You may also want to transform some of the variables. For example, instead of including "number of applications received" and "number of applicants accepted" in the model, you may want to convert to a variable stating number of applicants accepted divided by number of applications received. Try removing variables from the model that are not statistically significant, and see how this changes the results.

5 How Does Bivariate Regression Work?

If you've made it this far in the book, you already know much of what you need to be an intelligent user of multiple regression. If no one is making you read this book, you may decide that you don't want to go any further. Nevertheless, it's hard to learn too much about multiple regression: Regression is such a widely used technique that virtually any additional knowledge or understanding is bound to pay off at some point.

With that in mind, we begin a series of chapters that delve a little more deeply into the algebra, geometry, mechanics, and philosophy of least squares regression. This chapter deals with the simplest setting of a dependent variable y and a single independent variable x. That may seem a little too simple because the main attractiveness of regression lies in its ability to control for other variables. Unfortunately, both the algebra and geometry become much more complicated in the three-variable case, so it's essential to begin a step lower. Even in bivariate regression, we'll have to introduce some mathematical notation.

5.1. How Do We Picture the Least Squares Regression Problem?

To understand bivariate regression, it's helpful to have a picture of the data. Figure 5.1 shows a scatterplot of income by age for the data in Table 1.1. Each little square on this graph represents one of the 35 observations in the data set. For example, the square nearest the lower right-hand corner corresponds to a person 76 years old who had an annual income of $6,000. Although there appears to be

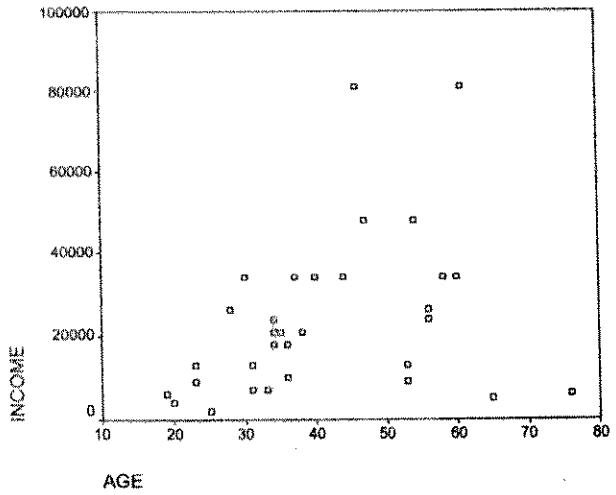

Figure 5.1. Scatterplot of Income on Age for Data in Table 1.1

a general trend of income increasing with age, the relationship is far from perfect.

If we think that the relationship between age and income is best described by a straight line, our problem is to find a line that comes close to the points on the graph. We could, of course, put a ruler on the graph and draw a line that looks like it comes close to many of the points, but someone else could come along and draw a quite different line. How can we tell whose line is better?

In Chapter 1, we introduced the least squares criterion: Choose a line that minimizes the sum of squared prediction errors. Figure 5.2 shows the same scatterplot, but with the least squares regression line added. As you can see, the line passes directly through some of the points but is quite distant from some of the others.

Where did the line come from? Well, I simply asked my graphics program to add a least squares regression line. The program produced the line by performing some simple calculations that we'll discuss in a moment. Those calculations minimize the sum of squared prediction errors (also called residuals). Figure 5.3 shows the prediction errors for this regression line. The errors are represented by vertical lines from each data point to the regression line. The length of each line corresponds to the size of the error. Errors above the regression line are positive; errors below the line are negative.

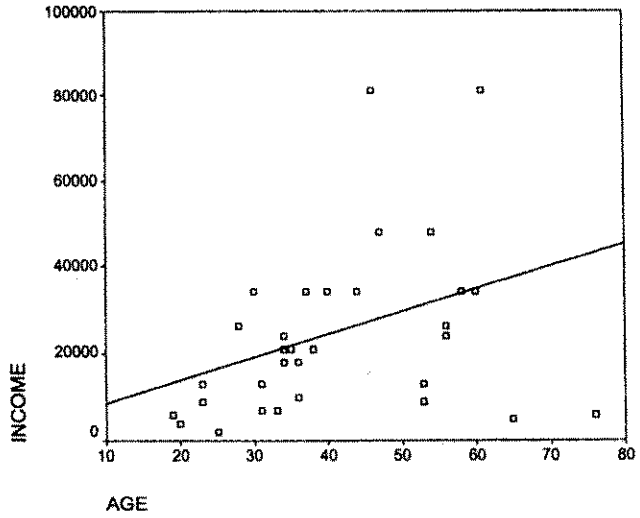

Figure 5.2. Scatterplot With Least Squares Regression Line

Before discussing how the least squares line is calculated, a few things should be noted about this graph.

- As explained in Chapter 1, the regression line corresponds to the general equation $y = a + bx$,

where a and b are numbers to be calculated. Choosing a line is equivalent to choosing a and b.

- The value of a corresponds to the point on the vertical axis where it is crossed by the regression line. We see from the graph that it's around $3,000.
- The slope of the line corresponds to b. In this case, it tells us how much income increases for each 1-year increase in age.
- Although it may not be readily apparent, the line passes through the point $(\bar{x}, \bar{y})$, the intersection of the mean of y and the mean of x. For any bivariate regression, the least squares line will pass through the two means.

5.2. How Is the Least Squares Regression Line Calculated?

To get a regression line, we must find values for a and b in the linear equation $y = a + bx$. The least squares method says to choose

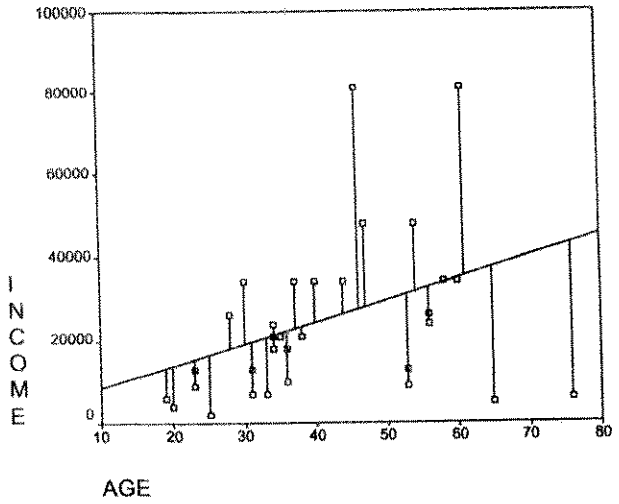

Figure 5.3. Scatterplot With Regression Line and Error Lines

values that minimize the sum of squared residuals. To see how this is accomplished, we need some algebraic notation. Suppose we label each individual in the sample by a number, starting at 1 and going up to n (the number of cases in the sample). The order of the numbers doesn't matter, just so each individual has a unique number. Let i be a variable corresponding to the number. For individual i, let y_i be that individual's value of y and let x_i be that individual's value of x. Thus, x_1 is the value of x for individual 1.

If a and b are possible values of the intercept and slope, the predicted value of y for individual i is just $a + bx_i$. For each individual in the sample, the residual is the observed value of y minus the predicted value of y. The residual, then, is $y_i - (a + bx_i)$. We can therefore represent the sum of *squared* residuals as

$$\sum_i [y_i - (a + bx_i)]^2$$

The capital Greek sigma (Σ) is a common symbol, used to represent summation. The i underneath it means to sum over all the possible values of i, that is, all the individuals in the sample.

As noted in Chapter 1, we could simply try out different values of a and b until we found the ones that produced the smallest sum

of squared residuals. That would be a rather inefficient and time-consuming process, however. Instead, differential calculus can be used to derive formulas for the values of a and b that minimize the sum of squared residuals. These formulas can be used in any sample. The formula for the slope b is

$$b = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

In words, this formula says to take each value of y and subtract the mean of y. This is sometimes called a *deviation score*. Similarly, take each value of x and subtract the mean of x. To get the numerator in the formula for b, we multiply the two deviation scores for each individual, then add them up for all individuals. To get the denominator, we square the deviation score for x for each individual, then add them all together.

The denominator in this formula may look familiar. The formula for the sample variance of x is

$$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n - 1}$$

It follows that if we divide the denominator in the formula for b by $n - 1$, we get the sample variance of x. If we divide the numerator in the formula for b by $n - 1$, we get something called the covariance of x and y:

$$s_{xy} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{n - 1}$$

The covariance is a statistic that is not often interpreted itself, but is typically used as an intermediate step in getting other statistics. Note that the covariance of a variable with itself is just the variance of that variable, that is, $s_{xx} = s_x^2$.

We can put these two results together to get a more compact formula for b. If we divide the covariance of x and y by the variance of x, the $n - 1$ cancels out of the numerator and the denominator, and we're left with our formula for b. We can say, then, that the least

squares slope coefficient is equal to the sample covariance of x and y divided by the sample variance of x, or $b = s_{xy}/s_x^2$.

This is the formula for b that I find easiest to remember.

We'll meet the covariance again when we talk about correlation. Right now, we still need a formula for a, the least squares intercept. Earlier, I said that the least squares line must always pass through the point $(\bar{x}, \bar{y})$. That implies that we can write the equation

$$\bar{y} = a + b\bar{x}.$$

If we solve this equation for a, we get

$$a = \bar{y} - b\bar{x},$$

and this becomes our formula for the intercept. This formula tells us to first calculate b, then multiply b by the mean of x, and then subtract the result from the mean of y.

Table 5.1 displays the intermediate calculations necessary for computing a and b for our age and income example. The second and third columns are just the data for income and age. We first get the means for these two variables by dividing the totals by the sample size, 35:

$$\bar{x} = 1{,}462/35 = 41.77$$

$$\bar{y} = 882{,}000/35 = 25{,}200.$$

The next two columns simply subtract these means from the original data to get the deviation scores. Notice that the deviation scores necessarily sum to 0. The sixth column multiplies the x and y deviation scores together, and the last column is the square of the deviation scores for x.

To get b, we simply take the ratio of the totals for the last two columns:

$$b = 3{,}559{,}600/6{,}796.17 = 523.76.$$

This tells us that each additional year of age is associated with a $524 increase in income. To get a, we calculate

$$a = 25{,}200 - (523.76 \times 41.77) = 3{,}323.$$

Our final regression equation is then

$$y = 3{,}323 + 523.76x.$$

TABLE 5.1 Calculations for Regression of Income on Age

i	*Income* y_i	*Age* x_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	48,000	54	22,800	12.23	278,844	149.57
2	26,000	28	800	−13.77	−11,016	189.61
3	26,000	56	800	14.23	113,840	202.49
4	48,000	47	22,800	5.23	119,244	27.35
5	13,000	23	−12,200	−18.77	228,994	352.31
6	34,000	60	8,800	18.23	160,424	332.33
7	18,000	36	−7,200	−5.77	415,440	33.29
8	24,000	34	−1,200	−7.77	93,240	60.37
9	81,000	61	55,800	19.23	1,073,034	369.79
10	21,000	38	−4,200	−3.77	158,340	14.21
11	9,000	53	−16,200	11.23	−181,926	126.11
12	18,000	34	−7,200	−7.77	559,440	60.37
13	34,000	58	8,800	16.23	142,824	263.41
14	21,000	38	−4,200	−3.77	158,340	14.21
15	81,000	46	55,800	4.23	236,034	17.89
16	48,000	54	22,800	12.23	278,844	149.57
17	6,000	76	−19,200	34.23	−657,216	1,171.69
18	21,000	35	−4,200	−6.77	284,340	45.83
19	21,000	34	−4,200	−7.77	326,340	60.37
20	9,000	23	−16,200	−18.77	304,074	352.31
21	34,000	44	8,800	2.23	196,240	4.97
22	7,000	31	−18,200	−10.77	196,014	115.99
23	24,000	56	−1,200	14.23	−17,076	202.49
24	34,000	37	8,800	−4.77	−41,976	22.75
25	34,000	40	8,800	−1.77	−15,576	3.13
26	4,000	20	−21,200	−21.77	461,524	473.93
27	5,000	65	−20,200	23.23	−469,246	539.63
28	13,000	53	−12,200	11.23	−137,006	126.11
29	7,000	33	−18,200	−8.77	159,614	76.91
30	13,000	31	−12,200	−10.77	131,394	115.99
31	34,000	30	8,800	−11.77	−103,576	138.53
32	10,000	36	−15,200	−5.77	877,040	33.29
33	48,000	54	22,800	12.23	278,844	149.57
34	6,000	19	−19,200	−22.77	437,184	518.47
35	2,000	25	−23,200	−16.77	389,064	281.23
Total	882,000	1,462	0	0.0	3,559,600	6,796.17

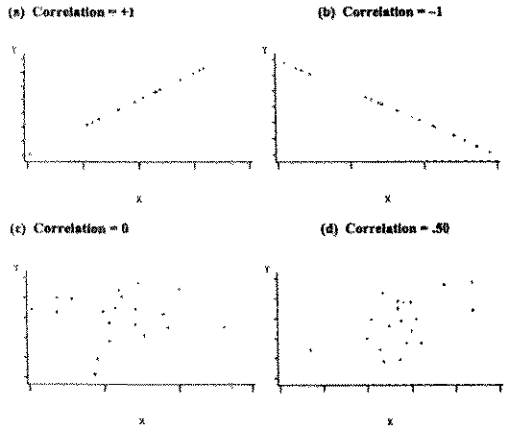

Figure 5.4. Scatterplots for Correlations of Various Sizes

5.3. How Is Regression Related to Correlation?

The regression line is a way of describing the relationship between two variables x and y. Another popular way of describing the relationship between two variables is the correlation coefficient. More precisely, it's called the *Pearson product-moment correlation coefficient*. It turns out that there is a very close relationship between the regression line and the correlation coefficient. In fact, they can be thought of as two alternative ways of describing the same thing. Let's see how.

The correlation coefficient can have any value between −1 and +1. If the correlation between x and y is +1, then there is a perfect linear relationship between the two variables. That means that if we draw a scatterplot for data on x and y, the points will all lie exactly on a straight line, as in Figure 5.4a. In Figure 5.4b, we see a scatterplot for a correlation of −1. Again there is a perfect linear relationship between x and y, but now it's an *inverse* relationship: As x increases, y decreases.

In Figure 5.4c, we see a scatterplot for two variables that have a correlation of 0. In this case, there is *no* linear relationship between x and y. If we were to draw a least squares regression line for Figure 5.4c, it would be a horizontal line passing through the mean of y. In other words, the slope would be 0.

Finally, Figure 5.4d shows a scatterplot corresponding to a correlation of .50. In this graph, there is a clear tendency for y to go up as x goes up, but if we drew a least squares regression line, there would be a good deal of scatter around the line. This plot looks vaguely similar to the plot for age and income in Figure 5.1. In fact, the correlation for that plot is .38, not very far from .50.

In general, we say that the correlation coefficient measures the *degree of scatter* around a regression line. The slope coefficient b, on the other hand, measures the steepness of the regression line. With one exception, the magnitude of the slope has nothing to do with the degree of scatter. It's possible to have a very steep line with a lot of scatter or a very shallow line with little scatter. The exception is that when the slope is 0 the correlation must also be 0, and vice versa. This fact can be seen by comparing the formulas for the two statistics. As we saw earlier, one simple formula for the slope is $b = s_{xy}/s_x^2$, the covariance of x and y divided by the variance of x. The correlation coefficient, often denoted by the letter r, can be written as

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

In words, the correlation is equal to the covariance of x and y divided by the standard deviation of x multiplied by the standard deviation of y. What's important here is that both r and b have the same numerator—the covariance. It follows that if one of them is equal to zero, the other must also be zero. Although that may seem like just a mathematical curiosity, it also has a major practical implication: Testing the hypothesis that $r = 0$ is equivalent to testing whether $b = 0$. There's no need for two distinct tests.

There's another way in which the correlation coefficient is related to regression. In Chapter 1, we discussed the statistic R^2 as a measure of how well we could predict the dependent variable from knowledge of the independent variables. In a bivariate regression, it turns out that the R^2 is equal to the squared correlation between x and y, that is, r^2. In fact, that's how R^2 got its name (or symbol). This can be generalized by the following principle: For any multiple regression, the R^2 is equal to the squared correlation between the observed values of the dependent variable and the predicted values of the dependent variable.

There is one final point to keep in mind about regression and correlation. Because both statistics measure the *linear* relationship

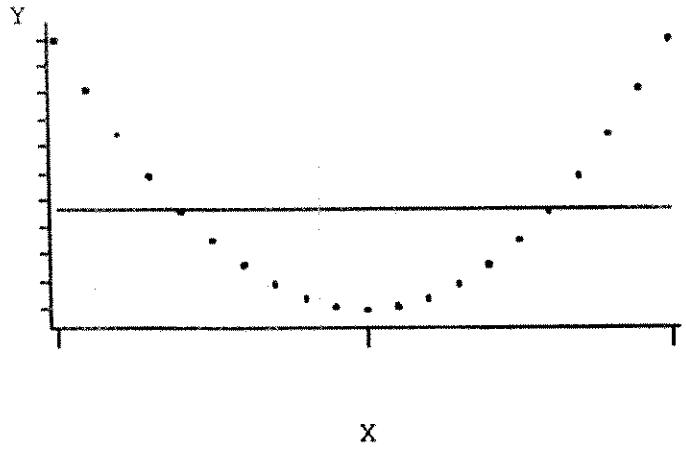

Figure 5.5. A Nonlinear Relationship Between Two Variables

between two variables, they can be highly misleading if the true relationship is very nonlinear. Consider the graph in Figure 5.5. As the scatterplot shows, y can be perfectly predicted from knowledge of x, but the relationship is *quadratic* rather than linear. As x increases, y sometimes decreases and sometimes increases. The net result is that the regression line (the solid, horizontal line) has a slope of 0. Naturally, the correlation coefficient is also zero. If all we knew about the relationship between x and y were the correlation and the slope, we would be seriously mistaken in our assessment of the relationship between these two variables.

5.4. How Is the Standard Error Calculated?

Another important statistic in any regression application is the standard error of the slope coefficient. (The intercept also has a standard error, but we're usually not very interested in that.) We need the standard error of the slope so that we can construct a confidence interval or a test statistic. There are a variety of equivalent formulas for the standard error. I like the one below because it is expressed in terms of statistics we've already encountered and because it is easily interpreted:

$$s.e.\ (b) = \frac{s_y}{s_x}\sqrt{\frac{1-r^2_{xy}}{n-2}}$$

In this formula, s_y and s_x are the standard deviations of y and x. The sample size is n, and r^2_{xy} is the squared correlation between x and y (equivalently, the R^2). For the age and income regression, the standard deviation of income is 19,514 and the standard deviation of age is 14.13. The R^2 is .144 and the sample size is 35. Putting them all together, we have

$$\frac{19{,}514}{14.13}\sqrt{\frac{1-.144}{35-2}} = 222$$

A closer look at the formula for the standard error can tell us some interesting things. First, the fact that n is in the denominator means that the standard error gets smaller as the sample gets larger. Of course, this is true for the standard error of almost any statistic: The bigger the sample, the more precisely we can estimate the statistic. Second, although the squared correlation is in the numerator, it's subtracted from 1. Consequently, as the correlation gets larger in magnitude, the standard error gets smaller. Again, this makes sense. When all the data points are close to the line, we can get very good estimates of the slope, but when there's a large amount of scatter around the line, our estimates of the slope become very imprecise. Finally, note that the standard deviation of x is in the denominator. This means that we get better estimates when there is a lot of variation in the *independent variable*. This can be an important consideration in the design of empirical studies.

Now that we've calculated our standard error of 222, let's see what we can do with it. Recall that the slope coefficient was 524 (rounded). To test the null hypothesis that the true slope is zero, we simply take the ratio of the coefficient to its standard error, 524/222 = 2.36. If the sample had more than 100 cases or so, we could simply treat this statistic as a *standard normal variable*. In that case, anything over 2 would be considered statistically significant, and we could conclude that the true slope is *not* zero. Because our sample is only 35 cases, we need to use a t table (found in any introductory statistics textbook). To use a t table, you must know the *degrees of freedom (df)*.

For a bivariate regression, the number of degrees of freedom is $n - 2$ or, in our case, 33. Most t tables give values for 30 *df* and 40 *df*, but not for the values in between. We'll use 30 because it's both closer and more conservative. For a two-tailed test, at the .05 level and 30 *df*, the *critical value* for statistical significance is 2.04. Anything greater than 2.04 or less than –2.04 is said to be statistically significant. So we can say that the coefficient for age is significantly different from zero, which means that there's some evidence that age is related to income.

The other thing we can do with the standard error is construct a confidence interval. To get a 95% confidence interval, we first multiply the standard error by the critical t value for a two-tailed, .05 test. As we just saw, that critical value is 2.04. So $2.04 \times 222 = 453$. Then we both add and subtract this number from the slope coefficient:

$$524 - 453 = 71$$

$$524 + 453 = 977$$

Then, we can say with 95% confidence that the true slope lies somewhere between 71 and 977. That's a pretty wide interval, but, of course, the sample is rather small. The interpretation of the confidence interval is this: If we repeated this procedure many times (with a different sample of 35 cases each time), about 95% of the confidence intervals that we calculated would contain the true value. The accuracy of that statement depends on several assumptions that are discussed in the next chapter.

5.5. How Is Bivariate Regression Related to Trivariate Regression?

As I mentioned at the beginning of this chapter, it's a shame that we have to focus on bivariate regression because regression doesn't really get interesting until we have at least two independent variables. But the mathematics gets rather complicated with multiple independent variables. Nevertheless, when there are just two independent variables, it turns out that the mathematics can be translated into things we've already learned about bivariate regression. This can help us understand in what sense regression allows us to statistically control for variables.

In this chapter we've been looking at the relationship between income and age. Now we'll add the third variable in Table 1.1, years of schooling. Let y be income, x be age, and z be years of schooling. We want to estimate a linear equation of the form

$$y = a + b_1x + b_2z.$$

Let's focus on b_1, the effect of x on y, controlling for z. As already noted, many hand calculators have special functions for doing bivariate regression. I'm now going to explain how to use such a calculator to calculate b_1. Before I do so, I want to make it clear that I am not recommending this method for routine use. If you want to do a multiple regression, do it on a computer with an appropriate regression program. The reason for considering this alternative method is to get some insight into how multiple regression works.

We can get b_1 by doing three bivariate regressions. I'll list them first, then explain the details.

1. Regress y on z (that is, take y as the dependent variable and z as the independent variable). Calculate the *residuals* from this regression and call them y^*.
2. Regress x on z. Calculate the *residuals* from this regression and call them x^*.
3. Regress y^* on x^*.

The slope from the third regression will be b_1, exactly the same as if we had used a full-scale multiple regression program.

Now let's look at the details. Step 1 produces the regression equation

$$\hat{y} = 3{,}732 + 1{,}692z_i.$$

The "hat" over the y means that this is the *predicted* value of y. To get the residuals, we subtract the predicted value of y from the observed value of y for each individual:

$$y_i^* = y_i - \hat{y}_i.$$

These residuals are shown in the "Income Residual" column of Table 5.2.

Step 2 produces the regression equation

$$\hat{x}_i = 49.48 - .608z_i.$$

TABLE 5.2 Income, Age, and Schooling, With Residuals

Income	*Schooling*	*Age*	*Age Residual*	*Income Residual*
48,000	12	54	11.8	23,960
26,000	12	28	–14.2	1,960
26,000	7	56	10.8	10,422
48,000	14	47	6.0	20,576
13,000	14	23	–18.0	–14,424
34,000	12	60	17.8	9,960
18,000	11	36	–6.8	–4,347
24,000	16	34	–5.8	–6,809
81,000	16	61	21.2	50,191
21,000	12	38	–4.2	–3,040
9,000	6	53	7.2	–4,886
18,000	12	34	–8.2	–6,040
34,000	13	58	16.4	8,268
21,000	14	38	–3.0	–6,424
81,000	12	46	3.8	56,960
48,000	20	54	16.7	10,422
6,000	7	76	30.8	–9,578
21,000	14	35	–6.0	–6,424
21,000	12	34	–8.2	–3,040
9,000	14	23	–18.0	–18,424
34,000	14	44	3.0	6,576
7,000	9	31	–13.0	–11,963
24,000	8	56	11.4	6,729
34,000	16	37	–2.8	3,191
34,000	17	40	.8	1,499
4,000	12	20	–22.2	–20,040
5,000	9	65	21.0	–13,963
13,000	14	53	12.0	–14,424
7,000	20	33	–4.3	–30,578
13,000	12	31	–11.2	–11,040
34,000	7	30	–15.2	18,422
10,000	16	36	–3.8	–20,809
48,000	18	54	15.5	13,807
6,000	12	19	–23.2	–18,040
2,000	10	25	–18.4	–18,655

The predicted values from this equation are then used to calculate the residuals

$$x_i^* = x_i - \hat{x}_i.$$

These are shown in the "Age Residual" column of Table 5.2.

Why do we want the residuals? Our aim is to estimate the effect of x on y, controlling for z. By calculating residuals, x^* and y^*, we get new versions of x and y that remove or "purge" any relationship with z. Specifically, because independent variables are always uncorrelated with the residuals, the correlation of the age residual with schooling is necessarily 0, and the correlation of the income residual with schooling is also 0. Now we can examine the relationship between x^* and y^* without having to be concerned about any mutual dependence on z.

Step 3 produces the equation

$$\hat{y}_i^* = 600x_i^*$$

which is exactly the slope we got in Chapter 1 using a regular regression program. Notice that there's no intercept in this equation, which is characteristic of this multistep process.

The standard error given by the hand calculator is 207, which is a little smaller than the 210 produced by a multiple regression program. The reason is this. The calculator uses the formula for the standard error that I gave in Section 5.4 but that is appropriate only for a *bivariate* regression. For a trivariate regression, the $n-2$ in the denominator must be changed to $n-3$. The hand calculator doesn't know to do this, but we can correct it by multiplying its standard error by $\sqrt{(n-2)/(n-3)}$. When we multiply 207 by $\sqrt{33/32}$ we get exactly the right standard error of 210.

Now we have the slope and standard error for x, controlling for z. If we also want the slope and standard error for z controlling for x, we have to repeat the three steps with z and x interchanged:

1. Regress y on x. Calculate the *residuals* from this regression and call them y^{**}.
2. Regress z on x. Calculate the *residuals* from this regression and call them z^*.
3. Regress y^{**} on z^*.

Again, this gives us exactly the same slope coefficient for z that we would get in a multiple regression program, but it still doesn't give us the intercept in the full trivariate equation. To get that, we must calculate

$$\bar{a} = y - b_1\bar{x} - b_2\bar{z}.$$

As you can see, the process gets a bit tedious, which is why I can't recommend it for routine use. The important thing to remember is that each slope in the trivariate regression can be interpreted as a bivariate slope where the two variables have been purged of their relationship with the third variable. For more details, see McClendon (1994).

One final point is that we can also calculate the correlation between x^* and y^*, the residuals from the bivariate regressions on z. This correlation is called the *partial correlation* and it is often written as $r_{xy.z}$. For our example, the partial correlation between age and income controlling for schooling is .45. The partial correlation between schooling and income controlling for age is .39.

5.6. Why Are There Two Bivariate Regression Lines?

In Section 5.2, we calculated the least squares regression line as

$$y = 3{,}323 + 524x$$

where y is income and x is age. Suppose we turn the problem around and take age as the dependent variable and income as the independent variable. Although that may not make sense from a causal point of view, there's nothing in the mathematics to prevent it. If we then calculate the least squares regression line for predicting x from y, we get

$$x = 34.84 + .000275y.$$

Do these two equations represent the same line? To find out, we can solve the new equation for y. A little algebra produces

$$y = -126{,}691 + 3{,}636x.$$

Clearly the lines are not the same. The second line has a slope that is nearly seven times the size of the slope in the first equation. This is also evident if we graph the two regression lines, as in Figure 5.6. The line with the steeper slope corresponds to the least squares regression with age as the dependent variable and income as the independent variable.

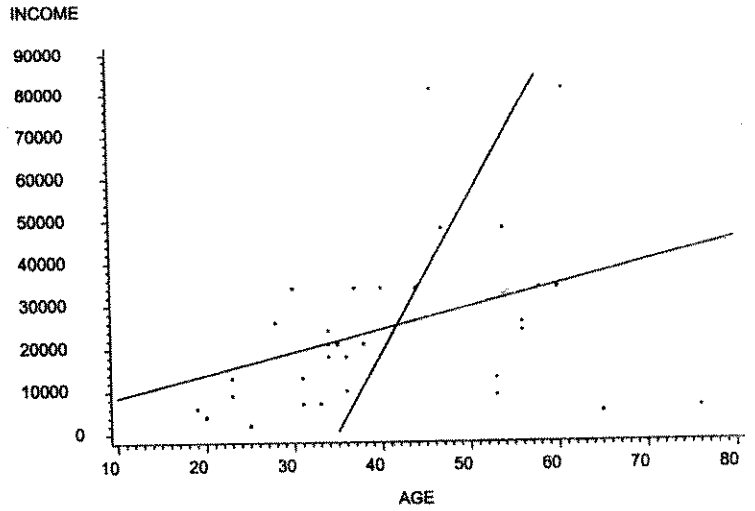

Figure 5.6. Regressions of Income on Age and Age on Income

There's nothing unusual about this. For any two variables x and y, the least squares regression of y on x will produce a different line from the least squares regression of x on y. There's one exception: If the correlation between x and y is 1 or –1, the two regression lines will coincide. The opposite extreme is when the correlation between x and y is 0. Then one regression line will be exactly vertical (at the mean of x), and the other will be exactly horizontal (at the mean of y), forming a cross. Regardless of the correlation, the point where the two lines intersect is always $(\bar{x}, \bar{y})$.

Why are there two different regression lines? Remember that least squares regression is always trying to minimize the sum of squared prediction errors in the dependent variable. The basic fact is that prediction errors in one variable aren't the same as prediction errors in another variable. In Figure 5.6, prediction errors in income would be represented by vertical lines from the data points to the regression line. Prediction errors in age would be represented by *horizontal* lines from the data points to the regression line. Minimizing one kind of prediction error does not produce the same result as minimizing prediction errors in a different variable.

Unfortunately, some people mistakenly believe that the two regression lines can represent reciprocal causation, with x causing y and y causing x. One regression line is supposed to show the effect of x on y, whereas the other line shows the effect of y on x. Although reciprocal causation is often a realistic possibility, the two regression lines do not help us distinguish the two causal directions. One indication of this is that the two slope coefficients must have the same sign. This is easily seen from their formulas. Letting b_{yx} be the slope from regressing y on x, and b_{xy} be the slope from regressing x on y, we have

$$b_{yx} = \frac{s_{xy}}{s_x^2}, \quad b_{xy} = \frac{s_{xy}}{s_y^2}$$

The denominators in both formulas must be greater than 0, so the signs will depend on the numerators, which are the same in both formulas. Why is this important? Well, when we think about reciprocal causation, we usually want to allow for the possibility that the signs may be different. For example, deviant behavior is likely to increase punishment, but punishment may *decrease* deviant behavior. That's not possible if the effects are measured by the two least squares regression lines. If you really want to study reciprocal causation, you must use much more complex methods that also require more elaborate data.

Chapter Highlights

1. In a bivariate regression, the least squares regression line always passes through the intersection of the means of the two variables.
2. The slope of the least squares regression line of y on x is given by s_{xy}/s_x^2. In this formula, the numerator is the covariance of x and y, and the denominator is the variance of x.
3. The intercept of the least squares regression line of y on x is given by $\bar{y} - b\bar{x}$, where b is the least squares slope.
4. The correlation coefficient is calculated as $s_{xy}/(s_x s_y)$, where the numerator is the covariance of x and y and the denominator is the product of their standard deviations.
5. The correlation measures the degree of scatter around a regression line. If it's 1 or –1, all the data points lie on a straight line.

6. The sign of the correlation coefficient is always the same as the sign of the regression slope; therefore, if one is 0, the other is also 0. Testing the hypothesis that the correlation is 0 is always equivalent to testing the hypothesis that the slope is 0.
7. In both bivariate and multiple regression, the R^2 is equal to the squared correlation between the observed values of the dependent variable and the predicted values of the dependent variable.
8. Both the regression slope and the correlation coefficient measure the *linear* relationship between two variables. If the relationship is nonlinear, y could be a perfect function of x, yet the correlation could be small or even 0.
9. The standard error of the regression slope is used to calculate confidence intervals and test statistics. A large standard error means an unreliable estimate of the coefficient.
10. The standard error goes up with the variance in y. The standard error goes down with the sample size, the variance in x, and the R^2.
11. The ratio of the slope coefficient to its standard error is a t statistic, which can be used to test the null hypothesis that the coefficient is 0. With samples larger than 100, a t statistic greater than 2 (or less than –2) means that the coefficient is statistically significant (.05 level, two-tailed test).
12. With samples larger than 100, we can get a confidence interval around a regression coefficient by adding and subtracting twice its standard error.
13. For a trivariate regression of y on x and z, the slope for x can be thought of as arising from a bivariate regression of y^* on x^*, where y^* represents residuals from regressing y on z and x^* represents the residuals from regressing x on z. This helps us understand what it means to "control" for z.
14. The regression of y on x produces a different regression line from the regression of x on y. The two lines cross at the intersection of the two means. The two slopes always have the same sign. The two lines coincide when x and y are perfectly correlated.

Questions to Think About

1. Suppose you have a sample in which the mean income is $32,000 and the mean age is 44. You estimate a bivariate regression of

income on age. Based on the estimated regression model, what's the predicted income for a 44-year-old?

2. For the sample described in question 1, the standard deviation of income is 8,000, the standard deviation of age is 10, and the correlation between age and income is .30. What is the least squares regression line for income on age? (Hint: The correlation can be transformed into the regression slope by appropriate multiplication and division by the standard deviations.)
3. Table 5.3, taken from *USA Today* (December 28, 1984), reports findings by N. Glenn and B. A. Shelton. Regress the divorce rate on the mobility rate. Do it first by hand and then on a computer or calculator. What is the least squares regression line? What is the R^2? Is the coefficient for mobility significantly different from 0? Do you think it's appropriate to interpret the coefficient for mobility as a causal effect on divorce? Why or why not? What important variables might have been left out of this regression? Is reverse causation possible?

TABLE 5.3

Region	*Mobility Rate*	*Divorce Rate*
New England	41	4.0
Middle Atlantic	37	3.4
East North Central	44	5.1
West North Central	46	4.6
South Atlantic	47	5.6
East South Central	44	6.0
West South Central	50	6.5
Mountain	57	7.6
Pacific	56	5.9

4. A researcher is studying the relationship between IQ and GPA for a sample of college students. She finds a correlation of .20 among men and .30 among women. Does this mean that the regression lines are different for men and women?
5. Dr. Jones believes that college students who have a lot of friends are more likely to be satisfied with their college experience. Based on a telephone survey at his college, he finds that the correlation between self-reported number of friends and college satisfaction is .35, which is statistically significant at the .01 level. Dr. Smith says that Jones should also do a linear regression to test his hypothesis. Is Smith right?
6. Using data from the past 2 years, a weather forecaster estimates a multiple regression model in which the dependent variable is the amount of rainfall on a given day and the independent variables include many climatological measurements taken on the previous day. His regression model has an R^2 of .64. Another forecaster claims that he can make "better" predictions by applying a completely different technique known as "neural networks" to the same data. How can the predictions be compared? Can you think of a way to calculate an R^2 for the predictions based on neural networks?
7. For a sample of 1,000 dry cleaning establishments, a marketing researcher finds that the correlation between advertising expenditures and net profit is only .02, which is not statistically significant. Is he justified in concluding that advertising is useless?
8. Suppose an educational researcher wants to estimate a regression model for the relationship between age and hours per day spent on homework. He has a choice between doing the study in a middle school or in a school with kindergarten through 12th grade. On purely statistical grounds, why is he better off with the comprehensive school?
9. Based on a sample of size 250, a regression coefficient of −.753 has a standard error of .310. Is this statistically significant? Calculate a 95% confidence interval for this coefficient.
10. Section 5.5 explained how a trivariate regression could be calculated by running multiple bivariate regressions. Suppose that for the regression of x on z, the R^2 is 1.0. What problem does this create?