Chapter 13

Linear Regression

13.1 Least-Squares Regression

One of the most important problems in data analysis, particularly when all its generalizations are included, is curve-fitting. In the univariate case, involving a single explanatory variable $x$, curve-fitting results in the summary of data pairs $(x, y)$ as roughly following $y = f(x)$ for some suitable function $f(x)$. When $f$ is a linear function of $x$, we may write it in the form $f(x) = \beta_0 + \beta_1 x$, and the statistical problem becomes one of estimating the coefficients $\beta_0$ and $\beta_1$ and, of course, representing the uncertainty in the estimates. There are many ways to fit lines to data, but the usual solution to the problem is in terms of the least squares estimates, $\hat{\beta}_0$ and $\hat{\beta}_1$, and their standard errors. The resulting fitted line is usually called the (linear) regression line.

There are numerous examples of linear regression in the neuroscience and psychology literature. For instance, Platt and Glimcher (1999, *Nature*) 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; Behrmann et al. (2002, J. Cognitive Neuroscience) used linear regression in examining the modulation of saccadic reaction time as a function of angle to target by eye, head, or trunk orientation among patients with hemispatial neglect. Let us return to neural conduction velocity example, presented in Chapter 1.

**Neural conduction velocity** Recall that Hursh (1939) presented data on the relationship between a neuron’s conduction velocity and its axonal diameter, in adult cats. Hursh measured maximal velocity among fibers in several nerve bundles, and then also measured the diameter of the largest fiber in the bundle. The resulting data together with the least-squares regression line, are shown in Figure 13.1. The regression line has intercept $\hat{\beta}_0 = -3.8(\pm 1.4)$ meters per second and slope $\hat{\beta}_1 = 6.07(\pm .14)$ meters per second per micron (the ± figures being standard errors). The relationship is quite close to linear, with relatively little noise: the relative proportion of signal-to-noise variability is $R^2 = .97$, which will be discussed below; the residual noise standard deviation is about 5.9 meters per second, which may be interpreted by saying that a prediction of velocity based on diameter would incur an error due to noise having about the same magnitude as mismeasuring diameter by 1 micron. We may summarize the data by saying that, in the cat, action potential velocity increases by very nearly 6 meters per second ($6.07 \pm .14$ according to our re-analysis of the data) for every micron increase in diameter of a neuron. \hfill $\square$

### 13.1.1 Least-squares estimates are found by minimizing the sum of the squared residuals.

Suppose we have a line $y = \beta_0 + \beta_1 x$ that is fitted by some method, possibly least-squares (in which case we would write $\beta_0 = \hat{\beta}_0$ and $\beta_1 = \hat{\beta}_1$) or possibly another method. We may think of the value $\beta_0 + \beta_1 x_i$ as predicting $y_i$, and then define the $i$-th residual as

$$ e_i = y_i - (\beta_0 + \beta_1 x_i) $$
13.1. LEAST-SQUARES REGRESSION

Figure 13.1: Conduction velocity of action potentials, as a function of diameter. The $x$-axis is diameter in microns; the $y$-axis is velocity in meters per second. Based on Hursh (1939, Figure 2). Also shown is the least-squares regression line.

The value $e_i$ is the error of prediction, i.e., the vertical distance between the observation $y_i$ and the line at the corresponding $x_i$. See Figure 13.2. To judge the quality of the fit of the line, we examine the $e_i$’s. An overall assessment of the fit must somehow combine the magnitudes of the $e_i$’s. The method of least squares uses the criterion $\sum e_i^2$. Specifically, for each possible choice of $\beta_0^*$ and $\beta_1^*$, we may compute the sum of squares $\sum e_i^2$ then choose the values of $\beta_0^*$ and $\beta_1^*$ that minimizes this sum of squares. A relatively easy way to minimize the sum of squares is to apply calculus: we differentiate $\sum e_i^2$ with respect to each of $\beta_0^*$ and $\beta_1^*$, set the derivatives equal to 0, and solve the resulting pair of equations. We omit the details here.
CHAPTER 13. LINEAR REGRESSION

Figure 13.2: Plot of the Hursh data set, with data points in grey except for a particular point \((x_i, y_i)\) which is shown in black to identify the corresponding fitted value \(\hat{y}_i\). The regression line also passes through the point \((\bar{x}, \bar{y})\), as indicated on the plot.

The least-squares estimates \(\hat{\beta}_0\) and \(\hat{\beta}_1\) are the values of \(\beta_0^*\) and \(\beta_1^*\) that minimize \(\sum e_i^2\). The least-squares line is then

\[
y = \hat{\beta}_0 + \hat{\beta}_1 x.
\]
13.1. LEAST-SQUARES REGRESSION

13.1.2 Linear regression assumes linearity of \( f(x) \) and independence of the noise contributions at the various observed \( x \) values.

Least-squares may be considered an algorithm for fitting a line to data. To get standard errors more structure is necessary: we need to have a statistical model. The starting point is to describe the deviations of the data from the line in terms of random variables. The fundamental assumption in linear regression is

\[
Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{13.1}
\]

for \( i = 1, \ldots, n \), where \( \epsilon_i \) is a random variable. This says that there is a theoretical linear relationship between \( x \) and \( y \) that holds except for “errors” \( \epsilon_i \). The \( \epsilon_i \)'s are assumed to be zero, on average, i.e., \( E(\epsilon_i) = 0 \) for all \( i \), so that \( E(Y_i) = \beta_0 + \beta_1 x_i \). Thus, 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; \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) are considered estimates of \( \beta_0 \) and \( \beta_1 \). The uncertainty arising from the \( \epsilon_i \)'s is used to quantify uncertainty in the least-squares estimates.

This linear model is the first crucial assumption. Another is that the errors \( \epsilon_i \) are independent of each other. The independence assumption is often violated when there is a time trend in the observations (as when they are recorded sequentially), in which case more elaborate methods are needed. Relationships may be examined in the presence of serial dependence using time series methods.

Important, though less potentially problematic, additional assumptions are that the variances of the \( \epsilon_i \)'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 are sufficient to guarantee that the 95% confidence intervals (see below) do have probability .95 of covering the coefficients. In sufficiently large samples these two assumptions become unnecessary, though in smaller samples they require attention, with substantial departures from the equality of variance assumption tending to be more problematic than departures from the Normality assumption.

The assumptions may be enumerated as follows:
(i) the linear regression model (13.1) holds;
(ii) the errors satisfy $E(\epsilon_i) = 0$ for all $i$;
(iii) the errors $\epsilon_i$ 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).

13.1.3 In reporting results, standard errors should be included along with coefficient values.

When reporting least-squares estimates, standard errors should also be supplied. Thus, in the conduction velocity example: we found that, on average, action potential velocity increases by $6.07 \pm .14$ meters per second for every micron increase in diameter of a neuron.

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

In the conduction velocity example the least-squares line provides an excellent 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 relationship is not so close to being linear. There are two ways to summarize this overall tendency of the line to fit, or fail to fit the data: we use $s$ and $R^2$.

The error standard deviation $\sigma$ (see item (iii) in the assumptions above) represents the average size of the error, in the sense that it is an average amount of deviation of each $\epsilon_i$ from zero. Thus, $\sigma$ 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

$$s^2 = \frac{1}{n - 2} \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))^2$$
is usually called the “residual mean squared error.” 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 \( \beta_0 \) and \( \beta_1 \) 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. Its definition is

\[
R^2 = 1 - \frac{SSE}{SST}
\]

where \( SSE = (n - 2)s^2 \) and \( SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \). Here, \( SSE \) stands for “sum of squares due to error” and \( SST \) stands for “total sum of squares”. 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. It is often interpreted as “the proportion of variability of \( Y \) that is explained by \( X \).” In different terminology, we may think of \( SSR = SST - SSE \) 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 would be \( SSR/SSE \).\(^1\)

The decomposition of total variability (leading to \( SST \)) into the regression and error components (leading to \( SSR \) and \( SSE \)) begins with consideration of the values \( y_i, \hat{y}_i, \) and \( \bar{y}, \) as shown in Figure 13.2, 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

\[
\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) + \sum (\hat{y}_i - \bar{y})^2
\]

but the cross-product term vanishes and, defining \( SSR = \sum (\hat{y}_i - \bar{y})^2 \), we have

\[
SST = SSR + SSE.
\]

\(^1\)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; it translates most immediately into a ratio of signal variance to noise variance, which is not quite what is being examined here, but is very closely related.
A demonstration of the vanishing of the cross-product term will be provided in our discussion of multiple regression.

The estimated standard deviation $s$ has the units of $Y$ and is therefore always immediately 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 engineering applications one expects an $R^2$ of at least .99 while in many social science applications an $R^2$ of .40 would be considered quite substantial. One gets a feeling for the size of $R^2$ mainly by examining, and thinking about, many specific examples.

### 13.1.5 For large samples, if the model is correct, the least-squares estimate is likely to be accurate.

### 13.2 Checking Assumptions

#### 13.2.1 Residual analysis is helpful because residuals should represent unstructured noise.

In examining single batches of data we have seen how the data may be used not only to estimate unknown quantities (there, an unknown mean $\mu$) 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 (the $e_i$s). Sometimes the residuals are replaced by standardized residuals. The $i$th standardized residual is simply $e_i/SD(e_i)$, where $SD(e_i)$ is the standard deviation of $e_i$ (as estimated from the data). 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.

Two kinds of plots are used. Residual versus fit plots are supposed to reveal (i) nonlinearity, (ii) inhomogeneity variances, or (iii) outliers. Plots
13.2. CHECKING ASSUMPTIONS

Figure 13.3: Residual plots: the top left plot depicts unstructured noise while the latter three reveal structure, and thus deviations from the assumptions.

having structure of the kind that would indicate these problems are shown in the Figure 13.3. 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 Normal distribution.
13.2.2 Graphical examination of \((x, y)\) data can yield crucial information.

Figure 13.4: 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, F.J. (1973), Graphs in statistical analysis, American Statistician, 27: 17-21.

Figure 13.4 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. Taken together the main point of the example is that until one plots the data and/or the residuals it may not be clear how or whether the
regression line summarizes the relationship.

13.3 Evidence of a Linear Trend

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

We again use the general form

\[ \text{estimate} \pm 2 \cdot SE \]

as an approximate 95% CI. Thus, for the conduction velocity example we have that an approximate 95% CI for the slope of the regression line is 6.07 ± 2(.14) or (5.79, 6.35).

An alternative in small samples, which is very much analogous to the small sample procedure in estimating a population mean, is to substitute for \(SE\) the value \(t_{0.025,\nu}\), 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 \(\hat{\beta}_1 \pm t_{0.025,n-2} \cdot SE(\hat{\beta}_1)\) as our 95% CI.

13.3.2 Evidence in favor of a linear trend can be obtained from a \(t\)-test concerning the slope.

In the conduction velocity example it is obvious that there is a linear trend in the data. This kind of increasing or decreasing tendency is sometimes a central issue in an analysis. For example, it is of interest to know whether the temperature in North America, and elsewhere, has had a tendency to increase over the years 1880-1980; this is the so-called phenomenon of “global warming.” Here, detailed study would show that temperature has not increased linearly over this period of time, but there is a tendency toward increase that may be summarized with a line. Similarly, it was not crucial to the results of Platt and Glimcher (1999) that the relationship between firing rate and volume of juice received was exactly linear; the most important thing was
to establish there was some upward tendency. 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 \( H_0 : \beta_1 = 0 \).

When the 95% CI \( \hat{\beta}_1 \pm t_{0.025,n-2} \cdot SE(\hat{\beta}_1) \) provides an interval that omits the value 0, as in the conduction velocity case where the approximate CI is \((5.79, 6.35)\), we may regard this as evidence against \( H_0 : \beta_1 = 0 \). We may, instead test this hypothesis by computing a \( p \)-value based on the \( t \)-ratio:

\[
t\text{-ratio} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}.\]

For large samples, under \( H_0 \), this statistic has a \( N(0,1) \) distribution. For small samples, if the assumptions of linear regression are satisfied, under \( H_0 \) the \( t \)-ratio has a \( t \) distribution on \( \nu = n - 2 \) degrees of freedom (which for large \( n \) is approximately Normal). Thus, generally speaking, when 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. Again, it is conventional to consider \( p < .05 \) to indicate mild evidence against \( H_0 \) and \( p < .01 \) to indicate strong evidence against \( H_0 \). And, again, it is important to remember that departures from \( H_0 \) that are so small as to be irrelevant are very likely to be identified as significant in large samples. In the case of the conduction velocity example we obtained \( p < 10^{-15} \); if it were necessary to report this, we would remember that such small \( p \)-values are sensitive to distributional form and we might instead say simply that the result is clearly statistically significant and report \( p << .0001 \).

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

An interesting related issue arises in the conduction velocity example. 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 in 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 (roughly) the range of values of diameters actually examined in the data.

### 13.4 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 the joint distribution of \(X\) and \(Y\). Recall that the correlation coefficient \(\rho\) is a measure of association between the random variables \(X\) and \(Y\). In this situation, as we discussed in Chapter 2, the theoretical regression of \(Y\) on \(X\) is then defined 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 2 we noted that the regression is, in fact, linear when \((X, Y)\) has a bivariate Normal distribution and we furthermore have

\[
\beta_1 = \frac{\sigma_Y}{\sigma_X} \cdot \rho.
\]

This linearity, and its interpretation, was illustrated in Figure 13.5. However, the right-hand plot in Figure 13.5 concerns data, rather than a theoretical distribution, and there is an analogous formula and interpretation using the sample correlation, which is defined by

\[
r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{s_xs_y}
\]

where \(s_x\) and \(s_y\) are the sample standard deviations among the \(x_i\)s and \(y_i\)s. The sample correlation \(r\) may be considered an estimate of \(\rho\).

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:
Figure 13.5: Conditional expectation for bivariate Normal data mimicking Pearson and Lee’s data on heights of fathers and sons. Left panel shows contours of the bivariate Normal distribution based on the means, standard deviations, and correlation in Pearson and Lee’s data. The dashed vertical lines indicate the averaging process used in computing the conditional expectation when \( X = 64 \) or \( X = 72 \) inches: we average \( y \) using the probability \( f_{Y|X}(y|x) \), which is the probability, roughly, in between the dashed vertical lines, integrating across \( y \). In the right panel we generated a sample of 1,078 points (the sample size in Pearson and Lee’s data set) from the bivariate Normal distribution pictured in the left panel. We then, again, illustrate the averaging process: when we average the values of \( y \) within the dashed vertical lines we obtain the two values indicated by the red \( x \). These fall very close to the least-squares regression line (the solid line).

- \(-1 \leq r \leq 1\) 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, and

- the value of \( r \) does not depend on the units in which the two variables are measured.

There are no general guidelines as to what constitutes a “large” value of the correlation coefficient. Interpretation depends on the application.
13.4. CORRELATION AND REGRESSION

13.4.1 The correlation coefficient is determined by the regression coefficient and the standard deviations of \( x \) and \( y \).

The sample correlation is closely related to the fitted regression slope. In Chapter 2 we stated the theoretical relationship, for bivariate Normal random variables,

\[
\beta = \frac{\rho \sigma_Y}{\sigma_X}.
\]

For data pairs \((x_i, y_i)\) we have

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

Thus, 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.

13.4.2 Association is not causation.

There are two very important points to keep in mind. First, correlation is a measure of linear association; for instance, a correlation can be zero even if \( x \) perfectly, but nonlinearly, predicts \( y \). Second, 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 example, one author\(^2\) 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\(^3\), can one be confident there are no


\(^3\)Randomization refers to the random assignment of treatments to subjects, and to the process of randomly ordering treatment conditions.
confounding factors.

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

Let $X$ and $Y$ be random variables. Suppose we define new random variables $U = X + \varepsilon_x$ and $V = Y + \varepsilon_y$ with $\varepsilon_x$ and $\varepsilon_y$ being independent random variables that are also both independent of $X$ and $Y$. The interpretation is that $\varepsilon_x$ and $\varepsilon_y$ represent “noise” being added to $X$ and $Y$. We have

$$\text{Cov}(U, V) = \text{Cov}(X, Y) + \text{Cov}(X, \varepsilon_y) + \text{Cov}(\varepsilon_x, Y) + \text{Cov}(\varepsilon_x, \varepsilon_y)$$

but, because of independence, the last 3 terms are all 0 so that

$$\text{Cov}(U, V) = \text{Cov}(X, Y).$$

Now, again by independence, $V(U) = V(X) + V(\varepsilon_x)$ and $V(V) = V(Y) + V(\varepsilon_y)$ and, assuming $V(\varepsilon_x) > 0$ and $V(\varepsilon_y) > 0$ we have

$$V(U) > V(X)$$

and

$$V(V) > V(Y).$$

Combining these results and recalling the formula

$$\text{Cor}(W, Z) = \frac{\text{Cov}(W, Z)}{\sqrt{V(W)V(Z)}}$$

we have

$$\text{Cor}(U, V) = \frac{\text{Cov}(U, V)}{\sqrt{V(U)V(W)}} < \frac{\text{Cov}(X, Y)}{\sqrt{V(X)V(Y)}} = \text{Cor}(X, Y).$$

In words, when noise is added to a pair of random variables their correlation is decreased. This is an important fact to keep in mind: it often happens that measurements imperfectly reflect the underlying variables being assessed. If the imperfection may be considered to result from additive independent noise, the resulting correlation is necessarily less than it would have been if better measurements had been available. Sometimes it is possible to correct for this “attenuation” in the correlation.