# 5 Linear Least-Squares Regression

On several occasions in the first part of the text, I emphasized the limitations of linear least-squares regression. Despite these limitations, linear least squares lies at the very heart of applied statistics:[1]

- Some data are adequately summarized by linear least-squares regression.
- The effective application of linear regression is considerably expanded through data transformations and techniques for diagnosing problems such as nonlinearity and overly influential data.
- As we will see, the general linear model—a direct extension of linear least-squares regression—is able to accommodate a very broad class of specifications, including, for example, qualitative explanatory variables and polynomial functions of quantitative explanatory variables.
- Linear least-squares regression provides a computational basis for a variety of generalizations, including weighted regression, robust regression, nonparametric regression, and generalized linear models.

Linear least-squares regression, and the closely related topic of linear statistical models, are developed in this chapter and in Chapters 6 through 10:

- The current chapter describes the mechanics of linear least-squares regression. That is, I will explain how the method of least squares can be employed to fit a line to a bivariate scatterplot, a plane to a three-dimensional scatterplot, and a general linear surface to multivariate data (which, of course, cannot be directly visualized).
- Chapter 6 develops general and flexible methods of statistical inference for linear models.
- Chapters 7 and 8 extend linear models to situations in which some or all of the explanatory variables are qualitative and categorical rather than quantitative.
- Chapter 9 casts the linear model in matrix form and describes the statistical theory of linear models more formally and more generally.
- Chapter 10 introduces the vector geometry of linear models, a powerful tool for conceptualizing linear models and least-squares estimation.[2]

[1] The extensions of linear least-squares regression mentioned here are the subject of subsequent chapters.

[2] Chapters 9 and 10 are "starred" (i.e., marked with asterisks), and therefore are more difficult; like all starred material in this book, these chapters can be skipped without loss of continuity, although some of the later starred material can depend on earlier starred text.

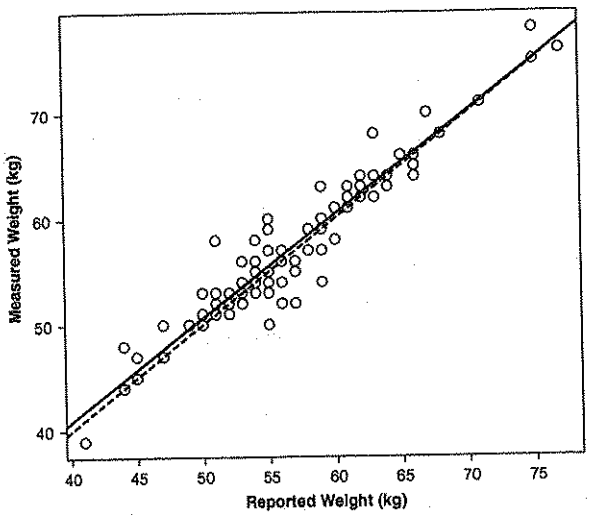


**Figure 5.1** Scatterplot of Davis's data on the measured and reported weight of 101 women. The solid line gives the least-squares fit; the broken line is $Y = X$. Because weight is given to the nearest kilogram, both variables are discrete, and some points are overplotted.

## 5.1 Simple Regression

### 5.1.1 Least-Squares Fit

Figure 5.1 shows Davis's data, introduced in Chapter 2, on the measured and reported weight in kilograms of 101 women who were engaged in regular exercise.[3] The relationship between measured and reported weight appears to be linear, so it is reasonable to fit a line to the plot. A line will help us determine whether the subjects in Davis's study were accurate and unbiased reporters of their weights; and it can provide a basis for predicting the measured weight of similar women for whom only reported weight is available.

Denoting measured weight by $Y$ and reported weight by $X$, a line relating the two variables has the equation $Y = A + BX$.[4] It is obvious, however, that no line can pass perfectly through all the data points, despite the strong linear relationship between these two variables. We introduce a *residual*, $E$, into the regression equation to reflect this fact; writing the regression equation for the $i$th of the $n = 101$ observations:

$$\begin{aligned} Y_i &= A + BX_i + E_i \\ &= \widehat{Y}_i + E_i \end{aligned} \tag{5.1}$$

where $\widehat{Y}_i = A + BX_i$ is the *fitted value* for observation $i$. The essential geometry is shown in Figure 5.2, which reveals that the residual

$$E_i = Y_i - \widehat{Y}_i = Y_i - (A + BX_i)$$

is the signed vertical distance between the point and the line—that is, the residual is negative when the point lies below the line and positive when the point is above the line [as is the point $(X_i, Y_i)$ in Figure 5.2].

[3] The misrecorded data value that produced an outlier in Figure 2.5 has been corrected.

[4] See Appendix C for a review of the geometry of lines and planes.

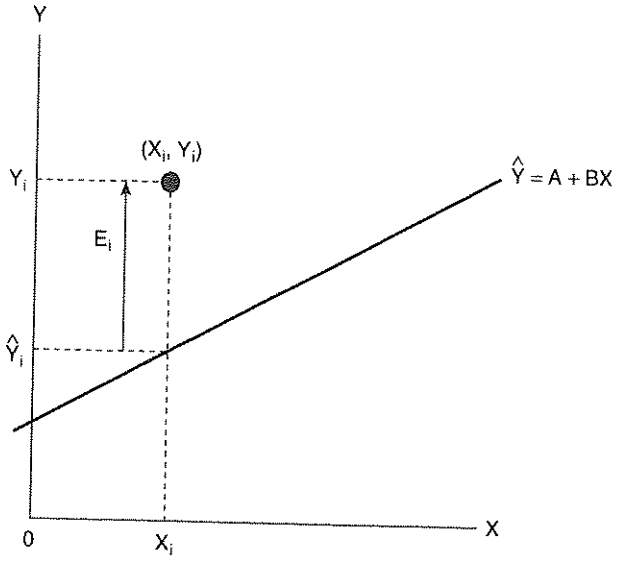


**Figure 5.2** Linear regression of $Y$ on $X$, showing the residual $E_i$ for the $i$th observation.

A line that fits the data well therefore makes the residuals small, but to determine a line analytically we need to be more precise about what we mean by "small." First of all, we want residuals that are small in magnitude, because large negative residuals are as offensive as large positive ones. For example, simply requiring that the sum of residuals, $\sum_{i=1}^{n} E_i$, be small is futile, because large negative residuals can offset large positive ones.

Indeed, any line through the means of the variables—the point $(\overline{X}, \overline{Y})$—has $\sum E_i = 0$. Such a line satisfies the equation $\overline{Y} = A + B\overline{X}$. Subtracting this equation from Equation 5.1 produces

$$Y_i - \overline{Y} = B(X_i - \overline{X}) + E_i$$

Then, summing over all observations,

$$\sum_{i=1}^{n} E_i = \sum (Y_i - \overline{Y}) - B \sum (X_i - \overline{X}) = 0 - B \times 0 = 0 \tag{5.2}$$

Two possibilities immediately present themselves: We can employ the unsigned vertical distances between the points and the line, that is, the absolute values of the residuals; or we can employ the squares of the residuals. The first possibility leads to *least-absolute-value (LAV) regression*:

Find $A$ and $B$ to minimize the sum of the absolute residuals, $\sum |E_i|$.

The second possibility leads to the *least-squares criterion*:

Find $A$ and $B$ to minimize the sum of squared residuals, $\sum E_i^2$.

Squares are more tractable mathematically than absolute values, so we will focus on least squares here, but LAV regression should not be rejected out of hand, because it provides greater resistance to outlying observations.[5]

[5]We will return to LAV regression in Chapter 19, which discusses robust regression.

We need to consider the residuals in the aggregate, because it is no trick to produce a 0 residual for an individual point simply by placing the line directly through the point. The least-squares criterion therefore minimizes the *sum* of squared residuals over all observations; that is, we seek the values of $A$ and $B$ that minimize:

$$S(A, B) = \sum_{i=1}^{n} E_i^2 = \sum (Y_i - A - BX_i)^2$$

I have written this expression as a *function* $S(A, B)$ of the regression coefficients $A$ and $B$ to emphasize the dependence of the sum of squared residuals on the coefficients: For a fixed set of data $\{X_i, Y_i\}$, each possible choice of values for $A$ and $B$ corresponds to a specific residual sum of squares, $\sum E_i^2$; we want the pair of values for the regression coefficients that makes this sum of squares as small as possible.

*The most direct approach to finding the least-squares coefficients is to take the partial derivatives of the sum-of-squares function with respect to the coefficients:[6]

$$\frac{\partial S(A, B)}{\partial A} = \sum (-1)(2)(Y_i - A - BX_i)$$
$$\frac{\partial S(A, B)}{\partial B} = \sum (-X_i)(2)(Y_i - A - BX_i)$$

Setting the partial derivatives to 0 yields simultaneous linear equations for the least-squares coefficients, $A$ and $B$.[7]

Simultaneous linear equations for the least-squares coefficients $A$ and $B$, the so-called *normal equations*[8] for simple regression, are

$$An + B\sum X_i = \sum Y_i$$
$$A\sum X_i + B\sum X_i^2 = \sum X_i Y_i$$

where $n$ is the number of observations. Solving the normal equations produces the least-squares coefficients:

$$A = \overline{Y} - B\overline{X}$$
$$B = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} \tag{5.3}$$

The formula for $A$ implies that the least-squares line passes through the point of means of the two variables. By Equation 5.2, therefore, the least-squares residuals sum to 0. The second normal equation implies that $\sum X_i E_i = 0$, for

$$\sum X_i E_i = \sum X_i (Y_i - A - BX_i) = \sum X_i Y_i - A\sum X_i - B\sum X_i^2 = 0$$

Similarly, $\sum \widehat{Y}_i E_i = 0$.[9] These properties, which will be useful to us below, imply that the least-squares residuals are uncorrelated with both the explanatory variable $X$ and the fitted values $\widehat{Y}$.[10]

---

[6] In Chapter 10, I will derive the least-squares solution by an alternative geometric approach.

[7] As a formal matter, it remains to be shown that the solution of the normal equations *minimizes* the least-squares function $S(A, B)$. See Section 9.2.

[8] The term *normal* here refers not to the normal distribution but to orthogonality (perpendicularity); see Chapter 10 on the vector geometry of regression.

[9] See Exercise 5.1.

[10] See the next section for a definition of correlation.

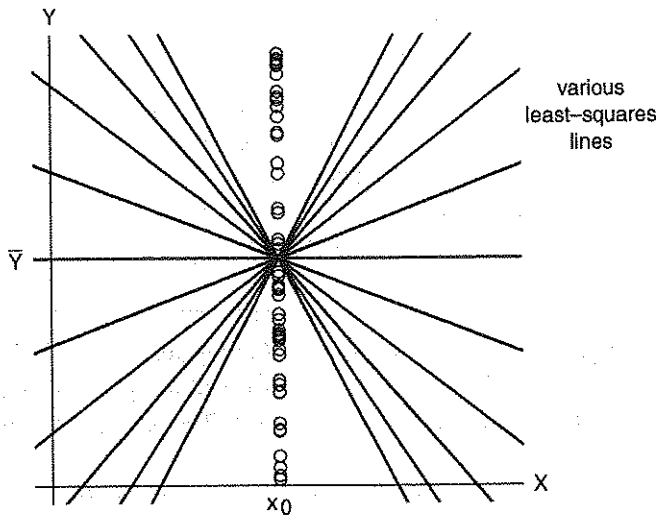


**Figure 5.3** When all values of $X$ are the same ($x_0$), any line through the point ($x_0$, $\overline{Y}$) is a least-squares line.

It is clear from Equation 5.3 that the least-squares coefficients are uniquely defined as long as the explanatory-variable values are not all identical, for when there is no variation in $X$, the denominator of $B$ vanishes. This result is intuitively plausible: Only if the explanatory-variable scores are spread out can we hope to fit a (unique) line to the $X, Y$ scatter; if, alternatively, all the $X$ values are the same (say, equal to $x_0$), then, as is shown in Figure 5.3, any line through the point ($x_0$, $\overline{Y}$) is a least-squares line.

I will illustrate the least-squares calculations using Davis's data on measured weight ($Y$) and reported weight ($X$), for which

$$
\begin{aligned}
n &= 101 \\
\overline{Y} &= \frac{5780}{101} = 57.228 \\
\overline{X} &= \frac{5731}{101} = 56.743 \\
\sum (X_i - \overline{X})(Y_i - \overline{Y}) &= 4435.9 \\
\sum (X_i - \overline{X})^2 &= 4539.3 \\
B &= \frac{4435.9}{4539.3} = 0.97722 \\
A &= 57.228 - 0.97722 \times 56.743 = 1.7776
\end{aligned}
$$

Thus, the least-squares regression equation is

$$\widehat{\text{Measured weight}} = 1.78 + 0.977 \times \text{Reported weight}$$

Interpretation of the least-squares slope coefficient is straightforward: $B = 0.977$ indicates that a 1-kg increase in reported weight is associated, on average, with just under a 1-kg increase in measured weight. Because the data are not longitudinal, the phrase "a unit increase" here implies

not a literal change over time, but rather a notional static comparison between two individuals who differ by 1 kg in their reported weights.

Ordinarily, we may interpret the intercept $A$ as the fitted value associated with $X = 0$, but it is, of course, impossible for an individual to have a reported weight equal to 0. The intercept $A$ is usually of little direct interest, because the fitted value above $X = 0$ is rarely important. Here, however, if individuals' reports are unbiased predictions of their actual weights, then we should have the equation $\widehat{Y} = X$—that is, an intercept of 0 and a slope of 1. The intercept $A = 1.78$ is indeed close to 0, and the slope $B = 0.977$ is close to 1.

> In simple linear regression, the least-squares coefficients are given by $A = \overline{Y} - B\overline{X}$ and $B = \sum(X_i - \overline{X})(Y_i - \overline{Y})/\sum(X_i - \overline{X})^2$. The slope coefficient $B$ represents the average change in $Y$ associated with a one-unit increase in $X$. The intercept $A$ is the fitted value of $Y$ when $X = 0$.

## 5.1.2 Simple Correlation

Having calculated the least-squares line, it is of interest to determine how closely the line fits the scatter of points. This is a vague question, which may be answered in a variety of ways. The standard deviation of the residuals, $S_E$, often called the *standard error of the regression* or the *residual standard error*, provides one sort of answer.[11] Because of estimation considerations, the variance of the residuals is defined using *degrees of freedom* $n - 2$, rather than the sample size $n$, in the denominator:[12]

$$S_E^2 = \frac{\sum E_i^2}{n-2}$$

The residual standard error is, therefore,

$$S_E = \sqrt{\frac{\sum E_i^2}{n-2}}$$

Because it is measured in the units of the response variable, and represents a type of "average" residual, the standard error is simple to interpret. For example, for Davis's regression of measured weight on reported weight, the sum of squared residuals is $\sum E_i^2 = 418.87$, and thus the standard error of the regression is

$$S_E = \sqrt{\frac{418.87}{101-2}} = 2.0569 \text{ kg}$$

On average, then, using the least-squares regression line to predict measured weight from reported weight results in an error of about 2 kg, which is small, but perhaps not negligible. Moreover, if the residuals are approximately normally distributed, then about 2/3 of them are in the range $\pm 2$, and about 95% are in the range $\pm 4$. I believe that social scientists overemphasize correlation

[11] The term *standard error* is usually used for the estimated standard deviation of the sampling distribution of a statistic, and so the use here to denote the standard deviation of the residuals is potentially misleading. This usage is common, however, and I therefore adopt it.

[12] Estimation is discussed in the next chapter. Also see the discussion in Section 10.3.

(described immediately below) and pay insufficient attention to the standard error of the regression as an index of fit.

In contrast to the standard error of the regression, the *correlation coefficient* provides a *relative* measure of fit: To what degree do our predictions of $Y$ improve when we base these predictions on the linear relationship between $Y$ and $X$? A relative index of fit requires a baseline—how well can $Y$ be predicted if $X$ is disregarded?

To disregard the explanatory variable is implicitly to fit the equation $\widehat{Y}_i' = A'$, or, equivalently,

$$Y_i = A' + E_i'$$

By ignoring the explanatory variable, we lose our ability to differentiate among the observations; as a result, the fitted values are constant. The constant $A'$ is generally different from the intercept $A$ of the least-squares line, and the residuals $E_i'$ are different from the least-squares residuals $E_i$.

How should we find the best constant $A'$? An obvious approach is to employ a least-squares fit—that is, to minimize

$$S(A') = \sum E_i'^2 = \sum (Y_i - A')^2$$

As you may be aware, the value of $A'$ that minimizes this sum of squares is simply the response-variable mean, $\overline{Y}$.[13]

The residuals $E_i = Y_i - \widehat{Y}_i$ from the linear regression of $Y$ on $X$ will mostly be smaller in magnitude than the residuals $E_i' = Y_i - \overline{Y}$, and it is necessarily the case that

$$\sum (Y_i - \widehat{Y}_i)^2 \leq \sum (Y_i - \overline{Y})^2$$

This inequality holds because the "null model," $Y_i = A' + E_i'$, specifying no relationship between $Y$ and $X$, is a special case of the more general linear regression "model," $Y_i = A + BX_i + E_i$: The two models are the same when $B = 0$.[14] The null model therefore cannot have a smaller sum of squared residuals. After all, the least-squares coefficients $A$ and $B$ are selected precisely to minimize $\sum E_i^2$, so constraining $B = 0$ cannot improve the fit and will usually make it worse.

We call

$$\sum E_i'^2 = \sum (Y_i - \overline{Y})^2$$

the *total sum of squares* for $Y$, abbreviated TSS, while

$$\sum E_i^2 = \sum (Y_i - \widehat{Y}_i)^2$$

is called the *residual sum of squares*, and is abbreviated RSS. The difference between the two, termed the *regression sum of squares*,

$$\text{RegSS} \equiv \text{TSS} - \text{RSS}$$

gives the reduction in squared error due to the linear regression. The ratio of RegSS to TSS, the proportional reduction in squared error, defines the square of the correlation coefficient:

$$r^2 \equiv \frac{\text{RegSS}}{\text{TSS}}$$

[13] See Exercise 5.3.

[14] A formal statistical model for linear regression is introduced in the next chapter.

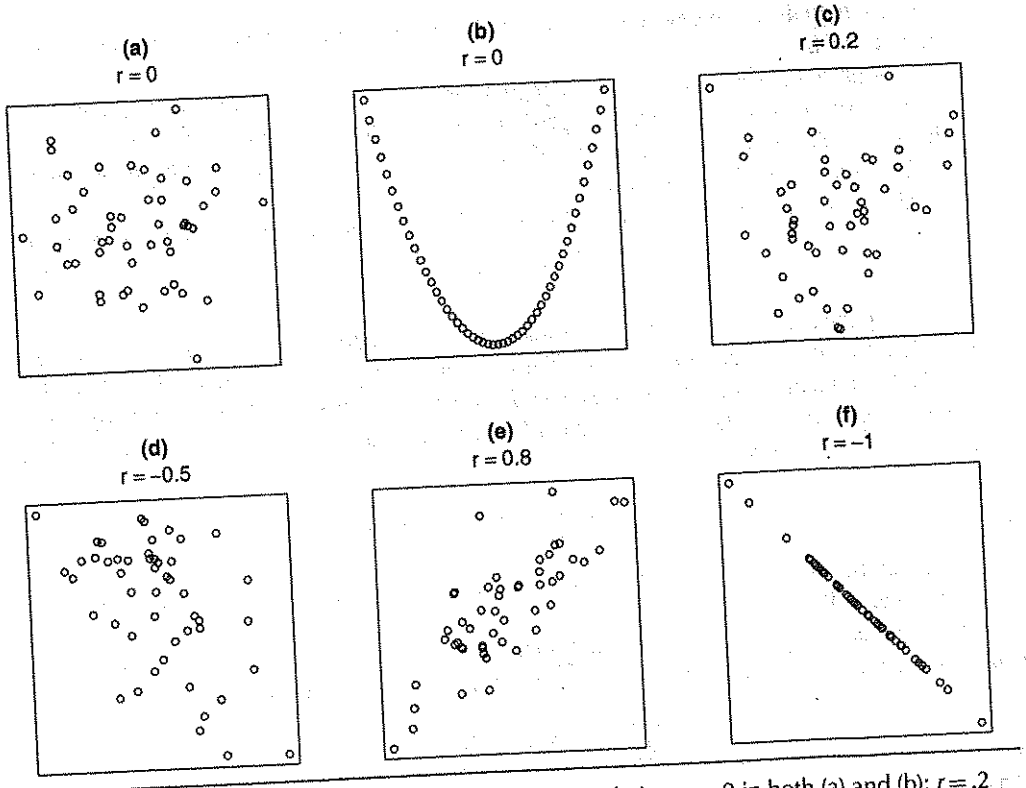


**Figure 5.4** Scatterplots illustrating different levels of correlation: $r = 0$ in both (a) and (b); $r = .2$ in (c); $r = -.5$ in (d); $r = .8$ in (e); and $r = -1$ in (f). All the data sets have $n = 50$ observations. Except in panel (b), the data were generated by sampling from bivariate normal distributions.

To find the *correlation coefficient* $r$, we take the positive square root of $r^2$ when the simple-regression slope $B$ is positive and the negative square root when $B$ is negative.

Thus, if there is a perfect positive linear relationship between $Y$ and $X$ (i.e., if all of the residuals are 0 and $B > 0$), then $r = 1$. A perfect negative linear relationship corresponds to $r = -1$. If there is no linear relationship between $Y$ and $X$, then RSS = TSS, RegSS = 0, and $r = 0$. Between these extremes, $r$ gives the direction of the linear relationship between the two variables, and $r^2$ can be interpreted as the proportion of the total variation of $Y$ that is "captured" by its linear regression on $X$. Figure 5.4 illustrates several levels of correlation. As is clear in Figure 5.4(b), where $r = 0$, the correlation can be small even when there is a strong *nonlinear* relationship between $X$ and $Y$.

It is instructive to examine the three sums of squares more closely: Starting with an individual observation, we have the identity

$$Y_i - \overline{Y} = (Y_i - \widehat{Y}_i) + (\widehat{Y}_i - \overline{Y})$$

This equation is interpreted geometrically in Figure 5.5. Squaring both sides of the equation and summing over observations produces

$$\sum (Y_i - \overline{Y})^2 = \sum (Y_i - \widehat{Y}_i)^2 + \sum (\widehat{Y}_i - \overline{Y})^2 + 2\sum (Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y})$$

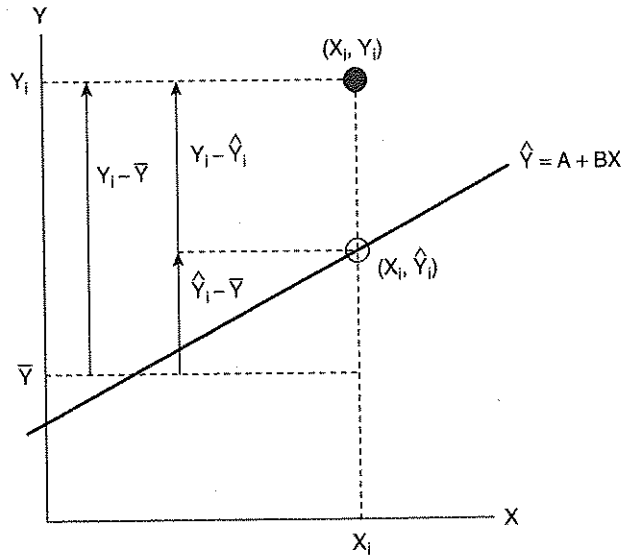


**Figure 5.5** Decomposition of the total deviation $Y_i - \overline{Y}$ into components $Y_i - \widehat{Y}_i$ and $\widehat{Y}_i - \overline{Y}$.

The last term in this equation is 0,[15] and thus the regression sum of squares, which I previously defined as the difference TSS − RSS, may also be written directly as

$$\text{RegSS} = \sum(\widehat{Y}_i - \overline{Y})^2$$

This decomposition of total variation into "explained" and "unexplained" components, paralleling the decomposition of each observation into a fitted value and a residual, is typical of linear models. The decomposition is called the *analysis of variance* for the regression: TSS = RegSS + RSS.

Although I have developed the correlation coefficient from the regression of $Y$ on $X$, it is also possible to define $r$ by analogy with the correlation $\rho = \sigma_{XY}/\sigma_X\sigma_Y$ between two random variables (where $\sigma_{XY}$ is the covariance of the random variables $X$ and $Y$, $\sigma_X$ is the standard deviation of $X$, and $\sigma_Y$ is the standard deviation of $Y$).[16] First defining the *sample covariance* between $X$ and $Y$,

$$S_{XY} \equiv \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{n - 1}$$

we may then write

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum(X_i - \overline{X})^2 \sum(Y_i - \overline{Y})^2}} \tag{5.4}$$

where $S_X$ and $S_Y$ are, respectively, the sample standard deviations of $X$ and $Y$.[17]

It is immediately apparent from the symmetry of Equation 5.4 that the correlation does not depend on which of the two variables is treated as the response variable. This property of $r$ is

[15] See Exercise 5.1 and Section 10.1.

[16] See Appendix D on probability and estimation.

[17] The equivalence of the two formulas for $r$ is established in Section 10.1 on the geometry of simple regression analysis.

surprising in light of the *asymmetry* of the regression equation used to define the sums of squares: Unless there is a perfect correlation between the two variables, the least-squares line for the regression of $Y$ on $X$ differs from the line for the regression of $X$ on $Y$.[18]

There is another central property, aside from symmetry, that distinguishes the correlation coefficient $r$ from the regression slope $B$. The slope coefficient $B$ is measured in the units of the response variable per unit of the explanatory variable. For example, if dollars of income are regressed on years of education, then the units of $B$ are dollars/year. The correlation coefficient $r$, however, is unitless, as can be seen from either of its definitions. As a consequence, a change in scale of $Y$ or $X$ produces a compensating change in $B$, but does not affect $r$. If, for example, income is measured in thousands of dollars rather than in dollars, the units of the slope become \$1,000s/year, and the value of the slope decreases by a factor of 1,000, but $r$ remains the same.

For Davis's regression of measured on reported weight,

$$\begin{aligned}\text{TSS} &= 4753.8\\ \text{RSS} &= 418.87\\ \text{RegSS} &= 4334.9\end{aligned}$$

Thus,

$$r^2 = \frac{4334.9}{4753.8} = .91188$$

and, because $B$ is positive, $r = +\sqrt{.91188} = .9549$. The linear regression of measured on reported weight, therefore, captures 91% of the variation in measured weight. Equivalently,

$$\begin{aligned}S_{XY} &= \frac{4435.9}{101-1} = 44.359\\ S_X^2 &= \frac{4539.3}{101-1} = 45.393\\ S_Y^2 &= \frac{4753.8}{101-1} = 47.538\\ r &= \frac{44.359}{\sqrt{45.393 \times 47.538}} = .9549\end{aligned}$$

## 5.2 Multiple Regression

### 5.2.1 Two Explanatory Variables

The linear multiple-regression equation

$$\widehat{Y} = A + B_1X_1 + B_2X_2$$

for two explanatory variables, $X_1$ and $X_2$, describes a plane in the three-dimensional $\{X_1, X_2, Y\}$ space, as shown in Figure 5.6. As in the case of simple regression, it is unreasonable to expect that the regression plane will pass precisely through every point, so the fitted value for observation $i$ in general differs from the observed value. The residual is the signed vertical distance from the point to the plane:

$$E_i = Y_i - \widehat{Y}_i = Y_i - (A + B_1X_{i1} + B_2X_{i2})$$

[18] See Exercise 5.2.

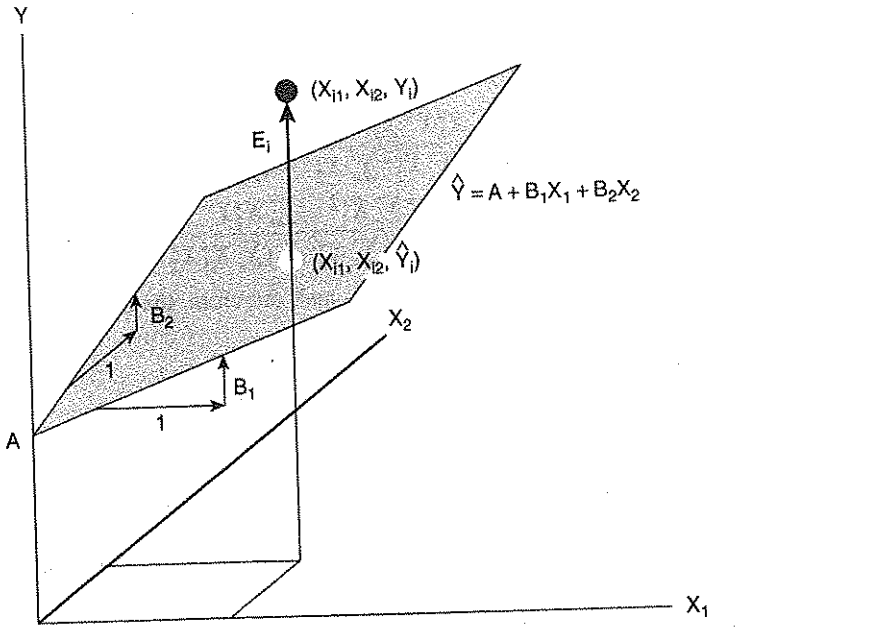


**Figure 5.6** The multiple-regression plane, showing the partial slopes $B_1$ and $B_2$ and the residual $E_i$ for the $i$th observation. The white dot in the regression plane represents the fitted value. Compare this graph with Figure 5.2 for simple regression.

To make the plane come as close as possible to the points in the aggregate, we want the values of $A$, $B_1$, and $B_2$ that minimize the sum of squared residuals:

$$S(A, B_1, B_2) = \sum E_i^2 = \sum (Y_i - A - B_1X_{i1} - B_2X_{i2})^2$$

*As in simple regression, we can proceed by differentiating the sum-of-squares function with respect to the regression coefficients:

$$\frac{\partial S(A, B_1, B_2)}{\partial A} = \sum (-1)(2)(Y_i - A - B_1X_{i1} - B_2X_{i2})$$
$$\frac{\partial S(A, B_1, B_2)}{\partial B_1} = \sum (-X_{i1})(2)(Y_i - A - B_1X_{i1} - B_2X_{i2})$$
$$\frac{\partial S(A, B_1, B_2)}{\partial B_2} = \sum (-X_{i2})(2)(Y_i - A - B_1X_{i1} - B_2X_{i2})$$

Setting the partial derivatives to 0 and rearranging terms produces the normal equations for the regression coefficients $A$, $B_1$, and $B_2$.

The normal equations for the regression coefficients $A$, $B_1$, and $B_2$ are

$$\begin{aligned} An + B_1 \sum X_{i1} + B_2 \sum X_{i2} &= \sum Y_i \\ A \sum X_{i1} + B_1 \sum X_{i1}^2 + B_2 \sum X_{i1}X_{i2} &= \sum X_{i1}Y_i \\ A \sum X_{i2} + B_1 \sum X_{i2}X_{i1} + B_2 \sum X_{i2}^2 &= \sum X_{i2}Y_i \end{aligned} \tag{5.5}$$

Because Equation 5.5 is a system of three linear equations in three unknowns, it usually provides a unique solution for the least-squares regression coefficients $A$, $B_1$, and $B_2$. We can write out the

solution explicitly, if somewhat tediously: Dropping the subscript $i$ for observations, and using asterisks to denote variables in mean-deviation form (e.g., $Y^* \equiv Y_i - \overline{Y}$),

$$\begin{aligned} A &= \overline{Y} - B_1\overline{X}_1 - B_2\overline{X}_2 \\ B_1 &= \frac{\sum X_1^* Y^* \sum X_2^{*2} - \sum X_2^* Y^* \sum X_1^* X_2^*}{\sum X_1^{*2} \sum X_2^{*2} - (\sum X_1^* X_2^*)^2} \\ B_2 &= \frac{\sum X_2^* Y^* \sum X_1^{*2} - \sum X_1^* Y^* \sum X_1^* X_2^*}{\sum X_1^{*2} \sum X_2^{*2} - (\sum X_1^* X_2^*)^2} \end{aligned} \tag{5.6}$$

The denominator of $B_1$ and $B_2$ is nonzero—and, therefore, the least-squares coefficients are uniquely defined—as long as

$$\sum X_1^{*2} \sum X_2^{*2} \neq \left(\sum X_1^* X_2^*\right)^2$$

This condition is satisfied unless $X_1$ and $X_2$ are perfectly correlated or unless one of the explanatory variables is invariant.[19] If $X_1$ and $X_2$ are perfectly correlated, then they are said to be *collinear*.

To illustrate the computation of multiple-regression coefficients, I will employ Duncan's occupational prestige data, which were introduced in Chapter 3. I will, for the time being, disregard the problems with these data that were revealed by graphical analysis. Recall that Duncan wished to predict the prestige of occupations ($Y$) from their educational and income levels ($X_1$ and $X_2$, respectively). I calculated the following quantities from Duncan's data:

$$\begin{aligned} n &= 45 \\ \overline{Y} &= \frac{2146}{45} = 47.689 \\ \overline{X}_1 &= \frac{2365}{45} = 52.556 \\ \overline{X}_2 &= \frac{1884}{45} = 41.867 \\ \sum X_1^{*2} &= 38,971 \\ \sum X_2^{*2} &= 26,271 \\ \sum X_1^* X_2^* &= 23,182 \\ \sum X_1^* Y^* &= 35,152 \\ \sum X_2^* Y^* &= 28,383 \end{aligned}$$

Substituting these values into Equation 5.6 produces $A = -6.0647$, $B_1 = 0.54583$, and $B_2 = 0.59873$. The fitted least-squares regression equation is, therefore,

$$\widehat{\text{Prestige}} = -6.065 + 0.5458 \times \text{Education} + 0.5987 \times \text{Income}$$

Although the development of least-squares linear regression for two explanatory variables is very similar to the development for simple regression, there is this important difference in

[19] The correlation between $X_1$ and $X_2$ is, in the current notation,

$$r_{12} = \frac{\sum X_1^* X_2^*}{\sqrt{\sum X_1^{*2} \sum X_2^{*2}}}$$

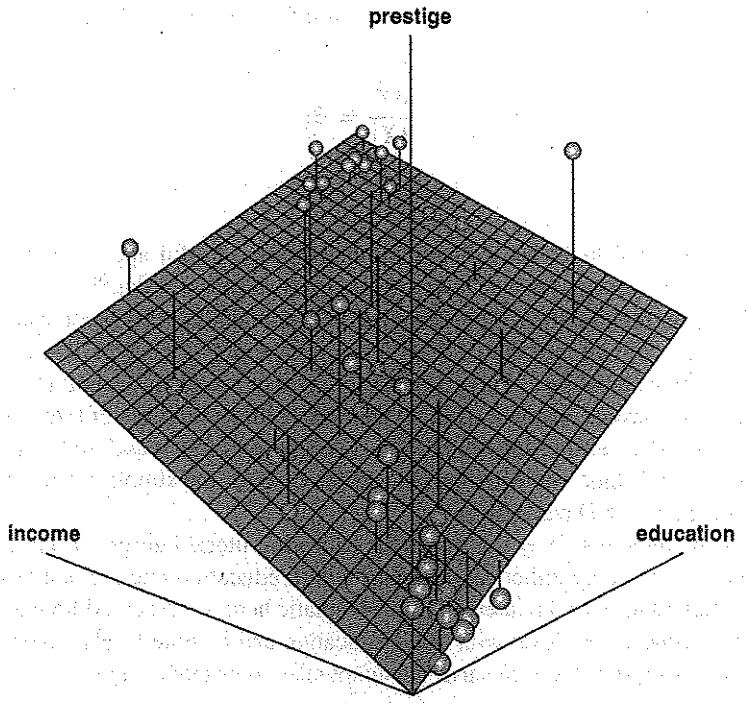


**Figure 5.7** The multiple-regression plane in Duncan's regression of prestige on education and income. The two sets of parallel lines on the regression plane represent the partial relationship of prestige to each explanatory variable holding the other explanatory variable at particular values.

interpretation: The slope coefficients for the explanatory variables in multiple regression are *partial* coefficients, while the slope coefficient in simple regression gives the *marginal* relationship between the response variable and a single explanatory variable. That is, each slope in multiple regression represents the "effect" on the response variable of a one-unit increment in the corresponding explanatory variable *holding constant* the value of the other explanatory variable. The simple-regression slope effectively ignores the other explanatory variable.

This interpretation of the multiple-regression slope is apparent in Figure 5.7, which shows the multiple-regression plane for Duncan's regression of prestige on education and income (also see Figure 5.6). Because the regression plane is flat, its slope ($B_1$) in the direction of education, holding income constant, does not depend on the specific value at which income is fixed. Likewise, the slope in the direction of income, fixing the value of education, is always $B_2$.

Algebraically, let us fix $X_2$ to the specific value $x_2$ and see how $\widehat{Y}$ changes as $X_1$ is increased by 1, from some specific value $x_1$ to $x_1 + 1$:

$$[A + B_1(x_1 + 1) + B_2x_2] - (A + B_1x_1 + B_2x_2) = B_1$$

Similarly, increasing $X_2$ by 1, fixing $X_1$ produces

$$[A + B_1x_1 + B_2(x_2 + 1)] - (A + B_1x_1 + B_2x_2) = B_2$$

*Because the regression surface

$$\widehat{Y} = A + B_1X_1 + B_2X_2$$

is a plane, precisely the same results follow from differentiating the regression equation with respect to each of $X_1$ and $X_2$:

$$\frac{\partial \widehat{Y}}{\partial X_1} = B_1$$
$$\frac{\partial \widehat{Y}}{\partial X_2} = B_2$$

Nothing new is learned here, but differentiation is often a useful approach for understanding *nonlinear* statistical models, for which the regression surface is not flat.[20]

For Duncan's regression, then, a unit increase in education, holding income constant, is associated, on average, with an increase of 0.55 units in prestige (which, recall, is the percentage of respondents rating the prestige of the occupation as good or excellent). A unit increase in income, holding education constant, is associated, on average, with an increase of 0.60 units in prestige. Because Duncan's data are not longitudinal, this language of "increase" or "change" is a shorthand for hypothetical static comparisons (as was the case for the simple regression of measured on reported weight using Davis's data).

The regression intercept, $A = -6.1$, has the following literal interpretation: The fitted value of prestige is $-6.1$ for a hypothetical occupation with education and income levels both equal to 0. Literal interpretation of the intercept is problematic here, however. Although there are some observations in Duncan's data set with small education and income levels, no occupations have levels of 0. Moreover, the response variable cannot take on negative values.

### 5.2.2 Several Explanatory Variables

The extension of linear least-squares regression to several explanatory variables is straightforward. For the general case of $k$ explanatory variables, the multiple-regression equation is

$$\begin{aligned} Y_i &= A + B_1 X_{i1} + B_2 X_{i2} + \cdots + B_k X_{ik} + E_i \\ &= \widehat{Y}_i + E_i \end{aligned}$$

It is, of course, not possible to visualize the point cloud of the data directly when $k > 2$, but it is a relatively simple matter to find the values of $A$ and the $B$s that minimize the sum of squared residuals:

$$S(A, B_1, B_2, \ldots, B_k) = \sum_{i=1}^{n} [Y_i - (A + B_1 X_{i1} + B_2 X_{i2} + \cdots + B_k X_{ik})]^2$$

Minimization of the sum-of-squares function produces the normal equations for general multiple regression:[21]

$$\begin{aligned} An + B_1 \sum X_{i1} + B_2 \sum X_{i2} + \cdots + B_k \sum X_{ik} &= \sum Y_i \\ A \sum X_{i1} + B_1 \sum X_{i1}^2 + B_2 \sum X_{i1} X_{i2} + \cdots + B_k \sum X_{i1} X_{ik} &= \sum X_{i1} Y_i \\ A \sum X_{i2} + B_1 \sum X_{i2} X_{i1} + B_2 \sum X_{i2}^2 + \cdots + B_k \sum X_{i2} X_{ik} &= \sum X_{i2} Y_i \\ \vdots \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ A \sum X_{ik} + B_1 \sum X_{ik} X_{i1} + B_2 \sum X_{ik} X_{i2} + \cdots + B_k \sum X_{ik}^2 &= \sum X_{ik} Y_i \end{aligned} \tag{5.7}$$

[20] For example, see the discussion of quadratic surfaces in Section 17.1.1.

[21] See Exercise 5.5.

We cannot write out a general solution to the normal equations without specifying the number of explanatory variables $k$, and even for $k$ as small as 3, an explicit solution would be very complicated.[22] Nevertheless, because the normal equations are linear, and because there are as many equations as unknown regression coefficients $(k + 1)$, there is usually a unique solution for the coefficients $A, B_1, B_2, \ldots, B_k$. Only when one explanatory variable is a perfect linear function of others, or when one or more explanatory variables are invariant, will the normal equations not have a unique solution. Dividing the first normal equation through by $n$ reveals that the least-squares surface passes through the point of means $(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_k, \overline{Y})$.

> The least-squares coefficients in multiple linear regression are found by solving the normal equations for the intercept $A$ and the slope coefficients $B_1, B_2, \ldots, B_k$. The slope coefficient $B_1$ represents the average change in $Y$ associated with a one-unit increase in $X_1$ when the other $X$s are held constant.

To illustrate the solution of the normal equations, let us return to the Canadian occupational prestige data, regressing the prestige of the occupations on average education, average income, and the percentage of women in each occupation. Recall that our graphical analysis of the data in Chapter 3 cast doubt on the appropriateness of the linear regression, but I will disregard this problem for now.

The various sums, sums of squares, and sums of products that are required are given in Table 5.1. Notice that the sums of squares and products are very large, especially for income, which is scaled in small units (dollars). Substituting these values into the four normal equations and solving for the regression coefficients produces

$$\begin{aligned} A &= -6.7943 \\ B_1 &= 4.1866 \\ B_2 &= 0.0013136 \\ B_3 &= -0.0089052 \end{aligned}$$

The fitted regression equation is, therefore,

$$\begin{aligned} \widehat{\text{Prestige}} = &-6.794 + 4.187 \times \text{Education} + 0.001314 \times \text{Income} \\ &- 0.008905 \times \text{Percent women} \end{aligned}$$

In interpreting the regression coefficients, we need to keep in mind the units of each variable. Prestige scores are arbitrarily scaled, and range from a minimum of 14.8 to a maximum of 87.2 for these 102 occupations; the interquartile range of prestige is 24.1 points. Education is measured in years, and hence the impact of education on prestige is considerable—a little more than four points, on average, for each year of education, holding income and gender composition constant. Likewise, despite the small absolute size of its coefficient, the partial effect of income is also fairly large—more than 0.001 points, on average, for an additional dollar of income, or more than 1 point for each $1,000. In contrast, the impact of gender composition, holding education and income constant, is very small—an average decline of about 0.01 points for each 1% increase in the percentage of women in an occupation.

[22] As I will show in Section 9.2, however, it is simple to write out a general solution to the normal equations using matrices.

**Table 5.1** Sums of Squares (Diagonal), Sums of Products (Off Diagonal), and Sums (Last Row) for the Canadian Occupational Prestige Data

| *Variable* | *Prestige* | *Education* | *Income* | *Percentage of women* |
|---|---|---|---|---|
| Prestige | **253,618.** | 55,326. | 37,748,108. | 131,909. |
| Education | 55,326. | **12,513.** | 8,121,410. | 32,281. |
| Income | 37,748,108. | 8,121,410. | **6,534,383,460.** | 14,093,097. |
| Percentage of women | 131,909. | 32,281. | 14,093,097. | **187,312.** |
| Sum | 4,777. | 1,095. | 693,386. | 2,956. |

### 5.2.3 Multiple Correlation

As in simple regression, the residual standard error in multiple regression measures the "average" size of the residuals. As before, we divide by degrees of freedom, here $n-(k+1) = n-k-1$, rather than by the sample size $n$, to calculate the variance of the residuals; thus, the standard error of the regression is

$$S_E = \sqrt{\frac{\sum E_i^2}{n-k-1}}$$

Heuristically, we "lose" $k+1$ degrees of freedom by calculating the $k+1$ regression coefficients, $A, B_1, \ldots, B_k$.[23]

For Duncan's regression of occupational prestige on the income and educational levels of occupations, the standard error is

$$S_E = \sqrt{\frac{7506.7}{45-2-1}} = 13.37$$

Recall that the response variable here is the percentage of raters classifying the occupation as good or excellent in prestige; an average prediction error of 13 is substantial given Duncan's purpose, which was to use the regression equation to calculate substitute prestige scores for occupations for which direct ratings were unavailable. For the Canadian occupational prestige data, regressing prestige scores on average education, average income, and gender composition, the standard error is

$$S_E = \sqrt{\frac{6033.6}{102-3-1}} = 7.846$$

which is also a substantial figure.

The sums of squares in multiple regression are defined in the same manner as in simple regression:

$$\text{TSS} = \sum (Y_i - \overline{Y})^2$$
$$\text{RegSS} = \sum (\widehat{Y}_i - \overline{Y})^2$$
$$\text{RSS} = \sum (Y_i - \widehat{Y}_i)^2 = \sum E_i^2$$

[23] A deeper understanding of the central concept of degrees of freedom is developed in Chapter 10.

Of course, the fitted values $\widehat{Y}_i$ and residuals $E_i$ now come from the multiple-regression equation. Moreover, we have a similar analysis of variance for the regression:

$$\text{TSS} = \text{RegSS} + \text{RSS}$$

The least-squares residuals are uncorrelated with the fitted values and with each of the $X$s.[24]

> The linear regression decomposes the variation in $Y$ into "explained" and "unexplained" components: TSS = RegSS + RSS. The least-squares residuals, $E$, are uncorrelated with the fitted values, $\widehat{Y}$, and with the explanatory variables, $X_1, \ldots, X_k$.

The squared multiple correlation $R^2$, representing the proportion of variation in the response variable captured by the regression, is defined in terms of the sums of squares:

$$R^2 \equiv \frac{\text{RegSS}}{\text{TSS}}$$

Because there are now several slope coefficients, potentially with different signs, the *multiple correlation coefficient* is, by convention, the positive square root of $R^2$. The multiple correlation is also interpretable as the simple correlation between the fitted and observed $Y$ values—that is, $r_{\widehat{Y}Y}$.

> The standard error of the regression, $S_E = \sqrt{\sum E_i^2/(n-k-1)}$, gives the "average" size of the regression residuals; the squared multiple correlation, $R^2 = \text{RegSS/TSS}$, indicates the proportion of the variation in $Y$ that is captured by its linear regression on the $X$s.

For Duncan's regression, we have the following sums of squares:

$$\begin{aligned} \text{TSS} &= 43{,}688. \\ \text{RegSS} &= 36{,}181. \\ \text{RSS} &= 7506.7 \end{aligned}$$

The squared multiple correlation,

$$R^2 = \frac{36{,}181}{43{,}688} = .8282$$

indicates that more than 80% of the variation in prestige among the 45 occupations is accounted for by the linear regression of prestige on the income and educational levels of the occupations. For the Canadian prestige regression, the sums of squares and $R^2$ are as follows:

$$\begin{aligned} \text{TSS} &= 29{,}895 \\ \text{RegSS} &= 23{,}862 \\ \text{RSS} &= 6033.6 \\ R^2 &= \frac{23{,}862}{29{,}895} = .7982 \end{aligned}$$

[24] These and other properties of the least-squares fit are derived in Chapters 9 and 10.

Because the multiple correlation can only rise, never decline, when explanatory variables are added to the regression equation,[25] investigators sometimes penalize the value of $R^2$ by a "correction" for degrees of freedom. The corrected (or "adjusted") $R^2$ is defined as

$$\widetilde{R}^2 \equiv 1 - \frac{S_E^2}{S_Y^2} = 1 - \frac{\frac{\text{RSS}}{n-k-1}}{\frac{\text{TSS}}{n-1}}$$

Unless the sample size is very small, however, $\widetilde{R}^2$ will differ little from $R^2$. For Duncan's regression, for example,

$$\widetilde{R}^2 = 1 - \frac{\frac{7506.7}{45-2-1}}{\frac{43{,}688}{45-1}} = .8200$$

### 5.2.4 Standardized Regression Coefficients

Social researchers often wish to compare the coefficients of different explanatory variables in a regression analysis. When the explanatory variables are commensurable (i.e., measured in the same units on the same scale), or when they can be reduced to a common standard, comparison is straightforward. In most instances, however, explanatory variables are not commensurable. Standardized regression coefficients permit a limited assessment of the relative effects of incommensurable explanatory variables.

To place standardized coefficients in perspective, let us first consider an example in which the explanatory variables are measured in the same units. Imagine that the annual dollar income of wage workers is regressed on their years of education, years of labor force experience, and some other explanatory variables, producing the fitted regression equation

$$\widehat{\text{Income}} = A + B_1 \times \text{Education} + B_2 \times \text{Experience} + \cdots$$

Because education and experience are each measured in years, the coefficients $B_1$ and $B_2$ are both expressed in dollars/year and, consequently, can be directly compared. If, for example, $B_1$ is larger than $B_2$, then (disregarding issues arising from sampling variation) a year's increment in education yields a greater average return in income than a year's increment in labor force experience, holding constant the other factors in the regression equation.

It is, as I have mentioned, much more common for explanatory variables to be measured in different units. In the Canadian occupational prestige regression, for example, the coefficient for education is expressed in points (of prestige) per year; the coefficient for income is expressed in points per dollar; and the coefficient of gender composition in points per percentage of women. I have already pointed out that the income coefficient (0.001314) is much smaller than the education coefficient (4.187) not because income is a much less important determinant of prestige, but because the unit of income (the dollar) is small, while the unit of education (the year) is relatively large. If we were to reexpress income in \$1,000s, then we would multiply the income coefficient by 1,000.

By the very meaning of the term, *incommensurable* quantities cannot be directly compared. Still, in certain circumstances, incommensurables can be reduced to a common (e.g., monetary) standard. In most cases, however—as in the prestige regression—there is no obvious basis for this sort of reduction.

In the absence of a theoretically meaningful basis for comparison, an empirical comparison can be made by rescaling regression coefficients according to a measure of explanatory-variable

[25] See Exercise 5.6.

spread. We can, for example, multiply each regression coefficient by the interquartile range of the corresponding explanatory variable. For the Canadian prestige data, the interquartile range of education is 4.2025 years; of income, 4081.3 dollars; and of gender composition, 48.610%. When each explanatory variable is manipulated over this range, holding the other explanatory variables constant, the corresponding average changes in prestige are

$$\begin{aligned} \text{Education:}\ & 4.2025 \times 4.1866 && = 17.59 \\ \text{Income:}\ & 4081.3 \times 0.0013136 && = 5.361 \\ \text{Gender:}\ & 48.610 \times -0.0089052 && = -0.4329 \end{aligned}$$

Thus, education has a larger effect than income over the central half of scores observed in the data, and the effect of gender is very small. Note that this conclusion is distinctly circumscribed: For other data, where the variation in education and income may be different, the relative impact of the variables may also differ, even if the regression coefficients are unchanged.

There is really no deep reason for equating the interquartile range of one explanatory variable to that of another, as we have done here implicitly in calculating the relative "effect" of each. Indeed, the following observation should give you pause: If two explanatory variables are commensurable, and if their interquartile ranges differ, then performing this calculation is, in effect, to adopt a rubber ruler. If expressing coefficients relative to a measure of spread potentially distorts their comparison when explanatory variables are commensurable, then why should the procedure magically allow us to compare coefficients that are measured in different units?

It is much more common to standardize regression coefficients using the standard deviations of the explanatory variables rather than their interquartile ranges. Although I will proceed to explain this procedure, keep in mind that the standard deviation is not a good measure of spread when the distributions of the explanatory variables depart considerably from normality. The usual practice standardizes the response variable as well, but this is an inessential element of the computation of standardized coefficients, because the *relative* size of the slope coefficients does not change when $Y$ is rescaled.

Beginning with the fitted multiple-regression equation

$$Y_i = A + B_1 X_{i1} + \cdots + B_k X_{ik} + E_i$$

let us eliminate the regression constant $A$, expressing all the variables in mean-deviation form by subtracting[26]

$$\overline{Y} = A + B_1 \overline{X}_1 + \cdots + B_k \overline{X}_k$$

which produces

$$Y_i - \overline{Y} = B_1(X_{i1} - \overline{X}_1) + \cdots + B_k(X_{ik} - \overline{X}_k) + E_i$$

Then divide both sides of the equation by the standard deviation of the response variable $S_Y$; simultaneously multiply and divide the $j$th term on the right-hand side of the equation by the standard deviation $S_j$ of $X_j$. These operations serve to standardize each variable in the regression equation:

$$\frac{Y_i - \overline{Y}}{S_Y} = \left(B_1 \frac{S_1}{S_Y}\right) \frac{X_{i1} - \overline{X}_1}{S_1} + \cdots + \left(B_k \frac{S_k}{S_Y}\right) \frac{X_{ik} - \overline{X}_k}{S_k} + \frac{E_i}{S_Y}$$

$$Z_{iY} = B_1^* Z_{i1} + \cdots + B_k^* Z_{ik} + E_i^*$$

[26]Recall that the least-squares regression surface passes through the point of means for the $k + 1$ variables.

In this equation, $Z_{iY} \equiv (Y_i - \overline{Y})/S_Y$ is the standardized response variable, linearly transformed to a mean of 0 and a standard deviation of 1; $Z_{i1}, \ldots, Z_{ik}$ are the explanatory variables, similarly standardized; $E_i^* \equiv E_i/S_Y$ is the transformed residual, which, note, *does not* have a standard deviation of 1; and $B_j^* \equiv B_j(S_j/S_Y)$ is the *standardized partial regression coefficient* for the $j$th explanatory variable. The standardized coefficient is interpretable as the average change in $Y$, in standard-deviation units, for a one standard-deviation increase in $X_j$, holding constant the other explanatory variables.

> By rescaling regression coefficients in relation to a measure of variation—such as the interquartile range or the standard deviation—standardized regression coefficients permit a limited comparison of the relative impact of incommensurable explanatory variables.

For the Canadian prestige regression, we have the following calculations:

$$\begin{array}{lll} \text{Education:} & 4.1866 \times 2.7284/17.204 & = 0.6640 \\ \text{Income:} & 0.0013136 \times 4245.9/17.204 & = 0.3242 \\ \text{Gender:} & -0.0089052 \times 31.725/17.204 & = -0.01642 \end{array}$$

Because both income and gender composition have substantially non-normal distributions, however, the use of standard deviations here is difficult to justify.

I have stressed the restricted extent to which standardization permits the comparison of coefficients for incommensurable explanatory variables. A common misuse of standardized coefficients is to employ them to make comparisons of the effects of the *same* explanatory variable in two or more samples drawn from different populations. If the explanatory variable in question has different spreads in these samples, then spurious differences between coefficients may result, even when *unstandardized* coefficients are similar; on the other hand, differences in unstandardized coefficients can be masked by compensating differences in dispersion.

## Exercises

**Exercise 5.1.** *Prove that the least-squares fit in simple-regression analysis has the following properties:

(a) $\sum \widehat{Y}_i E_i = 0$.

(b) $\sum (Y_i - \widehat{Y}_i)(\widehat{Y}_i - \overline{Y}) = \sum E_i(\widehat{Y}_i - \overline{Y}) = 0$.

**Exercise 5.2.** *Suppose that the means and standard deviations of $Y$ and $X$ are the same: $\overline{Y} = \overline{X}$ and $S_Y = S_X$.

(a) Show that, under these circumstances,

$$B_{Y|X} = B_{X|Y} = r_{XY}$$

where $B_{Y|X}$ is the least-squares slope for the simple regression of $Y$ on $X$; $B_{X|Y}$ is the least-squares slope for the simple regression of $X$ on $Y$; and $r_{XY}$ is the correlation between the two variables. Show that the intercepts are also the same, $A_{Y|X} = A_{X|Y}$.

(b) Why, if $A_{Y|X} = A_{X|Y}$ and $B_{Y|X} = B_{X|Y}$, is the least-squares line for the regression of $Y$ on $X$ different from the line for the regression of $X$ on $Y$ (as long as $r^2 < 1$)?

(c) "Regression toward the mean" (the original sense of the term "regression"): Imagine that $X$ is father's height and $Y$ is son's height for a sample of father-son pairs. Suppose that $S_Y = S_X$, that $\overline{Y} = \overline{X}$, and that the regression of sons' heights on fathers' heights is linear. Finally, suppose that $0 < r_{XY} < 1$ (i.e., fathers' and sons' heights are positively correlated, but not perfectly so). Show that the expected height of a son whose father is shorter than average is also less than average, but to a smaller extent; likewise, the expected height of a son whose father is taller than average is also greater than average, but to a smaller extent. Does this result imply a contradiction—that the standard deviation of son's height is in fact less than that of father's height?

(d) What is the expected height for a father whose son is shorter than average? Of a father whose son is taller than average?

(e) Regression effects in research design: Imagine that educational researchers wish to assess the efficacy of a new program to improve the reading performance of children. To test the program, they recruit a group of children who are reading substantially below grade level; after a year in the program, the researchers observe that the children, on average, have improved their reading performance. Why is this a weak research design? How could it be improved?

**Exercise 5.3.** *Show that $A' = \overline{Y}$ minimizes the sum of squares

$$S(A') = \sum_{i=1}^{n} (Y_i - A')^2$$

**Exercise 5.4.** Linear transformation of $X$ and $Y$:

(a) Suppose that the explanatory-variable values in Davis's regression are transformed according to the equation $X' = X - 10$ and that $Y$ is regressed on $X'$. Without redoing the regression calculations in detail, find $A'$, $B'$, $S'_E$, and $r'$. What happens to these quantities when $X' = 10X$? When $X' = 10(X - 1) = 10X - 10$?

(b) Now suppose that the response variable scores are transformed according to the formula $Y'' = Y + 10$ and that $Y''$ is regressed on $X$. Find $A''$, $B''$, $S''_E$, and $r''$. What happens to these quantities when $Y'' = 5Y$? When $Y'' = 5(Y + 2) = 5Y + 10$?

(c) In general, how are the results of a simple-regression analysis affected by linear transformations of $X$ and $Y$?

**Exercise 5.5.** *Derive the normal equations (Equation 5.7) for the least-squares coefficients of the general multiple-regression model with $k$ explanatory variables. [*Hint*: Differentiate the sum-of-squares function $S(A, B_1, \ldots, B_k)$ with respect to the regression coefficients, and set the partial derivatives to 0.]

**Exercise 5.6.** Why is it the case that the multiple-correlation coefficient $R^2$ can never get smaller when an explanatory variable is added to the regression equation? (*Hint*: Recall that the regression equation is fit by minimizing the residual sum of squares.)

**Exercise 5.7.** Consider the general multiple-regression equation

$$Y = A + B_1X_1 + B_2X_2 + \cdots + B_kX_k + E$$

An alternative procedure for calculating the least-squares coefficient $B_1$ is as follows:

1. Regress $Y$ on $X_2$ through $X_k$, obtaining residuals $E_{Y|2\ldots k}$.
2. Regress $X_1$ on $X_2$ through $X_k$, obtaining residuals $E_{1|2\ldots k}$.
3. Regress the residuals $E_{Y|2\ldots k}$ on the residuals $E_{1|2\ldots k}$. The slope for this simple regression is the multiple-regression slope for $X_1$, that is, $B_1$.
   (a) Apply this procedure to the multiple regression of prestige on education, income, and percentage of women in the Canadian occupational prestige data, confirming that the coefficient for education is properly recovered.
   (b) Note that the intercept for the simple regression in Step 3 is 0. Why is this the case?
   (c) In light of this procedure, is it reasonable to describe $B_1$ as the "effect of $X_1$ on $Y$ when the influence of $X_2, \ldots, X_k$ is removed from both $X_1$ and $Y$"?
   (d) The procedure in this problem reduces the multiple regression to a series of simple regressions (in Step 3). Can you see any practical application for this procedure? (See the discussion of added-variable plots in Section 11.6.1.)

**Exercise 5.8.** Partial correlation: The *partial correlation* between $X_1$ and $Y$ "controlling for" $X_2$ through $X_k$ is defined as the simple correlation between the residuals $E_{Y|2\ldots k}$ and $E_{1|2\ldots k}$, given in the previous exercise. The partial correlation is denoted $r_{Y1|2\ldots k}$.

(a) Using the Canadian occupational prestige data, calculate the partial correlation between prestige and education controlling for income and percentage women (see the previous exercise).
(b) In light of the interpretation of a partial regression coefficient developed in the previous exercise, why is $r_{Y1|2\ldots k} = 0$ if and only if $B_1$ (from the multiple regression of $Y$ on $X_1$ through $X_k$) is 0?

**Exercise 5.9.** *Show that in simple-regression analysis, the standardized slope coefficient $B^*$ is equal to the correlation coefficient $r$. (In general, however, standardized slope coefficients *are not* correlations and can be outside of the range [0, 1].)

## Summary

- In simple linear regression, the least-squares coefficients are given by

$$B = \frac{\sum(X_i - \overline{X})(Y_i - \overline{Y})}{\sum(X_i - \overline{X})^2}$$
$$A = \overline{Y} - B\overline{X}$$

  The slope coefficient $B$ represents the average change in $Y$ associated with a one-unit increase in $X$. The intercept $A$ is the fitted value of $Y$ when $X = 0$.
- The least-squares coefficients in multiple linear regression are found by solving the normal equations for the intercept $A$ and the slope coefficients $B_1, B_2, \ldots, B_k$. The slope coefficient $B_1$ represents the average change in $Y$ associated with a one-unit increase in $X_1$ when the other $X$s are held constant.

- The least-squares residuals, $E$, are uncorrelated with the fitted values, $\widehat{Y}$, and with the explanatory variables, $X_1, \ldots, X_k$.
- The linear regression decomposes the variation in $Y$ into "explained" and "unexplained" components: $\text{TSS} = \text{RegSS} + \text{RSS}$. This decomposition is called the analysis of variance for the regression.
- The standard error of the regression, $S_E = \sqrt{\sum E_i^2/(n-k-1)}$, gives the "average" size of the regression residuals; the squared multiple correlation, $R^2 = \text{RegSS}/\text{TSS}$, indicates the proportion of the variation in $Y$ that is captured by its linear regression on the $X$s.
- By rescaling regression coefficients in relation to a measure of variation—such as the interquartile range or the standard deviation—standardized regression coefficients permit a limited comparison of the relative impact of incommensurable explanatory variables.