2

Fundamentals of Multiple Regression

In this chapter, we present some basic ideas about *multiple*, or *multivariate*, regression analysis, including an introduction to multiple regression focusing on the difference between bivariate (simple) and multivariate regression, and interpretation of multiple regression results. We discuss predicting Y via a multiple regression equation and also the problem of *collinearity*. In addition, there are several examples of multiple regression analysis, as well as homework exercises. The chapter's Appendix A also provides guidance on how to start a research project involving multiple regression analysis, how to evaluate research hypotheses, and how to organize a quantitative research paper using multiple regression.

2.1 Introduction to Multiple Regression

Bivariate, or simple, regression examines the effect of an independent variable (X) on the dependent variable (Y). Multiple regression extends this idea by considering the effects of multiple independent variables (X's) on the dependent variable (Y). It is almost always more realistic for there to be multiple influences on a dependent variable than to suppose that truly only a single factor influences Y. For example, criminal behavior might be influenced by many factors such as economic hardship, lack of informal social control, and likelihood of getting caught and punished. Similarly, a person's income could be influenced by multiple factors such as age, gender, race/ethnicity, education, and work experience. In some analyses we cannot include all factors that might plausibly influence Y, as there are technical reasons for preferring that the number of X's

be relatively small compared to the number of observations (N) in our data, or because we simply do not have all those measures available in our sample. Usually, though, this restriction is not very confining, and we are able to consider quite a few independent variables; we discuss the choice of variables later in this chapter.

Note that in both bivariate and multivariate regression, we decide whether an X influences Y, rather than the other way around. In making this decision we draw on theory, past research, or our common sense. We use our existing knowledge to decide whether to proceed as if "variable one" influences "variable two" or as if "variable two" influences "variable one." We label the variables X and Y based on this decision: in the first case, variable one is X and variable two is Y, while in the second case variable two is X and variable one is Y. Sometimes this can be quite difficult to decide, but in any case the direction of this influence will be assumed, not actually tested, in the methods we examine here.

Another point worth mentioning again is that we will often use the language of "effects" or "influence" of independent variables on the dependent variable even when, as noted in Chapter 1, the nature of the research design does not allow us to genuinely identify causal relations among the variables. Unless we are analyzing data from a true experiment, we are typically uncovering associations among variables rather than actual causal effects. Some formal methods attempt to draw *causal inferences* from nonexperimental data, but, except for a brief overview in Chapter 9 (Section 9.7, Causal Inference), those approaches are beyond the scope of this book. It is most convenient to simply speak of the effect of an X on Y, or X influencing Y, when discussing multiple regression results, but we should keep in mind that this is shorthand language, not necessarily an indication of a true causal relationship.

A single multivariate regression analysis includes multiple X's that might influence Y, and multiple regression aims to separate, or single out, the effect of each X on Y. Thus, the b (slope or *coefficient*) for a particular X in a multiple regression is interpreted as the effect of X on Y, expressed as how many units the prediction Y increases or decreases for each additional unit of X, while *holding other X's constant* or, in slightly different language, while *controlling for other X's*. The idea of "holding other X's constant" is the key conceptual element that distinguishes this interpretation of b from the interpretation made in bivariate regression. This is also why we cannot achieve the same results as a multiple regression by repeatedly applying bivariate regression (once for each independent variable). A series of bivariate regressions will not incorporate this idea of control/holding constant, so we need to consider all of the independent variables together in the same analysis.

In a bivariate regression, the apparent effect of X on Y may actually also incorporate effects of other X's that are related to it but are not included in the regression; although such a situation is also possible in multiple regression, it is especially likely in a bivariate regression that, by definition, includes only a single X. Suppose you are interested in finding the effect of age on income. It is easy to imagine that the variable "years in the labor force" also influences a person's income. Those two possible influences on income (age and years in labor force) are closely related conceptually, and will be correlated in any realistic data set: older people (those with higher ages) tend to also have more years in

the labor force. Thus, the effect of age on income that is obtained in a bivariate regression probably also reflects, to some extent at least, the influence of years in the labor force.

A multiple regression can include both age (X_1) and years in the labor force (X_2) in one analysis, and attempt to statistically separate the effects of age (X_1) and years in the labor force (X_2) on income (Y). Thus, the b for age (X_1) from such a multiple regression analysis is interpreted as the effect of age (X_1) on predicted income ($\hat{Y}$), holding years in the labor force (X_2) constant (or, controlling for years in the labor force). In other words, we are trying to imagine what would happen if everyone in our data set had the same years in the labor force (X_2)—which of course will not be true in our actual sample—but varied in age. (If somehow everyone in the sample did have exactly the same years in the labor force, then we would not be able to include that as an X in our multiple regression. There must be at least some variation in an X for it to be useful in predicting Y.) In that imaginary scenario, how would differences in age (X_1) be reflected in differences in income (Y)? Similarly, the b for years in the labor force (X_2) from this multivariate regression analysis is interpreted as the effect of an additional year in the labor force (X_2) on predicted income ($\hat{Y}$), holding age (X_1) constant. (Note that we will use "effect of X on Y" and "effect of X on $\hat{Y}$" interchangeably; the first may sound more natural, while the second is arguably more precise.) That is, if we imagine that everyone in the sample had the same age (X_1), then b_2 indicates how differences in years in the labor force would be reflected in differences in predicted income (Y). This statistical separation of the independent variables becomes more difficult the more closely they are correlated; we will return to this concern later.

A multiple regression equation with three X's can be written in symbols as

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3, \text{ where}$$

- $\hat{Y}$ represents the predicted value of Y;
- a represents the Y intercept;
- b_1 represents the effect (slope) of X_1 on Y, holding the other X's (X_2 and X_3) constant;
- b_2 represents the effect (slope) of X_2 on Y, holding the other X's (X_1 and X_3) constant; and
- b_3 represents the effect (slope) of X_3 on Y, holding the other X's (X_1 and X_2) constant.

We often refer to the regression equation as a regression *model*, because it embodies some social scientific hypotheses about which factors affect Y, and we recognize that as a model we do not necessarily expect it to capture all of the particular nuances of the data in our sample. In this book we will not focus on the technical details of actually calculating a and the b's from a particular data set; those details are covered in more advanced texts, and we will be relying on statistical software to do the calculations. It is enough to say that, as with simple regression in Chapter 1, these values are chosen under the least squares

principle. That is, the values of a and the b's reported by our statistical software are chosen so as to make the predicted values $\hat{Y}$ in the sample as close as possible to the actual values Y, in the sense of minimizing the sum of squared errors (or residuals, meaning differences between the actual and predicted values of Y among the N observations in our sample). We will revisit the sum of squared errors in Chapter 3.

The above regression equation with three X's helps show why we interpret b_1 as the effect of X_1 on Y (or the change in $\hat{Y}$ as X_1 increases by one unit) while holding other X's constant. If we compare the predicted Y for a case calculated before and after increasing its value of X_1 by one unit, while holding its values of X_2 and X_3 constant, it is as if we changed the regression equation for that case from

$$\hat{Y} = a + b_1X_1 + b_2X_2 + b_3X_3$$

to

$$\hat{Y} = a + b_1(X_1 + 1) + b_2X_2 + b_3X_3, \text{ or } \hat{Y} = a + b_1X_1 + b_1 + b_2X_2 + b_3X_3.$$

We can see, then, that the value of $\hat{Y}$ changed by b_1 as X_1 increased by one unit and other X's were held constant. This leads to the interpretation of b_1 as the effect of X_1 on Y while holding other X's constant. (Again, we will use the term "effect" for convenience even if our research design does not permit a genuinely causal interpretation.) If b_1 is a positive number, then $\hat{Y}$ increases when X_1 increases; if b_1 is negative, then $\hat{Y}$ decreases when X_1 increases. b_1, b_2, and b_3 are often called *regression coefficients*, because they multiply the values of the X's in the regression equation.

As in simple regression, we do not mean that we literally change our data by adding to the X values; our sample data do not change. Instead, we use this idea of a one-unit increase as a means of interpreting the results. Also as in simple regression, it may help to think of b_1 as the difference in $\hat{Y}$ between two cases that are equivalent except for a one-unit *difference* in their values of X_1. In many contexts this will seem more natural than thinking about a one-unit increase in X_1 for one case.

In a well-controlled lab experiment, the techniques of multiple regression would usually be less necessary. In that setting, it may be possible to focus on how changing a single factor affects Y, while controlling the research environment to such an extent that we are literally holding all the other influences on Y constant. For example, if we want to see how the amount of a specific chemical (X) in a solution influences the size of an explosion (Y), we can run the experiment with varying amounts of the chemical while, for instance, keeping the temperature, size of the container, and other important factors the same every time. Then, to the extent that we have successfully held these other factors constant, observed differences in the results (analogous to our Y) can logically be attributed to the changes that we made in the amount of the chemical.

However, in the social sciences there are a variety of practical and ethical reasons why it will be rare to have this degree of experimental control. Therefore multiple regression attempts to achieve statistically what we might be able to literally do in a science lab: distinguish the separate effects of different independent variables by seeing how Y changes when only a single X changes. The

ability to do this when a lab experiment is impractical, or even logically impossible, and when the independent variables correlated with each other in real data, makes multiple regression extremely important in a wide range of social science fields. (We will also examine difficulties that can arise when X's are too highly correlated with each other; this is the problem of *collinearity*.)

2.2 Interpretation of Multiple Regression Results

We now know what multivariate regression is and how it differs from bivariate regression. In this section, we look at the core interpretations of multiple regression analysis results. We will need to pay attention to (a) *R-squared (R^2)*, (b) statistical significance (obtained from p-values), and (c) slopes, regression coefficients, or effects (b's). Note again that scientific conclusions drawn from the analysis of just one single data set are inherently somewhat tentative. We should keep that in mind as we focus on the technical interpretations of the results.

We should also be aware that researchers use a variety of terms to refer to the act of carrying out a multiple regression analysis. "Run a regression" is common but informal, while "estimate a regression model" highlights the fact that our analysis of the sample data is meant to estimate what we would find if we could actually analyze data on the entire population from which we drew the sample. (We discuss below how this sort of thinking can also apply to situations in which we have data on the whole population.) In any case, these different terms do not imply differences in the actual analysis being done.

2.2.1 R-squared—Overall Performance of Multiple Regression

One of the main purposes in carrying out a regression analysis is to predict values of Y. Therefore, it is important to know how well the regression is actually doing at predicting Y in our sample. How close are the predicted values of Y to the actual values for the sample cases? If the predicted values of Y closely match the actual values, the regression is performing well in one important respect, and the regression model has "good fit" to our data. If the predicted values of Y do not closely match the actual values, the regression is not performing well in this respect, and the regression model has "poor fit" to our data.

In bivariate regression, we can look at the scatterplot of X against Y and, more formally, the correlation r_{xy} between X and Y to indicate the fit of the regression. If points in the scatterplot are generally close to the bivariate regression line, then the bivariate regression model has a good fit to the sample data. Remember that the correlation between X and Y will be high (in absolute value) in this situation, so r_{xy} indicates the bivariate regression's fit. As r_{xy} gets closer to the extremes of 1 or –1, the fit of the regression gets better; as r_{xy} gets closer to 0, the fit of the regression gets worse. It also can be shown that the absolute value of r_{xy} is exactly equal to the correlation between Y and $\hat{Y}$ (from the bivariate regression), which is another justification for interpreting r_{xy} as a measure of how closely predicted and actual values of Y match in the sample. For example, if $r_{xy} = 0.90$, the bivariate regression has a very good fit: in that

sample, predicted values of Y from the regression will be in general quite close to actual values of Y.

For multivariate regression, the situation is a little more complicated. Because we have to consider multiple X's together when determining the fit of the multiple regression model, the correlation between any single X and Y is not by itself adequate for assessing the fit. But we can still use the idea of the correlation between actual and predicted values of Y in the multiple regression context, with the only difference from the bivariate case being that the predicted values are based on the multiple regression. The square of this correlation between Y and $\hat{Y}$ is called "R-squared," written as R^2. It is the main indicator of the fit of a multiple regression model and is included in the regression output from any statistical software. R^2 will range between 0 and 1: it is the square of a correlation, so a squared value cannot be negative, and with a maximum possible correlation of 1 (in absolute value), the square cannot exceed 1 either. If R^2 is high (close to 1), then the multiple regression is predicting Y well, and the regression model has a good fit. If R^2 is low (close to 0), then the multiple regression is not predicting Y well and has a poor fit.

Poor fit likely means that the multiple regression model is missing important X's that are also related to Y and would help with predicting it. Therefore, a low R^2 leads us to think about the possibility of including additional (or different) X's in the regression model. Poor fit could also indicate that a straight-line regression does not effectively capture the relationship between the X's and Y, but we will wait until later chapters to consider that sort of situation.

It is difficult to set hard cutoffs for what is a high or low value of R^2, because the value of R^2 is affected by many aspects of the data we are analyzing. For instance, in general it will be easier to achieve a high R^2 in a small data set (that is, with a small N) than in a large data set. The same R^2 value may therefore give a different impression depending on the sample size. Still, when R^2 is not fairly close to 1, even in a large data set, we want to think about what important X's may be left out. Note that R^2 will always go up, at least a little, when we add any X's to the regression. But if the added X's do not really help much in predicting Y, the improvement in R^2 will be small.

R^2 is sometimes described as the proportion of variance (or variation) in Y that is explained, or accounted for, by the X's in the multiple regression. If, for example, the value of R^2 is 0.60, then we can say that 60% of the variance in Y is explained or accounted for by the set of X's that we included in the multiple regression. We can think of the variance of Y as referring to the overall pattern of cases in our sample with high or low values of Y, and the regression model for Y estimates how the cases' X values lead to these varying Y values. That means that the predicted values of Y reflect the X values through the regression equation. If the predictions generally match the actual Y values poorly, leading to a low R^2, then there is a good deal of case-to-case variability in Y that is not represented in, or accounted for by, the regression equation. On the other hand, generally close matches between predicted and actual values of Y—giving a high R^2—indicate that the regression equation does seem to represent the key sources of case-to-case variability in Y.

2.2.2 p-Values: Statistical Significance of Each X's Effect on Y

R^2 indicates the multiple regression's overall fit or performance in predicting Y in the sample data. The assessment of *statistical significance*, from p-values, is the first step in investigating the importance of each X's effect on Y (again, controlling for the other X's that also appear in the regression model and not necessarily meaning a genuinely causal effect; we also could refer here to $\hat{Y}$ instead of Y). In this section, we discuss how the general idea of statistical hypothesis testing applies to statistical significance in regression. As mentioned in the Preface, some aspects of the role of statistical significance in the interpretation of regression results, and social science more broadly, are becoming more controversial. However, here we simply present it as typically used by social scientists now and in the published literature from recent decades.

In adapting statistical hypothesis testing to the situation of multiple regression, we are interested in the question of what the effect of an X on Y would be in the entire population, rather than what effect we find in the particular sample we have. Of course, sometimes we actually do have data on an entire population, for instance when we are doing an aggregate-level analysis using data from all 50 states. In such a situation, we can still view the data as being like a sample from a theoretical "superpopulation" that reflects the many ways that history could have played out differently for these 50 states. Although this image can sometimes seem a bit farfetched, the practical result is that we will treat such data as if it were a sample from a large population.

If we had the entire population or superpopulation, we could find the *"true" effect* of an X on Y, holding other X's constant, and write it as β to distinguish it from b, the effect of this X on Y in our sample. That is, b is the estimate we obtain from our sample, while β is the value we would obtain if we could analyze data from the entire population. The true regression model would therefore be expressed with β's, and the estimated model would be expressed with b's. As usual in statistical hypothesis testing, we can frame our thinking in terms of two competing hypotheses about the population: the null hypothesis H_0 and the alternative hypothesis H_1.

The most common null hypothesis in regression analysis, and the one that is assumed in the output from standard software packages, is "H_0: X has no effect on Y in the *true model*," or, stated in symbols, "H_0: $\beta = 0$." It is possible to consider other null hypotheses that specify some other value of β, and those can be used in the general hypothesis test format that was discussed in Chapter 1, but other null hypotheses are not so common in social science research. In practice, most applications use a two-sided alternative hypothesis, and this will generally be the default for statistical software's calculation of p-values. The alternative "H_1: X has some effect on Y in the true model," or equivalently "H_1: $\beta \neq 0$," will then be paired with the usual null, so that the typical hypothesis test is set up as

H_0: (null hypothesis): X has no true effect on Y, or $\beta = 0$;

H_1: (alternative hypothesis): X has some true effect on Y, or $\beta \neq 0$.

After setting up H_0 and H_1 this way, we need to calculate a test statistic—in this situation, it will be a *t-statistic*—from the sample information and use it

to determine a p-value and decide whether we should reject or not reject H_0. Rejecting H_0 would mean that H_1, the hypothesis that X does have some effect on Y, appears more reasonable from our sample data, so in that case the sample leads us to believe that there really is some effect of X on Y in the entire population. If we do not reject H_0, then H_0—the hypothesis that X has no effect on Y—still appears plausible in light of the sample data. That is, if we do not reject H_0 we are saying that the sample information does not allow us to confidently conclude that there actually is an effect of X on Y in the population. Of course we must always remember that, in any particular analysis, it is possible that we reached an incorrect conclusion of rejecting H_0 or not, which is one reason why we are reluctant to rely too heavily on a single study of some research question.

The t-statistic follows the general format shown in Chapter 1 and can be calculated as follows:

$$t = \frac{b - \beta_0}{\text{standard error of b}}$$

The standard error of b is an estimate of sampling variability in b; that is, it gives us a sense of how much b might vary across samples that we could have drawn from the population. Because we are most commonly testing the null hypothesis that $\beta = 0$, the above equation usually can be rewritten:

$$t = \frac{b - 0}{\text{standard error of b}} = \frac{b}{\text{standard error of b}}$$

We can calculate this t-statistic for each of the X's in our regression and then find p-values for the t-statistics from a t-table or an online t-calculator. Either will require us to know the proper degrees of freedom (df), and for this test df are calculated from the sample size as df = N – (the number of X's in the regression) – 1. With a p-value (here two-sided) in hand, we can decide whether or not to reject H_0, using the traditional (though again ultimately rather arbitrary) p-value cutoff of 0.05. With this cutoff, we do not reject H_0 if the p-value is > 0.05. In that case we still believe that it is plausible that, in the entire population, X has no effect on Y, and we will say that "X does not have a statistically significant effect on Y." On the other hand, if the p-value is less than (or exactly equal to, which could occur due to rounding) 0.05, we reject H_0 in favor of H_1, believing that we would find an effect of X on Y if we were to analyze the whole population. In this case ($p \leq 0.05$), we say that "X has a statistically significant effect on Y." The language of statistical significance is therefore a summary of the results of this particular hypothesis test. When we are working with a multiple regression, it is good to also include "controlling for other X's" or "holding other X's constant" in our statement of the results.

Fortunately, much of the actual arithmetic here is done for us when we use statistical software, because the software will report each b's standard error and t-statistic, calculate the df, and find the p-value that corresponds to the value of the t-statistic. Our only work is to then use the p-values to decide whether each X has a statistically significant effect on Y. If the effect of an X on Y is

statistically significant, then we will go on to interpret the slope b for that X. If the effect is not statistically significant, then we generally do not want to interpret b further, as we are not sure that there is any true effect of that X at all. Again, though, 0.05 is a rather arbitrary cutoff, so it is hard to justify why we would interpret b when the p-value is just under 0.05, but not when the p-value is just over 0.05. When the p-value is above but close to the cutoff, we may want to consider this an instance of "borderline," or "marginal," significance. We probably want to go ahead with interpreting b for an X that shows a borderline significant effect, but we need to make clear that in doing so we are not strictly applying the usual cutoff and that our conclusion is therefore even more tentative than usual.

2.2.3 Interpreting b in the Multiple Regression Context

When the p-value indicates that an X has a statistically significant effect on Y, we next want to interpret its slope b. b could be positive or negative, with a positive b indicating a positive relationship between that X and Y, controlling for other X's; as in the bivariate case, this means that X and Y tend to move in the same direction—as X increases, Y tends to also increase, and as X decreases, Y tends to also decrease—but now we also think of the other X's as being held constant when making this interpretation. Likewise, a negative b indicates a negative relationship between X and Y, controlling for other X's; as X increases (holding other X's constant), Y tends to decrease, and as X decreases, Y tends to increase. Also as in the bivariate case, this logic applies to comparisons of Y values for cases that differ in their X values, and this will often be a more helpful way to understand a positive or negative relationship between X and Y.

As discussed in Section 2.1, the specific numerical value of b indicates the effect of that X on $\hat{Y}$ (showing the increase in $\hat{Y}$ associated with an additional unit of X), controlling for, or holding constant, the other X's in the regression. Suppose that b = 3. Then we interpret the result as "$\hat{Y}$ increases by three units as X increases by one unit, controlling for other X's." If b = −4, we could try to use similar language: "$\hat{Y}$ increases by negative four units as X increases one unit, controlling for other X's." But this "negative increase" sounds very awkward. An increase of −4 is equivalent to a decrease of 4, so it is much more natural to say "$\hat{Y}$ decreases by four units as X increases by one unit, controlling for other X's." We can also think of b as reflecting the difference in $\hat{Y}$ between two cases that are identical in their values of other variables but differ by one unit on this X.

Let us consider a social science example. Suppose that we wanted to examine the influence of a person's education and criminal history on their income, with education measured by years of schooling (X_1), criminal history measured by lifetime number of arrests (X_2), and income measured as annual income in dollars (Y). We then collected information on these variables from a sample of adults and analyzed the resulting data.

Suppose that the multiple regression output from our statistical software indicated that the effects of both education (X_1) and arrests (X_2) on income (Y) are statistically significant ($p \leq 0.05$), with b_1 (the slope for X_1) = 2,580, and b_2 (the slope for X_2) = −5,890. Then our interpretation of b_1 is "predicted annual income ($\hat{Y}$) increases by $2,580 as years of education (X_1) increase by 1 year,

controlling for the number of arrests (X_2)." b_2 is negative, so we could say "For each additional arrest (X_2), predicted annual income ($\hat{Y}$) decreases by $5,890, holding years of education (X_1) constant."

Although it is accurate to say "predicted" annual income as we have here, and helpful to focus attention on the fact that these figures are estimates from a model for the predicted Y ($\hat{Y}$), in practice we may assume that our audience understands that the b's refer to the predicted Y ($\hat{Y}$). If so, for convenience we could drop the "predicted" from this language, or use Y instead of $\hat{Y}$. Also, the data here came from a snapshot of the sample members at one point in time, rather than tracking individuals through time. Therefore this is an instance in which it may be more natural to view statements like "as education increases by 1 year, holding number of arrests constant" as referring to a comparison between people who differ by one year of education while having the same number of arrests.

2.2.4 Real-Life (Substantive) Significance

Real-life significance refers to an evaluation of whether the size of an X's effect on Y is large enough to be meaningful. That is, when considered in light of the definitions and observed values of the X and Y variables in our sample, does the effect seem important, or rather trivial? We always need to consider this question of real-life significance, because sometimes an X's effect is statistically significant (indicated by $p < 0.05$), but the size of the b is not large enough to suggest any meaningful impact on Y in real terms. Suppose that X_1 is years of education and Y is annual income, the p-value indicates that years of education has a statistically significant effect on Y, and b_1 is 25. This would mean that predicted annual income increases only $25 for each additional year of education, holding other X's constant. Even 4 additional years of education would only increase predicted annual income by $100: each additional year of education increases predicted income by $25, so 4 additional years increase the predicted income by $4 \times 25 = \$100$. If the data are contemporary rather than historical, and so come from a context in which full-time workers have annual incomes in the tens of thousands of dollars, this result would suggest that, in practical terms, there is no meaningful impact of education on income. This tiny difference in predicted income between people differing by 4 years of education would mean that income is effectively unrelated to education. In that case, the effect of education on income would have statistical significance, but not real-life significance. Alternative terms for "*real-life*" here include *real-world,* practical, and *substantive*.

There is no automatic numerical cutoff for determining real-life significance. We might simply use our knowledge of social science, or our common sense, in deciding whether the effect of X on Y is large in real terms. However, it is better to examine descriptive statistics such as the mean, standard deviation, and range (minimum and maximum values) of the variables to determine whether the effect is large enough to have real-life significance. This requires examination of both X and Y. For X, it may be that the usual "one-unit change" is not the most suitable for this real-life assessment. For example, if X is public school spending per student, measured in dollars, a

one-unit change in X refers to a single dollar. If the average public school spending per student in our sample were $11,762, with a standard deviation of $5,891, a minimum value of $4,152, and a maximum value of $32,366, a change of $1 would appear extremely small. One additional dollar of spending per student will almost surely have virtually no effect on whatever Y we are studying, making it hard to assess real-life significance. In this case a larger change in X would be more interpretable, perhaps an increase of $500 or even $1,000 per student. Then the change in $\hat{Y}$ would be 500 × b, or 1,000 × b, and we will be in a better position to assess whether the resulting change in $\hat{Y}$ is large in real terms.

There could also be cases in which the usual one-unit change in X is too large. For example, the Gini index of income inequality in a society is usually presented as a number between 0 and 1. A "one-unit change" would be too large to sensibly consider for that variable, as it would be equivalent to a change from the theoretical minimum value to the theoretical maximum value of this index. Again the resulting change in $\hat{Y}$ would be calculated as b multiplied by the change in the Gini index, and we might consider something like a change of 0.10 instead of 1.

When we are convinced that we are examining an appropriately sized change in X, we next assess the resulting change in $\hat{Y}$, based on Y's descriptive statistics. To illustrate this process, suppose that, for aggregate-level data, X_1 is unemployment rate (percentage unemployed) and Y is the suicide rate per 100,000 population. Suppose too that, from our software's regression output, the p-value indicates that the unemployment rate has a statistically significant effect on suicide rate, and b_1 is 4. If our descriptive statistics for percentage unemployment (X_1) showed a mean of 5.5, a standard deviation of 0.9, and a minimum to maximum range of 3.8 to 6.9 in our sample, we would probably decide that a 1% change in unemployment is indeed appropriate for evaluating real-life significance. The value of b_1 means that the predicted suicide rate per 100,000 increases by 4 as the unemployment percentage increases by one unit (here, 1%). If suicide rates in the sample have mean 13.4 and range between 10 (minimum) and 80 (maximum), an increase of 4 in the suicide rate seems to be large enough to suggest real-life significance. On the other hand, the same estimated effect ($b_1 = 4$) may not seem to have real-life significance for a different dependent variable that had mean 1,200 and range 1,000 to 1,500.

We can revisit the earlier example of public school spending per student as X. Based on the descriptive statistics for public school spending per student (X) mean $11,762 with a standard deviation of $5,891, minimum of $4,152, and maximum of $32,366), it seems that a $1 increase in X is too small to be easily interpretable. We can instead consider the predicted change in Y when public school spending per student (X) increases by $1,000. In this case, the change in $\hat{Y}$ is b × 1,000. If b is 0.003 and Y is high school graduation rate (%), then we can say that the predicted high school graduation rate increases 0.003 × 1,000 = 3% as public school spending (X) increases by $1,000, holding other independent variables constant. Then we can use the descriptive statistics for graduation rate (Y) to assess whether a 3% increase in graduation rate should be considered large enough to have real-life significance.

We always need to consider this question of real-life significance before deciding that a statistically significant X has a meaningful impact on Y, so we should not be too impressed by the words "statistically significant" alone. We still need to check real-life significance by looking at the size of b and the descriptive statistics for X and Y to reach a judgment as to whether the effect of this X is large enough to have some real impact. Of course different researchers may not always agree on whether the magnitude of a particular X's effect on Y is large enough to be called real-life significant, and there will be instances in which it is quite difficult to make this determination. Even so, attempts to assess the real size of any statistically significant effects are an important element in the full interpretation of regression results.

This is especially true when the sample size is very large. A large sample size (N) will tend to make the standard errors of b small, which will tend to make the t-statistics large and in turn tend to produce small p-values. It is, therefore, generally easier to find statistically significant effects of X's in large data sets, even if those X's actually have very small real-life impacts on Y. So it is especially important not to simply take statistical significance at face value when working with very large samples.

2.2.5 Other Notes on Interpretation

As in bivariate regression, we are usually not so interested in interpreting a, the Y-intercept. In the next section we look at calculation of the predicted value of Y, and a is certainly necessary for that. But in social science we rarely make a direct interpretation of a and usually do not discuss its value when interpreting results. We discuss this issue further in Section 2.3. Remember that, as we mentioned in Chapter 1, most software's regression output will provide the value of a in the "b" column, with a term like "intercept" or "constant" distinguishing a from the b's.

Note that along with "unstandardized coefficients," which are the b_1, b_2, and b_3 that we are discussing here, your statistical software may also present "*standardized coefficients*." Throughout this text, we focus on unstandardized coefficients, and b always refers to those. You may encounter standardized coefficients when reading journal articles or other research reports, however, and they can be thought of as giving another approach to the assessment of real-life significance. Therefore Appendix B to this chapter includes a brief discussion (see Section 2.8.1).

Research reports and articles sometimes present *confidence intervals* for regression coefficients. A confidence interval takes the general form of a sample estimate (of some population value) plus or minus a "margin of error" for that estimate. The confidence interval is reported with a numerical level of confidence, typically 95%. The interpretation of the 95% confidence interval is that the method for constructing the interval means that 95% of the time the interval will include the true population value. The analyst is therefore "95% confident" that the true value is somewhere in the interval.

A 95% confidence interval for a true regression coefficient like β_1 is found by b_1 ± margin of error. The margin of error reflects sampling variability in the estimate b_1, so it is based on the standard error of b_1. In particular, the margin of error is

calculated by $t_{.025}$ (SE of b_1). The value of $t_{.025}$ is found in a t-table, using the usual df of t for assessing statistical significance of a regression coefficient (N – the number of X's – 1), and is the value that leaves a tail of the t-curve with 2.5% of the total probability. For instance, if df = 30, then $t_{.025} = 2.042$. Together, then, the 95% confidence interval for $\beta_1 = b_1 \pm t_{.025}$ (SE of b_1), or if written to highlight the two endpoints, as the interval ($b_1 - t_{.025}$ SE, $b_1 + t_{.025}$ SE). Note that $t_{.025}$ gets closer to 1.96 as df increase, so a quick approximation for typical sample sizes is that $t_{.025}$ is about 2, and the margin of error is roughly twice the standard error.

This confidence interval is closely related to the hypothesis test. If the β hypothesized by H_0 is within the 95% confidence interval, then the (two-sided) p-value for that H_0 is > 0.05. If that hypothesized value is outside the 95% confidence interval, then the p-value is < 0.05. (A value precisely on the boundary of the 95% confidence interval corresponds to p = 0.05, but that will be unlikely when results are given with several decimal places.) So for the usual regression situation in which H_0: $\beta = 0$, looking for 0 in the 95% confidence interval will give the same conclusion as assessing statistical significance in the usual way. But because the confidence interval does not require the notion of statistical significance or hypothesis testing to make sense, it may be appealing to researchers who are uncomfortable with the usual approach to statistical significance, even if it is closely related in a technical sense.

Confidence intervals are more popular in some fields than others, so we do not report them in the examples here. But the formula above is easy to execute, so confidence intervals can be constructed for any of the regression coefficients shown in the examples or exercises. The main difficulty with the confidence interval is that the commonsense interpretation that many people would make, that there is a "95% probability that β lies in the interval," is not really justified. In particular, the statistical framework that underlies the analyses here views β as fixed, even if it is unknown. That is, the framework does not include a notion of probability of different values of β; randomness enters via sampling variability in b_1, so we really want to say something like there is "a 95% probability that we obtain a value of b such that the resulting confidence interval includes β." This is certainly less intuitive, but the more natural statement will be sensible only if we have adopted a "Bayesian" framework in which there explicitly is a probability distribution for β. We do not explore that framework here.

2.3 Prediction in Multiple Regression

2.3.1 Calculating the Predicted Value of Y From the Values of X's

The process of obtaining predicted values of Y in multivariate regression is a straightforward extension of that in bivariate regression: to predict Y, we simply plug into the regression equation the values of the X's that we are interested in using. For example, suppose we wanted to investigate the effects of age (X_1), years of work experience (X_2), and years of education (X_3) on annual income (Y), and our analysis found a = –27,710, b_1 = –975.85, b_2 = 2,114.29, and b_3 = 5,580.44. Then the regression equation would be:

$$\hat{Y} = -27,710 - 975.85\,X_1 + 2,114.29\,X_2 + 5,580.44\,X_3$$

If we want to know the predicted annual income for a person who is 26 years old ($X_1 = 26$), has 4 years of work experience ($X_2 = 4$), and 16 years of education ($X_3 = 16$), we simply plug these values for X_1, X_2, and X_3 into the above regression and do the arithmetic:

$$\hat{Y} = -27,710 - 975.85(26) + 2,114.29(4) + 5,580.44(16)$$
$$\hat{Y} = -27,710 - 25,372.1 + 8,457.16 + 89,287.04$$
$$\hat{Y} = 44,662.1$$

Based on this regression equation, we can say that our analysis of the sample data indicates a predicted annual income of $44,662.10 for a person who is 26 years old with 4 years of work experience and 16 years of education.

Because this prediction uses estimates from the sample—the b's—rather than the true effects from the population—the β's—we could also think about sampling variability in the predicted annual incomes. That is, we could have a standard error for the prediction itself. Here we do not explore the calculations to estimate that standard error, but those can be found in more advanced texts.

2.3.2 Not Trusting the Results of Prediction

When we have a low R^2, we should not trust the prediction from the regression as much as we would when R^2 is high. Remember that a low R^2 indicates that the predicted values of Y are not in general very close to the actual values of Y for the cases in our sample, so a low R^2 will reduce our confidence in the quality of predictions in general. Also, we are uncomfortable with predicting Y based on a value of X that is well outside the range of X that was seen in the sample used to obtain the regression equation. Suppose that we are interested in the effect of age on some Y and that we obtained the regression equation from a sample in which all persons were 18 to 49 years old. Even if the resulting R^2 were high, we probably should not try to use the equation to predict Y for a 93-year-old person. As 93 is far outside the actual range of age in the data (18–49), we would not be too confident that the prediction for such a person is meaningful.

2.3.3 Why Not Interpret a?

We saw above that we use the value of a (the Y-intercept) in the arithmetic needed to predict Y. Earlier we said that we are rarely interested in directly interpreting a. We mentioned in Chapter 1 that in a bivariate regression, a can be interpreted as the predicted value of Y when X is 0; for multiple regression, a will be the predicted value of Y for the situation in which every X has value 0. If we plug 0 in for every X in the regression equation, the only non-zero part is a, so $\hat{Y} = a$. Social science applications seldom involve a situation in which it would be realistic for every numerical independent variable to equal 0. Then the scenario in which every X equals 0 is usually not very helpful to explore, and interpreting a as the predicted value of Y under this scenario usually will not

lead to any scientific insight. (However, a situation like this can be interesting when all measures of independent variables are categorical; we discuss categorical independent variables in Chapter 3.)

2.4 Collinearity

With these fundamental interpretations of regression results in hand, we can begin to introduce some extensions of this core. *Collinearity* is, broadly speaking, a situation in which high negative or positive correlations among the independent variables inhibit our regression analysis. This does not require that all pairs of X's be highly correlated, so collinearity may be present if just some X or X's are highly correlated with one or more other X's. (There are also more complicated situations beyond high pairwise correlations in which collinearity can exist; advanced texts discuss this in detail.) Collinearity is potentially a concern because it causes increased uncertainty in the regression results.

In practice, one reason for highly correlated X's may be that a researcher is trying to measure one concept by multiple X's. For example, a survey respondent's social class background might be measured by father's and mother's education. In actual data, however, father's and mother's education are likely to be highly (and positively) correlated. We can imagine that people with a highly educated mother are likely to also have a highly educated father, as many couples meet in educational or occupational settings that tend to bring similarly educated people together. Or, at the aggregate level, consider the level of economic deprivation in a geographical unit like a city or census tract. Both the unit's poverty rate (perhaps measured as the percentage of households with incomes below an official poverty line) and its median income would be sensible measures of this concept. Again, though, these two measures are likely to be highly (and negatively) correlated: cities with a high poverty rate are likely to have low median income. Still, it is important to realize that there is not always an obvious theoretical reason for two independent variables to be highly correlated, so we will need to examine correlations among all the X's in our data, rather than just think about which pairs might be anticipated to be highly correlated.

When there is a strong correlation between two X's, it becomes more difficult to envision changing one while holding the other constant. More practically, it will be challenging for the regression analysis to determine the separate effects of two different X's if the cases in the data with a high value of one X almost always have a high (when there is a strong positive correlation) or low (strong negative correlation) value of the other. That is, if survey respondents with highly educated mothers tend to also have highly educated fathers, it will be hard to untangle whether mother's education, father's education, or both are actually influencing the dependent variable. In this situation, there will be more uncertainty in the estimated b's than there would be without such high correlations, and this uncertainty is reflected in larger standard errors. A tendency toward larger standard errors will mean a tendency toward smaller t-statistics (remember that the standard error is in the denominator of the formula for

the t-statistics shown on the regression output: $t = \frac{b}{\text{standard error of } b}$), and in turn larger p-values.

Collinearity, then, will tend to make us less apt to find statistically significant effects of X's on Y. In the extreme, a classic symptom of collinearity is the seeming paradox of a high R^2, suggesting that the set of X's does quite well at predicting Y, yet with no or almost no statistically significant effects of the X's. This odd situation is possible under collinearity. The X's are together doing a good job of explaining the variance in Y, but high correlations among the X's mean that there is great uncertainty as to how, or which, particular independent variables are related to Y.

2.4.1 Diagnosing and Addressing Collinearity

The most basic approach to detecting collinearity begins by checking for any high positive or negative correlations among the X's. Although there are no strict cutoffs, correlations that are greater than about 0.70 in absolute value are especially apt to create problems in our regression. If two variables have an extremely high correlation (> 0.95, say) we surely will not want to include both in our regression. Otherwise, though, we can first run our analysis with all X's that we had originally identified as belonging in the regression, even those that are strongly correlated with each other. Perhaps the output will still show statistical significance for the effects of X's that are involved in the strong correlations, and we can conclude that these correlations are not really causing difficulties for our interpretation of results. But if some or all of the X's that are highly correlated do not show statistically significant effects, it may be that there is a collinearity problem.

It is not easy to "fix" collinearity. The simplest strategy is to identify pairs of highly correlated X's that did not exhibit statistical significance in the initial regression and then retain only one of each such pair when rerunning the regression. The choice of which of a pair to retain is somewhat arbitrary, as usually R^2 and the pattern of statistically significant effects will be quite similar for either choice. This is because two highly correlated X's are, in a sense, providing the same (or greatly overlapping) information in our analysis and will, therefore, lead to similar regression results. But if one of the pair seems most relevant to the theory and previous research that is guiding our analysis then that X can be chosen to remain in the regression.

There are other, more advanced methods for detecting and addressing collinearity; see Appendix B to this chapter for some discussion of these (Section 2.8.2). For beginning researchers, though, the simple strategy above is often quite effective. Of course even after removing X's from highly correlated pairs, some effects may not be statistically significant. If so, then the explanation for the nonsignificance is more likely the genuine absence of a relationship with Y, not collinearity. We clearly do not want to use collinearity as a catch-all "excuse" for nonsignificance. Usually nonsignificance really does mean no relationship between an X and Y, or at least not one that can be detected in the data we are analyzing. But in many instances, statistically significant effects will emerge once we address collinearity this way. When that happens, it is important to

remember that our choice of which X to retain from a highly correlated pair was mostly arbitrary. Probably the other member of the pair would have given similar results had it been chosen instead, and we should keep that in mind when making interpretations. One or both of the pair are affecting Y, but we cannot really be more specific than that. In this situation, we should think of the X that we decided to include as a representative of this highly correlated pair, rather than thinking we have found that it is important for Y while the other X—the one we dropped due to the high correlation—is not.

2.5 Examples

The following examples illustrate the concepts we have discussed in this chapter.

2.5.1 Example 1: Crime in Colorado

A criminologist believed that economic deprivation, residential instability, racial inequality, and young male population have positive effects on the level of crime in society and wanted to test her research hypothesis with a multiple regression analysis. Economic deprivation, residential instability, racial inequality, and young male population were her independent variables, and the dependent variable was crime. Because her research question concerned aggregate-level relationships, she used aggregate-level data, in this case data from 24 cities in Colorado.

Economic deprivation was measured as the percentage of the labor force that was unemployed; a high percentage of unemployed persons in a city indicates a high level of economic deprivation. Residential instability was measured as the percentage of renter-occupied housing units; a high percentage of renter-occupied housing units indicates a high level of residential instability in the city. Racial inequality was measured by the index of white-Black residential segregation, with possible values from 0 to 100; a high value of the racial segregation index in a city indicates a high level of racial inequality. Young male population was measured as the percentage of the population that was male and aged 15 to 24 years. Crime was measured as the city's total violent crime (murder, robbery, rape, and assault) rate per 100,000 population (Y). She obtained these measures for each city from the U.S. Census Bureau and the FBI's *Uniform Crime Reports* (from the Bureau of Justice Statistics).

Her study can be summarized as follows:

Research hypothesis: Economic deprivation, residential instability, racial inequality, and young male population positively influence the level of crime in society.

Units of analysis: 24 cities in Colorado (aggregate-level data).

Measurements of independent and dependent variables:

X_1: Percentage of the labor force that was unemployed (economic deprivation).

X_2: Percentage of renter-occupied housing units (residential instability).

X_3: Racial segregation index (racial inequality).

X_4: Percentage of the population that was male, aged 15 to 24 years (young male population).

Y: Violent crime rate per 100,000 population (crime).

It is helpful to rewrite the research hypothesis using the actual measurements for the independent and dependent variables. We can also separate the different parts of the hypothesis.

Research hypothesis with measurements of independent and dependent variables:

1. The unemployment percentage (X_1) has a positive effect on the violent crime rate (Y).
2. The percentage of renter-occupied housing units (X_2) has a positive effect on the violent crime rate (Y).
3. The racial segregation index (X_3) has a positive effect on the violent crime rate (Y).
4. The percentage of young male population (X_4) has a positive effect on the violent crime rate (Y).

Her data are as follows:

City ID	City Name	Violent Crime (Y)	Unemployment (X_1)	Renter-occupied Housing (X_2)	Racial Segregation (X_3)	Young Male (X_4)
1	Arvada	125.9	7.3	26.7	24.3	6.1
2	Aurora	446.1	7.7	40.1	28.7	6.8
3	Boulder	211.6	7.0	52.3	20.3	16.7
4	Brighton	193.0	6.1	30.4	28.0	7.2
5	Colorado Springs	491.9	7.6	39.9	34.7	7.4
6	Commerce City	252.3	7.0	30.2	17.6	6.1
7	Denver	542.1	7.8	50.0	54.7	6.6
8	Durango	445.6	6.1	52.0	17.7	13.3

City ID	City Name	Violent Crime (Y)	Unemployment (X_1)	Renter-occupied Housing (X_2)	Racial Segregation (X_3)	Young Male (X_4)
9	Englewood	446.2	8.5	50.9	17.3	6.3
10	Evans	171.2	6.6	39.5	12.8	8.5
11	Federal Heights	478.8	9.4	48.0	20.2	7.5
12	Fort Collins	315.8	7.4	44.9	19.0	7.4
13	Fort Morgan	233.9	8.7	39.3	22.7	7.2
14	Golden	116.1	7.6	41.6	36.2	16.8
15	Grand Junction	349.0	5.4	37.6	22.4	7.9
16	Lafayette	153.5	7.2	27.2	13.4	5.7
17	Lakewood	442.7	7.8	41.1	24.4	6.7
18	Littleton	130.7	6.6	33.1	22.5	6.1
19	Longmont	311.5	7.0	36.5	10.2	6.3
20	Loveland	191.5	7.1	34.1	12.4	6.0
21	Northglenn	255.3	10.5	41.5	19.7	7.3
22	Pueblo	854.2	11.1	39.8	22.0	7.5
23	Sterling	487.3	7.7	41.1	63.8	10.7
24	Wheat Ridge	570.6	6.4	45.4	20.8	5.2

Year of data: 2010
Data sources: U.S. Census Bureau and Bureau of Justice Statistics, FBI's *Uniform Crime Reports*

After entering the above data in statistical software, the researcher checked correlations among the X's. The results of this correlation analysis are below:

	Unemployment (X_1)	Renter-occupied Housing (X_2)	Racial Segregation (X_3)	Young Male (X_4)
Unemployment (X_1)	1.000	0.175	0.068	−0.105
Renter-occupied Housing (X_2)	0.175	1.000	0.181	0.465
Racial Segregation (X_3)	0.068	0.181	1.000	0.183
Young Male (X_4)	−0.105	0.465	0.183	1.000

The highest correlation is 0.465, between renter-occupied housing units (X_2) and young male population (X_3). This does not seem so high as to suggest a collinearity problem, and none of the other correlations are at all remarkable.

She then ran the multivariate regression analysis with X_1, X_2, X_3, and X_4 as independent variables. The multiple regression results from these data are below (with "SE" written for "standard error"):

$R^2 = 0.509$

df for t = 19

	b	SE of b	t	p-value
Intercept (constant)	–392.009	218.221	–1.796	0.088
Unemployment (X_1)	37.081	22.891	1.620	0.122
Renter-occupied housing (X_2)	13.945	4.685	2.976	0.008
Racial segregation (X_3)	3.598	2.385	1.508	0.148
Young male population (X_4)	–23.818	10.334	–2.305	0.033

R^2 was only moderately high, with a value of 0.509. Such an R^2 generally indicates a reasonably satisfactory fit of the regression model to the data, but here does not seem so high in light of the small sample size (N = 24). 50.9% of the variance in violent crime rates (Y) in this sample was explained, or accounted for, by this set of X's (X_1 through X_4), and for the most part Y and $\hat{Y}$ were somewhat close for cities in the sample. Still, it appears that there are additional independent variables that should be considered in order to provide a more complete explanation for differences in crime rates among Colorado cities.

Checking for p-values below the usual 0.05 cutoff indicates that the effects of two of the four X's were statistically significant. Renter-occupied housing (X_2) and young male population (X_4) had statistically significant effects on violent crime, controlling for other X's. p-values for unemployment (X_1) and the racial segregation index (X_3) were above the 0.05 cutoff. Thus, when controlling for other X's, the effects of unemployment (X_1) and racial segregation index (X_3) on violent crime (Y) were not statistically significant.

The regression equation based on this output is

$$\hat{Y} = -392.009 + (37.081)\,X_1 + (13.945)\,X_2 + (3.598)\,X_3 + (-23.818)\,X_4$$

or

$$\hat{Y} = -392.009 + 37.081\,X_1 + 13.945\,X_2 + 3.598\,X_3 - 23.818\,X_4.$$

For a and the b's we have retained three decimal places from the output, but it would also be fine if we reported more rounded-off figures, especially if we wanted to avoid giving an undue impression of precision in our results. Here the three decimal places are helpful for readers who run this analysis in their software and want to confirm that they obtained exactly the same results; we do this throughout the book, but normally more rounding is fine when presenting results.

Because the effects of renter-occupied housing (X_2) and young male population (X_4) were statistically significant, the researcher proceeded to interpret the slopes (b's) for renter-occupied housing (b_2) and young male population (b_4). b_2 was 13.945, indicating a positive effect of renter-occupied housing (X_2) on violent crime rate (Y). The predicted violent crime rate (per 100,000 population) increases by 13.945 as renter-occupied housing (X_2) increases by 1% (because this particular X_2 is measured as a percentage, in this instance "one unit" of X_2 means 1%), controlling for the other X's in the model.

To examine whether the increase of 13.945 is large enough to have a meaningful impact in real terms (real-life significance), the researcher should first examine the descriptive statistics for renter-occupied housing units (X_2) and violent crime rates (Y) produced in her statistical software. Renter-occupied housing's (X_2) sample mean is 40.34, with a standard deviation of 7.29 and a range of 26.70 to 52.30. This context suggests that it may be more informative to look at something more than a one-unit increase in renter-occupied housing (X_2). This does not fundamentally change the regression results but rather allows us to make a more reasoned interpretation, because the descriptive statistics for renter-occupied housing (X_2), including the standard deviation and range, suggest that a one-unit increase is rather small. A five-unit increase seems like a more meaningful change in renter-occupied housing (X_2). To interpret with a five-unit increase, she needs to multiply b by five: a five-unit increase in renter-occupied housing (X_2) increases the predicted violent crime rate ($\hat{Y}$) by (5 × 13.945) = 69.725. Next, in light of the mean (342.17) and the range (116.1–854.2) of violent crime rates (Y), an increase of 69.725 in predicted violent crime rate ($\hat{Y}$) for a 5% increase in the renter-occupied housing seems large enough to have a real impact. Therefore, she can say that the effect of renter-occupied housing (X_2) on violent crime rate (Y) has real-life significance. Although this is ultimately a judgment that the researcher is making, and can be challenged by other researchers, it is rooted in a close examination of the sample data.

Suppose that b_2 had been 1.395, rather than the actual figure of 13.945. Then a five-unit increase in renter-occupied housing would have implied an increase of 6.975 (from 5 × 1.395) in the predicted violent crime rate. Many analysts would view such a change as rather small in light of the mean (342.17) and range (116.1–854.2) of violent crime rates in the sample, and the researcher would likely then conclude that renter-occupied housing's effect did not have much real-life significance. More generally, we always should be mindful of the possibility that a statistically significant effect is not large enough to demonstrate real-life significance. As noted earlier, this is especially a concern in analyses of very large samples; because standard errors tend to be smaller (and, in turn, t-statistics larger and p-values smaller) in larger samples, some effects that are quite small, or even trivial, in real terms can be statistically significant.

b_4 was –23.818, indicating a negative effect of the young male population (X_4) on violent crime rate (Y). The predicted violent crime rate (per 100,000 population) decreases by 23.818 as young male population (X_4) increases by 1%, controlling for the other X's. Young male population's (X_4) mean is 8.26, and its range is 5.20 to 16.80, so a 1% increase in young male population does not seem so small as to be uninterpretable. Relative to the mean (342.17) and range (116.1 to 854.2) of violent crime rates (Y), a decrease of 23.818 in the violent crime rate

seemed to the researcher to be large enough to say that the effect of young male population (X_4) on violent crime rate (Y) has real-life significance.

With these results, the researcher was able to evaluate the research hypothesis. Based on the earlier examination of p-values and slopes (b's), renter-occupied housing (X_2) is the only X that showed a statistically significant (and also real-life significant) effect on crime (Y) in society in the direction—violent crime rate increases as renter-occupied housing increases—that was expected under the research hypothesis. The effect of young male population (X_4) was statistically significant (and also real-life significant), but in the negative direction—violent crime rate decreases as young male population increases—which is opposite of what the research hypothesis suggested. R^2 is moderately high, indicating that the set of X's is explaining crime (Y) fairly well, but it would be worthwhile for further analyses to explore additional X's that might also influence crime (Y). Note, though, that there is a practical constraint on the number of X's here because as a general rule one wants the number of X's to be relatively small compared to the sample size N. The small sample size here (N = 24) means that it will not be appropriate to include a very large number of X's in the regression.

Overall, then, the research hypothesis was not very well supported by the regression analysis results for these data. Of course it is possible that the measurements of variables, or level of analysis, are not ideal. For example, economic deprivation could instead be measured by the percentage of city residents living in poverty, or racial inequality could be measured as the difference between average incomes for white and nonwhite residents. It is also possible that a lower level of analysis, such as neighborhoods, would provide a better test of the hypothesis. In addition, the counterintuitive finding of a negative relationship between young male population (X_4) and violent crime rate (Y) may not have arisen had the analysis controlled for other important X's. Among these cities, Boulder, Ft. Collins, and Golden have exceptionally high young male populations because these cities are home to the University of Colorado, Colorado State University, and the Colorado School of Mines, respectively. None of these cities have especially high violent crime rates, so without controlling for other variables that could capture the distinctive "college town" character of these cities, the regression analysis takes the data for these cities as evidence for a negative relationship between young male population and violent crime rate.

The researcher found a positive and significant relationship between renter-occupied housing and crime from these city-level (aggregate-level) data. She can make only aggregate-level interpretations from these findings. That is, it is legitimate to say that the analysis indicates that, holding other X's constant, cities with a high proportion of renter-occupied housing units are predicted to have higher crime rates than cities with a low proportion of renter-occupied housing units. However, it is not legitimate to make an individual-level interpretation of these results. This city-level analysis does *not* show that renters are more likely to commit crime than home owners; that would be the ecological fallacy of drawing individual-level conclusions from the aggregate-level results.

The p-values in the table of regression results for this example are two-sided, as they refer to the alternative hypothesis $\beta \neq 0$; this usually is the case with the default output from statistical software. Note that the original research hypothesis specified positive signs for the relationships between independent variables

and the dependent variable. The researcher could appeal to various theories of crime to argue that the negative relationships between these independent variables and crime would be very unlikely. Recalling the review of hypothesis tests in Chapter 1, one might then suggest using one-sided rather than two-sided p-values in this situation, with H_1 specifying the sign of the effect if the null hypothesis of $\beta = 0$ is rejected. These one-sided p-values can be obtained by simply dividing the reported two-sided p-values by 2, which of course makes the one-sided p-values smaller. Statistically significant relationships with the two-sided p-value ≤ 0.05 would remain so in the one-sided approach, but some previously nonsignificant relationships might become statistically significant.

In the analysis here, the one-sided p-values for the previously nonsignificant independent variables unemployment (X_1) and racial segregation (X_3) would remain over the 0.05 cutoff, but close enough (0.122 / 2 = 0.061 and 0.148 / 2 = 0.074, respectively) to perhaps be viewed as borderline or marginally significant. However, this example also illustrates our caution about using the one-sided p-value in many research settings. The seemingly impossible negative effect of young male population on crime was in fact observed in the researcher's analysis of these data, despite the strong belief among researchers that societies with large young male populations tend to have high levels of violent crime. As discussed earlier, we can understand why this counterintuitive result was found in this particular sample and how it might change in the presence of other controls. Still, this illustrates how we may not always be so confident that we can absolutely rule out either a positive or negative effect when applying our understanding of theory or past research to new situations or data. We therefore will reserve the one-sided approach for analyses in which we have an unusually strong basis for ruling out one sign as impossible for the true effect of X on Y, even if our initial expression of the research hypothesis suggests a particular direction of the effect.

Finally, these data were obviously not derived from a true experiment that could be manipulated by the researcher, but instead came from observation of the world as it is. Of course, there would be no possible way for the researcher to randomly assign different levels of renter-occupied housing and the other independent variables to different cities. Therefore the effect of renter-occupied housing on crime found by the researcher here should be understood as indicating an association between these variables. This *could* reflect a causal effect of renter-occupied housing on crime, but the nature of the research design does not allow the researcher to conclude that definitively.

2.5.2 Example 2: Infant Mortality in the World

A researcher believed that greater economic and technological development, health care availability, and urbanization decrease infant mortality in the world's nations and wanted to test this hypothesis with a multiple regression analysis. Further, the researcher was interested in exploring this question via historical rather than contemporary data, as theoretical arguments suggested that these relationships could be different in different historical periods. For this analysis, independent variables were economic and technological development, health care availability, and urbanization, and the dependent variable

was infant mortality. He decided to use 28 randomly selected countries as the cases, so this research was at the country level (aggregate level of analysis). Data were obtained from a reference book that collected data from various original sources and reflected conditions around 1990.

Economic development was measured by gross domestic product (GDP, essentially the value of goods and services produced in the country's economy) per capita (or per person) expressed in U.S. dollars. A high value of GDP per capita in a country indicates a high level of economic activity and development. Technological development was measured by the number of people per telephone in the country; note that these data were collected before the cell phone era. A high number of people per telephone in a country indicates a low level of technological development, because more people "sharing" each phone means that there are, relative to the population, fewer phones and less access to technology. Health care availability was measured by the number of people per hospital bed. As with the telephone measure, a high number of people per hospital bed in a country indicates a low level of health care availability, as this means that there are few hospital beds relative to the population. Urbanization was measured by the percentage of population living in urban areas. It hardly needs to be said that a country with a high percentage of population living in urban areas is highly urbanized. Finally, infant mortality was measured by the number of infant deaths per 1,000 births.

The research hypothesis, units of analysis, measurements of independent and dependent variables, and raw data are as follows:

- *Research hypothesis*: Greater economic and technological development, health care availability, and urbanization decrease infant mortality in society.
- *Units of analysis*: 28 countries (aggregate-level data).

Measurements of independent and dependent variables:

- X_1: GDP per capita in U.S. dollars (economic development).
- X_2: Number of people per telephone (technological development).
- X_3: Number of people per hospital bed (health care availability).
- X_4: Percentage of population living in urban areas (urbanization).
- Y: infant mortality rate per 1,000 births (infant mortality).

Research hypothesis with measurements of independent and dependent variables:

1. GDP per capita (X_1) has a negative effect on infant mortality rate (Y).
2. The number of people per telephone (X_2) has a positive effect on infant mortality rate (Y). Countries with more people "sharing" a phone (X_2), reflecting less access to technology, will have higher infant mortality rates (Y).

3. The number of people per hospital bed (X_3) has a positive effect on infant mortality rate (Y). Countries with more people per hospital bed (X_3), indicating less health care availability, will have higher infant mortality rates (Y).

4. The percentage living in urban areas (X_4) has a negative effect on infant mortality rate (Y).

Data:

Country ID	Country Name	Infant Mortality (Y)	GDP per Capita (X_1)	People per Telephone (X_2)	People per Hospital Bed (X_3)	Percentage Living in Urban Areas (X_4)
1	Angola	151	950	132	645	29
2	Bangladesh	112	200	572	3,175	24
3	Bolivia	83	690	37	685	51
4	Burkina Faso	117	205	482	1,359	8
5	China	33	360	83	428	27
6	Cyprus	10	7,585	2	165	69
7	Ecuador	60	1,070	28	610	54
8	Ethiopia	113	130	320	3,873	11
9	Germany	7	14,600	1.5	95	86
10	Guyana	5	300	47	341	35
11	Indonesia	70	630	172	1,485	31
12	Jamaica	17	1,400	13	468	52
13	Liberia	119	440	278	800	46
14	Madagascar	93	200	239	600	22
15	Mauritius	22	2,300	15	364	41
16	Morocco	56	1,060	62	959	50
17	Netherlands	7	16,600	1.6	164	88
18	Pakistan	105	380	131	1,706	32
19	Poland	14	4,300	7.5	154	62
20	Saudi Arabia	69	5,800	13	406	78
21	South Korea	23	6,300	3.3	429	74
22	Spain	6	12,400	2.5	198	79

(*Continued*)

Country ID	Country Name	Infant Mortality (Y)	GDP per Capita (X_1)	People per Telephone (X_2)	People per Hospital Bed (X_3)	Percentage Living in Urban Areas (X_4)
23	Syria	45	2,300	17	840	50
24	Turkey	54	3,400	7	465	61
25	United Kingdom	8	15,900	1.9	138	90
26	United States	10	22,470	1.9	198	76
27	Venezuela	23	2,590	11	370	83
28	Zambia	77	380	78	311	4[illegible]

The researcher first ran a correlation analysis in his statistical software, to check correlations among the X's in order to detect potential collinearity problems. The results of the correlation analysis were:

	GDP per Capita (X_1)	People per Telephone (X_2)	People per Hospital Bed (X_3)	Percent Living in Urban Areas (X_4)
GDP per Capita (X_1)	1.000	−0.439	−0.431	0.755
People per Telephone (X_2)	−0.439	1.000	0.767	−0.728
People per Hospital Bed (X_3)	−0.431	0.767	1.000	−0.656
Percent Living in Urban Areas (X_4)	0.755	−0.728	−0.656	1.000

The correlation analysis indicated that percentage living in urban areas (X_4) is strongly correlated with GDP per capita (X_1) (r = 0.755), the number of people per telephone (X_2) (r = −0.728), and the number of people per hospital bed (X_3) (r = −0.656). The number of people per telephone (X_2) and the number of people per hospital bed (X_3) are also strongly correlated (r = 0.767). These high correlations suggest that collinearity may indeed be a problem in the regression analysis.

Results of the multiple regression with all X's included were as follows, with "SE" again written for "standard error":

$R^2 = 0.673$

df for t = 23

	b	SE of b	t	p-value
Intercept (constant)	67.718	24.807	2.730	0.012
GDP per capita (X_1)	−0.002	0.001	−1.556	0.133
People per telephone (X_2)	0.116	0.061	1.899	0.070
People per hospital bed (X_3)	0.004	0.009	0.386	0.703
Percent living in urban areas (X_4)	−0.333	0.444	−0.750	0.461

The results show the classic symptom of collinearity that we discussed earlier. The R^2 is fairly high (0.673), suggesting rather successful predictions of Y in the sample, but none of the effects of X on Y are statistically significant (although the p-value for X_2, number of people per telephone, is close to the 0.05 cutoff). The researcher then tried to address the collinearity problem by excluding an X from each of the highly correlated pairs and rerunning the regression analysis.

Because percentage living in urban areas (X_4) is strongly correlated with three other X's (GDP per capita [X_1], the number of people per telephone [X_2], and the number of people per hospital bed [X_3]), it seemed prudent to exclude percentage living in urban area (X_4) from the next regression model. With the number of people per telephone (X_2) and the number of people per hospital bed (X_3) also strongly correlated, it would be reasonable to exclude one of those two also. The researcher chose to exclude the number of people per hospital bed (X_3) along with percentage living in urban areas (X_4). His next regression model then included only GDP per capita (X_1) and the number of people per telephone (X_2).

This seems to restate the overall research hypothesis as "greater economic and technological development decrease infant mortality in society," and the specific hypotheses as "GDP per capita (X_1) has a negative effect on infant mortality rate (Y)," and "the number of people per telephone (X_2) has a positive effect on infant mortality rate (Y)." Note, however, that the choice of which X's to exclude in response to collinearity was rather arbitrary. With health care availability and urbanization measures excluded for this reason, the researcher must be careful with interpretation. If results show statistically significant effects of the economic and technological measures, that does not rule out health care availability and urbanization as factors affecting infant mortality. Rather collinearity is making it too hard to distinguish all these different effects and forcing the researcher to be a bit more modest in his research goals.

Regression results with GDP per capita (X_1) and the number of people per telephone (X_2) are as follows:

$R^2 = 0.661$

df for t = 25

	b	SE of b	t	p-value
Intercept (constant)	52.430	8.138	6.443	0.000 (<0.001)
GDP per capita (X_1)	−0.003	0.001	−3.098	0.005
People per telephone (X_2)	0.157	0.037	4.265	0.000 (<0.001)

R^2 is little changed, with 66.1% of the variance in infant mortality (Y) explained by GDP per capita (X_1) and the number of people per telephone (X_2). Excluding the number of people per hospital bed (X_3) and percentage living in urban areas (X_4) from the regression did not reduce R^2 very much. That is, the new regression model's fit to the sample data is only slightly worse than when all four X's were included.

With number of people per hospital bed (X_3) and percentage living in urban areas (X_4) removed, effects of GDP per capita (X_1) and the number of people per telephone (X_2) on the predicted infant mortality rate (Y) become statistically significant. The p-value for GDP per capita (X_1) is 0.005, and the p-value for the number of people per telephone (X_2) is shown by some statistical software as 0.000, which can be interpreted as < 0.001 (it is not literally zero, but more decimal places would be needed to show that). Both are well below the 0.05 cutoff.

With this statistical significance, we can interpret effects (slopes) for GDP per capita (X_1) and the number of people per telephone (X_2). For GDP per capita (X_1), the slope b_1 is −0.003. This indicates a negative effect of GDP per capita (X_1) on the predicted infant mortality rate (Y) and that the predicted infant mortality rate per 1,000 births (Y) decreases by 0.003 as GDP per capita (X_1) increases by $1, controlling for the number of people per telephone (X_2). Of course $1 represents a very small change in GDP per capita (X_1), given the mean ($4,462.14) and range ($130–$22,470) of this variable's values in this data set. A more helpful interpretation might be that the predicted infant mortality rate per 1,000 births (Y) decreases by 3 as GDP per capita (X_1) increases by $1,000. The value 3 was obtained by multiplying b_1 by the $1,000 change in X_1: if a $1 increase in GDP per capita decreases the predicted infant mortality by 0.003, then a $1,000 increase in GDP per capita will decrease the predicted infant mortality by 1,000 × 0.003, or 3. Put this way, the effect does not seem so tiny, considering the mean (55.54) and the range (6–151) of infant mortality rate (Y) in the sample, even if it is still not very large. Thus, the effect of GDP per capita (X_1) appears to be meaningful in real-world terms.

For number of people per telephone (X_2), the slope (b_2) is 0.157. This is a positive effect of the number of people per telephone (X_2) on the predicted infant mortality rate (Y), and the predicted infant mortality rate (Y) increases 0.157 per 1,000 births as the number of people per telephone (X_2) increases by one, controlling for GDP per capita (X_1). Again it is interesting to consider a somewhat larger change in the number of people per telephone (X_2), as there is a fairly large mean (98.76) and very wide range of values (1.5–572) of X_2 in the sample. If the number of people per telephone (X_2) increases by 10, the predicted infant mortality rate (Y) increases by 1.57 (calculated as 10 × 0.157) per 1,000 births,

still controlling for GDP per capita (X_1). As with the effect of GDP per capita, this change now does not seem so tiny relative to the sample mean (55.54) and range (6–151) of infant mortality rate (Y) and so might be assessed as having real-world significance. However, some researchers might look at that change as still being too small to indicate real-world significance. As discussed earlier, there can be ambiguity in the assessment of real-life significance.

The researcher then evaluated the new research hypotheses. With R^2 fairly high, the model does pretty well at predicting levels of infant mortality for the countries in the sample. R^2 is not so close to 1 as to suggest that all important variables have been included, but the model is reasonably successful. Further, both GDP per capita (X_1) and number of people per telephone (X_2) have statistically significant effects on the infant mortality rate (Y), and both effects have the expected signs: negative for GDP per capita, and positive for number of people per telephone. Both effects are also judged to be large enough to suggest real-life significance, though with some ambiguity for number of people per telephone. The revised research hypotheses thus seem to be supported in this analysis. Still, it is important to keep in mind the collinearity-induced choices of X's from the original set and realize that similar results likely would have been found if different choices of which independent variables to keep and drop had been made.

The impact of different choices of which X's to keep in the face of collinearity can be illustrated by considering these data further. Remember that the number of people per telephone (X_2) and the number of people per hospital bed (X_3) were strongly correlated (r = 0.767), and the researcher chose number of people per hospital bed (X_3) to exclude from the regression model. What if the other choice—excluding the number of people per telephone (X_2)—had been made? Here is the output from the regression model with independent variables GDP per capita (X_1) and number of people per hospital bed (X_3):

R^2= 0.572

df for t = 25

	b	SE of b	t	p-value
Intercept (constant)	53.224	10.043	5.300	0.000 (<0.001)
GDP per capita (X_1)	−0.003	0.001	−3.134	0.004
People per hospital bed (X_3)	0.021	0.007	3.027	0.006

R^2 is somewhat smaller than in the previous regression but still moderately high. And, as in the previous model, both X's show statistically significant effects. We can interpret the slopes for X_1 and X_3. Predicted infant mortality rate per 1,000 births (Y) decreases by 0.003 as GDP per capita (X_1) increases by $1, controlling for the number of people per hospital bed (X_3). (As before, a more helpful interpretation for assessing real-life significance might focus on an

increase of $1,000 in GDP per capita.) This is the same effect of GDP per capita (X_1) as was estimated in the previous regression.

For number of people per hospital bed (X_3), the slope b_3 is 0.021. That is, predicted infant mortality per 1,000 births increases by 0.021 as the number of people per hospital bed (X_3) increases by one, controlling for GDP per capita (X_1). Again it is interesting to consider a somewhat larger change in the number of people per hospital bed (X_3), as there is a large mean (773.25) and very wide range of values of X_3 (95–3,873) in the sample. For an increase of 100 in the number of people per hospital bed (X_3), the predicted infant mortality rate (Y) increases by 2.1 per 1,000 births, while holding GDP per capita (X_1) constant. When considered this way, a reasonable change in the number of people per hospital bed produces a meaningful change in predicted infant mortality, suggesting real-world significance of this effect. An increase in the number of people per hospital bed indicates a decrease in the availability of health care, so, controlling for economic development, predicted infant mortality increases as health care availability decreases. This is consistent with the original hypothesis about infant mortality and health care availability.

The regression results with GDP per capita (X_1) and number of people per telephone (X_2) are quite similar to those with GDP per capita (X_1) and the number of people per hospital bed (X_3). Although R^2 is somewhat higher in the regression using number of people per telephone (X_2), in each regression both independent variables have statistically significant effects. Also, the size and direction of the effect of GDP per capita (X_1) are the same in the two regressions, and whether number of people per telephone (X_2) or number of people per hospital bed (X_3) is used, the effect of this variable is statistically significant and positive.

This similarity in results is not surprising. Because highly correlated variables represent similar information, using one or the other in the regression will often produce similar results. This is why we should be cautious in discussing results for regressions carried out after removing variables due to collinearity. We should not give the impression that one or the other of these regressions definitively establishes that either technological development (represented by X_2) or health care availability (represented by X_3) is the factor that is truly related to infant mortality (Y). Instead, the collinearity limits us to deciding that one or both of these factors appear to be related to infant mortality, but the nature of the data prevents a more conclusive interpretation. (We will revisit this idea as one application of the F-statistic in Chapter 3.) And as before, the nonexperimental research design does not permit us to ascribe a definitive causal interpretation to these effects.

2.6 Exercises

2.6.1 Exercise 1: Median Income Among U.S. States

A researcher hypothesized that higher unemployment and a larger older adult population decrease the typical income in a given area. On the other hand, she hypothesized that greater educational attainment and urbanization increase an area's typical income. To test these research hypotheses, she collected data from 34 American states.

Typical income was measured as the state's median annual income in dollars, and unemployment was measured as the percentage of its labor force that is unemployed. Older adult population was measured as the percentage of population aged 65 years and over, with educational attainment measured as the percentage of those 25 years and older holding at least a bachelor's degree. Finally, urbanization was measured as population density (population per square mile). The following data were collected from the Census Bureau:

ID	State	Median Income	Unemployment	Older Adult Population	Educational Attainment	Urbanization
1	Alabama	48,123	5.8	16.5	25.5	96.3
2	Arizona	56,581	5.8	17.1	27.4	61.8
3	Arkansas	45,869	5.6	16.5	23.4	57.7
4	California	71,805	5.9	13.9	33.6	253.8
5	Colorado	69,117	4.2	13.9	41.2	54.1
6	Connecticut	74,168	6.1	16.8	38.7	741.1
7	Delaware	62,852	5.3	18.0	31.5	493.6
8	Florida	52,594	5.5	20.1	29.7	391.3
9	Georgia	56,183	5.8	13.4	30.9	181.3
10	Illinois	62,992	6.1	15.2	34.4	230.6
11	Indiana	54,181	4.7	15.4	26.8	186.1
12	Iowa	58,570	3.6	16.7	28.9	56.3
13	Kentucky	48,375	5.5	15.9	24.0	112.8
14	Louisiana	45,145	6.5	14.9	23.8	108.4
15	Maryland	80,776	5.2	14.9	39.7	623.5
16	Massachusetts	77,385	4.6	16.1	43.4	879.5
17	Michigan	54,909	5.9	16.7	29.1	176.2
18	Minnesota	68,388	3.6	15.4	36.1	70.0
19	Mississippi	43,529	7.0	15.6	21.9	63.6
20	Missouri	53,578	4.6	16.5	29.1	88.9
21	New Hampshire	73,381	3.8	17.6	36.9	150.0
22	New Jersey	80,088	5.3	15.7	39.7	1,224.6
23	New York	64,894	5.5	15.9	36.0	421.2
24	North Carolina	52,752	5.3	15.9	31.3	211.3

(*Continued*)

ID	State	Median Income	Unemployment	Older Adult Population	Educational Attainment	Urbanization
25	Ohio	54,021	5.2	16.6	28.0	285.3
26	Oklahoma	50,051	5.4	15.3	25.5	57.3
27	Pennsylvania	59,195	5.3	17.8	31.4	286.2
28	Rhode Island	63,870	5.7	16.7	33.5	1,024.8
29	South Carolina	50,570	5.8	17.2	28.0	167.1
30	Tennessee	51,340	4.9	15.9	27.3	162.9
31	Texas	59,206	5.1	12.2	29.6	108.4
32	Vermont	57,513	3.8	18.8	38.3	67.7
33	Virginia	71,535	4.6	15.0	38.7	214.5
34	Washington	70,979	4.9	15.1	35.1	111.4

Year of data: 2017
Data source: Census Bureau

1. What are the independent and dependent variables for this analysis?
2. After entering the above data in statistical software, check for collinearity and then run multiple regression analysis as needed to test the research hypothesis.
3. Interpret the results (output) as fully as possible, focusing on R^2, p-values, b's, and real-life significance (remember that descriptive statistics are helpful in making interpretations of real-life significance).
4. Evaluate the research hypothesis.
5. Use the results to predict median income for a state with 4.3% unemployment, 18.1% older adult population, 35.5% holding a bachelor's degree, and 260.3 people per square mile. Should you trust this predicted value? Explain.

2.6.2 Exercise 2: Predicting Educational Attainment

Researchers were interested in studying influences on educational attainment, and obtained data on a sample of American adults who had completed their formal education. They believed that an adult's education could be predicted from measurements taken at childhood on his or her reading test score, resource competition in the family, parental education, and household income. In the following hypothetical descriptive and regression analyses of the data from this sample, the dependent variable is years of education, and the independent variables are reading test score, number of siblings (measuring resource competition), parental education (whichever parent had the highest, in years), and household annual income for the respondent at age 15 (in dollars).

Descriptive Statistics:

	Mean	Minimum	Maximum
Respondent's education	14.3	9	21
Test score	100	82	128
Number of siblings	2.6	0	6
Parental education	12.2	8	18
Childhood household income	54,567.12	24,000	78,000

Multiple Regression Analysis Results:

$R^2 = 0.228$

df for t = 103

	b	SE of b	t	p-value
Intercept (constant)	−95.5			
Test score	0.028	0.017	(i)	0.103
Number of siblings	−0.121	0.058	(ii)	0.039
Parental education	1.122	0.580	(iii)	0.056
Childhood household income	0.002	0.001	2.000	0.048

1. Assuming no collinearity, interpret the results as fully as possible, focusing on R^2, p-values, b's, and statistical and real-life significance.
2. What are the values for the t-statistics (i), (ii), and (iii) not shown in the table?
3. What are the predicted years of education for a person with a test score of 105, three siblings, whose most-educated parent had 12 years of education, and whose household income was $46,000 when they were 15 years old? Should you trust the predicted value?
4. What is the total number of cases in the sample?
5. Is the research hypothesis supported? Explain.

2.6.3 Exercise 3

1. Suppose that our research hypothesis is represented by the set of independent variables in the regression. Explain why both R^2 and statistical significance of the b's are important in deciding whether the research hypothesis is supported.

2. If we believe that there are multiple factors that influence our dependent variable, why do we need to use multiple regression instead of repeatedly applying simple (bivariate, with just one independent variable) regression?

3. Explain how collinearity can lead to artificial findings of statistical nonsignificance of X's.

2.7 Appendix A: Beginning a Research Project Using Multiple Regression

2.7.1 Setting Up the Research Question and Hypothesis

This chapter presented core interpretations of multiple regression results. Here we step back and consider how to start a research project using multiple regression analysis. We begin with a research question and a research hypothesis about the expected influences of the multiple independent variables on the dependent variable. Social science theories often suggest such relationships, and so theories can help guide us in choosing the independent and dependent variables. Additional independent variables may be suggested by previous quantitative research, or even just common sense. Note that we are using "research hypothesis" to mean an overall statement of expected influences on the dependent variable. This is different from the specific statistical hypotheses (H_0 and H_1) that we discussed in the main text of this chapter. Of course specific statistical hypothesis tests will be helpful in evaluating the research hypothesis.

When a theory is guiding the analysis, independent variables that do not have a central role in the theory are sometimes called *"control" variables*. This distinguishes them from the variables that are the main elements of the theory and the focus of the research hypothesis. The researcher expects them to be related to Y and likely correlated with at least some of the independent variables that are the theory's main focus, and so wants to control for them when determining the effects of the independent variables of primary interest. Despite this conceptual distinction, however, the control variables are still treated the same as other independent variables in the regression analysis itself, and the control variables play the same part in predicting Y as do the other independent variables. It is just that results for the control variables may not be emphasized as much when writing a report or article on the research.

For example, in criminology Shaw and McKay's (1942) social disorganization theory suggests that greater poverty, residential instability, and cultural conflict increase crime in a community. In this case, the dependent variable is crime, and the independent variables suggested by the theory will measure poverty, residential instability, and cultural conflict. In addition, much previous research has found that areas with more young males—an especially crime-prone group—tend to have higher crime rates, and the size of the young male population may also be correlated with the independent variables suggested by the theory. Thus young male population should be included in the regression analysis as another independent variable, but with respect to the theory it is a

control variable. In this case, then, the research hypothesis would be "The levels of poverty, residential instability, cultural conflict, and young male population in a community all have positive effects on community crime rates." Here the research hypothesis specified the direction of the relationships (in this case, positive) between the independent variables and the dependent variable, but sometimes the research hypothesis only says that there is a relationship, leaving the direction unspecified.

2.7.2 Level of Analysis

Theories can also suggest whether an aggregate-level or an individual-level analysis is most appropriate for our research. As discussed in Chapter 1, level of analysis refers to the nature of the subjects or units we are studying. In the social disorganization example above, the theory argues that communities with high levels of poverty, residential instability, and cultural conflict are more likely to have high crime rates. This suggests the community as the appropriate unit to study, yielding an aggregate-level analysis. The theory does not necessarily imply that, within a community, crime is being committed by individuals who are poor or have recently moved; an individual-level theory of crime would be needed for that sort of investigation. (In fact theories sometimes encompass more than one level; we briefly discuss multilevel data in Chapter 9.)

In this way, social disorganization theory is tied to aggregate-level analysis, using data from communities such as neighborhoods, cities, counties, or states. An individual-level theory such as Hirschi's (1969) social control theory, on the other hand, should be examined with data on individuals. As we noted in Chapter 1, in practice it is usually the case that data from higher levels, such as states, are easier to obtain than data from lower levels, such as neighborhoods or individuals. This makes it especially important to be careful not to assume that results obtained from data at one level of analysis automatically apply to another level. Recall the "ecological fallacy" from Chapter 1, in which findings from aggregate-level research are inappropriately applied to individual-level phenomena. Because aggregate-level data are often more accessible, it can be very tempting to draw individual-level conclusions from aggregate-level data, but this temptation should be resisted. It is important to use data that match the level of analysis suggested by the theory we are investigating and to interpret results in terms of that level.

2.7.3 Measuring Independent and Dependent Variables

Our next step is to think how to actually measure the independent and dependent variables for the subjects or units we are studying. Again we can take advantage of previous quantitative research and common sense in coming up with measurements of the variables. For the social disorganization theory example, poverty is often measured by the percentage of households in a community whose total income falls below an official poverty line. Residential instability can be measured as the percentage of residents living in different housing than they were 5 years ago, or by the community's percentage of renter-occupied housing units. Cultural conflict is often measured by the community's

percentage of foreign-born population. Young male population is simply the percentage of the population that is made up of men in a particular age range. Crime can be measured as the total crimes (or a more specific crime type such as homicides or robberies) reported to police in a community, converted into a rate per 100,000 population so that values from communities of varying size are comparable.

In any case, we want measurements that closely match the theoretical concept for each variable while still being practical to obtain for all the subjects or units in our study. When there is no available variable that can directly measure a particular theoretical concept, or there is no practical way to carry out that measurement in our sample, we must either find the best available alternative or not include a variable measuring that concept at all. In either case, this will be a limitation of our research and will be important to note in any papers or reports describing our work.

2.7.4 Data Collection

After deciding on the level of analysis, and appropriate measurements of the independent variables and the dependent variable, we can collect data on these measurements for the subjects or units of analysis in our sample. For the social disorganization example, suppose that we have decided to use states as our units of analysis; remember that the theory requires aggregate-level analysis of some kind. We can then collect data from all 50 American states on the percentage of households below the poverty line (the poverty measure) the percentage of residents living in different housing than 5 years ago or the percentage of renter-occupied housing units (the residential instability measure), the foreign-born population percentage (the cultural conflict measure), the population percentage of young men aged 15 to 21 years (the young male population measure), and the total crimes reported to police per 100,000 population (the crime measure). For the United States, a great variety of national or community (aggregate-level) data can be found at the websites of various government agencies, including, among others, the Census Bureau and the Bureau of Justice Statistics. An excellent source for individual-level data that can be used to investigate a wide variety of research questions is the University of Michigan's Inter-University Consortium for Political and Social Research (ICPSR) data archive.

After data collection, we enter data in our statistical software, and use the software to run the multiple regression analysis that we have decided will best address our research question and hypothesis. The software output will give us the numerical results described above and allow us to make interpretations. In some cases, there will be an intermediate step in which some variables need to be transformed before being used in the regression analysis, but we will discuss that situation in subsequent chapters.

2.7.5 Evaluation of Research Hypothesis

We want to use the regression results to evaluate the research hypothesis that we developed from our research question. How well is the research hypothesis

supported in our data? To evaluate this, we go through the different elements of interpretation that we discussed in the preceding sections of this chapter. Recapping those sections, we need to consider (a) the value of R^2, (b) statistical significance of the effects of the X's (via the p-values), and (c) for those that are statistically significant, the slopes, or effects, of the X's (the b's) in terms of direction of the effects and their real-life significance. In general, we view the analysis as strongly supporting our research hypothesis if R^2 is high and all (or most of) the X's have statistically significant effects on Y that are in the expected (by the research hypothesis) direction, and, further, these effects are large enough to have real-life significance.

A high R^2 means that the set of X's suggested by the research hypothesis is predicting Y well and that we seem to be accounting for the main influences on Y. Statistical significance of every X's effect means that each of the hypothesized X's does appear to be related to Y, because for each X we are rejecting the statistical null hypothesis of no effect (H_0). For each X with a statistically significant effect we can check the effect's direction (positive or negative) from the slope b and see if it matches the direction suggested by the research hypothesis (though again sometimes that direction is not specified by the research hypothesis). We can also determine from each b if the magnitude of the effect is enough to believe that the corresponding X has a real impact on Y (still realizing that a genuinely causal interpretation of this effect usually cannot be made).

Even if most, or all, of the X's have statistically significant effects on Y in the direction suggested by the research hypothesis, there could still be a low value of R^2. This typically means that the regression model is incomplete, in the sense that our regression is missing some other X's that are important in predicting Y, so that the research hypothesis is incomplete as an explanation for Y. In that case we may want to seek other X's based on theory, previous research, or common sense that could also affect Y and rerun the regression with these additional X's to see if R^2 improves. (Note too that a low R^2 could be due, at least in part, to the presence of nonlinear relationships between some X's and Y. We discuss such relationships in later chapters.) The opposite situation, with a high R^2 but none or almost none of the effects statistically significant, often indicates the problem of collinearity, which is discussed in this chapter.

The situation of a low R^2 along with none or almost none of the effects of the X's being statistically significant is quite damaging for the research hypothesis. In that case, the research hypothesis seems to be both missing important influences on Y **and** incorrect in suggesting that the X's it named are actually related to Y. This circumstance may be unlikely when we are deriving the research hypothesis from a well-developed theory, but sometimes even a well-developed theory is not supported when confronted with real data.

Along with discussing results, a paper reporting on a quantitative research project should also compare the current findings to those of previous research, discuss limitations, and, if applicable, suggest policy implications. In many cases it is also helpful to indicate future research directions that could build on the work. In the following section, we discuss a typical framework for organizing the parts of a quantitative research paper.

2.7.6 Organization of Quantitative Research Paper Using Multiple Regression

To help in writing a complete quantitative research paper, we list each part of such a paper below, following a standard framework that is typical of published research papers in the social sciences: (1) abstract, (2) introduction, (3) literature review, (4) research question, (5) data and measurement, (6) analysis and results of analysis, (7) conclusion, (8) tables and figures, and (9) references. Each part is discussed below. Note that the organization of the paper is separate from the specific format in which it is written. By "format" we mean details of how to present elements such as sections, section headings, citations, references, tables, and so on. Commonly used styles include those of the American Psychological Association (APA), American Sociological Association (ASA), and the Chicago Manual. Note that the organization of the paper laid out here is not universal in the social sciences, and examples of many variations on this structure can be found in the published literature.

1. **Abstract**
    - A short overall description of the paper, including results.
2. **Introduction**
    - Present your research topic (focusing on the relationship between X[s] and Y).
    - Be sure to mention the level of analysis for your research question (individual or aggregate).
    - Discuss why it is important to study this relationship (or relationships).
3. **Literature Review (Review of Theory and Previous Research)**
    - In social science, researchers' interest in the relationship between X(s) and Y is typically inspired by theories and/or previous research findings on the same, or closely related, topics.
    - Summarize relevant theories as well as findings of previous quantitative research. When summarizing previous research, focus on the research questions, the units being studied, the data sources, the types of data analysis, the independent and dependent variables, and the results/findings.
4. **Research Question**
    - Discuss your research question (topic) and research hypothesis (expectations for findings). It is often helpful to note the main differences between previous research and what you are doing, which could include points related to the data (such as the data source or when the data were collected) or the analytic methods being applied. When the goal is to eventually publish the paper

in a scholarly journal, the "novelty" of the research is a key part of the evaluation of the paper's potential importance. In other settings, novelty may not be emphasized so much. Also, sometimes a researcher working on a new topic or question for which little research or theory exists will not really have expectations for how results will look and so cannot state a research hypothesis. Exploratory work, therefore, will not always have this element.

5. **Data and Measurement**

   - Describe your data in detail, focusing on when it was originally collected, units (such as persons, cities, states, or nations), and data sources.
   - List independent (including control) and dependent variables and discuss how they were measured. For example: "Economic hardship is measured by the unemployment rate, obtained by dividing the number of unemployed people by the total number of people in the civilian labor force, times 100. The level of crime is measured by the number of reported homicides per 100,000 population."

6. **Analysis and Results of Analysis**

   - Use statistical software to run appropriate descriptive statistics for the independent and dependent variables involved in the multiple regression analysis and briefly discuss the descriptive statistics.
   - Diagnose collinearity and, if present, try to address it.
   - Discuss the statistical techniques being used.
   - Use statistical software to run multiple regression analysis; interpret and discuss results of this analysis, focusing on R-squared, p-values, and b's (including real-life significance).

7. **Conclusion (Implications of Analysis for Research Question and Hypothesis)**

   - Discuss whether your findings support the theoretical perspectives that you presented earlier in the paper and the research hypothesis posed earlier.
   - Compare your results to the findings of previous research. If there are differences that appear to stem from differences in data sources or analytic methods, it can be useful to discuss those, but this is not always the case.
   - If applicable, discuss policy implications.
   - Discuss the limitations of your research and future research possibilities.

8. **Tables and Figures**

   - Following the specific style required by a course instructor or the applicable style manual, present detailed results of analysis (descriptive and multivariate) in tables.

9. **References**

   - Follow the specific format required by a course instructor or the applicable style manual.

2.8 Appendix B: Additional Issues in Multiple Regression

2.8.1 Standardized Coefficients

This chapter has focused on the usual unstandardized coefficients, interpreted in terms of the units of measurement for the X's and Y. That is, if X_1 is years of education, and Y is dollars of income, b_1 is interpreted as the change, in dollars, in predicted income for an additional year of education, or when education increases by 1 year (holding other independent variables constant). This is very natural, because we have an intuitive grasp of the meaning of a certain change in dollars of income and the magnitude of the change represented by an additional year of education.

However, the fact that the X's in the multiple regression are typically measured in various different units makes it hard to compare the effects of different X's on Y. Continuing the example, suppose that X_2 is father's income. Certainly in any realistic sample the variance in people's incomes (measured in dollars) will be much greater than the variance in people's years of education (measured in years), so "1 more year of education" and "1 more dollar of father's income" are not comparable in any obvious way. Therefore the corresponding effects on Y are difficult to compare, and we cannot easily say which effect is more "important" in predicting Y. Even when we make a careful assessment of each independent variable's real-life significance, it may not be very clear which of the real-life significant effects are the largest.

The standardized coefficient is one response to this problem. This coefficient is a transformation of the usual b, so that the interpretation of an effect is made with respect to standard deviations of X and Y, not the original measurement units. That is, the standardized coefficient for the effect of education would indicate how many standard deviations—not dollars—predicted income would increase for each additional standard deviation—not year—of education. Standardized coefficients seem to allow for better assessment of which X's have the biggest effects on Y because the shift to standard deviations means that the interpretations no longer involve all the different and incomparable units in which the different X's are measured. This sounds as if it also would enhance assessments of real-life significance. An important difficulty, though, is that it is hard for us to think in units of standard deviations. Even if it is advantageous to put every variable's effect into a common framework, real-world significance

still will often seem easier to assess in terms of the original units than in terms of standard deviations. We have a better grasp of the meaning of a change that is reported in the original units than of a change that is reported in units of standard deviations. The common framework of the standardized coefficients may assist with comparison of the effects' size but at the possible cost of less understanding of the effects' meaning.

There are also some technical objections to standardized coefficients, such as the possibility of the same b leading to very different standardized coefficients in different samples, due to differences across samples in the variables' variances. (More advanced texts discuss these technical points.) For these reasons, we will use unstandardized coefficients throughout the text. Still, it is valuable to understand what standardized coefficients are and how they are interpreted, as some researchers prefer to present multiple regression results in that form. Some articles and reports will show both the unstandardized and standardized coefficients in tables of regression results. The symbol β (or the word "beta") will sometimes be used to represent the standardized coefficient, but it is important to be clear that this is a different use than we are making of that symbol. As discussed above, we use β to symbolize a coefficient in the "true" regression model for the entire population, for which b is our estimate from the sample.

2.8.2 More on Diagnosing and Addressing Collinearity

Many statistical software packages include some more formal methods for diagnosing collinearity than simply examining correlations among the X's as we discussed above. For example, variance inflation factor (VIF) scores are an attempt to assess the extent to which collinearity is affecting standard errors of the regression coefficients by considering how strongly each X is related to the set of all other X's in the regression model. When high VIF scores are observed, collinearity is likely a problem, and there are various cutoffs in use for determining what is a high enough VIF score to indicate this.

When collinearity has been detected, one response is to create new variables that combine several highly correlated X's, using techniques such as *principal components* analysis or factor analysis. We discuss principal components a bit more in Chapter 9, but for now we just point out that the nature of the correlations among X's determines the construction of the new variables. Typically these new variables will, by virtue of the method used to construct them, be uncorrelated with each other, while still aiming to convey the information contained in the original X's. The new variables can then be used instead of the original X's in the regression, and by definition there will be no collinearity among independent variables in this new regression. The tradeoff for the desirable absence of collinearity when using these new variables is that b's for these constructed variables will be harder to interpret. The b's in the regression results will no longer refer to natural variables whose measurement and meaning feel intuitive to us, but instead to the constructed variables that have standardized scales. Advanced texts can be consulted for more guidance on this and other methods for detecting and addressing collinearity.