= Cov(U, W) + Cov(U, \delta) + Cov(W, \epsilon) + Cov(\epsilon, \delta). \end{aligned}$$

Because of independence the last 3 terms above are 0. Therefore, $Cov(X, Y) = Cov(U, W)$, which shows that ρ_{XY} and ρ_{UW} have the same sign. Suppose $\rho_{UW} > 0$, so that $Cov(U, W) > 0$. Then we have

$$\begin{aligned}\rho_{XY} &= Cor(U+\epsilon, W+\delta)\\ &= \frac{Cov(U,W)}{\sqrt{V(U+\epsilon)V(W+\delta)}}\\ &= \frac{Cov(U,W)}{\sqrt{(V(U)+V(\epsilon))(V(W)+V(\delta))}}\\ &< \frac{Cov(U,W)}{\sqrt{Var(U)Var(W)}}\\ &= \rho_{UW}.\end{aligned}$$

If $\rho_{UW} < 0$ then $Cov(U, W) < 0$ and the inequality above is reversed. □

The theorem above indicates that when measurements are subject to substantial noise a measured correlation will underestimate the strength of the actual correlation between two variables. In the notation above, we wish to find ρ_{UW} but the corrupted measurements we observe would be $(X_1, Y_1), \ldots, (X_n, Y_n)$, and if we compute the sample correlation r based on these observations it will tend to be smaller than ρ_{UW} even for large samples. Thus, it is often the case that the sample correlation will underestimate an underlying correlation between two variables. However, if the *likely magnitude* of the noise is known it becomes possible to correct the estimate. Such corrections for attenuation of the correlation can be consequential.

Example 12.3 Correction for attenuation of the correlation in SEF selectivity indices Behseta et al. (2009) reported analysis of data from an experiment on neural mechanisms of serial order performance. Monkeys were trained to perform eye movements in a given order signaled by a cue. For example, one cue carried the instruction: look up, then right, then left. Based on recordings of neural activity in frontal cortex (the supplementary eye field, SEF) during task performance, Behseta et al. reported that many neurons fire at different rates during different stages of the task, with some firing at the highest rate during the first, some during the second and some during the third stage. These rank-selective neurons might genuinely be sensitive to the monkey's stage in the sequence. Alternatively, they might be sensitive to some correlated factor. One such factor is expectation of reward. Reward (a drop of juice) was delivered only after all three movements had been completed. Thus as the stage of the trial progressed from one to three, the expectation of reward might have increased.

To see whether rank-selective neurons were sensitive to the size of the anticipated reward, the same monkeys were trained to perform a task in which a visual cue presented at the beginning of the trial signaled to the monkey whether he would receive one drop or three drops of juice after a fixed interval. The idea was that neuronal activity related to expectation of reward would be greater after the promise of three drops than after the promise of one. Spike counts from 54 neurons were collected during the performance of both the serial order task and the variable reward task, and selectivity indices for rank in the serial order task and size of the anticipated reward in the variable reward task were computed. The rank selectivity index was $I_{\text{rank}} =$

$\frac{(f_3-f_1)}{(f_3+f_1)}$, where f_1 and f_3 were the mean firing rates measured at the times of the first and third saccades respectively, the mean being taken across trials. Similarly, the reward selectivity index was $I_{\text{reward}} = \frac{(f_b-f_s)}{(f_b+f_s)}$ where f_b and f_s were the mean firing rates during the post-cue delay period on big-reward and small-reward trials respectively. The selectivity indices I_{rank} and I_{reward} turned out to be positively correlated, but the effect was smaller than expected, with $r = .49$. The correlation between the rank and reward indices was expected to be larger because, from previous research, it was known that (a) the expectation of reward increases over the course of a serial order trial and (b) neuronal activity in the SEF is affected by the expectation of reward. Behseta et al. speculated that the correlation between the two indices had been attenuated by noise arising from trial-to-trial variations in neural activity, and they applied a correction for attenuation discussed in Chapter 16. This gave a dramatically increased correlation, with the new estimate of correlation becoming .83. Results given by Behseta et al. showed that the new estimate may be considered much more reliable than the original $r = .49$. □

12.5 Multiple Linear Regression

The simple linear regression model (12.5) states that the response variable Y arises when a linear function of a single predictive variable x is subjected to additive noise ϵ. The idea is easily extended to two or more predictive variables. Let us write the ith observation of the jth predictive variable as x_{ji}. Then, for p predictive variables the linear regression model becomes

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i \tag{12.44}$$

where the ϵ_i's have the same assumptions as in (12.5).

Let us start with the case $p = 2$. Just as $y = \beta_0 + \beta_1 x_1$ describes a line, the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ describes a plane. When only a single variable x_1 is involved, the coefficient β_1 is the slope: $\beta_1 = \Delta y / \Delta x$. For example, if we increase x by $\Delta x = 2$ then we increase y by $\Delta y = 2\beta_1$. In the case of the equation $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, if we increase x_1 by $\Delta x_1 = 2$ and ask what happens to y, the answer will depend on how we change x_2. However, if we hold x_2 fixed while we increase x_1 by $\Delta x_1 = 2$ then we will increase y by $\Delta y = 2\beta_1$. When $p = 2$, β_1 is interpreted as the change in y for a one-unit change in x_1 *when* x_2 *is held fixed*. If $p > 2$ then β_1 becomes the change in y for a one-unit change in x_1 *when* $x_2, \ldots, x_p$ *are all held fixed*. Thus, linear regression is often used as a way of assessing what *might* happen if we *were* to hold one or more variables fixed while allowing a different variable to fluctuate. Put differently, regression allows us to assess the relationship between x_1 and y *after adjusting for* the variables $x_2, \ldots, x_p$. In this context $x_2, \ldots, x_p$ are often called *covariates*, because[9] they co-vary with x_1 and y.

[9] See also "analysis of covariance," mentioned in the footnote on p. 379.

Example 12.4 Developmental change in working memory from fMRI Many studies have documented the way visuospatial working memory (VSWM) changes during development. Kwon et al. (2002) used fMRI to examine neural correlates of these changes. These authors studied 34 children and young adults, ranging in age from 7 to 22. Each subject was given a VSWM task while being imaged. The task consisted of 12 alternating 36-s working memory (WM) and control epochs during which subjects viewed items on a screen. During both the WM and control versions of the task the subjects viewed the letter "O" once every 2 s at one of nine distinct locations on the screen. In the WM task the subjects responded when the current location was the same as it was when the symbol was presented two stimuli back. This required the subjects to engage their working memory. In the control condition the subjects responded when the "O" was in the center of the screen.

One of the y variables used in this study was the maximal BOLD activation (as a difference between WM and control) among voxels within the right prefrontal cortex. They were interested in the relationship of this variable with age (x_1). However, it is possible that Y would increase due to better performance of the task, and that this would increase with age. Therefore, in principle, the authors wanted to "hold fixed" the performance of task while age varied. This is, of course, impossible. What they did instead was to introduce two measures of task performance: the subjects' accuracy in performing the task (x_2) and their mean reaction time (x_3). □

Example 12.1 (continued, see p. 310) The firing rates in Fig. 12.1 appear clearly to increase with size of reward, and the analysis the authors reported (see p. 326) substantiated this impression. Platt and Glimcher also considered whether other variables might be contributing to firing rate by fitting a multiple regression model using, in addition to the normalized reward size, amplitude of each eye saccade, average velocity of saccade, and latency of saccade. This allowed them to check whether firing rate tended to increase with normalized reward size after accounting for these eye saccade variables. □

Equation (12.6) defined the least squares fit of a line. Let us rewrite it in the form

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \min_{\beta^*} \sum_{i=1}^{n}(y_i - y_i^*)^2 \tag{12.45}$$

where $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, $y_i^* = \beta_0^* + \beta_1^* x_i$ and $\beta^* = (\beta_0^*, \beta_1^*)$. If we now re-define y_i^* as

$$y_i^* = \beta_0^* + \beta_1^* x_{1i} + \cdots + \beta_p^* x_{pi}$$

with $\beta^* = (\beta_0^*, \beta_1^*, \ldots, \beta_p^*)$, Eq. (12.45) defines the least-squares multiple regression problem. We write the solution in vector form as

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p), \tag{12.46}$$

where the components satisfy (12.45) with the fitted values being

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \cdots + \hat{\beta}_p x_{pi}. \tag{12.47}$$

We interpret the multiple regression equation in Section 12.5.1 and discuss the decomposition of sums of squares in Section 12.5.2. In Section 12.5.3 we show how the multiple regression model may be written in matrix form, which helps in demonstrating how it includes ANOVA models as special cases, and in Section 12.5.4 we show that multiple regression also may be used to analyze certain nonlinear relationships. In Section 12.5.5 we issue an important caveat concerning correlated explanatory variables; in Section 12.5.6 we describe the way interaction effects are fitted by multiple regression; and in Section 12.5.7 we provide a brief overview of the way multiple regression is used when there are substantial numbers of alternative explanatory variables. We close our discussion of multiple regression in Section 12.5.8 with a few words of warning.

12.5.1 Multiple regression estimates the linear relationship of the response with each explanatory variable, while adjusting for the other explanatory variables.

To demonstrate multiple regression in action we consider a simple example.

Example 12.5 Toxicity as a function of dose and weight In many studies of toxicity, including neurotoxicity (Makris et al. 2009) a drug or other agent is given to an animal and toxicity is examined as a function of dose and animal weight. A relatively early example was the study of sodium arsenate (arsenic) in silkworm larvae (Bliss 1936). We reanalyzed data reported there. The response variable (y) was $\log(w/1{,}000)$ where w was minutes survived, and the two predictive variables were log weight, in log grams, and log dose, given in 1.5 plus log milligrams. A plot of log survival versus log dose is given in Fig. 12.8. Because there were two potential outliers that might affect the slope of the line fitted to the plotted data we have provided in the plot the fitted regression lines with and without those two data pairs. The results we discuss were based on the complete set of data.

The linear regression of log survival on log dose gave the fitted line

$$\text{log survival} = .140(\pm.057) - .704(\pm.078)\text{log dose}$$

which says that survival decreased roughly $.704(\pm.078)$ log 1,000 min for every log milligram increase in dose. The regression was very highly significant ($p = 10^{-12}$), consistently with the obvious downward trend.

The linear regression of log survival on both log dose and log weight gave the fitted line

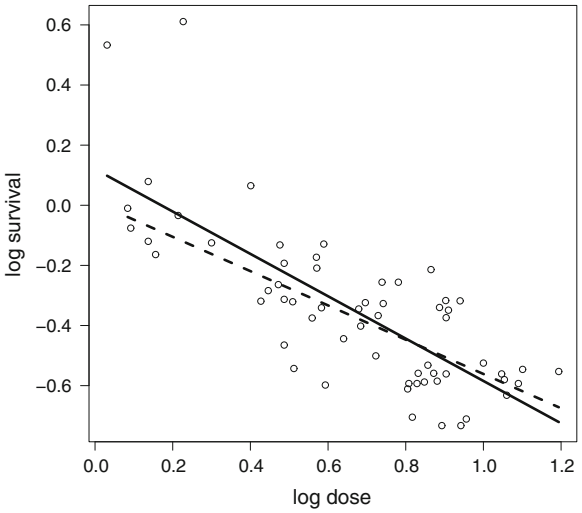

Fig. 12.8 Plot of log survival time (log(w/1,000) where w was minutes survived) versus log dose (1.5 plus log milligrams) of sodium arsenate in silkworm larvae; data from Bliss (1936). Lines are fits based on linear regression: *solid line* used the original data shown in plot; *dashed line* after removing the two high values of survival at low dose.

$$\text{log survival} = .140(\pm.057) - .734(\pm.058)\text{log dose} + 1.07(\pm.16)\text{log weight}.$$

In this case, including weight in the regression does not change very much the relationship between dose and survival: the slope is nearly the same in both cases. □

12.5.2 Response variation may be decomposed into signal and noise sums of squares.

As in simple linear regression we define the sums of squares *SSE* and *SSR*, again using (12.22) and (12.28) except that now $\hat{y}_i$ is defined by (12.47). If we continue to define the total sum of squares as in (12.24) we may again decompose it as

$$SST = SSR + SSE$$

and we may again define R^2 as in (12.25) or, equivalently, (12.27). In the multiple regression context R^2 is interpreted as a measure of the strength of the linear relationship between y and the multiple explanatory variables.

With p variables we may again use the sum of squares of the residuals to estimate the noise variation σ^2 but we must change the degrees of freedom appearing in (12.21). Because we again start with $n-1$ degrees of freedom in total, we subtract p to get $n-1-p$ degrees of freedom for error, and we have

$$s^2 = \frac{1}{n-1-p} SSE \tag{12.48}$$

where *SSE* is defined by (12.22). In multiple regression the hypothesis of no linear relationship between y and the x variables is H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. To test this hypothesis we define and compare suitable versions of *MSR* and *MSE*, the idea being that under H_0, with no linear relationship at all, *MSR* and *MSE* should be about the same size because both represent fluctuation due to noise. With p explanatory variables there are p degrees of freedom for regression. We therefore define the mean squared error for regression

$$MSR = \frac{SSR}{p}.$$

We use (12.48) in (12.23) for the mean squared error. We then form[10] the F-ratio

$$F = \frac{MSR}{MSE}. \tag{12.49}$$

In words, F is the ratio of the mean squared errors for regression and error, which are obtained by dividing the respective sums of squares by the appropriate degrees of freedom. Under the standard assumptions, if H_0 holds this F-ratio follows an F distribution, which will be centered near 1.

To state the result formally we must define a theoretical counterpart to (12.49). Let $\hat{Y}_i$ be the random variable representing the least-squares fit under the linear regression assumptions on p. 315, i.e., it is the theoretical counterpart of (12.47). We define

$$U_{MSE} = \frac{1}{p} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{12.50}$$

and

$$U_{MSR} = \frac{1}{n-1-p} \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y}_i)^2. \tag{12.51}$$

[10] The letter F was chosen (by George Snedecor in 1934) to honor Fisher, who had first suggested a log-transformed normalized ratio of sums of squares, and derived its distribution, in the context of ANOVA, which we discuss in Chapter 13.

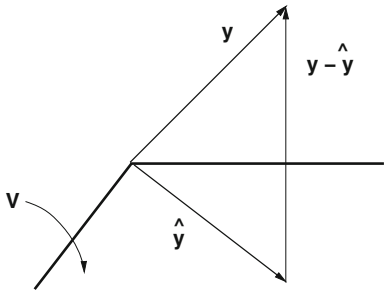

Fig. 12.9 Orthogonal projection of the vector y onto the vector subspace V resulting in the vector $\hat{y}$ in V. The residual vector $y - \hat{y}$ is orthogonal to $\hat{y}$, which gives the Pythagorean relationship (12.57). This corresponds to the total sum of squares (the squared length of y) equaling the sum of the regression sum of squares (the squared length of $\hat{y}$) and the error sum of squares (the squared length of $y - \hat{y}$).

Result: F-Test for Regression
Under the linear regression assumptions on p. 315, with (12.44) replacing (12.5), if H_0: $\beta_1 = \beta_2 = \cdots = \beta_p = 0$ holds then the F-statistic

$$F = \frac{U_{MSR}}{U_{MSE}} \tag{12.52}$$

follows an F_{ν_1,ν_2} distribution, where $\nu_1 = p$ and $\nu_2 = n - 1 - p$.

Proof outline: If H_0 is true, it may be shown that

$$\sum(\hat{Y}_i - \bar{Y})^2 \sim \chi^2_{\nu_1}$$

and

$$\sum(Y_i - \hat{Y}_i)^2 \sim \chi^2_{\nu_2}$$

where $\nu_2 = n - 1 - p$ is the degrees of freedom for error, and it may be shown that these are independent. Therefore, the random variable F defined by (12.52) is a ratio of independent chi-squared random variables divided by their degrees of freedom, which, by the definition on p. 129 has an F_{ν_1,ν_2} distribution. □

We provide a geometrical interpretation of the sum of squares decomposition below, in Fig. 12.9 and Eq. (12.57).

In simple linear regression, where there is only one explanatory variable, $\nu_1 = 1$ and F is equal to the square of the t-ratio. Because the square of a t_ν distributed random variable has an $F_{1,\nu}$ distribution, it follows that the t-test and the F-test of H_0: $\beta_1 = 0$ are identical. In multiple regression, hypotheses may also be tested about the individual coefficients, e.g., $H_0: \beta_2 = 0$, using t-tests.

Table 12.1 Simple linear regression results for Example 12.5.

Variable	Coefficient	SE	t_{obs}	p-value
(Intercept)	.120	.057	2.1	.038
Log dose	−.704	.078	−9.1	10^{-12}

Table 12.2 Multiple regression results for Example 12.5.

Variable	Coefficient	SE	t_{obs}	p-value
(Intercept)	−.140	.057	−2.49	.017
Log dose	−.734	.058	−12.6	2×10^{-16}
Log weight	1.07	.16	6.8	6×10^{-9}

Example 12.5 (continued) Returning to the toxicity data, the results for the regression of log survival on log dose are given in Table 12.1. We also obtained $s = .17$ and $R^2 = .59$. The F-statistic was $F = 82$ on 1 and 58 degrees of freedom, with $p = 10^{-12}$ in agreement with the p-value for the t-test in Table 12.1. The results for the regression of log survival on both log dose and log weight are in Table 12.2 and here $s = .13$ and $R^2 = .77$, which is a much better fit. The F-statistic was $F = 97$ on 2 and 57 degrees of freedom, with $p = 2 \times 10^{-16}$.

We would interpret the t ratios and F-statistics as follows: there is very strong evidence of a linear relationship between log survival and a linear combination of log dose and log weight ($F = 97, p << 10^{-5}$); given that log weight is included in the regression model, there is very strong evidence ($t = -12.6, p << 10^{-5}$) that log survival has a decreasing linear trend with log dose; similarly, given that log dose is in the model, there is very strong evidence ($t = 6.8, p << 10^{-5}$) that survival has an increasing linear trend with log weight. □

Example 12.4 (continued from p. 333) Recall that in one of their analyses Kwon et al. defined Y to be the maximal BOLD activation (as a difference between WM and control) among voxels within the right prefrontal cortex, and they considered its linear relationship with age (X_1), accuracy (X_2) and reaction time (X_3). They then performed multiple linear regression and found $R^2 = .53$ with $\beta_1 = .75(\pm.20)$, $p < .001$, $\beta_2 = -.21(\pm.19)$, $p = .28$, and $\beta_3 = -.15(\pm.17)$, $p = .37$. They interpreted the results as showing that the right PFC tends to become much more strongly activated in the VSWM task as the subjects' age increases, and that this is not due solely to improvement in performance of the task. □

Example 12.1 (continued from p. 333) Platt and Glimcher fit a multiple regression model to the firing rate data using as explanatory variables normalized reward size, amplitude of each eye saccade, average velocity of saccade, and latency of saccade. They reported the results of the t-test for the normalized reward size coefficient as $p < .05$, which indicates that firing rate tended to increase with normalized

reward size even after accounting for these eye saccade variables. A plot showing the coefficient with *SE* makes it appear that actually $p << .05$, which is much more convincing. □

The distributional results for the statistic F in (12.52) are based on the assumption of normality of the errors. For sufficiently large samples the p-values for the F-statistic, and the t-based p-values and confidence intervals, will be approximately correct even if the errors are non-normal. This is due to the theorems on consistency (p. 318) and approximate normality (p. 324), which extend to multiple regression (p. 344). However, the independence assumption is crucial. The standard errors and other distributional results generally may be trusted for reasonably large samples when the errors are independent, but they require correction otherwise. The assumptions should be examined using residual plots, as in simple linear regression.

12.5.3 Multiple regression may be formulated concisely using matrices.

Mathematical manipulations in multiple regression could get very complicated. A great simplification is to collect multiple equations together and write them as single equations in matrix form. We start by writing the n random variables Y_i as an $n \times 1$ random vector

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$$

and then similarly write

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

and

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

The linear regression model may then be written in the form

$$Y = X\beta + \epsilon \tag{12.53}$$

where it is quickly checked that both left-hand side and right-hand side are $n \times 1$ vectors. The usual assumptions may also be stated in matrix form. For example, we have

$$\epsilon \sim N_n(0, \sigma^2 \cdot I_n) \tag{12.54}$$

which says that ϵ has a multivariate normal distribution of dimension n, with mean equal to the zero vector and variance matrix equal to σ^2 times the n-dimensional identity matrix, i.e.,

$$V(\epsilon) = \sigma^2 \cdot \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & & & \ddots & 0 \\ 0 & \cdots & 0 & & 1 \end{pmatrix}.$$

Equation (12.53), together with the assumptions, is often called the *general linear model*. It accommodates not only multiple regression but also a large variety of models[11] that compare experimental conditions, which arise in analysis of variance (Chapter 13). For example, a standard approach to the analysis of fMRI data is based on a suitable linear model.

Example 12.2 (continued from p. 313) In Eq. (12.20) we defined a variable x_i that could be used with simple linear regression to analyze the BOLD response due to activity associated with a particular stimulus, according to an assumed form for the hemodynamic response function.[12] Suppose there are two stimuli with $u_j = 1$ corresponding to the first stimulus being on, with $u_j = 0$ otherwise, and $v_j = 1$ corresponding to the second stimulus being on, with $v_j = 0$ otherwise. We then define

[11] Sometimes when someone refers to the general linear model they may also allow the variance matrix to be different, or they may allow for non-normal errors.

[12] Before regression is applied various pre-processing steps are usually followed to make the assumptions of linear regression a reasonable representation of the variation in the fMRI data.

$$x_{i1} = \sum_{j<i} h(i-j)u_j$$
$$x_{i2} = \sum_{j<i} h(i-j)v_j$$

and set the X matrix equal to

$$X = \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix}.$$

If we apply (12.53) with $\beta = (\beta_0, \beta_1, \beta_2)^T$ the coefficient β_1 will represent the magnitude of the effect of the first stimulus on the BOLD response, the coefficient β_2 will represent the magnitude of the effect of the second stimulus on the BOLD response, and the coefficient β_0 will represent the baseline BOLD response. □

Because X often reflects the design of an experiment, as in Example 12.2 above, it is called the *design matrix*. The assumptions associated with (12.53) are essentially the same as those enumerated (i)–(v) for simple linear regression, where (i) becomes the validity of Eq. (12.53) and (ii)–(v) refer to the components of ϵ.

In matrix form we may write the least-squares fit as $\hat{y}$ according to

$$||y - \hat{y}||^2 = \min_{\beta^*} ||y - y^*||^2$$
$$y^* = X\beta^*$$

where $||w||$ is used to indicate the length of the vector w. We assume here that $X^T X$ is nonsingular (see the Appendix for a definition). The solution is found by solving the equations

$$X^T X\beta = X^T y \tag{12.55}$$

numerically (by numerically stable methods) and the solution may be written in the form[13]

$$\hat{\beta} = (X^T X)^{-1} X^T y. \tag{12.56}$$

Formula (12.56) may be obtained by a simple geometrical argument. We begin by thinking of y as a vector in n-dimensional space and we consider the subspace V consisting of all linear combinations of the columns of X. We say that V is the linear subspace spanned by the columns of X, which is the set of all vectors that may be written in the form $X\beta^*$ for some β^*, i.e.,

[13] The equations are *not* solved merely by inverting the matrix $X^T X$; this can lead to grossly incorrect answers due to seemingly innocuous round-off error. See Section 12.5.5.

$$V = \{X\beta^*,\ \beta^* \in R^{p+1}\}$$

(see the Appendix). The subspace V is the space of all possible fitted vectors. The problem of least squares, then, is to find the closest vector in V to the data vector y, i.e., the problem is to minimize the Euclidean distance between y and V. The solution to this minimization problem is the fitted vector $\hat{y} = X\hat{\beta}$. See Fig. 12.9. This geometry also gives us the Pythagorean relationship

$$||y||^2 = ||\hat{y}||^2 + ||y - \hat{y}||^2 \tag{12.57}$$

which is the basis for the decomposition $SST = SSR + SSE$.

Details: Euclidean geometry says that $\hat{y}$ must be obtained by orthogonal projection of y onto the subspace spanned by the columns of X and, as a result, the residual $y - \hat{y}$ must be orthogonal to the subspace spanned by the columns of X, which means that $y - \hat{y}$ must be orthogonal to $X\beta$ for every β. This, in turn, may be written in the following form: for all β,

$$\langle X\beta, y - \hat{y}\rangle = 0 \tag{12.58}$$

where $\langle u, v\rangle = u^T v$ is the inner product of u and v. Substituting $\hat{y} = X\hat{\beta}$ we have

$$\langle X\beta, y - X\hat{\beta}\rangle = 0$$

for all β, and rewriting this we find that

$$\beta^T X^T y = \beta^T X^T X\hat{\beta}$$

for all β, which gives us Eq. (12.55). Equation (12.55) is sometimes called the set of *normal equations* (presumably using "normal" in the sense of "orthogonal"; and plural because (12.55) is a vector equation and therefore a set of scalar equations). Because (12.58) holds for all β, it holds in particular for $\beta = \hat{\beta}$, i.e.,

$$\langle \hat{y}, y - \hat{y}\rangle = 0$$

which, as illustrated in Fig. 12.9, gives (12.57).
For the SST decomposition we introduce the $n \times 1$ vector having all of its elements equal to 1, which we write $1_{vec} = (1, 1, \ldots, 1)^T$. In the argument above we replace y by the residual following projection of y onto 1_{vec},

$$\begin{aligned}\tilde{y} &= y - \frac{< y, 1_{vec} >}{< 1_{vec}, 1_{vec} >} 1_{vec} \\ &= y - \bar{y} 1_{vec}\end{aligned}$$

(which is the vector of residuals found by regressing y on 1_{vec}) and similarly for all $j = 2, \ldots, p+1$ replace the j column of X by its residual following projection onto 1_{vec} (which produces the vectors of residuals found by regressing each x variable on 1_{vec}). When we repeat the argument with these new variables we get a new fitted vector $\hat{\tilde{y}}$ and everything goes through as before. We then obtain the version of (12.57) needed for the decomposition:

$$||\tilde{y}||^2 = ||\hat{\tilde{y}}||^2 + ||y - \hat{y}||^2.$$

It may be verified that this is the same as $SST = SSR + SSE$. For example, $||\tilde{y}||^2 = \sum(y_i - \bar{y})^2$. □

The variance matrix of the least-squares estimator is easy to calculate using a generalization of Eq. (4.26): with a little algebra it may be shown that if W is a $p \times 1$ random vector with variance matrix $V(W) = \Sigma$ and A is a $k \times p$ matrix, then the variance matrix of AW is

$$V(AW) = A\Sigma A^T. \tag{12.59}$$

Using (12.59) we obtain

$$\begin{aligned} V(\hat{\beta}) &= ((X^TX)^{-1}X^T)\sigma^2 I_n((X^TX)^{-1}X^T)^T \\ &= \sigma^2 \cdot (X^TX)^{-1}X^TX(X^TX)^{-1} \\ &= \sigma^2 \cdot (X^TX)^{-1}. \end{aligned}$$

This variance matrix summarizes the variability of $\hat{\beta}$. For instance, we have

$$V(\hat{\beta}_k) = \sigma^2 \cdot (X^TX)^{-1}_{kk},$$

which is the kth diagonal element of the variance matrix. To use such formulas with data, however, we must substitute s for σ. We then have the estimated variance matrix

$$\hat{V}(\hat{\beta}) = s^2 \cdot (X^TX)^{-1} \tag{12.60}$$

and the standard errors are given by

$$SE(\hat{\beta}_k) = \sqrt{s^2 \cdot (X^TX)^{-1}_{kk}}. \tag{12.61}$$

For example, (12.61) is the formula that was used to produce the standard errors in Table 12.2, and to get the standard errors and t-ratios, and thus the p-values, in Example 12.4 reported on p. 338. For problems involving propagation of uncertainty (Section 9.1) to a function of $\hat{\beta}$, the variance matrix in (12.60) would be used.

The estimator (12.60), and resulting inferences, may be justified by the analogue to (12.37).

Theorem: Asymptotic normality of least squares estimators For the linear regression model (12.53) suppose conditions (i)–(iv) hold and let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of design matrices such that

$$\frac{1}{n} X^T X \to C \tag{12.62}$$

for some positive definite matrix C, as $n \to \infty$. Then the least-squares estimator defined by (12.56) satisfies

$$\frac{1}{s}(X_n^T X_n)^{1/2}(\hat{\beta} - \beta) \xrightarrow{D} N_{p+1}(0, I_{p+1}). \tag{12.63}$$

Proof: See Wu (1981) for references. □

> *A Detail:* It is also possible to use the bootstrap in regression, but this requires some care because under the assumptions (i)–(iv) the random variables Y_i have distinct expected values,
>
> $$E(Y_i) = (1, x_{i1}, \ldots, x_{ip})\beta$$
>
> and so are not i.i.d. The usual approach is to resample the studentized residuals (see p. 319), which are approximately i.i.d. See Davison and Hinkley (1997, page 275). Alternatively, when each vector $x_i = (x_{i1}, \ldots, x_{ip})$ is observed, rather than chosen by the experimenter, it is possible to treat x_i as an observation from an unknown multivariate probability distribution, and thus (x_i, y_i) becomes an observation from an unknown distribution, and the data vectors $((x_1, y_1), \ldots, (x_n, y_n))$ may be resampled.[14] This was the bootstrap procedure mentioned in Example 8.2 on p. 241. For additional discussion see Davison and Hinkley (1997). □

There are many conveniences of the matrix formulation of multiple regression in (12.53) together with (12.54). One is that the independence and homogeneity assumptions in (12.54) may be replaced. Those assumptions imply

$$V(\epsilon) = \sigma^2 I_n,$$

as in (12.54). The analysis remains straightforward if we instead assume

$$V(\epsilon) = R \tag{12.64}$$

[14] Here, Eq. (9.27) becomes

$$\hat{F}_n(x, y) \xrightarrow{P} F_{(X,Y)}(x, y)$$

where $\hat{F}_n$ is the empirical cdf computed from the random vectors $((X_1, Y_1), \ldots, (X_n, Y_n))$.

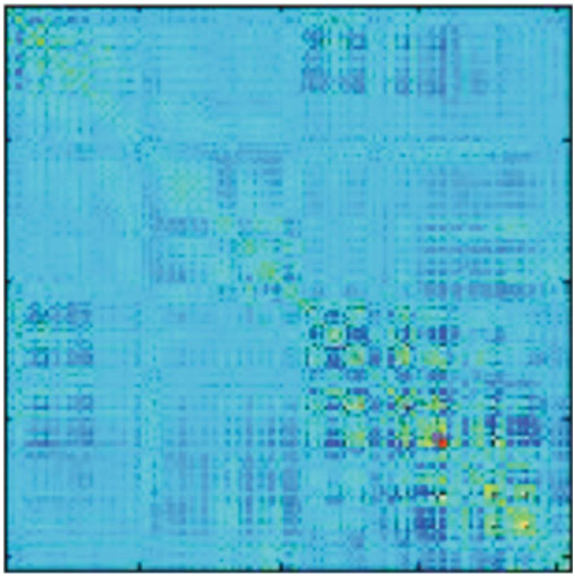

Fig. 12.10 MEG gradiometer background noise covariance matrix. The *light blue* corresponds to zero elements and *darker blue*, *yellow*, and *red* indicate non-zero elements. (Figure furnished by Gustavo Sudre.)

where R can be any $n \times n$ variance matrix (i.e., a positive definite symmetric matrix). **Example 1.2 (continued from p. 5)** We previously noted that MEG imaging requires sensor data to be obtained first from background scanner noise, meaning the sensor data must be obtained with nothing in the scanner. We displayed on p. 54 a histogram of such data, from a single sensor, as an example of a normal distribution. The separate sensor readings are not independent but are, instead, correlated. Figure 12.10 displays a representation of the background noise variance matrix from 204 gradiometer sensors in a MEG scanner. MEG analysis is based on (12.53) together with (12.64), with R being based on the background noise variance matrix. □

Given a matrix R in (12.64), and assuming it is positive definite, the least-squares problem may be reformulated. Letting $U = R^{-1/2}Y$ and $W = R^{-1/2}X$ we have

$$R^{-1/2}(Y - X\beta) = R^{-1/2}\epsilon \sim N_n(0, I_n),$$

so that the new model

$$U = W\beta + \delta,$$

where $\delta = R^{-1/2}\epsilon$, satisfies the usual assumptions in (12.53) together with (12.54). Therefore, to fit the model (12.53) with (12.64) we may first transform Y and X by premultiplying with $R^{-1/2}$ and then can apply ordinary least squares to the transformed variables. This is called *weighted least squares* and it arises in various extensions of

multiple regression. On p. 212 we showed that the least-squares estimator was also the MLE under the standard assumptions of regression, including normality of the errors. More generally, the weighted least squares estimator of β is the MLE under (12.53) with (12.64).

Example 1.2, above, provides a case in which the non-independence of the components of ϵ is due to the spatial layout of the sensors, and the resulting dependence among the magnetic field readings at different sensors. Neuroimaging also typically generates temporal correlation in the measurements, i.e., the measurements are time series with some dependence across time. Using auto-regressive time series models described in Section 18.2.3 the variance matrix may be determined from the data and this furnishes an R matrix in (12.64). The model (12.53) with (12.64) then leads to *regression with time series errors*.

12.5.4 The linear regression model applies to polynomial regression and cosine regression.

In many data sets the relationship of y and x is mildly nonlinear, and a quadratic in x may offer better results than a line. Even though a quadratic is nonlinear, a neat trick allows us to fit quadratic regression via multiple linear regression. The trick is to set $w_1 = x$ and to define a new variable $w_2 = x^2$. Then, when y is regressed on both w_1 and w_2 this amounts to fitting a general quadratic of the form $y = a + bx + cx^2$, where now $a = \beta_0$, $b = \beta_1$ and $c = \beta_2$. To be clear, we define the vector w_1 as

$$w_1 = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \tag{12.65}$$

and the vector w_2 as

$$w_2 = \begin{pmatrix} x_1^2 \\ x_2^2 \\ \vdots \\ x_n^2 \end{pmatrix} \tag{12.66}$$

and then we regress $y = (y_1, \ldots, y_n)$ on w_1 and w_2.

In quadratic regression there are several possibilities. First, there may be evidence of a linear association between y and x (from the simple linear regression), but the relationship appears nonlinear and there is also evidence of a linear association between y and both x and x^2 combined. This latter evidence would come from the combined regression output of (i) a statistically significant F-ratio and (ii) a

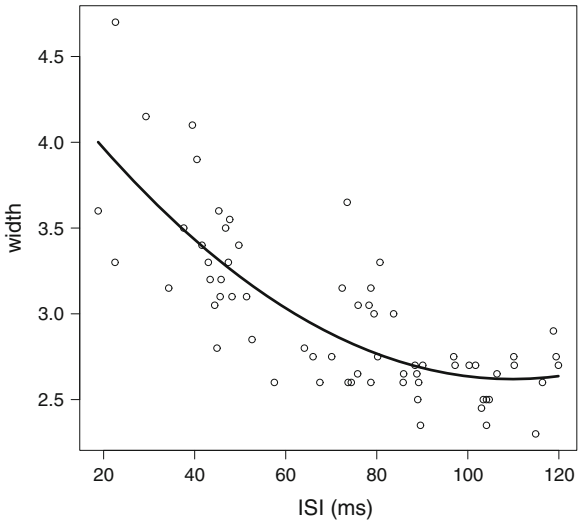

Fig. 12.11 Plot of action potential width against length of previous ISI, together with quadratic fitted by linear regression.

significant t-ratio for the coefficient of x^2. This case is illustrated below. Note that it is possible for the coefficient of x in the combined regression to be non-significant. This should not necessarily be taken to mean that there is no linear component to the relationship: it is generally preferable to use the general form $y = a + bx + cx^2$, which requires the bx term and thus the x variable. Actually, it is possible for the coefficients of *both* x and x^2 to be non-significant while the F-ratio is significant; this occurs when the two variables are themselves so highly correlated that neither adds anything to the regression when the other is already used.

As a second possibility, there may be evidence of a linear association between y and x (from the simple linear regression), but there is no evidence of a quadratic relationship. The latter would be apparent from (i) an OK (not curved) residual plot in the simple linear regression and (ii) a non-significant t-ratio for the coefficient of x^2. The third possibility is that there may be no evidence of a relationship between y and *either* x by itself or x combined with x^2. This would be evident from an insignificant t-ratio in the simple linear regression and an insignificant F-ratio in the combined regression.

Let us now turn to an example.

Example 8.2 (continued from p. 193) On p. 193 we examined spike train data recorded from a barrel cortex neuron in slice preparation, which was part of a study on the effects of seizure-induced neural activity. Figure 8.5 displayed the decreasing width of action potentials with increasing length of the interspike interval. Figure 12.11 shows a plot of many action potential widths against preceding interspike interval (ISI), where the data have been selected to include only ISIs of length

less than 120 ms. In the plot, the downward trend begins to level off near 100 ms, and a quadratic curve fitted by linear regression is able to capture the leveling off reasonably well within this range of ISI values. In this case the linear and quadratic regression coefficients were both highly significant ($p = 6 \times 10^{-6}$ and $p = .0017$, respectively, with the overall F-statistic giving $p = 8 \times 10^{-14}$) and $R^2 = .61$. □

In quadratic regression, illustrated in Example 8.2 above, we defined $w_1 = x$ and $w_2 = x^2$. To fit cubic and higher-order polynomials we may continue the process with $w_3 = x^3$, etc. An important caveat, however, is that the variables x_1, x_2, and x_3 defined in this way are likely to be highly correlated, which may cause difficulties in interpretation and, in extreme cases, may cause the matrix $X^T X$ to be singular (non-invertible), in which case least-squares software will fail to return a useful result. We discuss this issue further in Section 12.5.5.

A second nonlinear function that may be fitted with linear regression is the cosine.

Example 12.6 Directional Tuning in Motor Cortex In a well-known set of experiments, Georgopoulos, Schwartz and colleagues showed that motor cortex neurons are directionally "tuned." Figure 12.12 shows a set of raster plots for a "center-out" reaching task: the monkey reached to one of eight points on a circular image, and this neuron was much more active for reaches in some directions than for others. The bottom part of Fig. 12.12 shows a cosine function that has been fitted to the mean firing rate as a function of the angle around the circle, which indicates the direction of reach. For example (and as is also shown in the raster plots), reaches at angles near 180° from the x-axis produced high firing rates, while those at angles close to 0° (movement to the right) produced much lower firing rates. The angle at which the maximum firing rate occurs is called the "preferred direction" of the cell. It is obtained from the cosine function.

To fit a cosine to a set of spike counts, multiple linear regression is used. Let $v = (v_1, v_2)$ be the vector specifying the direction of movement and let $d = (d_1, d_2)$ be the preferred direction for the neuron. Both v and d are unit vectors. Assuming cosine tuning, the firing depends only on $\cos\theta$, where θ is the angle between v and d. We have

$$\cos\theta = v \cdot d = v_1 d_1 + v_2 d_2.$$

Letting $\mu(v)$ be the mean firing rate in a given interval of time when the movement is in direction v, if we let the minimal firing rate be B_{min} and the maximal firing rate be B_{max}, then cosine tuning may be written as the requirement that

$$\mu(v) = B_{min} + \frac{B_{max} - B_{min}}{2} + \frac{B_{max} - B_{min}}{2} \cos\theta.$$

(Recall that the minimal value of the cosine is -1, and its maximal value is 1.) If we now define $\beta_1 = \frac{B_{max} - B_{min}}{2} d_1$, $\beta_2 = \frac{B_{max} - B_{min}}{2} d_2$, and $\beta_0 = B_{min} + \frac{B_{max} - B_{min}}{2}$ we obtain the linear form

$$\mu(v) = \beta_0 + \beta_1 v_1 + \beta_2 v_2. \tag{12.67}$$

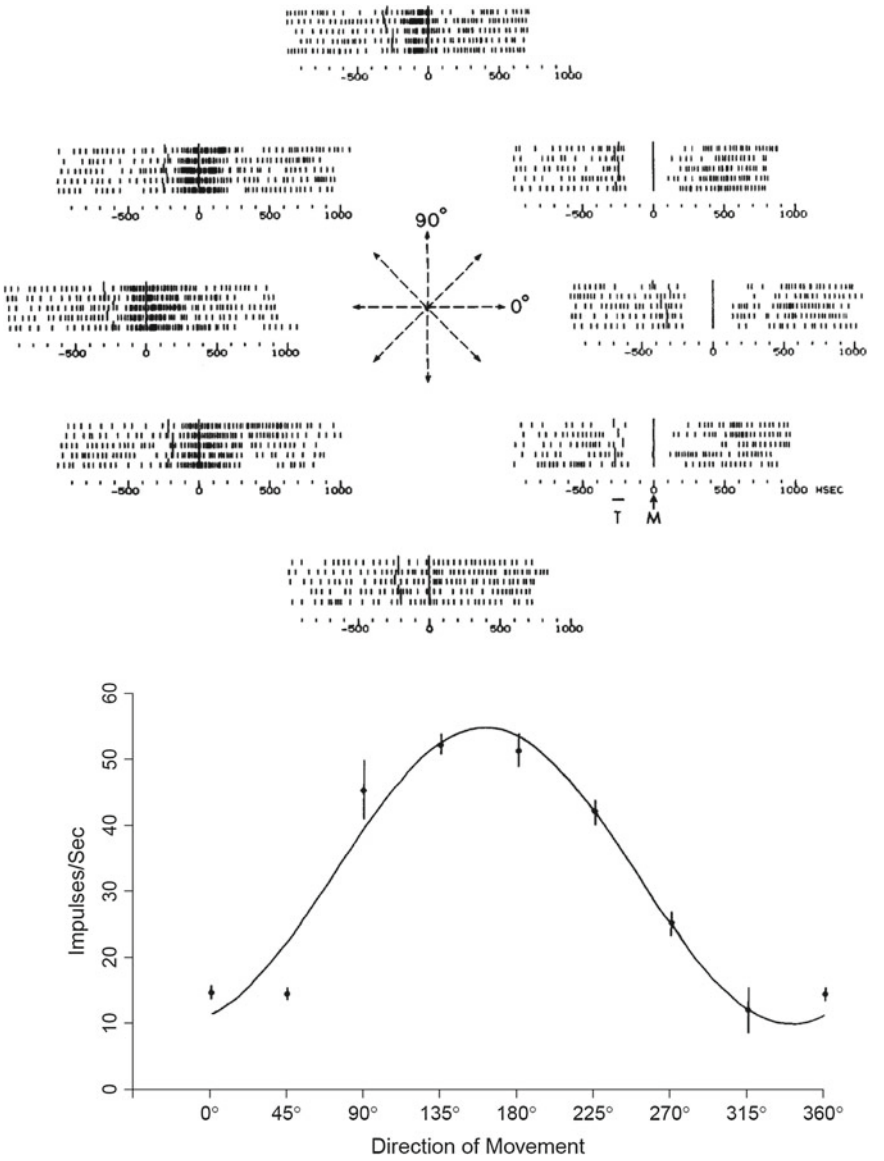

Fig. 12.12 Directional tuning of motor cortex neurons (adapted from Georgopoulos et al. 1982). *Top* displays raster plots (spike trains across five trials) for each of eight reaching directions. *Bottom* displays corresponding mean firing rates.

Taking $C_i(v)$ to be the spike count for the ith trial in direction v across a time interval of length T, the observed spike count per unit time is

$$Y_i(v) = \frac{1}{T} C_i(v).$$

and we have

$$Y_i(v) = \mu(v) + \epsilon_i(v). \tag{12.68}$$

Together, Eqs. (12.68) and (12.67) define a two-variable multiple linear regression model from which the tuning parameters may be obtained. □

12.5.5 Effects of correlated explanatory variables cannot be interpreted separately.

On p. 347 we used Example 8.2 to illustrate quadratic regression, and we then issued a note of caution that x and x^2 are often highly correlated. High correlation among explanatory variables may cause numerical and inferential difficulties. Let us first describe the numerical issue.

The least-squares solution (12.56) to Equation (12.55) results from multiplying both sides of Equation (12.55) by $(X^T X)^{-1}$, under the assumption that $X^T X$ is non-singular, i.e., that its inverse exists, which occurs when the columns of X are linearly independent (see the Appendix). Linear independence fails when it is possible to write some column of X as a linear combination of the other columns; in this case a regression of that dependent column on the other columns would produce $R^2 = 1$, i.e., perfect multiple correlation. When the columns of X are very highly correlated, even if they are mathematically linearly independent, they may be numerically essentially dependent; for example, a regression of any one column on all the others might produce R^2 that is very nearly equal to 1 (e.g., $R^2 = .999$). Because of this and related considerations the details of the methods used to compute the least-squares solution are important, as indicated in the footnote on p. 341. In the quadratic regression of Example 8.2 on p. 347, for instance, the correlation between *ISI* and its square was $r = .98$. An easy way to reduce correlation is to subtract the mean of the x variable before squaring, i.e., take $w_1 = x$ and $w_2 = (x - \bar{x})^2$. With w_1 and w_2 defined in this way for $x = ISI$ in Example 8.2 we obtained $r = -.08$. Good numerical methods use general procedures that effectively transform the x variables to reduce their correlations.

A deeper issue involves interpretation of results. The potential confusion caused by correlated explanatory variables may be appreciated from the following concocted illustration.

Illustration: Quadratic regression To demonstrate the interpretive subtlety when explanatory variables are correlated we set $x = (1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$ and then defined

$$y_i = x_i + u_i$$

Table 12.3 Quadratic regression results for the artificial data in the illustration.

Variable	Coefficient	SE	t_{obs}	p-value
(Intercept)	−2.4	2.5	−.95	.37
x	1.86	1.04	1.8	.12
x^2	−.067	.092	−.73	.487

where $u_i \sim N(0, 4)$. We then defined w_1 and w_2 using (12.65) and (12.66) and regressed $y = (y_1, \ldots, y_n)$ on both w_1 (representing x) and w_2 (representing x^2). We obtained the results shown in Table 12.3, with $R^2 = .77$, $s = 2.1$ and $F = 11.9$ on 2 and 7 degrees of freedom, yielding $p = .0056$. From Table 12.3 alone this regression might appear to provide no evidence that y was linearly related to either x or x^2. However, regressing y on either x or x^2 alone produces a highly significant linear regression. Furthermore, the F-statistic from the regression on both variables together is highly significant. These potentially puzzling results come from the high correlation of explanatory variables: the correlation between x and x^2 is $r = .975$. Keep in mind that the t-statistic for x^2 in Table 12.3 reflects the contribution of x^2 *after* the variable x has been used to explain y and likewise the t-statistic for x reflects the contribution of x after the variable x^2 has been used to explain y. □

Let us consider this phenomenon further. Suppose we want to use linear regression to say something about the degree to which a particular variable, say x_1, explains y (meaning the degree to which the variation in y is matched by the variation in the fit of x to y) but we are also considering other variables $x_2, \ldots, x_p$. We can regress y on x_1 by itself. Let us denote the resulting regression coefficient by b. Alternatively we can regress y on $x_1, \ldots, x_p$ and, after applying Eq. (12.56), the relevant regression coefficient would be $\hat{\beta}_1$, the first component of $\hat{\beta}$. When the explanatory variables are correlated, it is not generally true that $b = \hat{\beta}_1$ and, similarly, the quantities that determine the proportion of variability explained by x_1, the squared magnitudes of the fitted vectors, are not generally equal. Thus, when the explanatory variables are correlated, as is usually the case, it is impossible to supply a unique notion of the extent to which a particular variable explains the response—one must instead be careful to say which other variables were also included in the linear regression.

This lack of uniqueness in explanatory power of a particular variable may be considered a consequence of the geometry of least squares.

> *Details:* Let us return to the geometry depicted in Fig. 12.9. As in that figure we take V to be the linear subspace spanned by the columns of X. Because the columns of X are vectors, let us write them in the form $v_1, \ldots, v_p$, and let us ignore the intercept (effectively assuming it to be zero, as we did when we related the SST decomposition to the Pythagorean theorem). The observations on the first explanatory variable x_1 then make up the vector v_1. The extent to which x_1 "explains" the response vector y now becomes the proportion of y that

lies in the direction v_1. This is the length of the projection of y onto v_1 divided by the length of y. However, length of the projection of y onto v_1 depends on whether we do the calculation using v_1 by itself or together with $v_2, \ldots, v_p$. Let us write the projection as cv_1 for some constant c. If we consider v_1 in isolation, we find

$$c = \frac{\langle v_1, y \rangle}{\langle v_1, v_1 \rangle} = b. \tag{12.69}$$

If we consider v_1 together with $v_2, \ldots, v_p$, we must first project y onto V, and then find the component in the direction v_1. The result is $c = \hat{\beta}_1$. The exception to this bothersome reality occurs when v_1 is orthogonal to the span of $v_2, \ldots, v_p$ (i.e., $\langle v_1, v \rangle = 0$ for every vector v that is a linear combination of $v_2, \ldots, v_p$). In this special case of orthogonality we have $b = \hat{\beta}$, and we regain the interpretation that there is a proportion of y that lies in the direction of v_1. Specifically, in this orthogonal case we may write the projection of y onto V as $\hat{y} = c_1 v_1 + v$ for some v in the span of $v_2, \ldots, v_p$. We then have

$$\langle v_1, \hat{y} \rangle = \langle v_1, c_1 v_1 + v \rangle = c_1 \langle v_1, v_1 \rangle$$

so that the projection is $c_1 v_1$ where

$$c_1 = \frac{\langle v_1, \hat{y} \rangle}{\langle v_1, v_1 \rangle}.$$

On the other hand, we may reconsider the value c in (12.69). Because $y - \hat{y}$ is orthogonal to V when we write

$$\langle v_1, y \rangle = \langle v_1, \hat{y} + (y - \hat{y}) \rangle$$

we have $\langle v_1, y - \hat{y} \rangle = 0$. Therefore,

$$\langle v_1, \hat{y} \rangle = \langle v_1, y \rangle$$

so, in this case, $c = c_1$. Thus, in this orthogonal case, $b = \hat{\beta}_1$. □

12.5.6 In multiple linear regression interaction effects are often important.

We saw earlier that it is possible to fit a quadratic in a variable x using linear regression by defining a new variable x^2 and then performing multiple linear regression on x and x^2 simultaneously. Now suppose we have variables x_1 and x_2. The general quadratic

in these two variables would have the form

$$y = a + bx_1 + cx_2 + dx_1^2 + ex_1x_2 + fx_2^2.$$

Thus, we may again use multiple linear regression to fit a quadratic in these two variables if, in addition to defining new variables x_1^2 and x_2^2 we also define the new variable $x_1 \cdot x_2$. This latter variable is often called the *interaction* between x_1 and x_2. To see its effect consider the simpler equation

$$y = a + bx_1 + cx_2 + dx_1x_2. \tag{12.70}$$

Here, for instance, we have $\Delta y/\Delta x_1 = b + dx_2$. That is, the slope for the linear relationship between y and x_1 depends on the value of x_2 (and similarly the slope for x_2 depends on x_1). When $d = 0$ and we graph y versus x_1 for two different values of x_2 we get two parallel lines, but when $d \neq 0$ the two lines are no longer parallel.

Interaction effects are especially important in analysis of variance models, which we discuss in Chapter 13.

12.5.7 Regression models with many explanatory variables often can be simplified.

When one considers multiple explanatory variables it is possible that some of them will have very little predictive benefit beyond what the others offer. In that eventuality one typically removes from consideration the variables that seem redundant or irrelevant, and then proceeds to fit a model using only the variables that help predict the response. When the number of variables p is small it is not difficult to sort through such possibilities quickly, but sometimes there are much larger numbers of variables, particularly if combinations of them, defining interactions as described in Section 12.5.6, are considered. In this case choosing a suitable collection of variables to fit is called the problem of *model selection*, and is based on *model comparison* procedures such as those discussed in Section 11.1.6.

Example 12.7 Prediction of burden of disease in multiple sclerosis Li et al. (2006) investigated the relationship between a measure of severity of multiple sclerosis, known as burden of disease (BOD), and many clinical assessments. The response variable, BOD, was based on MRI scans, and 18 different clinical measurements were used as potential explanatory predictors, including such things as disease duration, age at onset, and symptom types, as well as an important variable of interest the Expanded Disability Status Scale (EDSS). One of their main analyses examined data from an initial set of 1,312 patients who had been entered into 11 clinical trials in multiple centers. The problem they faced was to determine the variables to use as predictors from among the 18, together with possible interactions. Note that there are $\binom{18}{2} = 153$ possible pairwise interaction terms. □

There is a huge literature on model selection in multiple regression. We very briefly describe the ideas behind a few of the major methods, and then offer some words of caution.

Let us begin with variables $x_1, x_2, \ldots, x_p$ and the aim of selecting some subset that predicts the response y well. Here, some of the x variables could be defined as interaction terms. For example, if we had variables $x_1, \ldots, x_k$ and wanted to consider all possible interaction effects, as defined in Section 12.5.6, then we would end up with $p = \binom{k}{2}$ variables in total. A very simple variable-selection algorithm is as follows:

1. Regress y on each single variable x_i and find the variable x_a that gives the best prediction (using R^2).
2. Regress y on all two-variable models that include x_a as one of the variables and find the variable x_b such that x_a together with x_b gives the best prediction.
3. Continue in this way: for $k \geq 3$ and some set of variables we label $x_{a_1}, x_{a_2}, \ldots, x_{a_{k-1}}$ that have already been selected in previous steps, consider all regression models that include, in addition, each of the remaining variables; find x_j such that (1) $x_{a_1}, x_{a_2}, \ldots, x_{a_{k-1}}, x_j$ gives the best prediction and (2) the coefficient of x_j is statistically significant.

   Note that criterion (2) provides a way of stopping the process with $k < p$.

This algorithm is an example of *forward selection*. It is also called a *greedy* algorithm (because at every step in the process it is taking an apparently best next step). In the form given above it is not yet completely specified because the level of significance, or the value of the t-ratio, must be chosen; this will determine the number of variables k that are selected. It is also possible to reverse the process by starting with a regression based on all variables $x_1, \ldots, x_p$ and then choosing, analogously to step 1 above, one variable to drop, and then repeatedly finding variables to drop until a satisfactory model is found in which all variables are statistically significant. This is called *backward elimination*. An algorithm that alternates between forward and backward steps is called *stepwise regression*.

Within model selection algorithms, including forward selection, backward elimination, or stepwise regression, it is also possible to use criteria such as AIC and BIC (see Section 11.1.6) to evaluate alternative regression models. (In regression, AIC is very similar to another popular criterion known as *Mallow's* C_p.) In principle, one would examine all possible models and pick the one that is optimal with respect to the chosen criterion, such as AIC. However, because each variable may be either included in a model, or excluded from the model, there are 2^p possible models and it quickly becomes prohibitive to examine all possible models as p grows. Model selection algorithms, therefore, provide search strategies but can not guarantee that the optimal model is found.

Example 12.7 (continued) In their study, Li et al. used a stepwise procedure based on AIC to select variables for predicting BOD. □

An additional, widely-used criterion for model selection is *cross-validation*. The idea begins by considering the prediction of y by each model. Let us define an observation from all the variables $x_1, \ldots, x_p$ to be a vector x. Then we are predicting y by some function $f(x)$. In the case of linear regression,

$$f(x) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

where each model fixes some of the coefficients β_j to be 0 (these are the coefficients corresponding to variables excluded from the model). The corresponding theoretical problem is to predict Y by some function $f(x)$ of a random vector X, and we may evaluate the prediction using mean squared error (MSE), $E((Y - f(X))^2)$. According to the prediction theorem on p. 89 the MSE is minimized by the conditional expectation $E(Y|X = x)$, and we would, in principle, find this conditional expectation through model selection and fitting. One possibility would be to attempt to choose the model that gives the smallest MSE. However, because the MSE will depend on unknown values of the coefficients, we must estimate it from the data. If we use the same data both to fit models and to evaluate how well the models fit, we necessarily obtain an overly optimistic answer for the MSE: we will have optimized the fit for the particular data values at hand; if we were to get new data we probably would not do as well. In other words, the estimated MSE will tend to be too small; it will be downwardly biased. Furthermore, the amount of downward bias in the estimated MSE will vary with the model, so the estimated MSE will not be a reliable model comparison procedure.

Cross-validation attempts to get around the problem of optimistic MSE assessment by splitting the n observations y_i into a set of K groups, each group having the same number of observations, or nearly the same number. Let us label the kth group G_k. Then, for $k = 1, \ldots, K$, we pick group G_k and call its observations "test data" and the remainder of the observations "training data." We use the training data to fit models and we use the test data to evaluate the fits. Specifically, an observation $y_i \in G_k$ is predicted by the fit from the training data in the $K - 1$ groups containing all $y_i \notin G_k$. Letting $\hat{y}_{i,CV}$ denote the fit of y_i based on the training data that excludes group G_k, the cross-validated estimate of MSE is

$$\widehat{MSE} = \frac{1}{n} \sum_{k=1}^{K} \sum_{y_i \in G_k} (y_i - \hat{y}_{i,CV})^2.$$

This represents the quality of "out of sample" fit; conceptually, MSE is the average squared error we would expect, theoretically, if we were to apply the fit on entirely new data collected under precisely the same conditions. The model with the best cross-validation performance $\widehat{MSE}$ is the model selected by *K-fold cross-validation*. Cross-validation should, in principle, provide good estimates of MSE as K gets large (so that the estimates of MSE will have good statistical properties). For any given

sample size n the largest possible value of K is $K = n$. This results in *leave-one-out cross validation*, a method recommended by Frederick Mosteller and John Tukey in an influential book (Mosteller and Tukey 1968). Here is an example.

Example 12.8 Prediction of fMRI face selectivity using anatomical connectivity Saygin et al. (2011) used anatomical connectivities established from diffusion-weighted imaging to predict differential responses to faces and objects in fMRI. It is highly intuitive that functional activity in the brain, as measured by fMRI, should depend on anatomical structure. Saygin et al. examined fMRI responses in the fusiform face area of the temporal lobe, an area known to respond more strongly when a subject is shown pictures of faces than when the same subject is shown pictures of objects. They considered the response to pictures of faces, and to objects, at every voxel in the fusiform face area and took as their y_i variable in regression analyses the normalized ratio of face response to object response for voxel i. The x_i vector of variables was made up of connectivities to 84 brain regions, which were found using diffusion weighted imaging. This constituted their "connectivity" model. Leave-one-out cross-validation was used across 23 subjects to compare this model with two other models that did not involve connectivity information. One model defined the x_i variables to be physical distances to the 84 brain regions. This was the "distance" model. The other used the group average among all the other subjects, as a single predictor x_i. This was the "group average" model. For each subject the authors fit these models to the other 22 subjects, then used the fits to predict the fMRI responses among all the voxels for each subject. These authors used mean absolute error instead of MSE. (We comment on this below.) Thus, they computed the sample mean absolute error across all voxels for each subject. The cross-validated estimate of mean absolute error was the sample mean[15] of these 23 values. The results were as follows: connectivity model, .65; distance model, 1.06; group average model, .78. This provided evidence that the connectivity model predicts fMRI activity better than either physical distances or group averaged responses. □

In some problems it is computationally expensive to obtain n distinct fits, one for each of the n training data sets needed for leave-one-out cross-validation. In such cases, K is chosen to be much smaller, so that only K fits need to be computed. The most popular value in this context is $K = 10$.

Cross-validation has been studied extensively (see Efron 2004; Arlot and Celisse 2010; and references therein). The argument that cross-validation should provide a correction for a downwardly biased estimate of MSE is reminiscent of the motivation for AIC given in Section 11.1.6. There, AIC was introduced to correct the bias in estimating the Kullback-Liebler discrepancy between fitted model and true model. In

[15] In K-fold cross-validation it is tempting to regard the average of the n MSE estimates as an ordinary mean, and to apply the usual standard error formula (7.17). This does not work correctly, however, because the n separate evaluations are not independent. Instead, the square of the standard error in (7.17) is an underestimate of the variance. In fact, it is not possible to provide a simple evaluation of the uncertainty attached to the cross-validation estimate of MSE, or risk (see Bengio and Granvalet 2004).

regression, minimizing the Kullback-Liebler discrepancy corresponds to minimizing MSE and, for large samples, AIC and leave-one-out cross-validation agree (Stone 1974). The great advantage of cross-validation is that it furnishes an estimate of MSE even if the relationship between Y and X does not follow the assumed linear model. On the other hand, if the linear model assumptions are roughly correct then AIC tends to outperform cross-validation (Efron 2004).

Let us make two additional remarks. First, we phrased our comments above in terms of MSE but, more generally, cross-validation provides an estimate of risk (see p. 102) using loss functions other than that defined by squared error. In Example 12.8 absolute error was used. Second, cross-validation is not a substitute for replication of experiments. Experimental replication provides much stronger evidence than any statistical manipulation can create: new data will inevitably involve both small and, sometimes, substantial changes in details of experimental design and data collection; to be trustworthy, findings should be robust to such modifications and should therefore be confirmed in subsequent investigations.

There is a different approach to the problem of using multiple regression in the presence of a large number of possible predictor variables. Instead of thinking that some variables are irrelevant, and trying to identify and remove them, one might say that the coefficients are noisy and, therefore, on aggregate, likely to be too large in magnitude. This suggests reducing the overall magnitude of the coefficients, a process usually called *shrinkage*. We replace the least squares criterion (12.45) with

$$\sum_{i=1}^{n}(y_i - \hat{y}_{i,p})^2 = \min_{\beta^*}\left(\sum_{i=1}^{n}(y_i - y_i^*)^2 + \lambda\,\text{magnitude}(\beta^*)\right) \tag{12.71}$$

where magnitude(β) is some measure of the overall size of β and is called a *penalty*. The number λ is an adjustable constant and is chosen based on the data, often by cross-validation (or, for some penalties, AIC or BIC). The criterion to be minimized in (12.71) is *penalized least squares* and the solution $\hat{y}_{i,p}$ is called *penalized regression*. The two most common penalties are

$$\text{magnitude}(\beta) = \sum_{j=1}^{p}\beta_j^2 \tag{12.72}$$

and

$$\text{magnitude}(\beta) = \sum_{j=1}^{p}|\beta_j|. \tag{12.73}$$

These penalties are also called, respectively, $L2$ and $L1$ penalties.[16] In the statistics literature $L2$-penalized regression is often called[17] *ridge regression* and $L1$-penalized regression is called the *LASSO* (see Tibshirani 2011, and references therein). We give a Bayesian interpretation of penalized regression in Section 16.2.3.

Example 12.9 MEG source localization In Example 1.2 we described, briefly, the way MEG signals are generated and detected, and we discussed an application in Example 4.7. There are 306 sensors and the sensor data may be analyzed directly or, alternatively, an attempt may be made to identify the brain sources that produce the sensor signals, a process known as *source localization*. One class of methods overlays a large grid of possible sources on a representation of the cortex, and then applies Maxwell's equations in what is known as a "forward solution" that predicts the sensor signals for any particular set of source activities. This results in a linear model of the form (12.53) where X is determined by Maxwell's equations and β represents the source activity. A typical number of sources might be 5,000, so this becomes a large problem. Furthermore, because $n = 306$ we have $p > n$ which makes the matrix $X^T X$ singular (non-invertible) and some alternative to least squares must be used. The most common solutions involve $L2$ and $L1$ penalized least squares,[18] which are used in the *minimum norm estimate* MNE and *minimum current estimate* MCE methods of source localization in MEG. □

12.5.8 Multiple regression can be treacherous.

Multiple linear regression is a wonderful technique, of wide-ranging applicability. It is important to bear in mind, however, the cautions we raised in the context of simple linear regression, especially in our discussion of Fig. 12.5. With many explanatory variables, the inadequacies of the linear model illustrated in Fig. 12.5 could be present for any of the y versus x_j relationships, for $j = 1, \ldots, p$, and there are similar but more complex possibilities when we use the multiple variables simultaneously. Furthermore, it is no longer possible to plot the data in the form y versus x when $x = (x_1, x_2, \ldots, x_p)$ and $p > 2$. The assumption of linearity of the relationship between y and x is crucial, and with multiple variables it is difficult to check.

An additional issue involves one of the most useful features of multiple regression, that it allows an investigator to examine the relationship of y versus x while adjusting for another variable u. This was discussed in Section 12.5.1 and its use in the interpretation of neural data was described in Examples 12.4 and 12.1. In this context, however, the phenomenon of attenuation of correlation, discussed in Section 12.4.4,

[16] The penalty in (12.72) may also be written $\text{magnitude}(\beta) = ||\beta||^2$ and in mathematical analysis the Euclidean length is called an $L2$ norm. The penalty (12.73) is called an $L1$ penalty because it is based, analogously, on the $L1$ norm.

[17] Strictly speaking ridge regression refers to $L2$-penalized regression after the x variables are normalized.

[18] Actually, the penalty is applied to weighted least squares as described on p. 345.

must be considered. In Example 12.4, for instance, the authors wanted to examine the effect of age on BOLD activity while adjusting for task performance. The variables used for adjustment were accuracy (x_2) and mean reaction time (x_3). For each subject, the numbers x_2 and x_3 obtained for these variables were based on limited data and therefore represent accuracy and reaction time with some uncertainty, which could be summarized by standard errors. These standard errors were not reported by the authors, and probably were small, but suppose, hypothetically, that the x_2 and x_3 measurements had large standard errors. In this case, according to the result in Section 12.5.1, the correlation of these noisy variables with BOLD activity would be less than it would have been if accuracy and reaction time had been measured perfectly. Therefore, the adjustment made with x_2 and x_3 would also be less than the adjustment that *would have been made* in the absence of noise.

A similar concern arises when the measured variables capture imperfectly the key features of the phenomenon they are supposed to represent. In Example 12.1, the authors wanted to adjust the effect of reward size on firing rate for relevant features of each eye saccade. They did this by introducing eye saccade amplitude, velocity, and latency. If, however, a different feature of eye saccades was crucial in determining firing rate (e.g., acceleration), then these measurements would only be correlated with the key feature and would represent it imperfectly. In this sense, the measured variables would again be noisy representations of the ideal variables. The fundamental issue for adjustment is whether the measured variables used in a regression analysis correctly represent the possible additional explanatory factors, which are often called *confounding* variables. We discuss confounding variables further in Section 13.4. The general problem of mismeasured explanatory variables is discussed in the statistics and epidemiology literature under the rubric of *errors in variables*. When multiple regression is used to provide statistical adjustments, the accuracy of explanatory variables should be considered.

Finally, in Section 12.5.7 we noted the many alternative regression models that present themselves when there are multiple possible explanatory variables, and we described very briefly some of the methods used for grappling with the problem of model determination. These approaches can be very successful in certain circumstances. However, there is often enormous uncertainty concerning the model that best represents the data. A careful analyst will consider whether interpretations are consistent across all plausible models. Furthermore, in assessing the relationship between the response y and one of the explanatory variables x_j, the process of model selection can spuriously inflate the magnitude of an estimated coefficient $\hat{\beta}_j$. See Kriegeskorte et al. (2010) for discussion.