CHAPTER 7

Multiple Regression Analysis with Qualitative Information: Binary (or Dummy) Variables

In previous chapters, the dependent and independent variables in our multiple regression models have had *quantitative* meaning. Just a few examples include hourly wage rate, years of education, college grade point average, amount of air pollution, level of firm sales, and number of arrests. In each case, the magnitude of the variable conveys useful information. In empirical work, we must also incorporate *qualitative* factors into regression models. The gender or race of an individual, the industry of a firm (manufacturing, retail, and so on), and the region in the United States where a city is located (South, North, West, and so on) are all considered to be qualitative factors.

Most of this chapter is dedicated to qualitative *independent* variables. After we discuss the appropriate ways to describe qualitative information in Section 7-1, we show how qualitative explanatory variables can be easily incorporated into multiple regression models in Sections 7-2, 7-3, and 7-4. These sections cover almost all of the popular ways that qualitative independent variables are used in cross-sectional regression analysis.

In Section 7-5, we discuss a binary dependent variable, which is a particular kind of qualitative dependent variable. The multiple regression model has an interesting interpretation in this case and is called the linear probability model. While much maligned by some econometricians, the simplicity of the linear probability model makes it useful in many empirical contexts. We will describe its drawbacks in Section 7-5, but they are often secondary in empirical work.

7-1 Describing Qualitative Information

Qualitative factors often come in the form of binary information: a person is female or male; a person does or does not own a personal computer; a firm offers a certain kind of employee pension plan or it does not; a state administers capital punishment or it does not. In all of these examples, the relevant

TABLE 7.1 A Partial Listing of the Data in WAGE1

person	*wage*	*educ*	*exper*	*female*	*married*
1	3.10	11	2	1	0
2	3.24	12	22	1	1
3	3.00	11	2	0	0
4	6.00	8	44	0	1
5	5.30	12	7	0	1
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
525	11.56	16	5	0	1
526	3.50	14	5	1	0

information can be captured by defining a **binary variable** or a **zero-one variable**. In econometrics, binary variables are most commonly called **dummy variables**, although this name is not especially descriptive.

EXPLORING FURTHER 7.1

Suppose that, in a study comparing election outcomes between Democratic and Republican candidates, you wish to indicate the party of each candidate. Is a name such as *party* a wise choice for a binary variable in this case? What would be a better name?

In defining a dummy variable, we must decide which event is assigned the value one and which is assigned the value zero. For example, in a study of individual wage determination, we might define *female* to be a binary variable taking on the value one for females and the value zero for males. The name in this case indicates the event with the value one. The same information is captured by defining *male* to be one if the person is male and zero if the person is female. Either of these is better than using *gender* because this name does not make it clear when the dummy variable is one: does $gender = 1$ correspond to male or female? What we call our variables is unimportant for getting regression results, but it always helps to choose names that clarify equations and expositions.

Suppose in the wage example that we have chosen the name *female* to indicate gender. Further, we define a binary variable *married* to equal one if a person is married and zero if otherwise. Table 7.1 gives a partial listing of a wage data set that might result. We see that Person 1 is female and not married, Person 2 is female and married, Person 3 is male and not married, and so on.

Why do we use the values zero and one to describe qualitative information? In a sense, these values are arbitrary: any two different values would do. The real benefit of capturing qualitative information using zero-one variables is that it leads to regression models where the parameters have very natural interpretations, as we will see now.

7-2 A Single Dummy Independent Variable

How do we incorporate binary information into regression models? In the simplest case, with only a single dummy explanatory variable, we just add it as an independent variable in the equation. For example, consider the following simple model of hourly wage determination:

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + u. \qquad [7.1]$$

We use δ_0 as the parameter on *female* in order to highlight the interpretation of the parameters multiplying dummy variables; later, we will use whatever notation is most convenient.

In model (7.1), only two observed factors affect wage: gender and education. Because $female = 1$ when the person is female, and $female = 0$ when the person is male, the parameter δ_0 has the following interpretation: δ_0 is the difference in hourly wage between females and males, *given* the same amount of education (and the same error term u). Thus, the coefficient δ_0 determines whether there is discrimination against women: if $\delta_0 < 0$, then for the same level of other factors, women earn less than men on average.

In terms of expectations, if we assume the zero conditional mean assumption $\mathrm{E}(u|female, educ) = 0$, then

$$\delta_0 = \mathrm{E}(wage|female = 1, educ) - \mathrm{E}(wage|female = 0, educ).$$

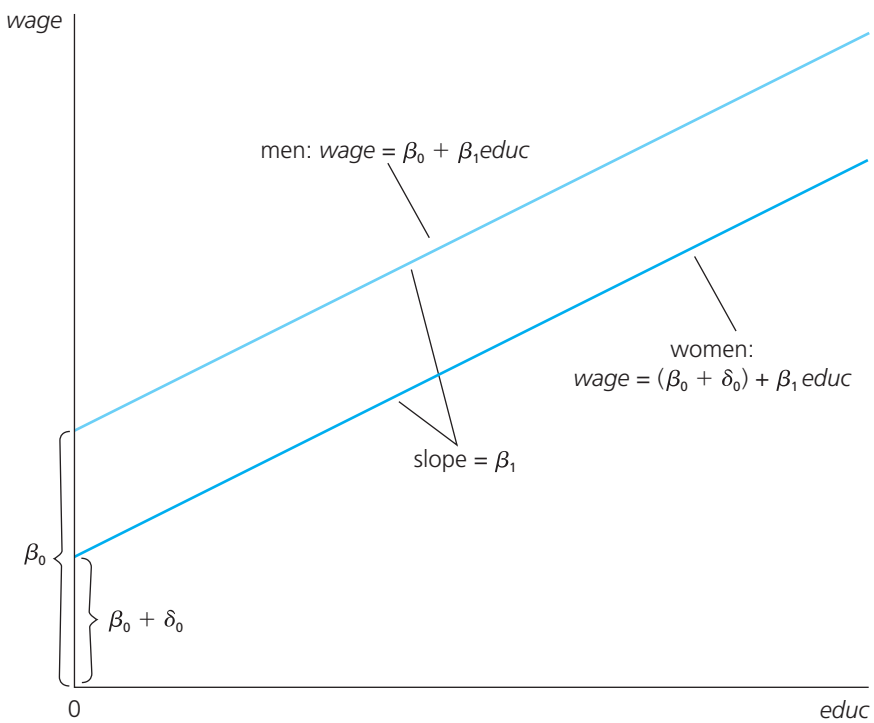
Because $female = 1$ corresponds to females and $female = 0$ corresponds to males, we can write this more simply as

$$\delta_0 = \mathrm{E}(wage|female, educ) - \mathrm{E}(wage|male, educ). \qquad [7.2]$$

The key here is that the level of education is the same in both expectations; the difference, δ_0, is due to gender only.

The situation can be depicted graphically as an **intercept shift** between males and females. In Figure 7.1, the case $\delta_0 < 0$ is shown, so that men earn a fixed amount more per hour than women. The difference does not depend on the amount of education, and this explains why the wage-education profiles for women and men are parallel.

FIGURE 7.1 Graph of $wage = \beta_0 + \delta_0 female + \beta_1 educ$ for $\delta_0 < 0$.



At this point, you may wonder why we do not also include in (7.1) a dummy variable, say *male*, which is one for males and zero for females. This would be redundant. In (7.1), the intercept for males is β_0, and the intercept for females is $\beta_0 + \delta_0$. Because there are just two groups, we only need two different intercepts. This means that, in addition to β_0, we need to use only *one* dummy variable; we have chosen to include the dummy variable for females. Using two dummy variables would introduce perfect collinearity because $female + male = 1$, which means that *male* is a perfect linear function of *female*. Including dummy variables for both genders is the simplest example of the so-called **dummy variable trap**, which arises when too many dummy variables describe a given number of groups. We will discuss this problem in detail later.

In (7.1), we have chosen males to be the **base group** or **benchmark group**, that is, the group against which comparisons are made. This is why β_0 is the intercept for males, and δ_0 is the *difference* in intercepts between females and males. We could choose females as the base group by writing the model as

$$wage = \alpha_0 + \gamma_0 male + \beta_1 educ + u,$$

where the intercept for females is α_0 and the intercept for males is $\alpha_0 + \gamma_0$; this implies that $\alpha_0 = \beta_0 + \delta_0$ and $\alpha_0 + \gamma_0 = \beta_0$. In any application, it does not matter how we choose the base group, but it is important to keep track of which group is the base group.

Some researchers prefer to drop the overall intercept in the model and to include dummy variables for each group. The equation would then be $wage = \beta_0 male + \alpha_0 female + \beta_1 educ + u$, where the intercept for men is β_0 and the intercept for women is α_0. There is no dummy variable trap in this case because we do not have an overall intercept. However, this formulation has little to offer, since testing for a difference in the intercepts is more difficult, and there is no generally agreed upon way to compute R-squared in regressions without an intercept. Therefore, we will always include an overall intercept for the base group.

Nothing much changes when more explanatory variables are involved. Taking males as the base group, a model that controls for experience and tenure in addition to education is

$$wage = \beta_0 + \delta_0 female + \beta_1 educ + \beta_2 exper + \beta_3 tenure + u. \qquad [7.3]$$

If *educ*, *exper*, and *tenure* are all relevant productivity characteristics, the null hypothesis of *no* difference between men and women is H_0: $\delta_0 = 0$. The alternative that there is discrimination against women is H_1: $\delta_0 < 0$.

How can we actually test for wage discrimination? The answer is simple: just estimate the model by OLS, *exactly* as before, and use the usual t statistic. Nothing changes about the mechanics of OLS or the statistical theory when some of the independent variables are defined as dummy variables. The only difference with what we have done up until now is in the interpretation of the coefficient on the dummy variable.

EXAMPLE 7.1 Hourly Wage Equation

Using the data in WAGE1, we estimate model (7.3). For now, we use *wage*, rather than log(*wage*), as the dependent variable:

$$\widehat{wage} = -1.57 - 1.81\, female + .572\, educ + 0.25\, exper + .141\, tenure \qquad [7.4]$$
$$(.72) \quad (.26) \qquad (.049) \qquad (.012) \qquad (.021)$$
$$n = 526, R^2 = .364.$$

The negative intercept—the intercept for men, in this case—is not very meaningful because no one has zero values for all of *educ*, *exper*, and *tenure* in the sample. The coefficient on *female* is interesting because it measures the average difference in hourly wage between a man and a woman who have the *same* levels of *educ*, *exper*, and *tenure*. If we take a woman and a man with the same levels of

education, experience, and tenure, the woman earns, on average, \$1.81 less per hour than the man. (Recall that these are 1976 wages.)

It is important to remember that, because we have performed multiple regression and controlled for *educ*, *exper*, and *tenure*, the \$1.81 wage differential cannot be explained by different average levels of education, experience, or tenure between men and women. We can conclude that the differential of \$1.81 is due to gender or factors associated with gender that we have not controlled for in the regression. [In 2013 dollars, the wage differential is about $4.09(1.81) \approx 7.40$.]

It is informative to compare the coefficient on *female* in equation (7.4) to the estimate we get when all other explanatory variables are dropped from the equation:

$$\widehat{wage} = 7.10 - 2.51\, female$$
$$(.21) \quad (.30) \qquad [7.5]$$
$$n = 526,\ R^2 = .116.$$

The coefficients in (7.5) have a simple interpretation. The intercept is the average wage for men in the sample (let $female = 0$), so men earn \$7.10 per hour on average. The coefficient on *female* is the difference in the average wage between women and men. Thus, the average wage for women in the sample is $7.10 - 2.51 = 4.59$, or \$4.59 per hour. (Incidentally, there are 274 men and 252 women in the sample.)

Equation (7.5) provides a simple way to carry out a *comparison-of-means test* between the two groups, which in this case are men and women. The estimated difference, -2.51, has a t statistic of -8.37, which is very statistically significant (and, of course, \$2.51 is economically large as well). Generally, simple regression on a constant and a dummy variable is a straightforward way to compare the means of two groups. For the usual t test to be valid, we must assume that the homoskedasticity assumption holds, which means that the population variance in wages for men is the same as that for women.

The estimated wage differential between men and women is larger in (7.5) than in (7.4) because (7.5) does not control for differences in education, experience, and tenure, and these are lower, on average, for women than for men in this sample. Equation (7.4) gives a more reliable estimate of the ceteris paribus gender wage gap; it still indicates a very large differential.

In many cases, dummy independent variables reflect choices of individuals or other economic units (as opposed to something predetermined, such as gender). In such situations, the matter of causality is again a central issue. In the following example, we would like to know whether personal computer ownership *causes* a higher college grade point average.

EXAMPLE 7.2 Effects of Computer Ownership on College GPA

In order to determine the effects of computer ownership on college grade point average, we estimate the model

$$colGPA = \beta_0 + \delta_0 PC + \beta_1 hsGPA + \beta_2 ACT + u,$$

where the dummy variable *PC* equals one if a student owns a personal computer and zero otherwise. There are various reasons PC ownership might have an effect on *colGPA*. A student's schoolwork might be of higher quality if it is done on a computer, and time can be saved by not having to wait at a computer lab. Of course, a student might be more inclined to play computer games or surf the Internet if he or she owns a PC, so it is not obvious that δ_0 is positive. The variables *hsGPA* (high school GPA) and *ACT* (achievement test score) are used as controls: it could be that stronger students, as measured

by high school *GPA* and *ACT* scores, are more likely to own computers. We control for these factors because we would like to know the average effect on *colGPA* if a student is picked at random and given a personal computer.

Using the data in GPA1, we obtain

$$\widehat{colGPA} = 1.26 + .157\,PC + .447\,hsGPA + .0087\,ACT \quad [7.6]$$
$$(.33)\ \ (.057)\qquad (.094)\qquad\ (.0105)$$
$$n = 141,\ R^2 = .219.$$

This equation implies that a student who owns a PC has a predicted GPA about .16 points higher than a comparable student without a PC (remember, both *colGPA* and *hsGPA* are on a four-point scale). The effect is also very statistically significant, with $t_{PC} = .157/.057 \approx 2.75$.

What happens if we drop *hsGPA* and *ACT* from the equation? Clearly, dropping the latter variable should have very little effect, as its coefficient and *t* statistic are very small. But *hsGPA* is very significant, and so dropping it could affect the estimate of β_{PC}. Regressing *colGPA* on *PC* gives an estimate on *PC* equal to about .170, with a standard error of .063; in this case, $\hat{\beta}_{PC}$ and its *t* statistic do not change by much.

In the exercises at the end of the chapter, you will be asked to control for other factors in the equation to see if the computer ownership effect disappears, or if it at least gets notably smaller.

Each of the previous examples can be viewed as having relevance for **policy analysis**. In the first example, we were interested in gender discrimination in the workforce. In the second example, we were concerned with the effect of computer ownership on college performance. A special case of policy analysis is **program evaluation**, where we would like to know the effect of economic or social programs on individuals, firms, neighborhoods, cities, and so on.

In the simplest case, there are two groups of subjects. The **control group** does not participate in the program. The **experimental group** or **treatment group** does take part in the program. These names come from literature in the experimental sciences, and they should not be taken literally. Except in rare cases, the choice of the control and treatment groups is not random. However, in some cases, multiple regression analysis can be used to control for enough other factors in order to estimate the causal effect of the program.

EXAMPLE 7.3 Effects of Training Grants on Hours of Training

Using the 1988 data for Michigan manufacturing firms in JTRAIN, we obtain the following estimated equation:

$$\widehat{hrsemp} = 46.67 + 26.25\,grant - .98\log(sales) - 6.07\log(employ) \quad [7.7]$$
$$(43.41)\ \ (5.59)\qquad (3.54)\qquad\ (3.88)$$
$$n = 105,\ R^2 = .237.$$

The dependent variable is hours of training per employee, at the firm level. The variable *grant* is a dummy variable equal to one if the firm received a job training grant for 1988, and zero otherwise. The variables *sales* and *employ* represent annual sales and number of employees, respectively. We cannot enter *hrsemp* in logarithmic form because *hrsemp* is zero for 29 of the 105 firms used in the regression.

The variable *grant* is very statistically significant, with $t_{grant} = 4.70$. Controlling for sales and employment, firms that received a grant trained each worker, on average, 26.25 hours more. Because

the average number of hours of per worker training in the sample is about 17, with a maximum value of 164, *grant* has a large effect on training, as is expected.

The coefficient on log(*sales*) is small and very insignificant. The coefficient on log(*employ*) means that, if a firm is 10% larger, it trains its workers about .61 hour less. Its t statistic is -1.56, which is only marginally statistically significant.

As with any other independent variable, we should ask whether the measured effect of a qualitative variable is causal. In equation (7.7), is the difference in training between firms that receive grants and those that do not due to the grant, or is grant receipt simply an indicator of something else? It might be that the firms receiving grants would have, on average, trained their workers more even in the absence of a grant. Nothing in this analysis tells us whether we have estimated a causal effect; we must know how the firms receiving grants were determined. We can only hope we have controlled for as many factors as possible that might be related to whether a firm received a grant and to its levels of training.

We will return to policy analysis with dummy variables in Section 7-6, as well as in later chapters.

7-2a Interpreting Coefficients on Dummy Explanatory Variables When the Dependent Variable Is log(*y*)

A common specification in applied work has the dependent variable appearing in logarithmic form, with one or more dummy variables appearing as independent variables. How do we interpret the dummy variable coefficients in this case? Not surprisingly, the coefficients have a *percentage* interpretation.

EXAMPLE 7.4 **Housing Price Regression**

Using the data in HPRICE1, we obtain the equation

$$\begin{aligned}\widehat{\log(price)} &= -1.35 + .168\log(lotsize) + .707\log(sqrft) \\ &\qquad (.65)\ \ (.038)\qquad\qquad (.093) \\ &\quad + .027\,bdrms + .054\,colonial \qquad [7.8] \\ &\qquad (.029)\qquad\ (.045) \\ n &= 88,\ R^2 = .649.\end{aligned}$$

All the variables are self-explanatory except colonial, which is a binary variable equal to one if the house is of the colonial style. What does the coefficient on *colonial* mean? For given levels of *lotsize*, *sqrft*, and *bdrms*, the difference in $\widehat{\log(price)}$ between a house of colonial style and that of another style is .054. This means that a colonial-style house is predicted to sell for about 5.4% more, holding other factors fixed.

This example shows that, when log(y) is the dependent variable in a model, the coefficient on a dummy variable, when multiplied by 100, is interpreted as the percentage difference in y, holding all other factors fixed. When the coefficient on a dummy variable suggests a large proportionate change in y, the exact percentage difference can be obtained exactly as with the semi-elasticity calculation in Section 6-2.

EXAMPLE 7.5 **Log Hourly Wage Equation**

Let us reestimate the wage equation from Example 7.1, using log(*wage*) as the dependent variable and adding quadratics in *exper* and *tenure*:

$$\begin{aligned}\widehat{\log(wage)} &= .417 - .297\,female + .080\,educ + .029\,exper \\ &\qquad (.099)\ \ (.036)\qquad\ (.007)\qquad (.005) \\ &\quad - .00058\,exper^2 + .032\,tenure - .00059\,tenure^2 \\ &\qquad (.00010)\qquad\ (.007)\qquad\ (.00023) \\ n &= 526,\ R^2 = .441.\end{aligned} \tag{7.9}$$

Using the same approximation as in Example 7.4, the coefficient on female implies that, for the same levels of *educ*, *exper*, and *tenure*, women earn about 100(.297) = 29.7% less than men. We can do better than this by computing the exact percentage difference in predicted wages. What we want is the proportionate difference in wages between females and males, holding other factors fixed: $(\widehat{wage_F} - \widehat{wage_M})/\widehat{wage_M}$. What we have from (7.9) is

$$\widehat{\log(wage_F)} - \widehat{\log(wage_M)} = -.297.$$

Exponentiating and subtracting one gives

$$(\widehat{wage_F} - \widehat{wage_M})/\widehat{wage_M} = \exp(-.297) - 1 \approx -.257.$$

This more accurate estimate implies that a woman's wage is, on average, 25.7% below a comparable man's wage.

If we had made the same correction in Example 7.4, we would have obtained $\exp(.054) - 1 \approx .0555$, or about 5.6%. The correction has a smaller effect in Example 7.4 than in the wage example because the magnitude of the coefficient on the dummy variable is much smaller in (7.8) than in (7.9).

Generally, if $\hat{\beta}_1$ is the coefficient on a dummy variable, say x_1, when log(y) is the dependent variable, the exact percentage difference in the predicted y when $x_1 = 1$ versus when $x_1 = 0$ is

$$100 \cdot [\exp(\hat{\beta}_1) - 1]. \tag{7.10}$$

The estimate $\hat{\beta}_1$ can be positive or negative, and it is important to preserve its sign in computing (7.10).

The logarithmic approximation has the advantage of providing an estimate between the magnitudes obtained by using each group as the base group. In particular, although equation (7.10) gives us a better estimate than $100 \cdot \hat{\beta}_1$ of the percentage by which y for $x_1 = 1$ is greater than y for $x_1 = 0$, (7.10) is not a good estimate if we switch the base group. In Example 7.5, we can estimate the percentage by which a man's wage exceeds a comparable woman's wage, and this estimate is $100 \cdot [\exp(-\hat{\beta}_1) - 1] = 100 \cdot [\exp(.297) - 1] \approx 34.6$. The approximation, based on $100 \cdot \hat{\beta}_1$, 29.7, is between 25.7 and 34.6 (and close to the middle). Therefore, it makes sense to report that "the difference in predicted wages between men and women is about 29.7%," without having to take a stand on which is the base group.

7-3 Using Dummy Variables for Multiple Categories

We can use several dummy independent variables in the same equation. For example, we could add the dummy variable *married* to equation (7.9). The coefficient on *married* gives the (approximate) proportional differential in wages between those who are and are not married, holding gender, *educ*, *exper*, and *tenure* fixed. When we estimate this model, the coefficient on *married* (with standard error

in parentheses) is .053 (.041), and the coefficient on *female* becomes $-.290(.036)$. Thus, the "marriage premium" is estimated to be about 5.3%, but it is not statistically different from zero ($t = 1.29$). An important limitation of this model is that the marriage premium is assumed to be the same for men and women; this is relaxed in the following example.

EXAMPLE 7.6 Log Hourly Wage Equation

Let us estimate a model that allows for wage differences among four groups: married men, married women, single men, and single women. To do this, we must select a base group; we choose single men. Then, we must define dummy variables for each of the remaining groups. Call these *marrmale*, *marrfem*, and *singfem*. Putting these three variables into (7.9) (and, of course, dropping *female*, since it is now redundant) gives

$$\begin{aligned}\widehat{\log(wage)} = {} & .321 + .213\,marrmale - .198\,marrfem \\ & \quad (.100)\ (.055)\qquad (.058) \\ & - .110\,singfem + .079\,educ + .027\,exper - .00054\,exper^2 \\ & \quad (.056)\qquad (.007)\qquad (.005)\qquad (.00011) \\ & + .029\,tenure - .00053\,tenure^2 \\ & \quad (.007)\qquad (.00023) \\ n = 526,\ & R^2 = .461.\end{aligned} \tag{7.11}$$

All of the coefficients, with the exception of *singfem*, have *t* statistics well above two in absolute value. The *t* statistic for *singfem* is about -1.96, which is just significant at the 5% level against a two-sided alternative.

To interpret the coefficients on the dummy variables, we must remember that the base group is single males. Thus, the estimates on the three dummy variables measure the proportionate difference in wage *relative* to single males. For example, married men are estimated to earn about 21.3% more than single men, holding levels of education, experience, and tenure fixed. [The more precise estimate from (7.10) is about 23.7%.] A married woman, on the other hand, earns a predicted 19.8% less than a single man with the same levels of the other variables.

Because the base group is represented by the intercept in (7.11), we have included dummy variables for only three of the four groups. If we were to add a dummy variable for single males to (7.11), we would fall into the dummy variable trap by introducing perfect collinearity. Some regression packages will automatically correct this mistake for you, while others will just tell you there is perfect collinearity. It is best to carefully specify the dummy variables because then we are forced to properly interpret the final model.

Even though single men is the base group in (7.11), we can use this equation to obtain the estimated difference between any two groups. Because the overall intercept is common to all groups, we can ignore that in finding differences. Thus, the estimated proportionate difference between single and married women is $-.110 - (-.198) = .088$, which means that single women earn about 8.8% more than married women. Unfortunately, we cannot use equation (7.11) for testing whether the estimated difference between single and married women is statistically significant. Knowing the standard errors on *marrfem* and *singfem* is not enough to carry out the test (see Section 4-4). The easiest thing to do is to choose one of these groups to be the base group and to reestimate the equation. Nothing substantive changes, but we get the needed estimate and its standard error directly. When we use married women as the base group, we obtain

$$\begin{aligned}\widehat{\log(wage)} = {} & .123 + .411\,marrmale + .198\,singmale + .088\,singfem + \ldots, \\ & \quad (.106)\ (.056)\qquad (.058)\qquad (.052)\end{aligned}$$

where, of course, none of the unreported coefficients or standard errors have changed. The estimate on *singfem* is, as expected, .088. Now, we have a standard error to go along with this estimate. The t statistic for the null that there is no difference in the population between married and single women is $t_{singfem} = .088/.052 \approx 1.69$. This is marginal evidence against the null hypothesis. We also see that the estimated difference between married men and married women is very statistically significant $(t_{marrmale} = 7.34)$.

The previous example illustrates a general principle for including dummy variables to indicate different groups: if the regression model is to have different intercepts for, say, g groups or categories, we need to include $g - 1$ dummy variables in the model along with an intercept. The intercept for the base group is the overall intercept in the model, and the dummy variable coefficient for a particular group represents the estimated difference in intercepts between that group and the base group. Including g dummy variables along with an intercept will result in the dummy variable trap. An alternative is to include g dummy variables and to exclude an overall intercept. Including g dummies without an overall intercept is sometimes useful, but it has two practical drawbacks. First, it makes it more cumbersome to test for differences relative to a base group. Second, regression packages usually change the way R-squared is computed when an overall intercept is not included. In particular, in the formula $R^2 = 1 - \text{SSR}/\text{SST}$, the total sum of squares, SST, is replaced with a total sum of squares that does not center y_i about its mean, say, $\text{SST}_0 = \sum_{i=1}^{n} y_i^2$. The resulting R-squared, say $R_0^2 = 1 - \text{SSR}/\text{SST}_0$, is sometimes called the **uncentered *R*-squared**. Unfortunately, R_0^2 is rarely suitable as a goodness-of-fit measure. It is always true that $\text{SST}_0 \geq \text{SST}$ with equality only if $\bar{y} = 0$. Often, SST_0 is much larger that SST, which means that R_0^2 is much larger than R^2. For example, if in the previous example we regress log(*wage*) on *marrmale*, *singmale*, *marrfem*, *singfem*, and the other explanatory variables—without an intercept—the reported R-squared from Stata, which is R_0^2, is .948. This high R-squared is an artifact of not centering the total sum of squares in the calculation. The correct R-squared is given in equation (7.11) as .461. Some regression packages, including Stata, have an option to force calculation of the centered R-squared even though an overall intercept has not been included, and using this option is generally a good idea. In the vast majority of cases, any R-squared based on comparing an SSR and SST should have SST computed by centering the y_i about $\bar{y}$. We can think of this SST as the sum of squared residuals obtained if we just use the sample average, $\bar{y}$, to predict each y_i. Surely we are setting the bar pretty low for any model if all we measure is its fit relative to using a constant predictor. For a model without an intercept that fits poorly, it is possible that $\text{SSR} > \text{SST}$, which means R^2 would be negative. The uncentered R-squared will always be between zero and one, which likely explains why it is usually the default when an intercept is not estimated in regression models.

EXPLORING FURTHER 7.2

In the baseball salary data found in MLB1, players are given one of six positions: *frstbase, scndbase, thrdbase, shrtstop, outfield,* or *catcher*. To allow for salary differentials across position, with outfielders as the base group, which dummy variables would you include as independent variables?

7-3a Incorporating Ordinal Information by Using Dummy Variables

Suppose that we would like to estimate the effect of city credit ratings on the municipal bond interest rate (*MBR*). Several financial companies, such as Moody's Investors Service and Standard and Poor's, rate the quality of debt for local governments, where the ratings depend on things like probability of default. (Local governments prefer lower interest rates in order to reduce their costs of borrowing.) For simplicity, suppose that rankings take on the integer values $\{0, 1, 2, 3, 4\}$, with zero being the worst credit rating and four being the best. This is an example of an **ordinal variable**. Call this

variable *CR* for concreteness. The question we need to address is: How do we incorporate the variable *CR* into a model to explain *MBR*?

One possibility is to just include *CR* as we would include any other explanatory variable:

$$MBR = \beta_0 + \beta_1 CR + other\ factors,$$

where we do not explicitly show what other factors are in the model. Then β_1 is the percentage point change in *MBR* when *CR* increases by one unit, holding other factors fixed. Unfortunately, it is rather hard to interpret a one-unit increase in *CR*. We know the quantitative meaning of another year of education, or another dollar spent per student, but things like credit ratings typically have only ordinal meaning. We know that a *CR* of four is better than a *CR* of three, but is the difference between four and three the same as the difference between one and zero? If not, then it might not make sense to assume that a one-unit increase in *CR* has a constant effect on *MBR*.

A better approach, which we can implement because *CR* takes on relatively few values, is to define dummy variables for each value of *CR*. Thus, let $CR_1 = 1$ if $CR = 1$, and $CR_1 = 0$ otherwise; $CR_2 = 1$ if $CR = 2$, and $CR_2 = 0$ otherwise; and so on. Effectively, we take the single credit rating and turn it into five categories. Then, we can estimate the model

$$MBR = \beta_0 + \delta_1 CR_1 + \delta_2 CR_2 + \delta_3 CR_3 + \delta_4 CR_4 + other\ factors. \qquad [7.12]$$

Following our rule for including dummy variables in a model, we include four dummy variables because we have five categories. The omitted category here is a credit rating of zero, and so it is the base group. (This is why we do not need to define a dummy variable for this category.) The coefficients are easy to interpret: δ_1 is the difference in *MBR* (other factors fixed) between a municipality with a credit rating of one and a municipality with a credit rating of zero; δ_2 is the difference in *MBR* between a municipality with a credit rating of two and a municipality with a credit rating of zero; and so on. The movement between each credit rating is allowed to have a different effect, so using (7.12) is much more flexible than simply putting *CR* in as a single variable. Once the dummy variables are defined, estimating (7.12) is straightforward.

EXPLORING FURTHER 7.3

In model (7.12), how would you test the null hypothesis that credit rating has no effect on *MBR*?

Equation (7.12) contains the model with a constant partial effect as a special case. One way to write the three restrictions that imply a constant partial effect is $\delta_2 = 2\delta_1$, $\delta_3 = 3\delta_1$, and $\delta_4 = 4\delta_1$. When we plug these into equation (7.12) and rearrange, we get $MBR = \beta_0 + \delta_1(CR_1 + 2CR_2 + 3CR_3 + 4CR_4) + other\ factors$. Now, the term multiplying δ_1 is simply the original credit rating variable, *CR*. To obtain the *F* statistic for testing the constant partial effect restrictions, we obtain the unrestricted *R*-squared from (7.12) and the restricted *R*-squared from the regression of *MBR* on *CR* and the other factors we have controlled for. The *F* statistic is obtained as in equation (4.41) with $q = 3$.

EXAMPLE 7.7 Effects of Physical Attractiveness on Wage

Hamermesh and Biddle (1994) used measures of physical attractiveness in a wage equation. (The file BEAUTY contains fewer variables but more observations than used by Hamermesh and Biddle. See Computer Exercise C12.) Each person in the sample was ranked by an interviewer for physical attractiveness, using five categories (homely, quite plain, average, good looking, and strikingly beautiful or handsome). Because there are so few people at the two extremes, the authors put people into one of three groups for the regression analysis: average, below average, and above average, where the base group is *average*. Using data from the 1977 Quality of Employment Survey, after

controlling for the usual productivity characteristics, Hamermesh and Biddle estimated an equation for men:

$$\widehat{\log(wage)} = \hat{\beta}_0 - \underset{(.046)}{.164}\, belavg + \underset{(.033)}{.016}\, abvavg + other\ factors$$
$$n = 700, \bar{R}^2 = .403$$

and an equation for women:

$$\widehat{\log(wage)} = \hat{\beta}_0 - \underset{(.066)}{.124}\, belavg + \underset{(.049)}{.035}\, abvavg + other\ factors$$
$$n = 409, \bar{R}^2 = .330.$$

The other factors controlled for in the regressions include education, experience, tenure, marital status, and race; see Table 3 in Hamermesh and Biddle's paper for a more complete list. In order to save space, the coefficients on the other variables are not reported in the paper and neither is the intercept.

For men, those with below average looks are estimated to earn about 16.4% less than an average-looking man who is the same in other respects (including education, experience, tenure, marital status, and race). The effect is statistically different from zero, with $t = -3.57$. Men with above average looks are estimated to earn only 1.6% more than men with average looks, and the effect is not statistically significant $(t < .5)$.

A woman with below average looks earns about 12.4% less than an otherwise comparable average-looking woman, with $t = -1.88$. As was the case for men, the estimate on *abvavg* is much smaller in magnitude and not statistically different from zero.

In related work, Biddle and Hamermesh (1998) revisit the effects of looks on earnings using a more homogeneous group: graduates of a particular law school. The authors continue to find that physical appearance has an effect on annual earnings, something that is perhaps not too surprising among people practicing law.

In some cases, the ordinal variable takes on too many values so that a dummy variable cannot be included for each value. For example, the file LAWSCH85 contains data on median starting salaries for law school graduates. One of the key explanatory variables is the rank of the law school. Because each law school has a different rank, we clearly cannot include a dummy variable for each rank. If we do not wish to put the rank directly in the equation, we can break it down into categories. The following example shows how this is done.

EXAMPLE 7.8 Effects of Law School Rankings on Starting Salaries

Define the dummy variables *top10, r11_25, r26_40, r41_60, r61_100* to take on the value unity when the variable *rank* falls into the appropriate range. We let schools ranked below 100 be the base group. The estimated equation is

$$\begin{aligned}\widehat{\log(salary)} = {} & 9.17 + \underset{(.41)}{.700}\, top10 + \underset{(.053)}{.594}\, r11\_25 + \underset{(.039)}{.375}\, r26\_40 \\ & + \underset{(.028)}{.263}\, r41\_60 + \underset{(.021)}{.132}\, r61\_100 + \underset{(.0031)}{.0057}\, LSAT \\ & + \underset{(.074)}{.041}\, GPA + \underset{(.026)}{.036}\log(libvol) + \underset{(.0251)}{.0008}\log(cost) \\ n = {} & 136, R^2 = .911, \bar{R}^2 = .905.\end{aligned} \qquad [7.13]$$

We see immediately that all of the dummy variables defining the different ranks are very statistically significant. The estimate on *r61_100* means that, holding *LSAT*, *GPA*, *libvol*, and *cost* fixed, the median salary at a law school ranked between 61 and 100 is about 13.2% higher than that at a law school ranked below 100. The difference between a top 10 school and a below 100 school is quite large. Using the exact calculation given in equation (7.10) gives $\exp(.700) - 1 \approx 1.014$, and so the predicted median salary is more than 100% higher at a top 10 school than it is at a below 100 school.

As an indication of whether breaking the rank into different groups is an improvement, we can compare the adjusted R-squared in (7.13) with the adjusted R-squared from including *rank* as a single variable: the former is .905 and the latter is .836, so the additional flexibility of (7.13) is warranted.

Interestingly, once the rank is put into the (admittedly somewhat arbitrary) given categories, all of the other variables become insignificant. In fact, a test for joint significance of *LSAT*, *GPA*, log(*libvol*), and log(*cost*) gives a p-value of .055, which is borderline significant. When *rank* is included in its original form, the p-value for joint significance is zero to four decimal places.

One final comment about this example: In deriving the properties of ordinary least squares, we assumed that we had a random sample. The current application violates that assumption because of the way *rank* is defined: a school's rank necessarily depends on the rank of the other schools in the sample, and so the data cannot represent independent draws from the population of all law schools. This does not cause any serious problems provided the error term is uncorrelated with the explanatory variables.

7-4 Interactions Involving Dummy Variables

7-4a Interactions among Dummy Variables

Just as variables with quantitative meaning can be interacted in regression models, so can dummy variables. We have effectively seen an example of this in Example 7.6, where we defined four categories based on marital status and gender. In fact, we can recast that model by adding an **interaction term** between *female* and *married* to the model where *female* and *married* appear separately. This allows the marriage premium to depend on gender, just as it did in equation (7.11). For purposes of comparison, the estimated model with the *female-married* interaction term is

$$\begin{aligned}\widehat{\log(wage)} = .321 &- .110\,female + .231\,married \\ &(.100)\ (.056)\qquad\ (.055) \\ &- .301\,female{\cdot}married + \ldots, \\ &(.072)\end{aligned} \qquad [7.14]$$

where the rest of the regression is necessarily identical to (7.11). Equation (7.14) shows explicitly that there is a statistically significant interaction between gender and marital status. This model also allows us to obtain the estimated wage differential among all four groups, but here we must be careful to plug in the correct combination of zeros and ones.

Setting $female = 0$ and $married = 0$ corresponds to the group single men, which is the base group, since this eliminates *female*, *married*, and $female \cdot married$. We can find the intercept for married men by setting $female = 0$ and $married = 1$ in (7.14); this gives an intercept of $.321 + .213 = .534$, and so on.

Equation (7.14) is just a different way of finding wage differentials across all gender–marital status combinations. It allows us to easily test the null hypothesis that the gender differential does not depend on marital status (equivalently, that the marriage differential does not depend on gender). Equation (7.11) is more convenient for testing for wage differentials between any group and the base group of single men.

EXAMPLE 7.9 **Effects of Computer Usage on Wages**

Krueger (1993) estimates the effects of computer usage on wages. He defines a dummy variable, which we call *compwork*, equal to one if an individual uses a computer at work. Another dummy variable, *comphome*, equals one if the person uses a computer at home. Using 13,379 people from the 1989 Current Population Survey, Krueger (1993, Table 4) obtains

$$\widehat{\log(wage)} = \hat{\beta}_0 + \underset{(.009)}{.177\ compwork} + \underset{(.019)}{.070\ comphome} + \underset{(.023)}{.017\ compwork{\cdot}comphome} + other\ factors. \qquad [7.15]$$

(The other factors are the standard ones for wage regressions, including education, experience, gender, and marital status; see Krueger's paper for the exact list.) Krueger does not report the intercept because it is not of any importance; all we need to know is that the base group consists of people who do not use a computer at home or at work. It is worth noticing that the estimated return to using a computer at work (but not at home) is about 17.7%. (The more precise estimate is 19.4%.) Similarly, people who use computers at home but not at work have about a 7% wage premium over those who do not use a computer at all. The differential between those who use a computer at both places, relative to those who use a computer in neither place, is about 26.4% (obtained by adding all three coefficients and multiplying by 100), or the more precise estimate 30.2% obtained from equation (7.10).

The interaction term in (7.15) is not statistically significant, nor is it very big economically. But it is causing little harm by being in the equation.

7-4b Allowing for Different Slopes

We have now seen several examples of how to allow different intercepts for any number of groups in a multiple regression model. There are also occasions for interacting dummy variables with explanatory variables that are not dummy variables to allow for **a difference in slopes**. Continuing with the wage example, suppose that we wish to test whether the return to education is the same for men and women, allowing for a constant wage differential between men and women (a differential for which we have already found evidence). For simplicity, we include only education and gender in the model. What kind of model allows for different returns to education? Consider the model

$$\log(wage) = (\beta_0 + \delta_0 female) + (\beta_1 + \delta_1 female)educ + u. \qquad [7.16]$$

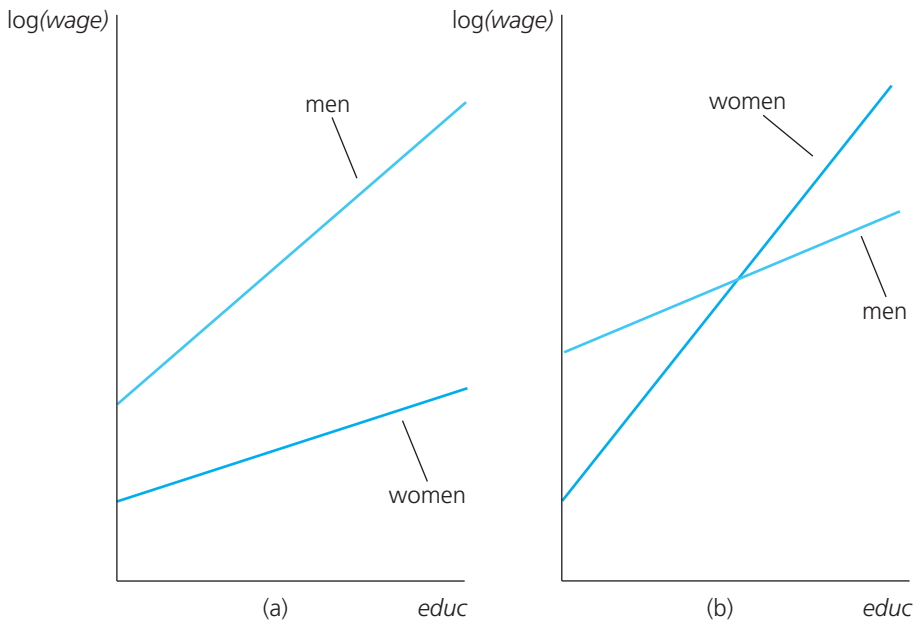
If we plug $female = 0$ into (7.16), then we find that the intercept for males is β_0, and the slope on education for males is β_1. For females, we plug in $female = 1$; thus, the intercept for females is $\beta_0 + \delta_0$, and the slope is $\beta_1 + \delta_1$. Therefore, δ_0 measures the difference in intercepts between women and men, and δ_1 measures the difference in the return to education between women and men. Two of the four cases for the signs of δ_0 and δ_1 are presented in Figure 7.2.

Graph (a) shows the case where the intercept for women is below that for men, and the slope of the line is smaller for women than for men. This means that women earn less than men at all levels of education, and the gap increases as *educ* gets larger. In graph (b), the intercept for women is below that for men, but the slope on education is larger for women. This means that women earn less than men at low levels of education, but the gap narrows as education increases. At some point, a woman earns more than a man with the same level of education, and this amount of education is easily found once we have the estimated equation.

How can we estimate model (7.16)? To apply OLS, we must write the model with an interaction between *female* and *educ*:

$$\log(wage) = \beta_0 + \delta_0 female + \beta_1 educ + \delta_1 female{\cdot}educ + u. \qquad [7.17]$$

FIGURE 7.2 Graphs of equation (7.16): (a) $\delta_0 < 0$, $\delta_1 < 0$; (b) $\delta_0 < 0$, $\delta_1 > 0$.



The parameters can now be estimated from the regression of log(*wage*) on *female*, *educ*, and *female*·*educ*. Obtaining the interaction term is easy in any regression package. Do not be daunted by the odd nature of *female*·*educ*, which is zero for any man in the sample and equal to the level of education for any woman in the sample.

An important hypothesis is that the return to education is the same for women and men. In terms of model (7.17), this is stated as H_0: $\delta_1 = 0$, which means that the slope of log(*wage*) with respect to *educ* is the same for men and women. Note that this hypothesis puts no restrictions on the difference in intercepts, δ_0. A wage differential between men and women is allowed under this null, but it must be the same at all levels of education. This situation is described by Figure 7.1.

We are also interested in the hypothesis that average wages are identical for men and women who have the same levels of education. This means that δ_0 and δ_1 must *both* be zero under the null hypothesis. In equation (7.17), we must use an F test to test H_0: $\delta_0 = 0$, $\delta_1 = 0$. In the model with just an intercept difference, we reject this hypothesis because H_0: $\delta_0 = 0$ is soundly rejected against H_1: $\delta_0 < 0$.

EXAMPLE 7.10 Log Hourly Wage Equation

We add quadratics in experience and tenure to (7.17):

$$\begin{aligned}\widehat{\log(wage)} &= .389 - .227\,female + .082\,educ \\ &\qquad (.119)\ \ (.168)\qquad\ (.008) \\ &\quad - .0056\,female{\cdot}educ + .029\,exper - .00058\,exper^2 \\ &\qquad (.0131)\qquad\ (.005)\qquad (.00011) \\ &\quad + .032\,tenure - .00059\,tenure^2 \\ &\qquad (.007)\qquad\ (.00024) \\ n &= 526,\ R^2 = .441.\end{aligned} \tag{7.18}$$

The estimated return to education for men in this equation is .082, or 8.2%. For women, it is .082 − .0056 = .0764, or about 7.6%. The difference, −.56%, or just over one-half a percentage point less for women, is not economically large nor statistically significant: the t statistic is $-.0056/.0131 \approx -.43$. Thus, we conclude that there is no evidence against the hypothesis that the return to education is the same for men and women.

The coefficient on *female*, while remaining economically large, is no longer significant at conventional levels $(t = -1.35)$. Its coefficient and t statistic in the equation without the interaction were −.297 and −8.25, respectively [see equation (7.9)]. Should we now conclude that there is no statistically significant evidence of lower pay for women at the same levels of *educ*, *exper*, and *tenure*? This would be a serious error. Because we have added the interaction *female·educ* to the equation, the coefficient on *female* is now estimated much less precisely than it was in equation (7.9): the standard error has increased by almost fivefold $(.168/.036 \approx 4.67)$. This occurs because *female* and *female·educ* are highly correlated in the sample. In this example, there is a useful way to think about the multicollinearity: in equation (7.17) and the more general equation estimated in (7.18), δ_0 measures the wage differential between women and men when $educ = 0$. Very few people in the sample have very low levels of education, so it is not surprising that we have a difficult time estimating the differential at $educ = 0$ (nor is the differential at zero years of education very informative). More interesting would be to estimate the gender differential at, say, the average education level in the sample (about 12.5). To do this, we would replace *female·educ* with *female·*$(educ - 12.5)$ and rerun the regression; this only changes the coefficient on *female* and its standard error. (See Computer Exercise C7.)

If we compute the F statistic for H_0: $\delta_0 = 0, \delta_1 = 0$, we obtain $F = 34.33$, which is a huge value for an F random variable with numerator $df = 2$ and denominator $df = 518$: the p-value is zero to four decimal places. In the end, we prefer model (7.9), which allows for a constant wage differential between women and men.

EXPLORING FURTHER 7.4

How would you augment the model estimated in (7.18) to allow the return to *tenure* to differ by gender?

As a more complicated example involving interactions, we now look at the effects of race and city racial composition on major league baseball player salaries.

EXAMPLE 7.11 Effects of Race on Baseball Player Salaries

Using MLB1, the following equation is estimated for the 330 major league baseball players for which city racial composition statistics are available. The variables *black* and *hispan* are binary indicators for the individual players. (The base group is white players.) The variable *percblck* is the percentage of the team's city that is black, and *perchisp* is the percentage of Hispanics. The other variables measure aspects of player productivity and longevity. Here, we are interested in race effects after controlling for these other factors.

In addition to including *black* and *hispan* in the equation, we add the interactions *black·percblck* and *hispan·perchisp*. The estimated equation is

$$\begin{aligned}\widehat{\log(salary)} = {} & 10.34 + .0673\,years + .0089\,gamesyr \\ & \quad (2.18)\ \ (.0129)\qquad\ (.0034) \\ & + .00095\,bavg + .0146\,hrunsyr + .0045\,rbisyr \\ & \quad (.00151)\qquad (.0164)\qquad\ (.0076) \\ & + .0072\,runsyr + .0011\,fldperc + .0075\,allstar \\ & \quad (.0046)\qquad\ (.0021)\qquad\ (.0029)\end{aligned}$$

$$\begin{aligned} &- .198\ black - .190\ hispan + .0125\ black{\cdot}percblck \\ &\quad (.125) \qquad (.153) \qquad (.0050) \\ &+ .0201\ hispan{\cdot}perchisp \\ &\quad (.0098) \\ &n = 330,\ R^2 = .638. \end{aligned} \tag{7.19}$$

First, we should test whether the four race variables, *black*, *hispan*, *black·percblck*, and *hispan·perchisp*, are jointly significant. Using the same 330 players, the *R*-squared when the four race variables are dropped is .626. Since there are four restrictions and $df = 330 - 13$ in the unrestricted model, the *F* statistic is about 2.63, which yields a *p*-value of .034. Thus, these variables are jointly significant at the 5% level (though not at the 1% level).

How do we interpret the coefficients on the race variables? In the following discussion, all productivity factors are held fixed. First, consider what happens for black players, holding *perchisp* fixed. The coefficient −.198 on *black* literally means that, if a black player is in a city with no blacks ($percblck = 0$), then the black player earns about 19.8% less than a comparable white player. As *percblck* increases—which means the white population decreases, since *perchisp* is held fixed—the salary of blacks increases relative to that for whites. In a city with 10% blacks, log(*salary*) for blacks compared to that for whites is $-.198 + .0125(10) = -.073$, so salary is about 7.3% less for blacks than for whites in such a city. When $percblck = 20$, blacks earn about 5.2% more than whites. The largest percentage of blacks in a city is about 74% (Detroit).

Similarly, Hispanics earn less than whites in cities with a low percentage of Hispanics. But we can easily find the value of *perchisp* that makes the differential between whites and Hispanics equal zero: it must make $-.190 + .0201\ perchisp = 0$, which gives $perchisp \approx 9.45$. For cities in which the percentage of Hispanics is less than 9.45%, Hispanics are predicted to earn less than whites (for a given black population), and the opposite is true if the percentage of Hispanics is above 9.45%. Twelve of the 22 cities represented in the sample have Hispanic populations that are less than 9.45% of the total population. The largest percentage of Hispanics is about 31%.

How do we interpret these findings? We cannot simply claim discrimination exists against blacks and Hispanics, because the estimates imply that whites earn less than blacks and Hispanics in cities heavily populated by minorities. The importance of city composition on salaries might be due to player preferences: perhaps the best black players live disproportionately in cities with more blacks and the best Hispanic players tend to be in cities with more Hispanics. The estimates in (7.19) allow us to determine that some relationship is present, but we cannot distinguish between these two hypotheses.

7-4c Testing for Differences in Regression Functions across Groups

The previous examples illustrate that interacting dummy variables with other independent variables can be a powerful tool. Sometimes, we wish to test the null hypothesis that two populations or groups follow the same regression function, against the alternative that one or more of the slopes differ across the groups. We will also see examples of this in Chapter 13, when we discuss pooling different cross sections over time.

Suppose we want to test whether the same regression model describes college grade point averages for male and female college athletes. The equation is

$$cumgpa = \beta_0 + \beta_1 sat + \beta_2 hsperc + \beta_3 tothrs + u,$$

where *sat* is SAT score, *hsperc* is high school rank percentile, and *tothrs* is total hours of college courses. We know that, to allow for an intercept difference, we can include a dummy variable for either males or females. If we want any of the slopes to depend on gender, we simply interact the appropriate variable with, say, *female*, and include it in the equation.

If we are interested in testing whether there is *any* difference between men and women, then we must allow a model where the intercept and all slopes can be different across the two groups:

$$cumgpa = \beta_0 + \delta_0 female + \beta_1 sat + \delta_1 female \cdot sat + \beta_2 hsperc + \delta_2 female \cdot hsperc + \beta_3 tothrs + \delta_3 female \cdot tothrs + u. \quad [7.20]$$

The parameter δ_0 is the difference in the intercept between women and men, δ_1 is the slope difference with respect to *sat* between women and men, and so on. The null hypothesis that *cumgpa* follows the same model for males and females is stated as

$$H_0: \delta_0 = 0, \delta_1 = 0, \delta_2 = 0, \delta_3 = 0. \quad [7.21]$$

If one of the δ_j is different from zero, then the model is different for men and women.

Using the spring semester data from the file GPA3, the full model is estimated as

$$\begin{aligned} \widehat{cumgpa} = {} & 1.48 - .353\, female + .0011\, sat + .00075\, female \cdot sat \\ & \quad (0.21)\ (.411) \qquad (.0002) \qquad (.00039) \\ & -.0085\, hsperc - .00055\, female \cdot hsperc + .0023\, tothrs \\ & \quad (.0014) \qquad (.00316) \qquad (.0009) \\ & -.00012\, female \cdot tothrs \\ & \quad (.00163) \\ n = 366, & \ R^2 = .406, \ \bar{R}^2 = .394. \end{aligned} \quad [7.22]$$

None of the four terms involving the female dummy variable is very statistically significant; only the *female·sat* interaction has a *t* statistic close to two. But we know better than to rely on the individual *t* statistics for testing a joint hypothesis such as (7.21). To compute the *F* statistic, we must estimate the restricted model, which results from dropping *female* and all of the interactions; this gives an R^2 (the restricted R^2) of about .352, so the *F* statistic is about 8.14; the *p*-value is zero to five decimal places, which causes us to soundly reject (7.21). Thus, men and women athletes do follow different GPA models, even though each term in (7.22) that allows women and men to be different is individually insignificant at the 5% level.

The large standard errors on *female* and the interaction terms make it difficult to tell exactly how men and women differ. We must be very careful in interpreting equation (7.22) because, in obtaining differences between women and men, the interaction terms must be taken into account. If we look only at the *female* variable, we would wrongly conclude that *cumgpa* is about .353 less for women than for men, holding other factors fixed. This is the estimated difference only when *sat*, *hsperc*, and *tothrs* are all set to zero, which is not close to being a possible scenario. At $sat = 1, 100$, $hsperc = 10$, and $tothrs = 50$, the predicted *difference* between a woman and a man is $-.353 + .00075(1,100) - .00055(10) - .00012(50) \approx .461$. That is, the female athlete is predicted to have a GPA that is almost one-half a point higher than the comparable male athlete.

In a model with three variables, *sat*, *hsperc*, and *tothrs*, it is pretty simple to add all of the interactions to test for group differences. In some cases, many more explanatory variables are involved, and then it is convenient to have a different way to compute the statistic. It turns out that the sum of squared residuals form of the *F* statistic can be computed easily even when many independent variables are involved.

In the general model with *k* explanatory variables and an intercept, suppose we have two groups; call them $g = 1$ and $g = 2$. We would like to test whether the intercept and all slopes are the same across the two groups. Write the model as

$$y = \beta_{g,0} + \beta_{g,1} x_1 + \beta_{g,2} x_2 + \ldots + \beta_{g,k} x_k + u, \quad [7.23]$$

for $g = 1$ and $g = 2$. The hypothesis that each beta in (7.23) is the same across the two groups involves $k + 1$ restrictions (in the GPA example, $k + 1 = 4$). The unrestricted model, which we can think of as having a group dummy variable and k interaction terms in addition to the intercept and variables themselves, has $n - 2(k + 1)$ degrees of freedom. [In the GPA example, $n - 2(k + 1) = 366 - 2(4) = 358$.] So far, there is nothing new. The key insight is that the sum of squared residuals from the unrestricted model can be obtained from two *separate* regressions, one for each group. Let SSR_1 be the sum of squared residuals obtained estimating (7.23) for the first group; this involves n_1 observations. Let SSR_2 be the sum of squared residuals obtained from estimating the model using the second group (n_2 observations). In the previous example, if group 1 is females, then $n_1 = 90$ and $n_2 = 276$. Now, the sum of squared residuals for the unrestricted model is simply $SSR_{ur} = SSR_1 + SSR_2$. The restricted sum of squared residuals is just the SSR from pooling the groups and estimating a single equation, say SSR_P. Once we have these, we compute the F statistic as usual:

$$F = \frac{[SSR_P - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k + 1)]}{k + 1}, \qquad [7.24]$$

where n is the *total* number of observations. This particular F statistic is usually called the **Chow statistic** in econometrics. Because the Chow test is just an F test, it is only valid under homoskedasticity. In particular, under the null hypothesis, the error variances for the two groups must be equal. As usual, normality is not needed for asymptotic analysis.

To apply the Chow statistic to the GPA example, we need the SSR from the regression that pooled the groups together: this is $SSR_P = 85.515$. The SSR for the 90 women in the sample is $SSR_1 = 19.603$, and the SSR for the men is $SSR_2 = 58.752$. Thus, $SSR_{ur} = 19.603 + 58.752 = 78.355$. The F statistic is $[(85.515 - 78.355)/78.355](358/4) \approx 8.18]$; of course, subject to rounding error, this is what we get using the R-squared form of the test in the models with and without the interaction terms. (A word of caution: there is no simple R-squared form of the test if separate regressions have been estimated for each group; the R-squared form of the test can be used only if interactions have been included to create the unrestricted model.)

One important limitation of the traditional Chow test, regardless of the method used to implement it, is that the null hypothesis allows for no differences at all between the groups. In many cases, it is more interesting to allow for an intercept difference between the groups and then to test for slope differences; we saw one example of this in the wage equation in Example 7.10. There are two ways to allow the intercepts to differ under the null hypothesis. One is to include the group dummy and all interaction terms, as in equation (7.22), but then test joint significance of the interaction terms only. The second approach, which produces an identical statistic, is to form a sum-of-squared-residuals F statistic, as in equation (7.24), but where the restricted SSR, called "SSR_P" in equation (7.24), is obtained using a regression that contains only an intercept shift. Because we are testing k restrictions, rather than $k + 1$, the F statistic becomes

$$F = \frac{[SSR_P - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \cdot \frac{[n - 2(k + 1)]}{k}.$$

Using this approach in the GPA example, SSR_P is obtained from the regression *cumgpa* on *female*, *sat*, *hsperc*, and *tothrs* using the data for both male and female student-athletes.

Because there are relatively few explanatory variables in the GPA example, it is easy to estimate (7.20) and test H_0: $\delta_1 = 0, \delta_2 = 0, \delta_3 = 0$ (with δ_0 unrestricted under the null). The F statistic for the three exclusion restrictions gives a p-value equal to .205, and so we do not reject the null hypothesis at even the 20% significance level.

Failure to reject the hypothesis that the parameters multiplying the interaction terms are all zero suggests that the best model allows for an intercept difference only:

$$\widehat{cumgpa} = 1.39 + .310\,female + .0012\,sat - .0084\,hsperc \\ (.18)\ (.059)\qquad (.0002)\qquad (.0012) \\ + .0025\,tothrs \\ (.0007) \\ n = 366,\ R^2 = .398,\ \overline{R}^2 = .392. \qquad [7.25]$$

The slope coefficients in (7.25) are close to those for the base group (males) in (7.22); dropping the interactions changes very little. However, *female* in (7.25) is highly significant: its t statistic is over 5, and the estimate implies that, at given levels of *sat*, *hsperc*, and *tothrs*, a female athlete has a predicted GPA that is .31 point higher than that of a male athlete. This is a practically important difference.

7-5 A Binary Dependent Variable: The Linear Probability Model

By now, we have learned much about the properties and applicability of the multiple linear regression model. In the last several sections, we studied how, through the use of binary independent variables, we can incorporate qualitative information as explanatory variables in a multiple regression model. In all of the models up until now, the dependent variable y has had *quantitative* meaning (for example, y is a dollar amount, a test score, a percentage, or the logs of these). What happens if we want to use multiple regression to *explain* a qualitative event?

In the simplest case, and one that often arises in practice, the event we would like to explain is a binary outcome. In other words, our dependent variable, y, takes on only two values: zero and one. For example, y can be defined to indicate whether an adult has a high school education; y can indicate whether a college student used illegal drugs during a given school year; or y can indicate whether a firm was taken over by another firm during a given year. In each of these examples, we can let $y = 1$ denote one of the outcomes and $y = 0$ the other outcome.

What does it mean to write down a multiple regression model, such as

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u, \qquad [7.26]$$

when y is a binary variable? Because y can take on only two values, β_j cannot be interpreted as the change in y given a one-unit increase in x_j, holding all other factors fixed: y either changes from zero to one or from one to zero (or does not change). Nevertheless, the β_j still have useful interpretations. If we assume that the zero conditional mean assumption MLR.4 holds, that is, $\mathrm{E}(u|x_1, \ldots, x_k) = 0$, then we have, as always,

$$\mathrm{E}(y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k,$$

where $\mathbf{x}$ is shorthand for all of the explanatory variables.

The key point is that when y is a binary variable taking on the values zero and one, it is always true that $\mathrm{P}(y = 1|\mathbf{x}) = E(y|\mathbf{x})$: the probability of "success"—that is, the probability that $y = 1$—is the same as the expected value of y. Thus, we have the important equation

$$\mathrm{P}(y = 1|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k, \qquad [7.27]$$

which says that the probability of success, say, $p(\mathbf{x}) = \mathrm{P}(y = 1|\mathbf{x})$, is a linear function of the x_j. Equation (7.27) is an example of a binary response model, and $\mathrm{P}(y = 1|\mathbf{x})$ is also called the **response**

probability. (We will cover other binary response models in Chapter 17.) Because probabilities must sum to one, $P(y = 0|\mathbf{x}) = 1 - P(y = 1|\mathbf{x})$ is also a linear function of the x_j.

The multiple linear regression model with a binary dependent variable is called the **linear probability model (LPM)** because the response probability is linear in the parameters β_j. In the LPM, β_j measures the change in the probability of success when x_j changes, holding other factors fixed:

$$\Delta P(y = 1|\mathbf{x}) = \beta_j \Delta x_j. \qquad [7.28]$$

With this in mind, the multiple regression model can allow us to estimate the effect of various explanatory variables on qualitative events. The mechanics of OLS are the same as before.

If we write the estimated equation as

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k,$$

we must now remember that $\hat{y}$ is the predicted probability of success. Therefore, $\hat{\beta}_0$ is the predicted probability of success when each x_j is set to zero, which may or may not be interesting. The slope coefficient $\hat{\beta}_1$ measures the predicted change in the probability of success when x_1 increases by one unit.

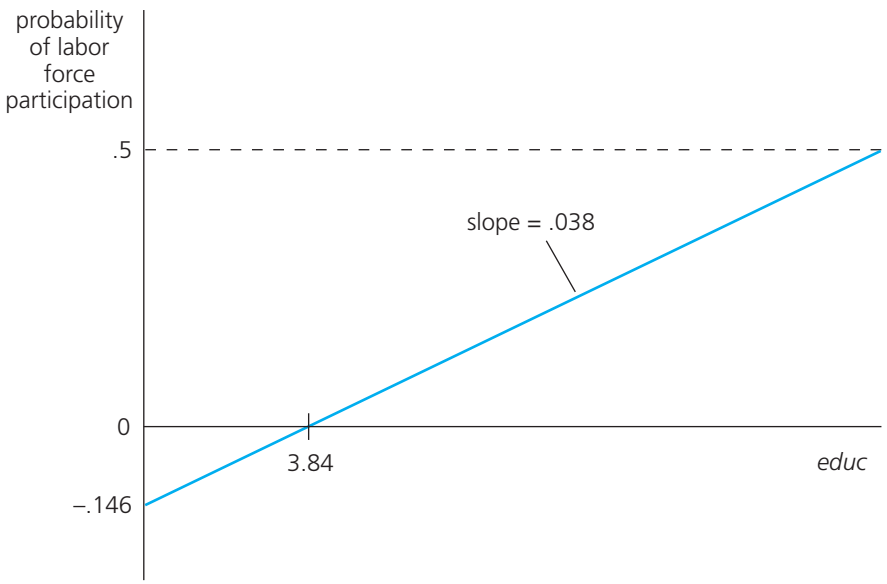
To correctly interpret a linear probability model, we must know what constitutes a "success." Thus, it is a good idea to give the dependent variable a name that describes the event $y = 1$. As an example, let *inlf* ("in the labor force") be a binary variable indicating labor force participation by a married woman during 1975: *inlf* = 1 if the woman reports working for a wage outside the home at some point during the year, and zero otherwise. We assume that labor force participation depends on other sources of income, including husband's earnings (*nwifeinc*, measured in thousands of dollars), years of education (*educ*), past years of labor market experience (*exper*), *age*, number of children less than six years old (*kidslt6*), and number of kids between 6 and 18 years of age (*kidsge6*). Using the data in MROZ from Mroz (1987), we estimate the following linear probability model, where 428 of the 753 women in the sample report being in the labor force at some point during 1975:

$$\begin{aligned} \widehat{inlf} &= .586 - .0034\, nwifeinc + .038\, educ + .039\, exper \\ &\quad\ (.154)\ (.0014)\qquad\ (.007)\qquad (.006) \\ &\quad - .00060\, exper^2 - .016\, age - .262\, kidslt6 + .013\, kidsge6 \\ &\quad\ (.00018)\qquad (.002)\qquad (.034)\qquad\ (.013) \\ n &= 753,\ R^2 = .264. \end{aligned} \qquad [7.29]$$

Using the usual t statistics, all variables in (7.29) except *kidsge6* are statistically significant, and all of the significant variables have the effects we would expect based on economic theory (or common sense).

To interpret the estimates, we must remember that a change in the independent variable changes the probability that *inlf* = 1. For example, the coefficient on *educ* means that, everything else in (7.29) held fixed, another year of education increases the probability of labor force participation by .038. If we take this equation literally, 10 more years of education increases the probability of being in the labor force by $.038(10) = .38$, which is a pretty large increase in a probability. The relationship between the probability of labor force participation and *educ* is plotted in Figure 7.3. The other independent variables are fixed at the values *nwifeinc* = 50, *exper* = 5, *age* = 30, *kidslt6* = 1, and *kidsge6* = 0 for illustration purposes. The predicted probability is negative until education equals 3.84 years. This should not cause too much concern because, in this sample, no woman has less than five years of education. The largest reported education is 17 years, and this leads to a predicted probability of .5. If we set the other independent variables at different values, the range of predicted probabilities would change. But the marginal effect of another year of education on the probability of labor force participation is always .038.

FIGURE 7.3 Estimated relationship between the probability of being in the labor force and years of education, with other explanatory variables fixed.



The coefficient on *nwifeinc* implies that, if $\Delta nwifeinc = 10$ (which means an increase of \$10,000), the probability that a woman is in the labor force falls by .034. This is not an especially large effect given that an increase in income of \$10,000 is substantial in terms of 1975 dollars. Experience has been entered as a quadratic to allow the effect of past experience to have a diminishing effect on the labor force participation probability. Holding other factors fixed, the estimated change in the probability is approximated as $.039 - 2(.0006)exper = .039 - .0012\, exper$. The point at which past experience has no effect on the probability of labor force participation is $.039/.0012 = 32.5$, which is a high level of experience: only 13 of the 753 women in the sample have more than 32 years of experience.

Unlike the number of older children, the number of young children has a huge impact on labor force participation. Having one additional child less than six years old reduces the probability of participation by $-.262$, at given levels of the other variables. In the sample, just under 20% of the women have at least one young child.

This example illustrates how easy linear probability models are to estimate and interpret, but it also highlights some shortcomings of the LPM. First, it is easy to see that, if we plug certain combinations of values for the independent variables into (7.29), we can get predictions either less than zero or greater than one. Since these are predicted probabilities, and probabilities must be between zero and one, this can be a little embarrassing. For example, what would it mean to predict that a woman is in the labor force with a probability of –.10? In fact, of the 753 women in the sample, 16 of the fitted values from (7.29) are less than zero, and 17 of the fitted values are greater than one.

A related problem is that a probability cannot be linearly related to the independent variables for all their possible values. For example, (7.29) predicts that the effect of going from zero children to one young child reduces the probability of working by .262. This is also the predicted drop if the woman goes from having one young child to two. It seems more realistic that the first small child would reduce the probability by a large amount, but subsequent children would have a smaller

marginal effect. In fact, when taken to the extreme, (7.29) implies that going from zero to four young children reduces the probability of working by $\Delta\widehat{inlf} = .262(\Delta kidslt6) = .262(4) = 1.048$, which is impossible.

Even with these problems, the linear probability model is useful and often applied in economics. It usually works well for values of the independent variables that are near the averages in the sample. In the labor force participation example, no women in the sample have four young children; in fact, only three women have three young children. Over 96% of the women have either no young children or one small child, and so we should probably restrict attention to this case when interpreting the estimated equation.

Predicted probabilities outside the unit interval are a little troubling when we want to make predictions. Still, there are ways to use the estimated probabilities (even if some are negative or greater than one) to predict a zero-one outcome. As before, let $\hat{y}_i$ denote the fitted values—which may not be bounded between zero and one. Define a predicted value as $\tilde{y}_i = 1$ if $\hat{y}_i \geq .5$ and $\tilde{y}_i = 0$ if $\hat{y}_i < .5$. Now we have a set of predicted values, $\tilde{y}_i$, $i = 1, \ldots, n$, that, like the y_i, are either zero or one. We can use the data on y_I and $\tilde{y}_i$ to obtain the frequencies with which we correctly predict $y_i = 1$ and $y_i = 0$, as well as the proportion of overall correct predictions. The latter measure, when turned into a percentage, is a widely used goodness-of-fit measure for binary dependent variables: the **percent correctly predicted**. An example is given in Computer Exercise C9(v), and further discussion, in the context of more advanced models, can be found in Section 17-1.

Due to the binary nature of y, the linear probability model does violate one of the Gauss-Markov assumptions. When y is a binary variable, its variance, conditional on $\mathbf{x}$, is

$$\text{Var}(y|\mathbf{x}) = p(\mathbf{x})[1 - p(\mathbf{x})], \qquad [7.30]$$

where $p(\mathbf{x})$ is shorthand for the probability of success: $p(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$. This means that, except in the case where the probability does not depend on any of the independent variables, there *must* be heteroskedasticity in a linear probability model. We know from Chapter 3 that this does not cause bias in the OLS estimators of the β_j. But we also know from Chapters 4 and 5 that homoskedasticity is crucial for justifying the usual t and F statistics, even in large samples. Because the standard errors in (7.29) are not generally valid, we should use them with caution. We will show how to correct the standard errors for heteroskedasticity in Chapter 8. It turns out that, in many applications, the usual OLS statistics are not far off, and it is still acceptable in applied work to present a standard OLS analysis of a linear probability model.

EXAMPLE 7.12 A Linear Probability Model of Arrests

Let *arr86* be a binary variable equal to unity if a man was arrested during 1986, and zero otherwise. The population is a group of young men in California born in 1960 or 1961 who have at least one arrest prior to 1986. A linear probability model for describing *arr86* is

$$arr86 = \beta_0 + \beta_1 pcnv + \beta_2 avgsen + \beta_3 tottime + \beta_4 ptime86 + \beta_5 qemp86 + u,$$

where

pcnv = the proportion of prior arrests that led to a conviction.
avgsen = the average sentence served from prior convictions (in months).
tottime = months spent in prison since age 18 prior to 1986.
ptime86 = months spent in prison in 1986.
qemp86 = the number of quarters (0 to 4) that the man was legally employed in 1986.

The data we use are in CRIME1, the same data set used for Example 3.5. Here, we use a binary dependent variable because only 7.2% of the men in the sample were arrested more than once. About 27.7% of the men were arrested at least once during 1986. The estimated equation is

$$\widehat{arr86} = .441 - .162\,pcnv + .0061\,avgsen - .0023\,tottime \\ (.017)\ (.021)\qquad (.0065)\qquad (.0050) \\ - .022\,ptime86 - .043\,qemp86 \qquad [7.31] \\ (.005)\qquad (.005) \\ n = 2{,}725,\ R^2 = .0474.$$

The intercept, .441, is the predicted probability of arrest for someone who has not been convicted (and so *pcnv* and *avgsen* are both zero), has spent no time in prison since age 18, spent no time in prison in 1986, and was unemployed during the entire year. The variables *avgsen* and *tottime* are insignificant both individually and jointly (the F test gives $p\text{-}value = .347$), and *avgsen* has a counterintuitive sign if longer sentences are supposed to deter crime. Grogger (1991), using a superset of these data and different econometric methods, found that *tottime* has a statistically significant *positive* effect on arrests and concluded that *tottime* is a measure of human capital built up in criminal activity.

Increasing the probability of conviction does lower the probability of arrest, but we must be careful when interpreting the magnitude of the coefficient. The variable *pcnv* is a proportion between zero and one; thus, changing *pcnv* from zero to one essentially means a change from no chance of being convicted to being convicted with certainty. Even this large change reduces the probability of arrest only by .162; increasing *pcnv* by .5 decreases the probability of arrest by .081.

The incarcerative effect is given by the coefficient on *ptime86*. If a man is in prison, he cannot be arrested. Since *ptime86* is measured in months, six more months in prison reduces the probability of arrest by $.022(6) = .132$. Equation (7.31) gives another example of where the linear probability model cannot be true over all ranges of the independent variables. If a man is in prison all 12 months of 1986, he cannot be arrested in 1986. Setting all other variables equal to zero, the predicted probability of arrest when $ptime86 = 12$ is $.441 - .022(12) = .177$, which is not zero. Nevertheless, if we start from the unconditional probability of arrest, .277, 12 months in prison reduces the probability to essentially zero: $.277 - .022(12) = .013$.

Finally, employment reduces the probability of arrest in a significant way. All other factors fixed, a man employed in all four quarters is .172 less likely to be arrested than a man who is not employed at all.

We can also include dummy independent variables in models with dummy dependent variables. The coefficient measures the predicted difference in probability relative to the base group. For example, if we add two race dummies, *black* and *hispan*, to the arrest equation, we obtain

$$\widehat{arr86} = .380 - .152\,pcnv + .0046\,avgsen - .0026\,tottime \\ (.019)\ (.021)\qquad (.0064)\qquad (.0049) \\ - .024\,ptime86 - .038\,qemp86 + .170\,black + .096\,hispan \qquad [7.32] \\ (.005)\qquad (.005)\qquad (.024)\qquad (.021) \\ n = 2{,}725,\ R^2 = .0682.$$

EXPLORING FURTHER 7.5

What is the predicted probability of arrest for a black man with no prior convictions—so that *pcnv, avgsen, tottime,* and *ptime86* are all zero—who was employed all four quarters in 1986? Does this seem reasonable?

The coefficient on *black* means that, all other factors being equal, a black man has a .17 higher chance of being arrested than a white man (the base group). Another way to say this is that the probability of arrest is 17 percentage points higher for blacks than for whites. The difference is statistically significant as well. Similarly, Hispanic men have a .096 higher chance of being arrested than white men.

7-6 More on Policy Analysis and Program Evaluation

We have seen some examples of models containing dummy variables that can be useful for evaluating policy. Example 7.3 gave an example of program evaluation, where some firms received job training grants and others did not.

As we mentioned earlier, we must be careful when evaluating programs because in most examples in the social sciences the control and treatment groups are not randomly assigned. Consider again the Holzer et al. (1993) study, where we are now interested in the effect of the job training grants on worker productivity (as opposed to amount of job training). The equation of interest is

$$\log(scrap) = \beta_0 + \beta_1 grant + \beta_2 \log(sales) + \beta_3 \log(employ) + u,$$

where *scrap* is the firm's scrap rate, and the latter two variables are included as controls. The binary variable *grant* indicates whether the firm received a grant in 1988 for job training.

Before we look at the estimates, we might be worried that the unobserved factors affecting worker productivity—such as average levels of education, ability, experience, and tenure—might be correlated with whether the firm receives a grant. Holzer et al. point out that grants were given on a first-come, first-served basis. But this is not the same as giving out grants randomly. It might be that firms with less productive workers saw an opportunity to improve productivity and therefore were more diligent in applying for the grants.

Using the data in JTRAIN for 1988—when firms actually were eligible to receive the grants—we obtain

$$\begin{aligned} \widehat{\log(scrap)} &= 4.99 - .052\, grant - .455 \log(sales) \\ &\qquad (4.66)\ \ (.431)\qquad\ (.373) \\ &\quad + .639 \log(employ) \\ &\qquad (.365) \\ n &= 50,\ R^2 = .072. \end{aligned} \tag{7.33}$$

(Seventeen out of the 50 firms received a training grant, and the average scrap rate is 3.47 across all firms.) The point estimate of $-.052$ on *grant* means that, for given *sales* and *employ*, firms receiving a grant have scrap rates about 5.2% lower than firms without grants. This is the direction of the expected effect if the training grants are effective, but the t statistic is very small. Thus, from this cross-sectional analysis, we must conclude that the grants had no effect on firm productivity. We will return to this example in Chapter 9 and show how adding information from a prior year leads to a much different conclusion.

Even in cases where the policy analysis does not involve assigning units to a control group and a treatment group, we must be careful to include factors that might be systematically related to the binary independent variable of interest. A good example of this is testing for racial discrimination. Race is something that is not determined by an individual or by government administrators. In fact,

race would appear to be the perfect example of an exogenous explanatory variable, given that it is determined at birth. However, for historical reasons, race is often related to other relevant factors: there are systematic differences in backgrounds across race, and these differences can be important in testing for *current* discrimination.

As an example, consider testing for discrimination in loan approvals. If we can collect data on, say, individual mortgage applications, then we can define the dummy dependent variable *approved* as equal to one if a mortgage application was approved, and zero otherwise. A systematic difference in approval rates across races is an indication of discrimination. However, since approval depends on many other factors, including income, wealth, credit ratings, and a general ability to pay back the loan, we must control for them *if* there are systematic differences in these factors across race. A linear probability model to test for discrimination might look like the following:

$$approved = \beta_0 + \beta_1 nonwhite + \beta_2 income + \beta_3 wealth + \beta_4 credrate + other\ factors.$$

Discrimination against minorities is indicated by a rejection of H_0: $\beta_1 = 0$ in favor of H_0: $\beta_1 < 0$, because β_1 is the amount by which the probability of a nonwhite getting an approval differs from the probability of a white getting an approval, given the same levels of other variables in the equation. If *income*, *wealth*, and so on are systematically different across races, then it is important to control for these factors in a multiple regression analysis.

Another problem that often arises in policy and program evaluation is that individuals (or firms or cities) choose whether or not to participate in certain behaviors or programs. For example, individuals choose to use illegal drugs or drink alcohol. If we want to examine the effects of such behaviors on unemployment status, earnings, or criminal behavior, we should be concerned that drug usage might be correlated with other factors that can affect employment and criminal outcomes. Children eligible for programs such as Head Start participate based on parental decisions. Since family background plays a role in Head Start decisions and affects student outcomes, we should control for these factors when examining the effects of Head Start [see, for example, Currie and Thomas (1995)]. Individuals selected by employers or government agencies to participate in job training programs can participate or not, and this decision is unlikely to be random [see, for example, Lynch (1992)]. Cities and states choose whether to implement certain gun control laws, and it is likely that this decision is systematically related to other factors that affect violent crime [see, for example, Kleck and Patterson (1993)].

The previous paragraph gives examples of what are generally known as **self-selection** problems in economics. Literally, the term comes from the fact that individuals self-select into certain behaviors or programs: participation is not randomly determined. The term is used generally when a binary indicator of participation might be systematically related to unobserved factors. Thus, if we write the simple model

$$y = \beta_0 + \beta_1 partic + u, \qquad [7.34]$$

where y is an outcome variable and *partic* is a binary variable equal to unity if the individual, firm, or city participates in a behavior or a program or has a certain kind of law, then we are worried that the average value of u depends on participation: $E(u|partic = 1) \neq E(u|partic = 0)$. As we know, this causes the simple regression estimator of β_1 to be biased, and so we will not uncover the true effect of participation. Thus, the self-selection problem is another way that an explanatory variable (*partic* in this case) can be endogenous.

By now, we know that multiple regression analysis can, to some degree, alleviate the self-selection problem. Factors in the error term in (7.34) that are correlated with *partic* can be included in a multiple regression equation, assuming, of course, that we can collect data on these factors. Unfortunately, in many cases, we are worried that unobserved factors are related to participation, in which case multiple regression produces biased estimators.

With standard multiple regression analysis using cross-sectional data, we must be aware of finding spurious effects of programs on outcome variables due to the self-selection problem. A good

example of this is contained in Currie and Cole (1993). These authors examine the effect of AFDC (Aid to Families with Dependent Children) participation on the birth weight of a child. Even after controlling for a variety of family and background characteristics, the authors obtain OLS estimates that imply participation in AFDC *lowers* birth weight. As the authors point out, it is hard to believe that AFDC participation itself *causes* lower birth weight. [See Currie (1995) for additional examples.] Using a different econometric method that we will discuss in Chapter 15, Currie and Cole find evidence for either no effect or a *positive* effect of AFDC participation on birth weight.

When the self-selection problem causes standard multiple regression analysis to be biased due to a lack of sufficient control variables, the more advanced methods covered in Chapters 13,14, and 15, can be used instead.

7-7 Interpreting Regression Results with Discrete Dependent Variables

A binary response is the most extreme form of a discrete random variable: it takes on only two values, zero and one. As we discussed in Section 7-5, the parameters in a linear probability model can be interpreted as measuring the change in the *probability* that $y = 1$ due to a one-unit increase in an explanatory variable. We also discussed that, because y is a zero-one outcome, $\mathrm{P}(y = 1) = \mathrm{E}(y)$, and this equality continues to hold when we condition on explanatory variables.

Other discrete dependent variables arise in practice, and we have already seen some examples, such as the number of times someone is arrested in a given year (Example 3.5). Studies on factors affecting fertility often use the number of living children as the dependent variable in a regression analysis. As with number of arrests, the number of living children takes on a small set of integer values, and zero is a common value. The data in FERTIL2, which contains information on a large sample of women in Botswana is one such example. Often demographers are interested in the effects of education on fertility, with special attention to trying to determine whether education has a causal effect on fertility. Such examples raise a question about how one interprets regression coefficients: after all, one cannot have a fraction of a child.

To illustrate the issues, the regression below uses the data in FERTIL2:

$$\begin{aligned} \widehat{children} &= -1.997 + .175\,age - .090\,educ \\ &\qquad (.094)\ \ (.003)\qquad (.006) \\ n &= 4{,}361,\ R^2 = .560. \end{aligned} \qquad [7.35]$$

At this time, we ignore the issue of whether this regression adequately controls for all factors that affect fertility. Instead we focus on interpreting the regression coefficients.

Consider the main coefficient of interest, $\hat{\beta}_{educ} = -.090$. If we take this estimate literally, it says that each additional year of education reduces the estimated number of children by .090—something obviously impossible for any particular woman. A similar problem arises when trying to interpret $\hat{\beta}_{age} = .175$. How can we make sense of these coefficients?

To interpret regression results generally, even in cases where y is discrete and takes on a small number of values, it is useful to remember the interpretation of OLS as estimating the effects of the x_j on the *expected* (or *average*) value of y. Generally, under Assumptions MLR.1 and MLR.4,

$$\mathrm{E}(y|x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k. \qquad [7.36]$$

Therefore, β_j is the effect of a ceteris paribus increase of x_j on the expected value of y. As we discussed in Section 6-4, for a given set of x_j values we interpret the predicted value, $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$, as an estimate of $\mathrm{E}(y|x_1, x_2, \ldots, x_k)$. Therefore, $\hat{\beta}_j$ is our estimate of how the *average* of y changes when $\Delta x_j = 1$ (keeping other factors fixed).

Seen in this light, we can now provide meaning to regression results as in equation (7.35). The coefficient $\hat{\beta}_{educ} = -.090$ means that we estimate that *average* fertility falls by .09 children given one more year of education. A nice way to summarize this interpretation is that if each woman in a group of 100 obtains another year of education, we estimate there will be nine fewer children among them.

Adding dummy variables to regressions when y is itself discrete causes no problems when we interpret the estimated effect in terms of average values. Using the data in FERTIL2 we get

$$\widehat{children} = -2.071 + .177\, age - .079\, educ - .362\, electric \qquad [7.37]$$
$$(.095)\ (.003)\qquad (.006)\qquad (.068)$$
$$n = 4{,}358,\ R^2 = .562,$$

where *electric* is a dummy variable equal to one if the woman lives in a home with electricity. Of course it cannot be true that a particular woman who has electricity has .362 less children than an otherwise comparable woman who does not. But we can say that when comparing 100 women with electricity to 100 women without—at the same age and level of education—we estimate the former group to have about 36 fewer children.

Incidentally, when y is discrete the linear model does not always provide the best estimates of partial effects on $\mathrm{E}(y|x_1, x_2, \ldots, x_k)$. Chapter 17 contains more advanced models and estimation methods that tend to fit the data better when the range of y is limited in some substantive way. Nevertheless, a linear model estimated by OLS often provides a good approximation to the true partial effects, at least on average.

Summary

In this chapter, we have learned how to use qualitative information in regression analysis. In the simplest case, a dummy variable is defined to distinguish between two groups, and the coefficient estimate on the dummy variable estimates the ceteris paribus difference between the two groups. Allowing for more than two groups is accomplished by defining a set of dummy variables: if there are g groups, then $g - 1$ dummy variables are included in the model. All estimates on the dummy variables are interpreted relative to the base or benchmark group (the group for which no dummy variable is included in the model).

Dummy variables are also useful for incorporating ordinal information, such as a credit or a beauty rating, in regression models. We simply define a set of dummy variables representing different outcomes of the ordinal variable, allowing one of the categories to be the base group.

Dummy variables can be interacted with quantitative variables to allow slope differences across different groups. In the extreme case, we can allow each group to have its own slope on every variable, as well as its own intercept. The Chow test can be used to detect whether there are any differences across groups. In many cases, it is more interesting to test whether, after allowing for an intercept difference, the slopes for two different groups are the same. A standard F test can be used for this purpose in an unrestricted model that includes interactions between the group dummy and all variables.

The linear probability model, which is simply estimated by OLS, allows us to explain a binary response using regression analysis. The OLS estimates are now interpreted as changes in the probability of "success" $(y = 1)$, given a one-unit increase in the corresponding explanatory variable. The LPM does have some drawbacks: it can produce predicted probabilities that are less than zero or greater than one, it implies a constant marginal effect of each explanatory variable that appears in its original form, and it contains heteroskedasticity. The first two problems are often not serious when we are obtaining estimates

of the partial effects of the explanatory variables for the middle ranges of the data. Heteroskedasticity does invalidate the usual OLS standard errors and test statistics, but, as we will see in the next chapter, this is easily fixed in large enough samples.

Section 7.6 provides a discussion of how binary variables are used to evaluate policies and programs. As in all regression analysis, we must remember that program participation, or some other binary regressor with policy implications, might be correlated with unobserved factors that affect the dependent variable, resulting in the usual omitted variables bias.

We ended this chapter with a general discussion of how to interpret regression equations when the dependent variable is discrete. The key is to remember that the coefficients can be interpreted as the effects on the expected value of the dependent variable.

Key Terms

Base Group
Benchmark Group
Binary Variable
Chow Statistic
Control Group
Difference in Slopes
Dummy Variable Trap
Dummy Variables
Experimental Group
Interaction Term
Intercept Shift
Linear Probability Model (LPM)
Ordinal Variable
Percent Correctly Predicted
Policy Analysis
Program Evaluation
Response Probability
Self-Selection
Treatment Group
Uncentered R-Squared
Zero-One Variable

Problems

1 Using the data in SLEEP75 (see also Problem 3 in Chapter 3), we obtain the estimated equation

$$\begin{aligned}\widehat{sleep} &= 3{,}840.83 - .163\,totwrk - 11.71\,educ - 8.70\,age \\ &\qquad (235.11)\ (.018)\qquad\quad (5.86)\qquad (11.21) \\ &\quad + .128\,age^2 + 87.75\,male \\ &\qquad (.134)\qquad (34.33) \\ n &= 706,\ R^2 = .123,\ \overline{R}^2 = .117.\end{aligned}$$

The variable *sleep* is total minutes per week spent sleeping at night, *totwrk* is total weekly minutes spent working, *educ* and *age* are measured in years, and *male* is a gender dummy.

(i) All other factors being equal, is there evidence that men sleep more than women? How strong is the evidence?

(ii) Is there a statistically significant tradeoff between working and sleeping? What is the estimated tradeoff?

(iii) What other regression do you need to run to test the null hypothesis that, holding other factors fixed, age has no effect on sleeping?

2 The following equations were estimated using the data in BWGHT:

$$\begin{aligned}\widehat{\log(bwght)} &= 4.66 - .0044\,cigs + .0093\log(faminc) + .016\,parity \\ &\qquad (.22)\ (.0009)\qquad (.0059)\qquad\quad (.006) \\ &\quad + .027\,male + .055\,white \\ &\qquad (.010)\qquad (.013) \\ n &= 1{,}388,\ R^2 = .0472\end{aligned}$$

and

$$\widehat{\log(bwght)} = 4.65 - .0052\,cigs + .0110\log(faminc) + .017\,parity$$
$$(.38)\ (.0010)\qquad (.0085)\qquad (.006)$$
$$+ .034\,male + .045\,white - .0030\,motheduc + .0032\,fatheduc$$
$$(.011)\qquad (.015)\qquad (.0030)\qquad (.0026)$$
$$n = 1{,}191,\ R^2 = .0493.$$

The variables are defined as in Example 4.9, but we have added a dummy variable for whether the child is male and a dummy variable indicating whether the child is classified as white.

(i) In the first equation, interpret the coefficient on the variable *cigs*. In particular, what is the effect on birth weight from smoking 10 more cigarettes per day?

(ii) How much more is a white child predicted to weigh than a nonwhite child, holding the other factors in the first equation fixed? Is the difference statistically significant?

(iii) Comment on the estimated effect and statistical significance of *motheduc*.

(iv) From the given information, why are you unable to compute the F statistic for joint significance of *motheduc* and *fatheduc*? What would you have to do to compute the F statistic?

3 Using the data in GPA2, the following equation was estimated:

$$\widehat{sat} = 1{,}028.10 + 19.30\,hsize - 2.19\,hsize^2 - 45.09\,female$$
$$(6.29)\ (3.83)\qquad (.53)\qquad (4.29)$$
$$- 169.81\,black + 62.31\,female{\cdot}black$$
$$(12.71)\qquad (18.15)$$
$$n = 4{,}137,\ R^2 = .0858.$$

The variable *sat* is the combined SAT score; *hsize* is size of the student's high school graduating class, in hundreds; *female* is a gender dummy variable; and *black* is a race dummy variable equal to one for blacks, and zero otherwise.

(i) Is there strong evidence that $hsize^2$ should be included in the model? From this equation, what is the optimal high school size?

(ii) Holding *hsize* fixed, what is the estimated difference in SAT score between nonblack females and nonblack males? How statistically significant is this estimated difference?

(iii) What is the estimated difference in SAT score between nonblack males and black males? Test the null hypothesis that there is no difference between their scores, against the alternative that there is a difference.

(iv) What is the estimated difference in SAT score between black females and nonblack females? What would you need to do to test whether the difference is statistically significant?

4 An equation explaining chief executive officer salary is

$$\widehat{\log(salary)} = 4.59 + .257\log(sales) + .011\,roe + .158\,finance$$
$$(.30)\ (.032)\qquad (.004)\qquad (.089)$$
$$+ .181\,consprod - .283\,utility$$
$$(.085)\qquad (.099)$$
$$n = 209,\ R^2 = .357.$$

The data used are in CEOSAL1, where *finance*, *consprod*, and *utility* are binary variables indicating the financial, consumer products, and utilities industries. The omitted industry is transportation.

(i) Compute the approximate percentage difference in estimated salary between the utility and transportation industries, holding *sales* and *roe* fixed. Is the difference statistically significant at the 1% level?

(ii) Use equation (7.10) to obtain the exact percentage difference in estimated salary between the utility and transportation industries and compare this with the answer obtained in part (i).

(iii) What is the approximate percentage difference in estimated salary between the consumer products and finance industries? Write an equation that would allow you to test whether the difference is statistically significant.

5 In Example 7.2, let *noPC* be a dummy variable equal to one if the student does not own a PC, and zero otherwise.

(i) If *noPC* is used in place of *PC* in equation (7.6), what happens to the intercept in the estimated equation? What will be the coefficient on *noPC*? (*Hint:* Write $PC = 1 - noPC$ and plug this into the equation $\widehat{colGPA} = \hat{\beta}_0 + \hat{\delta}_0 PC + \hat{\beta}_1 hsGPA + \hat{\beta}_2 ACT$.)

(ii) What will happen to the *R*-squared if *noPC* is used in place of *PC*?

(iii) Should *PC* and *noPC* both be included as independent variables in the model? Explain.

6 To test the effectiveness of a job training program on the subsequent wages of workers, we specify the model

$$\log(wage) = \beta_0 + \beta_1 train + \beta_2 educ + \beta_3 exper + u,$$

where *train* is a binary variable equal to unity if a worker participated in the program. Think of the error term u as containing unobserved worker ability. If less able workers have a greater chance of being selected for the program, and you use an OLS analysis, what can you say about the likely bias in the OLS estimator of β_1? (*Hint:* Refer back to Chapter 3.)

7 In the example in equation (7.29), suppose that we define *outlf* to be one if the woman is out of the labor force, and zero otherwise.

(i) If we regress *outlf* on all of the independent variables in equation (7.29), what will happen to the intercept and slope estimates? (*Hint: inlf* $= 1 -$ *outlf*. Plug this into the population equation $inlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \ldots$ and rearrange.)

(ii) What will happen to the standard errors on the intercept and slope estimates?

(iii) What will happen to the *R*-squared?

8 Suppose you collect data from a survey on wages, education, experience, and gender. In addition, you ask for information about marijuana usage. The original question is: "On how many separate occasions last month did you smoke marijuana?"

(i) Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, "Smoking marijuana five more times per month is estimated to change wage by x%."

(ii) Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?

(iii) Suppose you think it is better to measure marijuana usage by putting people into one of four categories: nonuser, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage.

(iv) Using the model in part (iii), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.

(v) What are some potential problems with drawing causal inference using the survey data that you collected?

9 Let d be a dummy (binary) variable and let z be a quantitative variable. Consider the model

$$y = \beta_0 + \delta_0 d + \beta_1 z + \delta_1 d \cdot z + u;$$

this is a general version of a model with an interaction between a dummy variable and a quantitative variable. [An example is in equation (7.17).]

(i) Since it changes nothing important, set the error to zero, $u = 0$. Then, when $d = 0$ we can write the relationship between y and z as the function $f_0(z) = \beta_0 + \beta_1 z$. Write the same relationship when $d = 1$, where you should use $f_1(z)$ on the left-hand side to denote the linear function of z.

(ii) Assuming that $\delta_1 \neq 0$ (which means the two lines are not parallel), show that the value of z^* such that $f_0(z^*) = f_1(z^*)$ is $z^* = -\delta_0/\delta_1$. This is the point at which the two lines intersect [as in Figure 7.2 (b)]. Argue that z^* is positive if and only if δ_0 and δ_1 have opposite signs.

(iii) Using the data in TWOYEAR, the following equation can be estimated:

$$\widehat{\log(wage)} = 2.289 - .357\, female + .50\, totcoll + .030\, female \cdot totcoll$$
$$(0.011)\ \ (.015)\qquad\quad (.003)\qquad\quad (.005)$$
$$n = 6{,}763,\ R^2 = .202,$$

where all coefficients and standard errors have been rounded to three decimal places. Using this equation, find the value of *totcoll* such that the predicted values of log(*wage*) are the same for men and women.

(iv) Based on the equation in part (iii), can women realistically get enough years of college so that their earnings catch up to those of men? Explain.

10 For a child i living in a particular school district, let $voucher_i$ be a dummy variable equal to one if a child is selected to participate in a school voucher program, and let $score_i$ be that child's score on a subsequent standardized exam. Suppose that the participation variable, $voucher_i$, is completely randomized in the sense that it is independent of both observed and unobserved factors that can affect the test score.

(i) If you run a simple regression $score_i$ on $voucher_i$ using a random sample of size n, does the OLS estimator provide an unbiased estimator of the effect of the voucher program?

(ii) Suppose you can collect additional background information, such as family income, family structure (e.g., whether the child lives with both parents), and parents' education levels. Do you need to control for these factors to obtain an unbiased estimator of the effects of the voucher program? Explain.

(iii) Why should you include the family background variables in the regression? Is there a situation in which you would not include the background variables?

11 The following equations were estimated using the data in ECONMATH, with standard errors reported under coefficients. The average class score, measured as a percentage, is about 72.2; exactly 50% of the students are male; and the average of *colgpa* (grade point average at the start of the term) is about 2.81.

$$\widehat{score} = 32.31 + 14.32\, colgpa$$
$$(2.00)\ \ (0.70)$$
$$n = 856,\ R^2 = .329,\ \overline{R}^2 = .328.$$

$$\widehat{score} = 29.66 + 3.83\, male + 14.57\, colgpa$$
$$(2.04)\ \ (0.74)\qquad\quad (0.69)$$
$$n = 856,\ R^2 = .349,\ \overline{R}^2 = .348.$$

$$\widehat{score} = 30.36 + 2.47\, male + 14.33\, colgpa + 0.479\, male \cdot colgpa$$
$$(2.86)\ (3.96)\qquad (0.98)\qquad (1.383)$$
$$n = 856,\ R^2 = .349,\ \overline{R}^2 = .347.$$

$$\widehat{score} = 30.36 + 3.82\, male + 14.33\, colgpa + 0.479\, male \cdot (colgpa - 2.81)$$
$$(2.86)\ (0.74)\qquad (0.98)\qquad (1.383)$$
$$n = 856,\ R^2 = .349,\ \overline{R}^2 = .347.$$

(i) Interpret the coefficient on *male* in the second equation and construct a 95% confidence interval for β_{male}. Does the confidence interval exclude zero?

(ii) In the second equation, how come the estimate on *male* is so imprecise? Should we now conclude that there are no gender differences in *score* after controlling for *colgpa*? [*Hint*: You might want to compute an F statistic for the null hypothesis that there is no gender difference in the model with the interaction.]

(iii) Compared with the third equation, how come the coefficient on *male* in the last equation is so much closer to that in the second equation and just as precisely estimated?

Computer Exercises

C1 Use the data in GPA1 for this exercise.

(i) Add the variables *mothcoll* and *fathcoll* to the equation estimated in (7.6) and report the results in the usual form. What happens to the estimated effect of PC ownership? Is *PC* still statistically significant?

(ii) Test for joint significance of *mothcoll* and *fathcoll* in the equation from part (i) and be sure to report the p-value.

(iii) Add $hsGPA^2$ to the model from part (i) and decide whether this generalization is needed.

C2 Use the data in WAGE2 for this exercise.

(i) Estimate the model

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 tenure + \beta_4 married + \beta_5 black + \beta_6 south + \beta_7 urban + u$$

and report the results in the usual form. Holding other factors fixed, what is the approximate difference in monthly salary between blacks and nonblacks? Is this difference statistically significant?

(ii) Add the variables $exper^2$ and $tenure^2$ to the equation and show that they are jointly insignificant at even the 20% level.

(iii) Extend the original model to allow the return to education to depend on race and test whether the return to education does depend on race.

(iv) Again, start with the original model, but now allow wages to differ across four groups of people: married and black, married and nonblack, single and black, and single and nonblack. What is the estimated wage differential between married blacks and married nonblacks?

C3 A model that allows major league baseball player salary to differ by position is

$$\log(salary) = \beta_0 + \beta_1 years + \beta_2 gamesyr + \beta_3 bavg + \beta_4 hrunsyr + \beta_5 rbisyr + \beta_6 runsyr + \beta_7 fldperc + \beta_8 allstar + \beta_9 frstbase + \beta_{10} scndbase + \beta_{11} thrdbase + \beta_{12} shrtstop + \beta_{13} catcher + u,$$

where outfield is the base group.

(i) State the null hypothesis that, controlling for other factors, catchers and outfielders earn, on average, the same amount. Test this hypothesis using the data in MLB1 and comment on the size of the estimated salary differential.

(ii) State and test the null hypothesis that there is no difference in average salary across positions, once other factors have been controlled for.

(iii) Are the results from parts (i) and (ii) consistent? If not, explain what is happening.

C4 Use the data in GPA2 for this exercise.

(i) Consider the equation

$$colgpa = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + \beta_3 hsperc + \beta_4 sat + \beta_5 female + \beta_6 athlete + u,$$

where *colgpa* is cumulative college grade point average; *hsize* is size of high school graduating class, in hundreds; *hsperc* is academic percentile in graduating class; *sat* is combined SAT score; *female* is a binary gender variable; and *athlete* is a binary variable, which is one for student-athletes. What are your expectations for the coefficients in this equation? Which ones are you unsure about?

(ii) Estimate the equation in part (i) and report the results in the usual form. What is the estimated GPA differential between athletes and nonathletes? Is it statistically significant?

(iii) Drop *sat* from the model and reestimate the equation. Now, what is the estimated effect of being an athlete? Discuss why the estimate is different than that obtained in part (ii).

(iv) In the model from part (i), allow the effect of being an athlete to differ by gender and test the null hypothesis that there is no ceteris paribus difference between women athletes and women nonathletes.

(v) Does the effect of *sat* on *colgpa* differ by gender? Justify your answer.

C5 In Problem 2 in Chapter 4, we added the return on the firm's stock, *ros*, to a model explaining CEO salary; *ros* turned out to be insignificant. Now, define a dummy variable, *rosneg*, which is equal to one if $ros < 0$ and equal to zero if $ros \geq 0$. Use CEOSAL1 to estimate the model

$$\log(salary) = \beta_0 + \beta_1 \log(sales) + \beta_2 roe + \beta_3 rosneg + u.$$

Discuss the interpretation and statistical significance of $\hat{\beta}_3$.

C6 Use the data in SLEEP75 for this exercise. The equation of interest is

$$sleep = \beta_0 + \beta_1 totwrk + \beta_2 educ + \beta_3 age + \beta_4 age^2 + \beta_5 yngkid + u.$$

(i) Estimate this equation separately for men and women and report the results in the usual form. Are there notable differences in the two estimated equations?

(ii) Compute the Chow test for equality of the parameters in the sleep equation for men and women. Use the form of the test that adds *male* and the interaction terms *male*·*totwrk*, …, *male*·*yngkid* and uses the full set of observations. What are the relevant *df* for the test? Should you reject the null at the 5% level?

(iii) Now, allow for a different intercept for males and females and determine whether the interaction terms involving *male* are jointly significant.

(iv) Given the results from parts (ii) and (iii), what would be your final model?

C7 Use the data in WAGE1 for this exercise.

(i) Use equation (7.18) to estimate the gender differential when $educ = 12.5$. Compare this with the estimated differential when $educ = 0$.

(ii) Run the regression used to obtain (7.18), but with $female \cdot (educ - 12.5)$ replacing $female \cdot educ$. How do you interpret the coefficient on *female* now?

(iii) Is the coefficient on *female* in part (ii) statistically significant? Compare this with (7.18) and comment.

C8 Use the data in LOANAPP for this exercise. The binary variable to be explained is *approve*, which is equal to one if a mortgage loan to an individual was approved. The key explanatory variable is *white*, a dummy variable equal to one if the applicant was white. The other applicants in the data set are black and Hispanic.

To test for discrimination in the mortgage loan market, a linear probability model can be used:

$$approve = \beta_0 + \beta_1 white + other\ factors.$$

(i) If there is discrimination against minorities, and the appropriate factors have been controlled for, what is the sign of β_1?

(ii) Regress *approve* on *white* and report the results in the usual form. Interpret the coefficient on *white*. Is it statistically significant? Is it practically large?

(iii) As controls, add the variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortlat1*, *mortlat2*, and *vr*. What happens to the coefficient on *white*? Is there still evidence of discrimination against nonwhites?

(iv) Now, allow the effect of race to interact with the variable measuring other obligations as a percentage of income (*obrat*). Is the interaction term significant?

(v) Using the model from part (iv), what is the effect of being white on the probability of approval when $obrat = 32$, which is roughly the mean value in the sample? Obtain a 95% confidence interval for this effect.

C9 There has been much interest in whether the presence of 401(k) pension plans, available to many U.S. workers, increases net savings. The data set 401KSUBS contains information on net financial assets (*nettfa*), family income (*inc*), a binary variable for eligibility in a 401(k) plan (*e401k*), and several other variables.

(i) What fraction of the families in the sample are eligible for participation in a 401(k) plan?

(ii) Estimate a linear probability model explaining 401(k) eligibility in terms of income, age, and gender. Include income and age in quadratic form, and report the results in the usual form.

(iii) Would you say that 401(k) eligibility is independent of income and age? What about gender? Explain.

(iv) Obtain the fitted values from the linear probability model estimated in part (ii). Are any fitted values negative or greater than one?

(v) Using the fitted values $\widehat{e401k}_i$ from part (iv), define $\widetilde{e401k}_i = 1$ if $\widehat{e401k}_i \geq .5$ and $\widetilde{e401k}_i = 0$ if $\widehat{e401k}_i < .5$. Out of 9,275 families, how many are predicted to be eligible for a 401(k) plan?

(vi) For the 5,638 families not eligible for a 401(k), what percentage of these are predicted not to have a 401(k), using the predictor $\widetilde{e401k}_i$? For the 3,637 families eligible for a 401(k) plan, what percentage are predicted to have one? (It is helpful if your econometrics package has a "tabulate" command.)

(vii) The overall percent correctly predicted is about 64.9%. Do you think this is a complete description of how well the model does, given your answers in part (vi)?

(viii) Add the variable *pira* as an explanatory variable to the linear probability model. Other things equal, if a family has someone with an individual retirement account, how much higher is the estimated probability that the family is eligible for a 401(k) plan? Is it statistically different from zero at the 10% level?

C10 Use the data in NBASAL for this exercise.

(i) Estimate a linear regression model relating points per game to experience in the league and position (guard, forward, or center). Include experience in quadratic form and use centers as the base group. Report the results in the usual form.

(ii) Why do you not include all three position dummy variables in part (i)?

(iii) Holding experience fixed, does a guard score more than a center? How much more? Is the difference statistically significant?

(iv) Now, add marital status to the equation. Holding position and experience fixed, are married players more productive (based on points per game)?

(v) Add interactions of marital status with both experience variables. In this expanded model, is there strong evidence that marital status affects points per game?

(vi) Estimate the model from part (iv) but use assists per game as the dependent variable. Are there any notable differences from part (iv)? Discuss.

C11 Use the data in 401KSUBS for this exercise.

(i) Compute the average, standard deviation, minimum, and maximum values of *nettfa* in the sample.

(ii) Test the hypothesis that average *nettfa* does not differ by 401(k) eligibility status; use a two-sided alternative. What is the dollar amount of the estimated difference?

(iii) From part (ii) of Computer Exercise C9, it is clear that *e401k* is not exogenous in a simple regression model; at a minimum, it changes by income and age. Estimate a multiple linear regression model for *nettfa* that includes income, age, and *e401k* as explanatory variables. The income and age variables should appear as quadratics. Now, what is the estimated dollar effect of 401(k) eligibility?

(iv) To the model estimated in part (iii), add the interactions $e401k \cdot (age - 41)$ and $e401k \cdot (age - 41)^2$. Note that the average age in the sample is about 41, so that in the new model, the coefficient on *e401k* is the estimated effect of 401(k) eligibility at the average age. Which interaction term is significant?

(v) Comparing the estimates from parts (iii) and (iv), do the estimated effects of 401(k) eligibility at age 41 differ much? Explain.

(vi) Now, drop the interaction terms from the model, but define five family size dummy variables: *fsize1, fsize2, fsize3, fsize4,* and *fsize5*. The variable *fsize5* is unity for families with five or more members. Include the family size dummies in the model estimated from part (iii); be sure to choose a base group. Are the family dummies significant at the 1% level?

(vii) Now, do a Chow test for the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 inc^2 + \beta_3 age + \beta_4 age^2 + \beta_5 e401k + u$$

across the five family size categories, allowing for intercept differences. The restricted sum of squared residuals, SSR_r, is obtained from part (vi) because that regression assumes all slopes are the same. The unrestricted sum of squared residuals is $SSR_{ur} = SSR_1 + SSR_2 + \ldots + SSR_5$, where SSR_f is the sum of squared residuals for the equation estimated using only family size *f*. You should convince yourself that there are 30 parameters in the unrestricted model (5 intercepts plus 25 slopes) and 10 parameters in the restricted model (5 intercepts plus 5 slopes). Therefore, the number of restrictions being tested is $q = 20$, and the *df* for the unrestricted model is $9{,}275 - 30 = 9{,}245$.

C12 Use the data set in BEAUTY, which contains a subset of the variables (but more usable observations than in the regressions) reported by Hamermesh and Biddle (1994).

(i) Find the separate fractions of men and women that are classified as having above average looks. Are more people rated as having above average or below average looks?

(ii) Test the null hypothesis that the population fractions of above-average-looking women and men are the same. Report the one-sided p-value that the fraction is higher for women. (*Hint:* Estimating a simple linear probability model is easiest.)

(iii) Now estimate the model

$$\log(wage) = \beta_0 + \beta_1 belavg + \beta_2 abvavg + u$$

separately for men and women, and report the results in the usual form. In both cases, interpret the coefficient on *belavg*. Explain in words what the hypothesis H_0: $\beta_1 = 0$ against H_1: $\beta_1 < 0$ means, and find the p-values for men and women.

(iv) Is there convincing evidence that women with above average looks earn more than women with average looks? Explain.

(v) For both men and women, add the explanatory variables *educ*, *exper*, $exper^2$, *union*, *goodhlth*, *black*, *married*, *south*, *bigcity*, *smllcity*, and *service*. Do the effects of the "looks" variables change in important ways?

(vi) Use the SSR form of the Chow F statistic to test whether the slopes of the regression functions in part (v) differ across men and women. Be sure to allow for an intercept shift under the null.

C13 Use the data in APPLE to answer this question.

(i) Define a binary variable as $ecobuy = 1$ if $ecolbs > 0$ and $ecobuy = 0$ if $ecolbs = 0$. In other words, *ecobuy* indicates whether, at the prices given, a family would buy any ecologically friendly apples. What fraction of families claim they would buy ecolabeled apples?

(ii) Estimate the linear probability model

$$ecobuy = \beta_0 + \beta_1 ecoprc + \beta_2 regprc + \beta_3 faminc + \beta_4 hhsize + \beta_5 educ + \beta_6 age + u,$$

and report the results in the usual form. Carefully interpret the coefficients on the price variables.

(iii) Are the nonprice variables jointly significant in the LPM? (Use the usual F statistic, even though it is not valid when there is heteroskedasticity.) Which explanatory variable other than the price variables seems to have the most important effect on the decision to buy ecolabeled apples? Does this make sense to you?

(iv) In the model from part (ii), replace *faminc* with log(*faminc*). Which model fits the data better, using *faminc* or log(*faminc*)? Interpret the coefficient on log(*faminc*).

(v) In the estimation in part (iv), how many estimated probabilities are negative? How many are bigger than one? Should you be concerned?

(vi) For the estimation in part (iv), compute the percent correctly predicted for each outcome, $ecobuy = 0$ and $ecobuy = 1$. Which outcome is best predicted by the model?

C14 Use the data in CHARITY to answer this question. The variable *respond* is a dummy variable equal to one if a person responded with a contribution on the most recent mailing sent by a charitable organization. The variable *resplast* is a dummy variable equal to one if the person responded to the previous mailing, *avggift* is the average of past gifts (in Dutch guilders), and *propresp* is the proportion of times the person has responded to past mailings.

(i) Estimate a linear probability model relating *respond* to *resplast* and *avggift*. Report the results in the usual form, and interpret the coefficient on *resplast*.

(ii) Does the average value of past gifts seem to affect the probability of responding?

(iii) Add the variable *propresp* to the model, and interpret its coefficient. (Be careful here: an increase of one in *propresp* is the largest possible change.)

(iv) What happened to the coefficient on *resplast* when *propresp* was added to the regression? Does this make sense?

(v) Add *mailsyear*, the number of mailings per year, to the model. How big is its estimated effect? Why might this not be a good estimate of the causal effect of mailings on responding?

C15 Use the data in FERTIL2 to answer this question.

(i) Find the smallest and largest values of *children* in the sample. What is the average of *children*? Does any woman have exactly the average number of children?

(ii) What percentage of women have electricity in the home?

(iii) Compute the average of *children* for those without electricity and do the same for those with electricity. Comment on what you find. Test whether the population means are the same using a simple regression.

(iv) From part (iii), can you infer that having electricity "causes" women to have fewer children? Explain.

(v) Estimate a multiple regression model of the kind reported in equation (7.37), but add age^2, *urban*, and the three religious affiliation dummies. How does the estimated effect of having electricity compare with that in part (iii)? Is it still statistically significant?

(vi) To the equation in part (v), add an interaction between *electric* and *educ*. Is its coefficient statistically significant? What happens to the coefficient on *electric*?

(vii) The median and mode value for *educ* is 7. In the equation from part (vi), use the centered interaction term $electric \cdot (educ - 7)$ in place of $electric \cdot educ$. What happens to the coefficient on *electric* compared with part (vi)? Why? How does the coefficient on *electric* compare with that in part (v)?

C16 Use the data in CATHOLIC to answer this question.

(i) In the entire sample, what percentage of the students attend a Catholic high school? What is the average of *math*12 in the entire sample?

(ii) Run a simple regression of *math*12 on *cathhs* and report the results in the usual way. Interpret what you have found.

(iii) Now add the variables *lfaminc*, *motheduc*, and *fatheduc* to the regression from part (ii). How many observations are used in the regression? What happens to the coefficient on *cathhs*, along with its statistical significance?

(iv) Return to the simple regression of *math*12 on *cathhs*, but restrict the regression to observations used in the multiple regression from part (iii). Do any important conclusions change?

(v) To the multiple regression in part (iii), add interactions between *cathhs* and each of the other explanatory variables. Are the interaction terms individually or jointly significant?

(vi) What happens to the coefficient on *cathhs* in the regression from part (v). Explain why this coefficient is not very interesting.

(vii) Compute the average partial effect of *cathhs* in the model estimated in part (v). How does it compare with the coefficients on *cathhs* in parts (iii) and (v)?