2 How Do I Interpret Multiple Regression Results?

If you read published articles in the social or biomedical sciences, you will inevitably encounter tables that report results from multiple regression analysis. Many of those tables are relatively simple, whereas others may be very complex. In this chapter, we examine multiple regression tables taken from five different articles that appeared in the *American Sociological Review*—arguably the leading journal in the field of sociology—and we shall see how to make sense of them. I chose these tables because they are fairly typical of what you'll find in social science journals, *not* because I thought they were examples of good research. In this chapter, we will be concerned only with what the tables mean, not whether they give correct answers to the research questions. Chapter 3 will be devoted to critiquing these and other studies.

2.1. How Does Education Promote Better Health?

Many studies have shown that more educated people tend to be in better health than less educated people. But why? Is it because educated people have better access to medical care, better work situations, or better lifestyles, or is there some other reason? Ross and Wu (1995) set out to answer this question by analyzing data from a 1990 telephone survey of 2,031 adults in the United States. Table 2.1 is one of several tables of multiple regression results that appeared in their article.

In Table 2.1, we see results from estimating two different regression equations. For both equations, the dependent variable, self-

TABLE 2.1 Self-Reported Health Regressed on Education, Controlling for Sociodemographic Characteristics (Equation 1) and Work and Economic Conditions (Equation 2)

	Equation 1		*Equation 2*	
Variable	*b*	*Beta*	*b*	*Beta*
Education	.076***	.220	.041***	.121
	(.007)		(.007)	
Sex (male = 1)	.114**	.062	.051	.028
	(.038)		(.038)	
Race (White = 1)	.239***	.089	.168**	.062
	(.056)		(.054)	
Age (in years)	−.013***	−.247	−.012***	−.226
	(.001)		(.001)	
Marital status	.105**	.058	.043	.024
(married = 1)	(.040)		(.037)	
Employed full-time[a]			.174***	.098
			(.044)	
Employed part-time[a]			.109	.038
			(.064)	
Unable to work[a]			−.835***	−.150
			(.114)	
Household income			.001	.040
			(.001)	
Economic hardship			−.172***	−.133
			(.028)	
Work fulfillment			.158***	.134
			(.024)	
Constant	3.413		3.729	
R^2	.152		.234	

SOURCE: Table adapted from Ross and Wu (1995). Reprinted by permission.
NOTE: N = 2,031; b = unstandardized regression coefficient with standard error in parentheses; Beta = standardized regression coefficient.
a. Compared to not employed (for reasons other than health).
*p < .05; **p < .01; ***p < .001 (two-tailed tests).

reported health, is "the respondent's subjective assessment of his or her general health" (coded 1 = *very poor*, 2 = *poor*, 3 = *satisfactory*, 4 = *good*, and 5 = *very good*). What distinguishes the two equations is the number of variables: Equation 1 has five independent variables, whereas Equation 2 has six additional variables. For each equation, two columns of numbers are shown, one labeled "*b*" and the other labeled "Beta." As stated in the footnote to the table, the "*b*" column contains the usual regression coefficients along with their estimated

standard errors in parentheses. The "Beta" column contains *standardized coefficients*, which I'll explain a little later in this section.

We'll discuss the reason for having two different equations in a moment. For now, let's focus on the variables and results in Equation 1. Education is simply the number of years of formal schooling completed by the respondent. Equation 1 also controls for four demographic characteristics, three of which are dummy variables. Sex is coded 1 for males and 0 for females; race is coded 1 for whites and 0 for nonwhites; and marital status is coded 1 for married (or cohabiting) and 0 for everyone else. The fourth demographic variable, age, is the reported age in years at the time of the survey.

The estimated regression coefficient for education in Equation 1 is .076. We can interpret that number like this: With each additional year of education, self-reported health goes up, on average, by .076 points. Is this a large or a small number? One way to answer that question is to determine whether or not the coefficient is statistically significant. You'll notice that many of the numbers have asterisks, or stars, after them. This is a common (but not universal) way to show which coefficients are significantly different from zero. As shown in the footnotes to the table, a single asterisk indicates a coefficient whose p value is less than .05, two asterisks signal a coefficient whose p value is less than .01, and three asterisks means that the p value is less than .001. In this table, we see that the coefficient for education has three stars in both Equations 1 and 2. The interpretation is this: If education really has no effect on self-rated health, the probability of finding a coefficient as large or larger than the least squares estimate is less than one in a thousand. All the variables in Equation 1 have statistically significant coefficients. In Equation 2, all the coefficients are significant except for sex, marital status, household income, and employed part-time.

If the stars were omitted from the table, we could still determine the statistical significance of the variables by simple hand calculations. Beneath each coefficient is the standard error (in parentheses). If we divide each coefficient by its standard error, we get the t statistic discussed in Chapter 1. For example, in Equation 2, the t statistic for sex is .051/.038 = 1.34. Because the sample size is so large (N = 2,031), the t statistic has a distribution that is essentially a standard normal distribution. Referring to a normal table (which can be found in any introductory statistics book), we find that the p value for the sex coefficient is .09. Because this is not markedly

higher than the criterion value of .05, it would be unwise to completely dismiss this variable.

There's more to life than statistical significance, however. To get a better idea of what the coefficient of .076 for education means, it's essential to have a clear understanding of the units of measurement for the dependent and independent variables. Recall that self-rated health is measured on a 5-point scale and education is measured in years of schooling. We say, then, that each 1-year increase in schooling is associated with an increase of .076 in self-rated health. This implies that it would take 13 = 1/.076 additional years of schooling to achieve a 1-point increase in self-rated health (e.g., moving from "satisfactory" to "good"). To me, this does not seem like a very large effect.

Remember that the estimate of .076 "controls" for sex, race, age, and marital status. Thus, it tells us what happens when years of schooling changes but all the other variables remain the same. In the same way, each of the coefficients represents the effect of that variable controlling for all the others. The coefficient for age of –.013 in Equation 1 tells us that each additional year of age reduces self-rated health by an average of .013 points on the 5-point scale. Again, although this coefficient is highly statistically significant, the effect is not very large in absolute terms. A 10-year increase in age yields only a tenth of a point reduction in self-rated health.

The other variables in the equation are dummy variables, coded as either 1 or 0. As we saw in Chapter 1, this is a standard way of incorporating dichotomous variables in a regression equation. The coefficient for a dummy variable is like any other coefficient in that it tells us how much the dependent variable changes for a 1-unit increase in the independent variable. For a dummy variable, however, the only way to get a 1-unit increase is to go from 0 to 1. That implies that the coefficient can be interpreted as the average value of the dependent variable for all the people with a value of 1 on the dummy variable minus the average value of the dependent variable for all the people with a 0 on the dummy variable, controlling for the other variables in the model. For this reason, such coefficients are sometimes described as adjusted differences in means. For example, males have self-rated health that averages .114 points higher than females, whites have self-rated health that's about .239 points higher than nonwhites, and married people have self-rated health that's .104 points higher than that of single people.

Equation 2 adds several more variables that describe work status and economic status. Household income is measured in thousands of dollars per year. Its coefficient is small and not statistically significant. Economic hardship is an index constructed by averaging responses to three questions, each measured on a 4-point scale. Not surprisingly, those who report a high degree of economic hardship have lower self-rated health. Work fulfillment is also an average of responses to three questions that are each measured on a 5-point scale. Those who report greater fulfillment have higher self-rated health. Because these two variables have units of measurement that are rather unfamiliar, it is difficult to interpret the numerical magnitudes of their coefficients.

There are three dummy variables in Equation 2 that describe work status: employed full-time, employed part-time, and unable to work. Each of these is coded 1 or 0 depending on whether or not the person is in that particular status or not. There are actually four categories of employment status that are mutually exclusive and exhaustive: employed full-time, employed part-time, unable to work, and not employed (for reasons other than health). As explained in Chapter 8, we can't put dummy variables for all four categories in the regression equation, because this would lead to a situation of *extreme multicollinearity*. Instead, we choose one of them to leave out of the model (in this case, not employed). We call the excluded category the *reference category* because each of the coefficients is a comparison between an included category and the reference category. For example, those employed full-time average .174 points higher in self-rated health than those who are unemployed. Those who are unable to work have self-rated health that is .835 points *lower*, on average, than those who are unemployed. This is hardly surprising because the main reason that people are unable to work is because of poor health. Those employed part-time have slightly better health (.109) than those who are unemployed, but the difference is not statistically significant.

Why did the authors of this study estimate two different regression equations for the same dependent variable? They were trying to answer this question: How much of the association between education and self-rated health can be "explained" by their relationships with other variables? By comparing the coefficients for education in Equation 1 and Equation 2, we can see how the effect of education changes when we control for employment status and

economic well-being. Specifically, we see that the education coefficient is cut almost in half when other variables are introduced into the equation. This tells us that the employment and income variables play an important role in mediating the impact of education on self-rated health. In the original article, the authors also reported two other equations for self-rated health that are not shown here. One equation added two variables measuring "social psychological resources," and the other added two more variables measuring "health lifestyle." When these variables were included, the coefficient for education declined even further, but the declines were not as big as when the employment and economic variables were added. In the model with all the independent variables, the coefficient for education was .031.

Now let's consider the standardized (Beta) coefficients that are reported in the second column for each of the two equations. These statistics are often reported in social science journals. As we've seen, *unstandardized* coefficients depend greatly on the units of measurement for the independent and dependent variables. That makes it hard to compare coefficients for two different variables that are measured in different ways. How can you compare, for example, a 1-unit increase on the economic hardship scale with a change from unmarried to married? Standardized coefficients solve that problem by putting everything into a common metric: standard deviation units. The standardized coefficients tell us how many standard deviations the dependent variable changes with an increase of one standard deviation in the independent variable. For example, in Equation 1 we see that an increase of one standard deviation in years of schooling produces an increase of .220 standard deviations in self-reported health. By comparing standardized coefficients across different variables, we can get some idea of which variables are more or less "important." In both equations, we see that age has the largest standardized coefficient. On the other hand, marital status has the smallest effect in both equations.

A few other things in Table 2.1 are worth noticing. The row labeled "Constant" gives estimates of the intercepts in the two equations. As usual, the intercepts don't tell us much that's interesting. In Equation 1, the intercept is 3.413. That's our estimate of the self-rated health of someone with a value of 0 on all the independent variables. In words, that person is female, nonwhite, and unmarried, with 0 years of schooling and age 0. The lesson here is that you shouldn't pay much attention to the intercept estimates.

The next row is R^2, which is our measure of how well we can predict the dependent variable knowing only the independent variables in the model. For Equation 1, the R^2 is only .15, whereas for Equation 2 it rises to .23. This means that the five variables in Equation 1 "explain" about 15% of the variation in self-rated health. When we add in the other six variables, we can explain about 23% of the variation in self-rated health. Thus, even though nearly all these variables have coefficients that are statistically significant, they account for only a small portion of the variation. Either many other factors also affect self-rated health, or else there's lots of random measurement error in self-rated health scores.

Unfortunately, many researchers put too much emphasis on R^2 as a measure of how "good" their models are. They feel terrific if they get an R^2 of .75, and they feel terrible if the R^2 is only .10. Although it is certainly true that higher is better, there's no reason to reject a model if the R^2 is small. As we see in this example, despite the small R^2 in Equation 1, we still got a clear confirmation that education affects self-rated health. On the other hand, it's quite possible that a model with a high R^2 could be a terrible model: The wrong variables could be in the model, or other assumptions could be violated (more on that in the next chapter).

2.2. Do Mental Patients Who Believe That People Discriminate Against Them Suffer Greater Demoralization?

Based on labeling theory, Link (1987) hypothesized that mental patients who believe that people generally devalue and discriminate against mental patients will experience greater levels of "demoralization." For a sample of 174 mental patients, he administered a 12-item devaluation-discrimination scale that included such statements as "Most people would willingly accept a former mental patient as a close friend." Scores on this scale ranged from 1 to 6, with 6 representing the most devaluation.

At the same time, the patients completed a 27-item scale designed to measure "demoralization." The questions dealt with such themes as low self-esteem, helplessness, hopelessness, confused thinking, and sadness. Several other variables were also measured so that they could be controlled in the regression analysis.

TABLE 2.2 Effect of the Devaluation and Discrimination Measure on Demoralization in First- and Repeat-Contact Patients

	Unstandardized Regression Coefficient	*Standardized Regression Coefficient*	*t Value*	*Significance*
Control variables				
Married (1) vs. unmarried (0)	–.322	–.163	–2.380	.018
Hispanic (1) vs. other (0)	.345	.173	2.515	.013
Hospitalized (1) vs. never hospitalized (0)	.377	.227	3.163	.002
Diagnosis of depression (1) vs. schizophrenia (0)	.486	.303	4.223	.000
Test Variable				
Devaluation-discrimination	.211	.232	3.35	.001
$R^2 = .232$				

SOURCE: Table adapted from Link (1987). Reprinted by permission.
NOTE: $N = 174$.

Table 2.2 gives results for a multiple regression in which demoralization was predicted by devaluation-discrimination and other variables. The table is presented almost exactly in the form it appeared in the *American Sociological Review*. The table divides the independent variables into "Control Variables" and "Test Variable." This merely reflects the author's primary interest in the effect of the devaluation-discrimination variable. The multiple regression procedure itself makes no distinction among the independent variables.

Let's look first at the last column, labeled "Significance." This is the p value for testing the hypothesis that a coefficient is equal to 0. Because all the p values are less than .05, we learn that all the coefficients in the table are statistically significant, some of them highly significant. Reporting the exact p value is considerably more informative than just putting asterisks next to the coefficients.

It's no accident that all the variables in the model are statistically significant. In a footnote, Link explains that he originally put three

other control variables in the regression equation: sex, years of education, and employment status. Because they were not statistically significant, he deleted them from the model and re-estimated the regression equation. This is a common, but not universal, practice among regression analysts. Those who do it argue that it simplifies the results and improves the estimates of the coefficients for the variables that remain in the equation. On the other side are those who claim that letting statistical significance determine which variables stay in the regression equation violates assumptions needed for calculating p values. In most cases, it doesn't make much difference one way or the other.

Link's hypothesis is supported by the positive, statistically significant coefficient for devaluation-discrimination. Those patients who believe more strongly that mental patients are discriminated against do seem to experience greater levels of demoralization. This relationship holds up even when we control for other factors that affect demoralization.

The column labeled "unstandardized regression coefficients" contains the usual regression coefficients. Devaluation-discrimination has a coefficient of .211, which tells us that each 1-unit increase on the devaluation-discrimination scale is associated with an increase of .211 units on the demoralization scale, controlling for the other variables in the model. Again, to get a sense of what this number means, we need to have some appreciation of the units of measurement for these two variables. As already noted, the devaluation-discrimination scale ranges from 1 to 6. Its mean is around 4, and its standard deviation is close to 1. The article doesn't indicate the range of possible values for the demoralization scale, but it does tell us that the mean is approximately 3 and the standard deviation is about .75. That tells us that about 95% of the sample is between 1.5 and 4.5 on this scale. For both these variables, then, a 1-unit change is pretty substantial. With this in mind, we interpret the coefficient of .211 as telling us that a moderately large change in devaluation-discrimination produces a modest, but not small, change in demoralization.

The four control variables in this regression model are all dummy variables, having values of either 0 or 1. The coefficient of –.322 for the dummy variable for marital status indicates that, on average, married patients score .322 points lower on the demoralization scale compared with those who are not married. The coefficients for other dummy variables have a similar interpretation. Hispanics average

.345 points higher on the demoralization scale than non-Hispanics. Those patients who had been hospitalized averaged .377 points higher than those never hospitalized. Finally, patients with a diagnosis of depression averaged .486 points higher on the demoralization scale than those diagnosed as schizophrenic. For any one of these variables, if we changed the 1s to 0s and the 0s to 1s, the sign of the coefficient would change, but the magnitude would be exactly the same.

The standardized coefficient of .232 for devaluation-discrimination tells us that an increase of one standard deviation in devaluation-discrimination produces an increase of .232 standard deviations on the demoralization scale. The standardized coefficient for diagnosis tells us that an increase of one standard deviation in the diagnosis variable produces an increase of .303 standard deviations in demoralization. We can therefore say that, for this sample, diagnosis is slightly more "important" than beliefs about discrimination in accounting for variation in demoralization. The smallest standardized coefficient belongs to marital status, so we can say that this variable is least important among the five variables in the model.

The column labeled t value is the test statistic for the null hypothesis that each coefficient is zero. As noted in Chapter 1, this is calculated by dividing each coefficient by its standard error. The results are then referred to a t table to determine the p values in the last column. Although this table does not report standard errors, they can easily be calculated from the t values. If you divide each (unstandardized) coefficient by its t value, you get the standard error. For devaluation-discrimination, we have .211/3.35 = .063. Multiplying this by 2, and then adding and subtracting the result from the coefficient (.211), produces a 95% confidence interval for the coefficient: .085 to .336. Thus, we can be 95% confident that the true coefficient for devaluation-discrimination lies somewhere between these two numbers.

2.3. Are People in Cities More Racially Tolerant?

In 1987, Tuch published a paper in the *American Sociological Review* which attempted to determine whether people who lived in large cities were less prejudiced toward blacks than people who

TABLE 2.3 Regression of Tolerance Index on Urbanism, Region, and Compositional Variables in Selected Years, 1972-1985 (Unstandardized Coefficients)

	Year			
Independent Variables	*1972 (N = 982)*	*1977 (N = 1,118)*	*1982 (N = 1,063)*	*1985 (N = 553)*
Children (1 = under 18)	−.11	−.08	.00	−.01
Income	.02	.02*	.02*	.00
Sex (1 = male)	−.01	−.03	−.07	−.03
Age	−.01**	−.01**	−.01**	−.01**
Marital status (1 = married)	−.17**	−.06	−.03	.05
Education	.07**	.08**	.07**	.09**
Urbanism (1 = urban)	.20**	.11*	.16**	.36**
Region (1 = non-South)	.53**	.49**	.49**	.30**
R^2	.28	.30	.27	.26

SOURCE: Table adapted from Tuch (1987). Reprinted by permission.
*$p < .05$; **$p < .01$.

lived in smaller towns or rural areas. He used data from several years of the General Social Survey, an annual survey of adults in the United States that is based on a large probability sample. His dependent variable was a tolerance index, which was a weighted sum of responses to five different questions. One question, for example, was "Do you think there should be laws against marriages between blacks and whites?"

Tuch divided the sample into two groups: those living in cities with 50,000 people or more (urban), and those living in cities or towns with less than 50,000 people (nonurban). He found that in 1985, the urbanites had a mean tolerance index of 13.95, whereas the nonurbanites had a mean tolerance index of 12.59, for a statistically significant difference of 1.36. Although that difference supports the hypothesis, we also know that people in urban areas differ in many ways from those in nonurban areas. There are more urban areas in the North than in the South, for example, so the urban-rural difference could merely be an artifact of the historically greater racial prejudice in the South. Tuch therefore did a multiple regression to control for region and many other factors. The results are shown in Table 2.3, which contains the information from the table in his

published article. The title of this table has a common form: "Regression of y on x and z." The convention is that the variable following "of" is the dependent variable, and the variables following "on" are the independent variables.

Table 2.3 displays results from four different regressions predicting the tolerance index. The same variables are in each regression, but they're measured for completely different samples of people in four different years. The reason for doing this is to see whether the results are stable over time.

The main variable of interest—urbanism—is a dummy variable that has a value of 1 for those in urban areas and a value of 0 for those in nonurban areas. There are four other dummy variables in these regression models: children (1 = respondent has children under the age of 18, 0 = no children under 18), sex (1 = male, 0 = female), marital status (1 = married, 0 = unmarried), and region (1 = non-South, 0 =South).

As in Table 2.1, one asterisk indicates a coefficient that is statistically significant at the .05 level and two asterisks indicate a coefficient that is statistically significant at the .01 level. We see that urbanism is highly significant in all four years, with a coefficient ranging from a low of .11 to a high of .36. All four coefficients are positive, which tells us that urban people express more tolerance than nonurban people. If we recoded this variable so that 0 meant urban and 1 meant nonurban, then the four coefficients for urbanism would have the same numerical value, but all would be negative.

What can we learn from the numerical values of these urbanism coefficients? Let's consider the 1985 coefficient of .36. As we saw in the previous example, the coefficient for a dummy variable can be interpreted as an adjusted difference in the mean value of the dependent variable for the two groups, controlling for the other variables in the model. On average, then, urban residents have tolerance scores that are .36 higher than nonurban residents, controlling for the other variables in the model. Earlier we saw that the unadjusted difference in the mean tolerance scores for urban and nonurban residents was 1.36. Now we see that the *adjusted* difference in the mean tolerance scores is only .36, still highly significant but greatly reduced when we control for the other variables. The conclusion is that even when we control for things like region and education, there's still a difference (although much smaller) between the tolerance levels of urban and rural residents.

That's the most important result in the table, but the coefficients for the other variables are also interesting. People outside the South express significantly more racial tolerance than Southerners and, except for 1985, the regional difference is larger than the urban/nonurban difference. In every year, more educated people express greater tolerance: Each additional year of schooling results in an increase of between .07 and .09 points in the tolerance index, depending on the year. Older people are less tolerant than younger people: Each year of age is associated with a decrease of .01 in the tolerance index. The coefficients of the other variables are either never statistically significant, or the results are inconsistent from one year to the next. There's little evidence here for an effect of sex, marital status, income, or children in the home.

A couple of other things are worth noting about this table. First, the R^2s range from .26 to .30, a fairly typical level for data of this sort. Second, unlike the two previous tables, this one reports neither the standard errors nor the t statistics. That's a fairly common way to save space but not a good one, in my opinion. Without standard errors or t statistics, it's impossible for the reader to calculate confidence intervals or exact p values.

2.4. What Determines How Much Time Married Couples Spend Together?

Kingston and Nock (1987) analyzed time diaries kept by 177 dual-earner married couples in 1981. Table 2.4 shows results of regressions in which the dependent variable is the number of minutes the couples spent together per day (excluding sleep time). There are two regression equations, one based on the husband's diary and the other based on the wife's diary. Most of the independent variables are self-explanatory, but two need some clarification. Off scheduling (minutes) is the total number of minutes per day in which one spouse is working while the other is not. Family prestige is an index based on both husband's and wife's occupation and education; it has a possible range from 0 to 100.

As with any statistical table, it's important to read the labels, legends, and footnotes carefully before trying to interpret the results. For example, unlike most tables in sociological journals, two stars after the coefficient mean that the p value is less than .10, not

TABLE 2.4 Regression of Minutes Together on Family Workday, Life Cycle, and Sociocultural Measures (*b*/standardized *b* in parentheses)

Independent Variables	*Husband's Account*	*Wife's Account*
Total minutes of work (husband + wife)	−0.151* (−0.239)	−0.115* (−0.187)
Off scheduling (minutes)	−0.190* (−0.309)	−.225* (−0.374)
Family income (1000s)	0.462 (0.043)	0.317 (0.030)
Family prestige	−1.447 (−0.102)	−2.148 (−0.156)
Respondent race (0 = nonwhite, 1 = white)	68.639 (0.072)	52.739 (0.057)
Number of children in household	−3.891 (−0.032)	−5.439 (−0.046)
Preschoolers? (0 = no, 1 = yes)	−7.959 (−0.022)	8.100 (0.023)
Years in present marriage	2.710** (0.159)	2.530** (0.152)
Constant	374.958	413.016
N	177	177
R^2	.213	.224
Significance (p value)	.000*	.000*

SOURCE: Table adapted from Kingston and Nock (1987). Reprinted by permission.
*$p < .05$; **$p < .10$.

.01. Many analysts would not even consider this to be statistically significant. Another difference is that the numbers in parentheses are not standard errors but standardized coefficients. If you missed this in the title, the fact that some of them are negative is a clear indication that they are not standard errors—standard errors can never be less than 0. Actually, the title is somewhat confusing because (*b*/standardized *b* in parentheses) does *not* mean that *b* is divided by standardized *b*. It simply means that the standardized coefficient lies immediately below the unstandardized coefficient in each case.

For both spouses, only two of the independent variables are statistically significant at the .05 level: total minutes of work and minutes of off scheduling. For the husband's report, each additional minute of total work time reduces the amount of time together by

.15 minutes. Each additional minute of off-schedule time reduces time together by .19 minutes. Based on the wife's report, the coefficient for total work time (−.12) is a little smaller in absolute value than the husband's, whereas the coefficient for off-schedule time (−.23) is a little larger. These small differences between husbands and wives are most likely due to random errors of one sort or another. There's some evidence that couples who've been married longer spend more time together, but the statistical significance of this result is debatable, at best.

It's not terribly surprising that work time and off-schedule time affect time couples spend together. What is surprising is that the other variables show little evidence of any effects. Most people would expect number of children and presence of a preschooler to have some impact. Actually, the coefficients for these variables are not all that small. Each additional child reduces the amount of time together by 4 to 5 minutes (depending on which spouse you listen to). Couples with a preschooler spend either 8 minutes less together (according to the husband) or 8 minutes more (according to the wife). Still, the fact that these coefficients are not statistically significant (even at the .10 level) means that we shouldn't put much stock in their values. There's also a whopping coefficient for race: White couples spend about an hour more together than nonwhite couples. Why isn't this statistically significant? Most likely because, according to the article, only 3% of the sample was nonwhite. That's only 5 couples out of 177, clearly not enough for a reliable comparison.

The last line in the table, labeled "Significance (p value)" is something we haven't seen before. It's the p value for testing the null hypothesis that *all* the coefficients are 0, sometimes referred to as a global test of significance. For both regressions, the p value is listed at .000. This doesn't mean that the p value is *exactly* 0, but it's 0 in the first three decimal places. In any case, a p value this small is statistically significant by anyone's criterion. We conclude, therefore, that the null hypothesis is false, and that *at least one* of the coefficients is not 0.

Why do we need this test, since we already have tests showing that some of the individual coefficients are not 0? When you're testing many coefficients for statistical significance, the p values for those tests can be misleading. If you have a 1 in 20 chance of being wrong on any one of those tests, the chance of making *at least* one mistake is substantially higher than that. The overall or global test

can help protect you against such errors. A good rule of thumb is that if the global test is not significant, then you shouldn't put much stock in the individual tests.

As we've seen in previous tables, the "Constant" or intercept in these regressions is uninteresting. The R^2 of a little over .20 tells us that this set of independent variables accounts for only about 20% of the variation in time spent together.

2.5. Do Male Workers Dislike Working With Women?

Wharton and Baron (1987) studied a national sample of 822 male workers to see if those who worked in all-male job settings were happier about their jobs than those who worked in mixed-sex settings. There were three dependent variables:

- Job satisfaction: a scale based on responses to five questions, such as "All in all, how satisfied would you say you are with your job?" The possible scores ranged from 10 to 50, with a mean of 38.
- Job-related depression: a scale based on responses to four questions about how workers felt on the job, such as "I feel downhearted and blue." The possible scores ranged from 10 to 40, with a mean of 20.
- Job-related self-esteem: a four-item scale asking respondents how they saw themselves in their work: happy/sad, successful/not successful, important/not important, doing their best/not doing their best. Possible scores ranged from 10 to 70, with a mean of 60.

Based on their occupations, the male workers were classified into five different types of work settings:

- All male ($N = 416$)
- Mixed but segregated ($N = 131$)
- Predominantly male ($N = 196$)
- Mixed ($N = 59$)
- Predominantly female ($N = 20$)

A substantial majority of the men worked in settings where they had little or no direct contact with female employees.

How do we handle an independent variable with five categories? As explained in Section 1.2, the trick is to create a dummy variable for each of the categories, except for one category that is called the

reference category. In this study, the reference category was All Male. One dummy variable was created for Mixed but Segregated so the 131 men who fell in this category got a 1 and everyone else got a 0. Another dummy variable was created for Predominantly Male: All 196 men in this category got a 1 and everyone else got a 0. Similar variables were created for Mixed and Predominantly Female.

Because workers and occupations differ in many ways besides sex composition, numerous other variables were included in the regressions as controls. Results for the three regressions are shown in Table 2.5. Notice that the sample size used for each regression is less than 700 even though the original sample had 822 men. That's because any man who had missing data on any of the independent variables was deleted from the sample. There are other ways of handling missing data, but this method (often called *listwise deletion*) is the simplest.

In the note at the bottom of the table, we are told that the coefficients are "metric," which is another way of saying that they are unstandardized rather than standardized. We also learn that a single asterisk means that a coefficient is significant at the .10 level, two asterisks mean significance at the .05 level, and three asterisks mean significant at the .01 level. The significance tests are said to be "two-tailed tests, except for effects pertaining to the mixed and predominantly male categories, which are one-tailed." You may recall from a previous statistics course that a one-tailed test is appropriate if you have a clear-cut expectation for the direction of the difference, whereas a two-tailed test is more appropriate if you have no such expectation. In this case, because the central hypothesis of the study was that men in mixed and predominantly male sex settings would be less happy with their jobs than men in all-male settings, the authors felt justified in choosing a one-tailed test. This is a more lenient test because the p value is obtained by dividing the two-tailed p value by 2. (Computer packages always give two-tailed p values.)

Our main interest is in the effects of the four dummy variables representing sex composition of the work setting. Whenever you have several dummy variables representing a single nominal variable, each coefficient is a comparison with the reference category, which, in this example, is the all-male setting. In the job satisfaction regression, we find a coefficient of –1.65 for predominantly male. This means that the average level of job satisfaction is about 1.65 points lower in predominantly male settings as compared with

TABLE 2.5 OLS Regression of Psychological Well-Being Variables on Gender Mix, Controlling for Individual, Job, and Organizational Characteristics

Independent Variables	*Job Satisfaction (N = 696)*	*Job-Related Depression (N = 695)*	*Job-Related Self-Esteem (N = 692)*
Mixed but segregated setting	.07	.40	.17
Predominantly male setting	−1.65**	1.30***	−1.95***
Mixed setting	−3.28**	1.59*	−3.09***
Predominantly female setting	.90	−2.01	.78
Race	.53	.86	.19
Age	.07	−.04	.07*
Age squared[a]	1.09E-03	−3.29E-03**	2.52E-04
High school graduate	−1.69	.34	−.84
Some college	−.46	.25	.39
College graduate	−.74	.55	−1.19
Unmarried	1.55	−1.11	.58
Married to housewife	3.70**	−1.90*	2.16
Lives in South	1.15	−.61	.85
Training time (SVP)	.55**	−.01	−.10
Danger	−2.27***	1.41***	−1.54**
Physical demands	.27	.16	.92**
Craft union	2.15*	.97	.35
Other union	.70	−.57	−.25
Job tenure	.26	.34	.05
Employer tenure	−.53	.13	.04
Steady	3.96***	.57	.69
Hours	−.03	.03	−.03
Organization scale	−.60***	.04	−.52***
Autonomy	.99**	−.14	1.43***
Authority	2.75***	−.61	2.76***
Log earnings	2.85***	−1.12*	1.78*
Log spouse earnings	.37**	−.15	.17
Status	.02	−9.71E-03	.02
R^2	.173	.072	.121
Adjusted R^2	.138	.033	.084
Standard error of estimate	9.304	5.727	9.269

SOURCE: Table adapted from Wharton and Baron (1987). Reprinted by permission.
NOTE: Metric coefficients are reported.
a. Squared deviation from average age in sample.
*$p \leq .10$; **$p \leq .05$; ***$p \leq .01$ (two-tailed tests, except for effects pertaining to the mixed and predominantly male categories, which are one-tailed).

all-male settings, controlling for other variables in the model. Similarly, the coefficient of −3.28 for mixed setting in the job satisfaction regression means that males working in mixed-sex settings average 3.28 points less on the job satisfaction scale than those working in all-male settings, controlling for other variables. These contrasts are statistically significant at the .05 level. On the other hand, job satisfaction in mixed but segregated settings and predominantly female settings are not significantly different from all-male settings. The results for these four variables are very similar in the job-related self-esteem regression.

There are also significant effects of predominantly male and mixed setting on job-related depression, but the coefficients have positive instead of negative signs. That's because higher values on the depression variable represent "worse" responses, whereas higher values on the satisfaction and self-esteem measures correspond to "better" responses. In all three regressions, the men in the predominantly female occupations are actually more positive about their job experiences than the men in the all-male settings, but the differences are not statistically significant, and it's plausible that the small number of men in these occupations may be a highly unusual group in other respects. In sum, the results seem to confirm the authors' hypothesis that men are happier when they don't have to work with women.

Although the results for the other variables are only incidental to the authors' main goal, we can learn something by examining them. Moving down the table, we see that race (1 = white, 0 = nonwhite) has little importance. Results for age are harder to interpret because "age squared" is also in the regression. This was done to allow for a possible nonlinear effect of age, a strategy we'll discuss in more detail in Chapter 8. The coefficients for the age squared term are reported in scientific notation: E-03 after a coefficient means that the decimal point should be moved three places to the left. The age squared term is not significant in the satisfaction and self-esteem regressions, indicating a lack of evidence for any departure from nonlinearity. This is one case where it would have been desirable to delete the nonsignificant variable from the regression to make the age effect more interpretable.

After the two age variables come three dummy variables representing education: high school graduate, some college, and college graduate. To interpret these coefficients, you have to know that the

omitted category is "did not graduate from high school." Thus, the coefficient of –1.69 for high school graduate in the job satisfaction regression tells us that high school graduates have an average satisfaction score that's about 1.69 points lower than the average score for those who didn't graduate, controlling for other variables. None of the education coefficients is statistically significant, however.

Next come two dummy variables representing three categories of marital status. Here the omitted category is "married to an employed woman." We see that those married to a housewife are significantly more satisfied with their jobs than those whose wife works, and significantly less depressed on the job. Southern residence doesn't have any effect. Training time ("specific vocational preparation") has a positive impact on job satisfaction but not on depression or self-esteem. Jobs involving danger (a dummy variable) are rated substantially worse on all three scales. Jobs involving more physical work actually seem to enhance self-esteem on the job.

The two dummy variables for union membership represent a three-category nominal variable. The reference category is "not a union member." We see that members of craft unions report more satisfaction than nonunion workers, although the difference is only marginally significant. Years on the job (tenure) and years with the employer have little or no effect. Workers who say their job is "steady" are much more satisfied than other workers. Number of hours worked in the preceding week has little or no effect. Workers in large organizations (organizational scale) are less satisfied and have less self-esteem than those in smaller organizations. Those who report more authority and autonomy are more satisfied and have higher self-esteem.

Earnings seems to have a substantial impact on satisfaction and more modest effects on depression and self-esteem, but note that the variable is the *logarithm* of earnings. Entering the variable in this way forces earnings to have diminishing returns: each additional dollar of income has a smaller effect on the dependent variable. We'll discuss this in more detail in Chapter 8. Spouse earnings (again in logarithmic form) has a significant effect on job satisfaction but not on the other two dependent variables.

The R^2s at the bottom of the table indicate only modest predictive power for these regression models, ranging from a low of .072 to a high of .173. Below the R^2 is something we haven't seen before,

called an adjusted R^2. The idea behind this adjustment is that regression models with lots of independent variables have a natural advantage over models with few independent variables in predicting the dependent variable. The adjusted R^2, explained more fully in Chapter 4, removes that advantage.

Just below that is the *standard error of the estimate*. This can be interpreted as the standard deviation of the dependent variable *after effects of the independent variables have been removed*. If we could perfectly predict the dependent variable from the set of independent variables (which would correspond to an R^2 of 1.0), the standard error of the estimate would be 0. On the other hand, when the R^2 is 0 (no predictive power), the standard error of the estimate is the same as the standard deviation of the dependent variable. This statistic is useful mainly in constructing confidence intervals around a predicted value of the dependent variable, which is largely irrelevant for this study. (It is not uncommon for journal articles to report statistics that are rather useless for the reader.)

Chapter Highlights

1. Asterisks after a regression coefficient usually indicate that the coefficient is significantly different from 0. The most common convention is one star for a p value below .05 and two stars for a p value below .01. This is not universal, however.
2. To interpret the numerical value of a regression coefficient, it's essential to understand the metrics of the dependent and independent variables.
3. Coefficients for dummy (0, 1) variables usually can be interpreted as differences in means on the dependent variable for the two categories of the independent variable, controlling for other variables in the regression model.
4. Standardized coefficients can be compared across independent variables with different units of measurement. They tell us how many standard deviations the dependent variable changes for an increase of one standard deviation in the independent variable.
5. The intercept (or constant) in a regression model rarely tells you anything interesting.

6. Don't exaggerate the importance of R^2 in evaluating a regression model. A model can still be worthwhile even if R^2 is low.
7. In computing a multiple regression, there is no distinction among different kinds of independent variables, nor do the results depend on the order which the variables appear in the model.
8. Varying the set of variables in the regression model can be helpful in understanding the causal relationships. If the coefficient for a variable x goes down when other variables are entered, it means that either (a) the other variables *mediate* the effect of x on the dependent variable y, or (b) the other variables affect both x and y and therefore the original coefficient of x was partly spurious.
9. If a dummy independent variable has nearly all the observations in one of its two categories, even large effects of this variable may not show up as statistically significant.
10. If the global F test for all the variables in a multiple regression model is not statistically significant, you should be very cautious about drawing any additional conclusions about the variables and their effects.
11. When multiple dummy variables are used to represent a categorical variable with more than two categories, it's crucial in interpreting the results to know which is the "omitted" category.

Questions to Think About

1. In a large organization, the average salary for men is $47,000 while the average salary for women is $30,000. A t test shows that this difference is significant at beyond the .001 level. When salary is regressed on years of work experience and a dummy variable for gender (male = 1, female = 0), the coefficient for sex is about $1,500 with a p value of .17. What would you conclude?
2. Dr. Smith regresses a measure of "happiness" on age, gender (male = 1, female = 0), and marital status (married = 1, unmarried = 0). All the coefficients are statistically significant at beyond the .05 level. Dr. Smith finds that the coefficient for gender is about five times as large as the coefficient for age. Can he legitimately conclude that gender is more important than age in predicting happiness? He also finds that the coefficient for gender is twice as large as the coefficient for marital status. What does this mean, if anything?
3. Hours per day spent watching TV is regressed on age (in years) and gender (female = 1, male = 0). The estimated intercept is 1.5. What does this number tell you, if anything?
4. The self-reported health study in Section 2.1 uses a dependent variable that is measured on a 5-point ordinal scale. Are there any potential problems with this?

TABLE 2.6 Regressions Predicting IQ Scores

	Model 1	*Model 2*
Brain size	.119*	.200**
	(.050)	(.066)
Gender (female = 1, male = 0)		−2.599
		(11.176)
Height (in.)		−2.767
		(1.447)
Weight (lbs.)		−.075
		(.220)
Constant	5.167	134.383
R^2	.13	.27
N	40	38

SOURCE: Table adapted from Willerman, Schultz, Rutledge, and Bigler (1987). Reprinted by permission.
NOTE: Standard errors are in parentheses.
*$p < .05$; **$p < .01$.

The remaining questions pertain to the regression table shown as Table 2.6, based on data reported in Willerman, Schultz, Rutledge, and Bigler (1991). Brain size was measured by an MRI scan. It was a count of the number of "pixels" (smallest unit of uniform shading) in the scan.

5. What conclusions, if any, would you draw about the relationship between brain size and intelligence?
6. How would you interpret the coefficients for gender and weight in Model 2? Are these coefficients worth paying attention to? Why or why not?
7. What's the point of having two different models? What can we learn from comparing them?

8. Compute a t statistic for the height coefficient. Then find its p value.
9. Why does the constant (intercept) change so dramatically from Model 1 to Model 2?
10. Model 2 has a larger R^2 than Model 1. Does this mean it's a better model? Why or why not?

3 What Can Go Wrong With Multiple Regression?

Any tool as widely used as multiple regression is bound to be frequently *mis*used. Nowadays, statistical packages are so user-friendly that anyone can perform a multiple regression with a few mouse clicks. As a result, many researchers apply multiple regression to their data with little understanding of the underlying assumptions or the possible pitfalls. Although the review process for scientific journals is supposed to weed out papers with incorrect or misleading statistical methods, it often happens that the referees themselves have insufficient statistical expertise or are simply too rushed to catch the more subtle errors. The upshot is that you need to cast a critical eye on the results of any multiple regression, especially those you run yourself.

Fortunately, the questions that you need to ask are neither extremely technical nor large in number. They do require careful thought, however, which explains why even experts occasionally make mistakes or overlook the obvious. Virtually all the questions have to do with situations where multiple regression is used to make causal inferences. The requirements for a model used only for predictions are much less severe. I'll start with the more important questions and progress to those of lesser importance. Each question will be illustrated with examples drawn from Chapter 2.

3.1. Are Important Independent Variables Left Out of the Model?

In evaluating any regression model, it's just as important to think about what's *not* in the model as what's in it. There are two possible reasons for putting a variable in a regression model: