CHAPTER 13

Dummy Dependent Variable Techniques

Until now, our discussion of dummy variables has been restricted to dummy independent variables. However, there are many important research topics for which the *dependent* variable is appropriately treated as a dummy, equal only to zero or one.

In particular, researchers analyzing consumer choice often must cope with dummy dependent variables (also called qualitative dependent variables). For example, how do high school students decide whether to go to college? What distinguishes Pepsi drinkers from Coke drinkers? How can we convince people to commute using public transportation instead of driving? For an econometric study of these topics, or of any topic that involves a *discrete* choice of some sort, the dependent variable is typically a dummy variable.

In the first two sections of this chapter, we'll present two frequently used ways to estimate equations that have dummy dependent variables: the linear probability method and the binomial logit model. In the last section, we'll briefly discuss two other useful dummy dependent variable techniques: the binomial probit model and the multinomial logit model.

13.1 The Linear Probability Model

13.1.1 What Is a Linear Probability Model?

The most obvious way to estimate a model with a dummy dependent variable is to run OLS on a typical linear econometric equation. A **linear proba-**

bility model is just that, a linear-in-the-coefficients equation used to explain a dummy dependent variable:

$$D_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{13.1}$$

where D_i is a dummy variable and the Xs, βs, and ε are typical independent variables, regression coefficients, and an error term, respectively.

For example, suppose you're interested in understanding why some state legislatures voted to ratify the Equal Rights Amendment (ERA) to the Constitution and others did not. In such a model, the appropriate dependent variable would be a dummy, for example D_i equal to one if the *i*th state ratified the ERA and equal to zero otherwise. If we hypothesize that states with a high percentage of females and a low percentage of Republicans would be likely to have approved the amendment, then a linear probability model of ERA voting by state would be:

$$D_i = \beta_0 + \beta_1 F_i + \beta_2 R_i + \epsilon_i$$

where: D_i = 1 if the *i*th state ratified the ERA, 0 otherwise
F_i = females as a percent of the *i*th state's population
R_i = Republicans as a percent of the *i*th state's registered voters

The term *linear probability model* comes from the fact that the right-hand side of the equation is linear, while the expected value of the left side is a probability. Let's discuss more thoroughly the concept that this equation measures a probability. It can be shown that the expected value of D_i equals the probability that D_i will equal one.[1] If we define P_i as the probability that D_i equals one, then this is the same as saying that the expected value of D_i equals P_i. Since Equation 13.1 specifies this choice as a function of X_{1i}, X_{2i}, this can be formally stated as:

$$E[D_i | X_{1i}, X_{2i}] = P_i \tag{13.2}$$

We can never observe the probability P_i, however, because it reflects the state of mind of a decision maker *before* a discrete choice is made. After a choice is made, we can observe only the outcome of that choice, and so the dependent

1. The expected value of a variable equals the sum of the products of each of the possible values the variable can take times the probability of that value occurring. If P_i is defined as the probability that D_i will equal one, then the probability that D_i will equal zero is $(1 - P_i)$, since D_i can take on only two values. Thus, the expected value of $D_i = P_i \cdot 1 + (1 - P_i) \cdot 0 = P_i$, the probability that D_i equals one.

variable D_i can take on the values of only zero or one. Thus, even though the expected value (P_i) can be anywhere from zero to one, we can only observe the two extremes (0 and 1) in our dependent variable (D_i).

13.1.2 Problems with the Linear Probability Model

Unfortunately, the use of OLS to estimate the coefficients of an equation with a dummy dependent variable encounters four major problems:

1. *The error term is not normally distributed.* Because the dependent variable takes on only two values, the error term is binomial, and Classical Assumption VII is violated. This flaw makes hypothesis testing unreliable.

2. *The error term is inherently heteroskedastic.* The variance of ϵ_i equals $P_i \cdot (1 - P_i)$, where P_i is the probability that D_i equals 1. Since P_i can vary from observation to observation, so too can the variance of ϵ_i. Thus the variance of ϵ_i is not constant, and Classical Assumption V is violated.

3. *$\bar{R}^2$ is not an accurate measure of overall fit.* For models with a dummy dependent variable, $\bar{R}^2$ tells us very little about how well the model explains the choices of the decision makers. To see why, take a look at Figure 13.1. D_i can equal only 1 or 0, but $\hat{D}_i$ must move in a continuous fashion from one extreme to the other. This means that $\hat{D}_i$ is likely to be quite different from D_i for some range of X_i. Thus, $\bar{R}^2$ is likely to be much lower than 1 even if the model actually does an exceptional job of explaining the choices involved. As a result, $\bar{R}^2$ (or R^2) should not be relied on as a measure of the overall fit of a model with a dummy dependent variable.

4. *$\hat{D}_i$ is not bounded by 0 and 1.* Since the expected value of D_i is a probability, we'd expect $\hat{D}_i$ to be limited to a range of 0 to 1. After all, the prediction that a probability equals 2.6 (or −2.6, for that matter) is almost meaningless. However, take another look at Equation 13.1. Depending on the values of the Xs and the $\hat{\beta}$s, the right-hand side might well be outside the meaningful range. For instance, if all the Xs and $\hat{\beta}$s in Equation 13.1 equal 2.0, then $\hat{D}_i$ equals 10.0, substantially greater than 1.0.

Luckily, there are potential solutions to the first three problems cited above. First, the nonnormality problem can be ignored in coefficient estimation

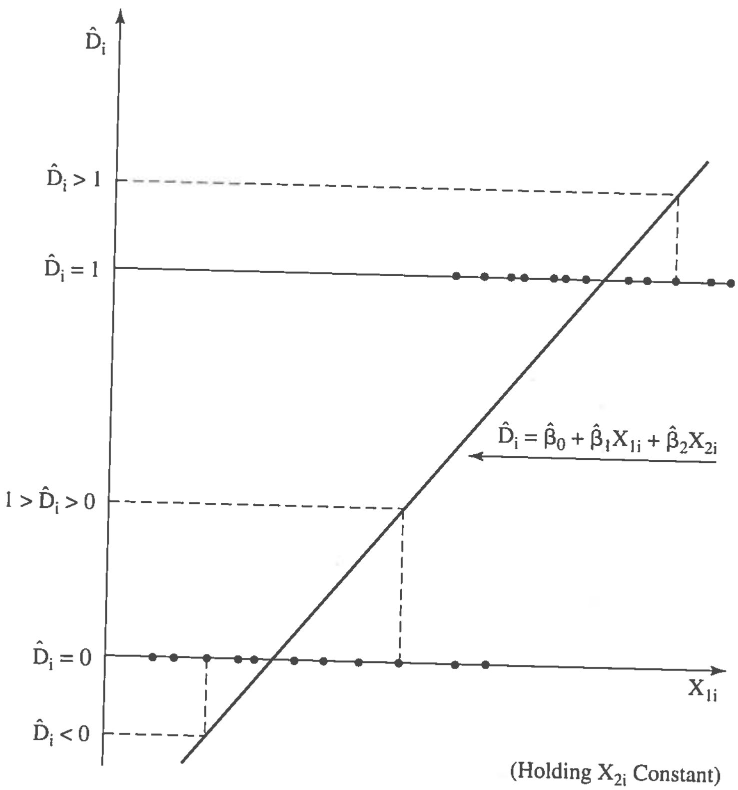

Figure 13.1 A Linear Probability Model

In a linear probability model, all the observed D_is equal either zero or one but $\hat{D}_i$ moves linearly from one extreme to the other. As a result, $\bar{R}^2$ is often quite low even if the model does an excellent job of explaining the decision maker's choice. In addition, exceptionally large or small values of X_{1i} (holding X_{2i} constant), can produce values of $\hat{D}_i$ outside the meaningful range of zero to one.

because Classical Assumption VII is not used to prove the Gauss–Markov Theorem.

Second, a solution to the heteroskedasticity problem is to use Weighted Least Squares. Recall that we know that the variance of ϵ_i equals $P_i \cdot (1 - P_i)$. As shown in Chapter 10, if we were to divide the equation through by $\sqrt{P_i \cdot (1 - P_i)}$, then the variance of the error term would no longer be heteroskedastic. Although we don't know the actual value of P_i, we do know that P_i equals the expected value of D_i. Thus, if we estimate Equation 13.1 and obtain $\hat{D}_i$, we can use $\hat{D}_i$ as an estimate of P_i. To run Weighted Least Squares, we'd then calculate:

$$Z_i = \sqrt{\hat{D}_i \cdot (1 - \hat{D}_i)} \tag{13.3}$$

divide Equation 13.1 by Z_i, and estimate the new equation with OLS.[2]

Third, an alternative to $\bar{R}^2$ is $\mathbf{R}_p^2$, the percentage of the observations in the sample that a particular estimated equation explains correctly. To use this approach, consider a $\hat{D}_i \geq .5$ to predict that $D_i = 1$ and a $\hat{D}_i < .5$ to predict that $D_i = 0$, compare this prediction with the actual D_i, and then compute:

$$R_p^2 = \frac{\text{number of observations "predicted" correctly}}{\text{total number of observations (n)}} \tag{13.4}$$

Since R_p^2 is not used universally, we'll calculate and discuss both $\bar{R}^2$ and R_p^2 throughout this chapter.

For most researchers, therefore, the major difficulty with the linear probability model is the unboundedness of the predicted D_is. Take another look at Figure 13.1 for a graphical interpretation of the situation. Because of the linear relationship between the X_is and $\hat{D}_i$, $\hat{D}_i$ can fall well outside the relevant range of 0 to 1. Using the linear probability model, despite this unboundedness problem, may not cause insurmountable difficulties. In particular, the signs and general significance levels of the estimated coefficients of the linear probability model are often similar to those of the alternatives we will discuss later in this chapter.

One simplistic way to get around the unboundedness problem is to assign $\hat{D}_i = 1.0$ to all values of $\hat{D}_i$ above one and $\hat{D}_i = 0.0$ to all negative values. This approach copes with the problem by ignoring it, since an observation for which the linear probability model predicts a probability of 2.0 has been judged to be more likely to be equal to 1.0 than an observation for which the model predicts a 1.0, and yet they are lumped together. What is needed is a systematic method of forcing the $\hat{D}_i$s to range from 0 to 1 in a smooth and meaningful fashion. We'll present such a method, the binomial logit, in Section 13.2.

13.1.3 An Example of a Linear Probability Model

Before moving on to investigate the logit, however, let's take a look at an example of a linear probability model: a disaggregate study of the labor force participation of women.

2. Note that when $\hat{D}_i$ is quite close to 0 or 1, $\hat{D}_i \cdot (1 - \hat{D}_i)$ is extremely small and X_i/Z_i is huge. Also note that when $\hat{D}_i$ is outside the 0–1 range, $\hat{D}_i \cdot (1 - \hat{D}_i)$ is negative and Z_i is undefined. See R. G. McGilvray, "Estimating the Linear Probability Function," *Econometrica*, 1970, pp. 775–776. Some researchers arbitrarily drop all such observations to avoid the resulting estimation problems. We think that a better alternative is to impose an arbitrary floor, say 0.02, on $\hat{D}_i \cdot (1 - \hat{D}_i)$. Either way, WLS is not efficient.

A person is defined as being in the labor force if she either has a job or is actively looking for a job. Thus, a disaggregate (cross-sectional by person) study of women's labor force participation is appropriately modeled with a dummy dependent variable:

$$D_i = 1 \text{ if the } i\text{th woman has or is looking for a job,}$$
$$0 \text{ otherwise (not in the labor force)}$$

A review of the literature[3] reveals that there are many potentially relevant independent variables. Two of the most important are the marital status and the number of years of schooling of the woman. The expected signs for the coefficients of these variables are fairly straightforward, since a woman who is unmarried and well educated is much more likely to be in the labor force than her opposite:

$$D_i = f(\overset{-}{M_i}, \overset{+}{S_i}) + \epsilon_i$$

where: M_i = 1 if the ith woman is married and 0 otherwise
S_i = the number of years of schooling of the ith woman

The data are presented in Table 13.1. The sample size is limited to 30 in order to make it easier for readers to estimate this example on their own. Unfortunately, such a small sample will make hypothesis testing fairly unreliable. Table 13.1 also includes the age of the ith woman for use in Exercises 8 and 9. Another typically used variable, O_i = other income available to the ith woman, is not available for this sample, introducing possible omitted variable bias.

If we choose a linear functional form for both independent variables, we've got a linear probability model:

$$D_i = \beta_0 + \beta_1 M_i + \beta_2 S_i + \epsilon_i \tag{13.5}$$

where ϵ_i is an inherently heteroskedastic error term with variance = $P_i \cdot (1 - P_i)$. If we now estimate Equation 13.5 with the data on the labor force participation of women from Table 13.1, we obtain (standard errors in parentheses):

3. See James P. Smith and Michael P. Ward, "Time-Series Growth in the Female Labor Force," *Journal of Labor Economics*, 1985, pp. 559–590. Smith and Ward include a number of estimated logits in their work.

TABLE 13.1 DATA ON THE LABOR FORCE PARTICIPATION OF WOMEN

Observation #	D_i	M_i	A_i	S_i	$\hat{D}_i$	$\hat{D}_i(1-\hat{D}_i)$	Z_i
1	1.0	0.0	31.0	16.0	1.20	0.020	0.141
2	1.0	1.0	34.0	14.0	0.63	0.231	0.481
3	1.0	1.0	41.0	16.0	0.82	0.146	0.382
4	0.0	0.0	67.0	9.0	0.55	0.247	0.497
5	1.0	0.0	25.0	12.0	0.83	0.139	0.374
6	0.0	1.0	58.0	12.0	0.45	0.247	0.497
7	1.0	0.0	45.0	14.0	1.01	0.020	0.141
8	1.0	0.0	55.0	10.0	0.64	0.228	0.478
9	0.0	0.0	43.0	12.0	0.83	0.139	0.374
10	1.0	0.0	55.0	8.0	0.45	0.248	0.498
11	1.0	0.0	25.0	11.0	0.73	0.192	0.439
12	1.0	0.0	41.0	14.0	1.01	0.020	0.141
13	0.0	1.0	62.0	12.0	0.45	0.247	0.497
14	1.0	1.0	51.0	13.0	0.54	0.248	0.498
15	0.0	1.0	39.0	9.0	0.17	0.141	0.376
16	1.0	0.0	35.0	10.0	0.64	0.228	0.478
17	1.0	1.0	40.0	14.0	0.63	0.231	0.481
18	0.0	1.0	43.0	10.0	0.26	0.194	0.440
19	0.0	1.0	37.0	12.0	0.45	0.247	0.497
20	1.0	0.0	27.0	13.0	0.92	0.069	0.263
21	1.0	0.0	28.0	14.0	1.01	0.020	0.141
22	1.0	1.0	48.0	12.0	0.45	0.247	0.497
23	0.0	1.0	66.0	7.0	−0.01	0.020	0.141
24	0.0	1.0	44.0	11.0	0.35	0.229	0.479
25	0.0	1.0	21.0	12.0	0.45	0.247	0.497
26	1.0	1.0	40.0	10.0	0.26	0.194	0.440
27	1.0	0.0	41.0	15.0	1.11	0.020	0.141
28	0.0	1.0	23.0	10.0	0.26	0.194	0.440
29	0.0	1.0	31.0	11.0	0.35	0.229	0.479
30	1.0	1.0	44.0	12.0	0.45	0.247	0.497

Note: $\hat{D}_i(1 - \hat{D}_i)$ has been set equal to 0.02 for all values of $\hat{D}_i$ less than 0.02 or greater than 0.98.
filename WOMEN13 (In this datafile D is represented by J.)

$$\hat{D}_i = -0.28 - \underset{(0.15)}{0.38M_i} + \underset{(0.03)}{0.09S_i} \tag{13.6}$$

$$n = 30 \qquad \bar{R}^2 = .32 \qquad R_p^2 = .80$$

How do these results look? At first glance, they look terrific. Despite the small sample and the possible bias due to omitting O_i, both independent variables have estimated coefficients that are significant in the expected direction. In addition, the $\bar{R}^2$ of .32 is fairly high for a linear probability model (since D_i equals only 0 or 1, it's almost impossible to get a $\bar{R}^2$ much higher than .70). Further evidence of good fit is the fairly high R^2_p of .80, meaning that 80 percent of the choices were predicted "correctly" by Equation 13.6.

We need to be careful when we interpret the estimated coefficients in Equation 13.6, however. The slope coefficient in a linear probability model represents the change in the probability that D_i equals one caused by a one-unit increase in the independent variable (holding the other independent variables constant). Viewed in this context, do the estimated coefficients still make economic sense? The answer is yes: the probability of a woman participating in the labor force falls by 38 percent if she is married (holding constant her schooling). In addition, each year of schooling increases the probability of labor force participation by 9 percent (holding constant marital status).

However, Equation 13.6 is far from perfect. Recall that the error term is inherently heteroskedastic, that hypothesis testing is unreliable in such a small sample, that $\bar{R}^2$ is not an accurate measure of fit, and that one or more relevant variables have been omitted. While we can do nothing about some of these problems, there is a solution to the heteroskedasticity problem: Weighted Least Squares (WLS).

To use WLS, we take the $\hat{D}_i$ from Equation 13.6 and calculate $Z_i = \sqrt{\hat{D}_i \cdot (1 - \hat{D}_i)}$, as in Equation 13.3 [taking care to impose a floor of 0.02 on $\hat{D}_i \cdot (1 - \hat{D}_i)$ as suggested in footnote 2.] We then divide Equation 13.5 through by Z_i, obtaining:

$$D_i/Z_i = \alpha_0 + \beta_0(1/Z_i) = \beta_1 M_i/Z_i + \beta_2 S_i/Z_i + u_i \qquad (13.7)$$

where u_i is a nonheteroskedastic error term $= \epsilon_i/Z_i$. Note that since Z_i is not an independent variable in Equation 13.6, we have chosen to add α_0, a constant term, to Equation 13.7 to avoid violating Classical Assumption II (as discussed in Chapter 9). If we now estimate Equation 13.7 with OLS, we obtain:

$$\widehat{D_i/Z_i} = 0.18 - 0.21(1/Z_i) - \underset{(0.15)}{0.39}M_i/Z_i + \underset{(0.02)}{0.08}S_i/Z_i \qquad (13.8)$$

$$n = 30 \qquad \bar{R}^2 = .86 \qquad R^2_p = .83$$

Let's compare Equations 13.8 and 13.6. Surprisingly, the estimated standard errors of the estimated coefficients are almost identical in the two equations, indicating that at least for this sample the impact of the heteroskedasticity is minimal. The high $\bar{R}^2$ comes about in part because dividing the entire equation by the same number (Z_i) causes some spurious correlation, especially when some of the Z_i values are quite small. As evidence, note that R_p^2 is only slightly higher in Equation 13.8 than in Equation 13.6 even though $\bar{R}^2$ jumped from .32 to .86.

To make it easier for the reader to reproduce the WLS procedure, the values for $\hat{D}_i \cdot (1 - \hat{D}_i)$ and Z_i have been included in Table 13.1. Also included are the $\hat{D}_i$s from Equation 13.6; note that $\hat{D}_i$ is indeed often outside the meaningful range of 0 and 1, causing most of the problems cited earlier. To attack this problem of the unboundedness of $\hat{D}_i$, however, we need a new estimation technique, so let's take a look at one.

13.2 The Binomial Logit Model

To avoid the possibility that a prediction of D_i might be outside the probability interval of 0 to 1, we no longer model D_i directly. Instead, we model the ratio $D_i/(1 - D_i)$. This ratio is the likelihood, or odds,[4] of obtaining a successful outcome ($D_i = 1$). If we take the log of this ratio, we have the left side of the equation that has become the standard approach to dummy dependent variable analysis: the binomial logit.

13.2.1 What Is the Binomial Logit?

The **binomial logit** is an estimation technique for equations with dummy dependent variables that avoids the unboundedness problem of the linear probability model by using a variant of the cumulative logistic function:

$$\ln\left(\frac{D_i}{[1 - D_i]}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \qquad (13.9)$$

where D_i is a dummy variable. The expected value of D_i continues to be P_i, the probability that the *i*th person will make the choice described by $D_i = 1$. Consequently, the dependent variable of Equation 13.9 can be thought of as the log of the odds that the choice in question will be made.

How does the logit avoid the unboundedness problem of the linear probability model? It turns out that *both* sides of Equation 13.9 are unbounded. To see this, note that if $D_i = 1$, then the left-side of Equation 13.9 becomes:

4. *Odds* refer to the ratio of the number of times a choice will be made divided by the number of times it will not. In today's world, odds are used most frequently with respect to sporting events, such as horse races, on which bets are made.

$$\ln\left(\frac{D_i}{[1 - D_i]}\right) = \ln\left(\frac{1}{0}\right) = \infty \tag{13.10}$$

Similarly, if $D_i = 0$:

$$\ln\left(\frac{D_i}{[1 - D_i]}\right) = \ln\left(\frac{0}{1}\right) = -\infty \tag{13.11}$$

because the log of zero approaches negative infinity.

Are the $\hat{D}_i$s produced by a logit now limited by zero and one? The answer is yes, but to see why we need to solve Equation 13.9 for D_i. It can be shown[5] that Equation 13.9 is equivalent to:

$$D_i = \frac{1}{1 + e^{-[\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i]}} \tag{13.12}$$

Take a close look at Equation 13.12. What is the largest that $\hat{D}_i$ can be? Well, if $\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$ equals infinity, then:

$$\hat{D}_i = \frac{1}{1 + e^{-\infty}} = \frac{1}{1} = 1 \tag{13.13}$$

because e to the minus infinity equals zero. What's the smallest that $\hat{D}_i$ can be? If $\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$ equals minus infinity, then:

$$\hat{D}_i = \frac{1}{1 + e^{\infty}} = \frac{1}{\infty} = 0 \tag{13.14}$$

Thus, $\hat{D}_i$ is bounded by one and zero. As can be seen in Figure 13.2, $\hat{D}_i$ approaches one and zero very slowly (asymptotically). The binomial logit model therefore avoids the major problem that the linear probability model encounters in dealing with dummy dependent variables. In addition, the logit is quite satisfying to most researchers because it turns out that real-world data often are described well by S-shape patterns like that in Figure 13.2.

Logits cannot be estimated using OLS. Instead, we use **maximum likelihood**, an iterative estimation technique that is especially useful for equations that are nonlinear in the coefficients. Maximum likelihood (ML) estimation is inherently different from least squares in that it chooses coefficient esti-

5. Those interested in this proof should see Exercise 4.

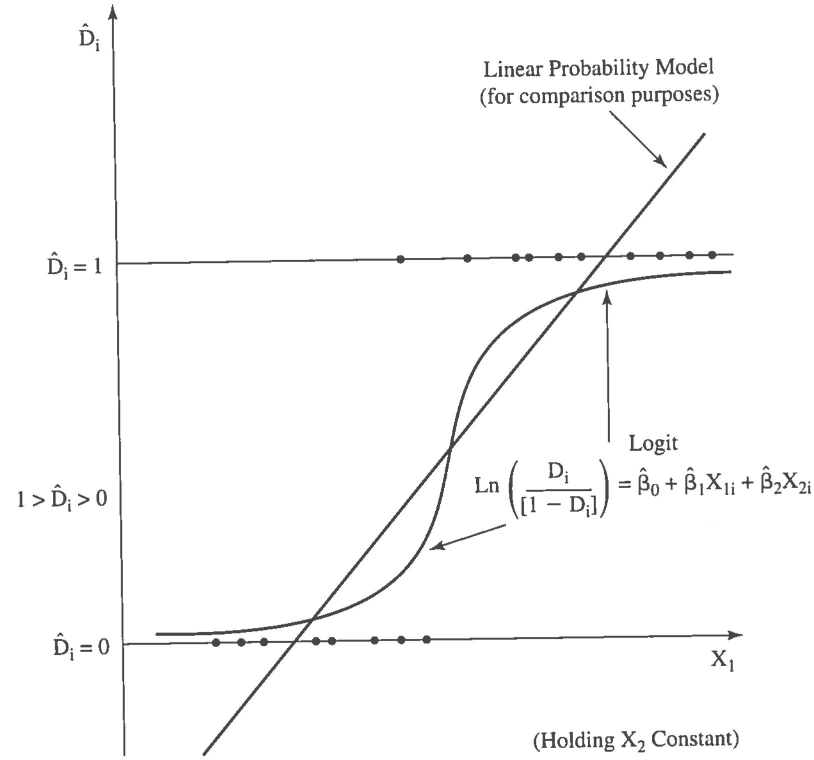

Figure 13.2 **$\hat{D}_i$ Is Bounded by Zero and One in a Binomial Logit Model**

In a binomial logit model, $\hat{D}_i$ is nonlinearly related to X_1, so even exceptionally large or small values of X_{1i}, holding X_{2i} constant, will not produce values of $\hat{D}_i$ outside the meaningful range of zero to one.

mates that *maximize the likelihood* of the sample data set being observed.[6] Interestingly, OLS and ML estimates are not necessarily different; for a linear equation that meets the Classical Assumptions (including the normality assumption), ML estimates are identical to the OLS ones.

One of the reasons that maximum likelihood is used is that ML has a number of desirable large sample properties; ML is consistent and asymptotically efficient (unbiased and minimum variance for large samples). With very

6. Actually, the ML program chooses coefficient estimates that maximize the log of the probability (or likelihood) of observing the particular set of values of the dependent variable in the sample ($Y_1, Y_2, \ldots, Y_n$) for a given set of Xs. For more on maximum likelihood, see Robert S. Pindyck and Daniel L. Rubinfeld, *Economic Models and Economic Forecasts* (New York: McGraw-Hill, 1998), pp. 51–53 and 329–330.

large samples, ML has the added advantage of producing normally distributed coefficient estimates, allowing the use of typical hypothesis testing techniques. As a result, sample sizes for logits should be substantially larger than for linear regressions. Some researchers aim for samples of 500 or more.

It's also important to make sure that a logit sample contains a reasonable representation of both alternative choices. For instance, if 98 percent of a sample chooses alternative A and 2 percent chooses B, a random sample of 500 would have only 10 observations that choose B. In such a situation, our estimated coefficients would be overly reliant on the characteristics of those 10 observations. A better technique would be to disproportionately sample from those who choose B. It turns out that using different sampling rates for subgroups within the sample does not cause bias in the slope coefficients of a logit model,[7] even though it might do so in a linear regression.

The maximum likelihood computer program is applied to a logit that has been solved for D_i (Equation 13.12), not to a logit solved for the log of the odds (Equation 13.9). This distinction is necessary because, as shown earlier, the left-hand side of Equation 13.9 can be observed only as infinity and negative infinity. Such infinite values make calculations quite difficult.

Once the binomial logit has been estimated, hypothesis testing and econometric analysis can be undertaken in much the same way as for linear equations. When interpreting coefficients, however, be careful to recall that they represent the impact of a one-unit increase in the independent variable in question, holding the other explanatory variables constant, on the log of the odds of a given choice, not on the probability itself (as was the case with the linear probability model).

For instance, β_1 in Equation 13.9 measures the impact of a one-unit increase in X_1 on the log of the odds of a given choice, holding X_2 constant. As a result, the absolute sizes of estimated logit coefficients tend to be quite different from the absolute sizes of estimated linear probability model coefficients for the same variables. Interestingly, as mentioned above, the signs and significance levels of the estimated coefficients from the two models often are similar.

Measuring the overall fit, however, is not quite as straightforward. Recall from Chapter 7 that since the functional form of the dependent variable has been changed, $\bar{R}^2$ cannot be used to compare the fit of a logit with an otherwise comparable linear probability model. One way around this difficulty is to use the quasi-R^2 approach of Chapter 7 (a nonlinear estimate of R^2) to com-

7. The constant term, however, needs to be adjusted. Multiply $\hat{\beta}_0$ by $[\ln(p_1) - \ln(p_2)]$, where p_1 is the proportion of the observations chosen if $D_i = 1$ and p_2 is the proportion of the observations chosen if $D_i = 0$. See G. S. Maddala, *Limited-Dependent and Qualitative Variables in Econometrics* (Cambridge: Cambridge University Press, 1983), pp. 90–91.

pare the two fits. However, this quasi-R^2 shares the general faults inherent in using $\bar{R}^2$ with equations with dummy dependent variables. A better approach might be to use the percentage of correct predictions, R_p^2, from Equation 13.4.

To allow a fairly simple comparison between the logit and the linear probability model, let's estimate a logit on the same women's labor force participation data that we used in the previous section. The OLS estimate of that model, Equation 13.6, was:

$$\begin{aligned} \hat{D}_i &= -0.28 - \underset{(0.15)}{0.38M_i} + \underset{(0.03)}{0.09S_i} \\ n &= 30 \qquad \bar{R}^2 = .32 \qquad R_p^2 = .80 \end{aligned} \tag{13.6}$$

where: D_i = 1 if the ith woman is in the labor force, 0 otherwise
M_i = 1 if the ith woman is married, 0 otherwise
S_i = the number of years of schooling of the ith woman

If we estimate a logit on the same data (from Table 13.1) and the same independent variables, we obtain[8]:

$$\begin{aligned} \widehat{\ln\left(\frac{D_i}{[1 - D_i]}\right)} &= -5.89 - \underset{(1.18)}{2.59M_i} + \underset{(0.31)}{0.69S_i} \\ &\qquad\quad t = -2.19 \qquad 2.19 \\ n = 30 \qquad R_p^2 &= .80 \qquad \text{iterations} = 5 \end{aligned} \tag{13.15}$$

Let's compare Equations 13.6 and 13.15. As expected, the signs and general significance of the slope coefficients are the same. Note, however, that the actual sizes of the coefficients are quite different because the dependent variable is different. The coefficient of M changes from −0.38 to −2.59! Despite these differences, the overall fits are roughly comparable, especially after taking account of the different dependent variables and estimation techniques. In this example, then, the two estimation procedures differ mainly in that the logit does not produce $\hat{D}_i$s outside the range of zero and one.

However, if the size of the sample in this example is too small for a linear probability model, it certainly is too small for a logit, making any in-depth analysis of Equation 13.15 problematic. Instead, we're better off finding an example with a much larger sample.

8. Equation 13.15 has the log of the odds as its dependent variable, but the maximum likelihood computer estimation program that produces the β estimates uses a functional form with D_i as the dependent variable (similar to Equation 13.12).

13.2.2 An Example of the Use of the Binomial Logit

For a more complete example of the binomial digit, let's look at a model of the probability of passing the California State Department of Motor Vehicles drivers' license test. To obtain a license, each driver must past a written and a behind-the-wheel test. Even though the tests are scored from 0 to 100, all that matters is that you pass and get your license.

Since the test requires some boning up on traffic and safety laws, driving students have to decide how much time to spend studying. If they don't study enough, they waste time because they have to retake the test. If they study too much, however, they also waste time, because there's no bonus for scoring above the minimum, especially since there is no evidence that doing well on the test has much to do with driving well after the test (this, of course, might be worth its own econometric study).

Recently, two students decided to collect data on test takers in order to build an equation explaining whether someone passed the Department of Motor Vehicles test. They hoped that the model, and in particular the estimated coefficient of study time, would help them decide how much time to spend studying for the test. (Of course, it took more time to collect the data and run the model than it would have taken to memorize the entire traffic code, but that's another story.)

After reviewing the literature, choosing variables, and hypothesizing signs, the students realized that the appropriate functional form was a binomial logit because their dependent variable was a dummy variable:

$$D_i = \begin{cases} 1 \text{ if the } i\text{th test taker passed the test on the first try} \\ 0 \text{ if the } i\text{th test taker failed the test on the first try} \end{cases}$$

Their hypothesized equation was:

$$D_i = f(\overset{+}{A_i}, \overset{+}{H_i}, \overset{+}{E_i}, \overset{+}{C_i}) + \epsilon_i$$

where:
A_i = the age of the ith test taker
H_i = the number of hours the ith test taker studied (usually less than one hour!)
E_i = a dummy variable equal to 1 if the ith test taker's primary language is English, 0 otherwise
C_i = a dummy variable equal to 1 if the ith test taker has any college experience, 0 otherwise

After collecting data from 480 test takers, the students estimated the following equation:

$$\widehat{\ln\left(\frac{D_i}{[1 - D_i]}\right)} = -1.18 + \underset{\substack{(0.009) \\ t = 1.23}}{0.011A_i} + \underset{\substack{(0.54) \\ 4.97}}{2.70H_i} + \underset{\substack{(0.34) \\ 4.65}}{1.62E_i} + \underset{\substack{(0.99) \\ 4.00}}{3.97C_i}$$

$$n = 480 \qquad R_p^2 = .74 \qquad \text{iterations} = 5 \qquad (13.16)$$

Note how similar these results look to a typical linear regression result. All the estimated coefficients have the expected signs, and all but one are significantly different from zero. Remember, though, that the coefficient estimates have different meanings than in a linear regression model. For example, 2.70 is the impact of an extra hour of studying on the log of the odds of passing the test, holding constant the other three independent variables. Note that R_p^2 is .74, indicating that the equation correctly "predicted" almost three quarters of the sample based on nothing but the four variables in Equation 13.16.

And what about the two students? Did the equation help them? How much did they end up deciding to study? They found that given their ages, their college experience, and their English-speaking backgrounds, the expected value of $\hat{D}_i$ for each of them was quite high, even if H_i was set equal to zero. So what did they actually do? They studied for a half hour "just to be on the safe side" and passed with flying colors, having devoted more time to passing the test than anyone else in the history of the state.

13.3 Other Dummy Dependent Variable Techniques

Although the binomial logit is the most frequently used estimation technique for equations with dummy dependent variables, it's by no means the only one. In this section, we'll mention two alternatives, the binomial probit and the multinomial logit, that are useful in particular circumstances. Our main goal is to briefly describe these estimation techniques, not to cover them in any detail.[9]

9. For more, see G. S. Maddala, *Limited Dependent Variables and Qualitative Variables in Econometrics* (Cambridge: Cambridge University Press, 1983) and T. Amemiya, "Qualitative Response Models: A Survey," *Journal of Economic Literature*, 1981, pp. 1483–1536. These surveys also cover additional techniques, like the Tobit model, that are useful with bounded dependent variables or other special situations.

13.3.1 The Binomal Probit Model

The **binomial probit model** is an estimation technique for equations with dummy dependent variables that avoids the unboundedness problem of the linear probability model by using a variant of the cumulative normal distribution.

$$P_i = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Z_i} e^{-s^2/2}\, ds \tag{13.17}$$

where: P_i = the probability that the dummy variable $D_i = 1$
$Z_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$
s = a standardized normal variable

As different as this probit looks from the logit that we examined in the previous section, it can be rewritten to look quite familiar:

$$Z_i = F^{-1}(P_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \tag{13.18}$$

where F^{-1} is the inverse of the normal cumulative distribution function. Probit models typically are estimated by applying maximum likelihood techniques to the model in the form of Equation 13.17, but the results often are presented in the format of Equation 13.18.

The fact that both the logit and the probit are cumulative distributive functions means that the two have similar properties. For example, a graph of the probit looks almost exactly like the logit in Figure 13.2. In addition, the probit has the same requirement of a fairly large sample before hypothesis testing becomes meaningful. Finally, $\bar{R}^2$ continues to be of questionable value as a measure of overall fit.

From a researcher's point of view, the biggest differences between the two models are that the probit is based on the cumulative normal distribution and that the probit estimation procedure uses more computer time than does the logit. Since the probit is similar to the logit and is more expensive to run, why would you ever estimate one? The answer is that since the probit is based on the normal distribution, it's quite theoretically appealing (because many economic variables are normally distributed). With extremely large samples, this advantage falls away, since maximum likelihood procedures can be shown to be asymptotically normal under fairly general conditions.

For an example of a probit, let's estimate one on the same women's labor force participation data we used in the previous logit and linear probability examples (standard errors in parentheses):

$$\hat{Z}_i = \widehat{F^{-1}(P_i)} = -3.44 - \underset{(0.62)}{1.44M_i} + \underset{(0.17)}{0.40S_i} \tag{13.19}$$
$$n = 30 \qquad R_p^2 = .80 \qquad \text{iterations} = 5$$

Compare this result with Equation 13.15 from the previous section. Note that except for a slight difference in the scale of the coefficients, the logit and probit models provide virtually identical results in this example.

13.3.2 The Multinomial Logit Model

In many cases, there are more than two qualitative choices available. In some cities, for instance, a commuter has a choice of car, bus, or subway for the trip to work. How could we build and estimate a model of choosing from more than two different alternatives?

One answer is to hypothesize that choices are made sequentially and to model a multichoice decision as a series of binary decisions. For example, we might hypothesize that the commuter would first decide whether or not to drive to work, and we could build a binary model of car versus public transportation. For those commuters who choose public transportation, the next step would be to choose whether to take the bus or the subway, and we could build a second binary model of that choice. This method, called a **sequential binary logit,** is cumbersome and at times unrealistic, but it does allow a researcher to use a binary technique to model an inherently multichoice decision.

If a decision between multiple alternatives is truly made simultaneously, a better approach is to build a multinomial logit model of the decision. A **multinomial logit model** is an extension of the binomial logit technique that allows several discrete alternatives to be considered at the same time. If there are n different alternatives, we need n − 1 dummy variables to describe the choice, with each dummy equalling one only when that particular alternative is chosen. For example, D_{1i} would equal one if the *i*th person chose alternative number 1 and would equal zero otherwise. As before, the probability that D_{1i} is equal to one, P_{1i}, cannot be observed.

In a multinomial logit, one alternative is selected as the "base" alternative, and then each other possible choice is compared to this base alternative with a logit equation. A key distinction is that the dependent variable of these equations is the log of the odds of the *i*th alternative being chosen *compared to the base alternative:*

$$\ln\left(\frac{P_{1i}}{P_{bi}}\right)$$

where: P_{1i} = the probability of the ith person choosing the first alternative
P_{bi} = the probability of the ith person choosing the base alternative

If there are n alternatives, there should be n − 1 different logit equations in the multinomial logit model system, because the coefficients of the last equation can be calculated from the coefficients of the first n − 1 equations. (If you know that A/C = 6 and B/C = 2, then you can calculate that A/B = 3.) For example, if n = 3, as in the commuter-work-trip example cited above, and the base alternative is taking the bus, then a multinomial logit model would have a system of two equations:

$$\ln\left(\frac{P_{si}}{P_{bi}}\right) = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} \tag{13.20}$$

$$\ln\left(\frac{P_{ci}}{P_{bi}}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{3i} \tag{13.21}$$

where s = subway, c = car, and b = bus.

The definitions of the independent variables (and therefore the meanings of their coefficients) are unusual in a multinomial logit. Some of the Xs are characteristics of the decision maker (like the income of the ith commuter). The coefficients of these variables represent the *difference* between the impact of income on the probability of choosing one mode and the impact of income on the probability of choosing the base mode. For example, in Equation 13.20, if X_1 is income, the coefficient α_1 is the impact of an extra dollar of income on the probability of taking the subway to work *minus* the impact of an extra dollar of income on the probability of taking the bus to work (holding X_2 constant).

Xs that aren't characteristics of the decision maker are usually characteristics of the alternative (like travel time for one of the possible modes of travel). A variable that measures a characteristic of an alternative in a multinomial logit model should be *defined* as the difference between the characteristics for the two modes. For example, if the second independent variable in our model is travel time to work, X_2 should be defined as the travel time to work by subway *minus* the travel time to work by bus. The coefficients of such characteristics of the alternatives measure the impact of a unit of time on the ratio of the probabilities (holding X_1 constant). For practice with the meanings of the independent variables and their coefficients in a multinomial logit model, see Exercise 11.

The multinomial logit system has all the basic properties of the binomial logit but with two additional complications in estimation. First, Equations

13.20 and 13.21 are estimated simultaneously,[10] so the iterative nonlinear maximum likelihood procedure used to estimate the system is more costly than for the binomial logit. Second, the relationship between the error terms in the equations (ϵ_{si} and ϵ_{ci}) must be strictly accounted for by using a GLS procedure, a factor that also complicates the estimation procedure.[11]

13.4 Summary

1. A linear probability model is a linear-in-the-coefficients equation used to explain a dummy dependent variable (D_i). The expected value of D_i is the probability that D_i equals one (P_i).

2. The estimation of a linear probability model with OLS encounters four major problems:
   a. The error term is not normally distributed.
   b. The error term is inherently heteroskedastic.
   c. $\bar{R}^2$ is not an accurate measure of overall fit.
   d. The expected value of D_i is not limited by 0 and 1.

3. When measuring the overall fit of equations with dummy dependent variables, an alternative to $\bar{R}^2$ is R^2_p, the percentage of the observations in the sample that a particular estimated equation would have explained correctly.

4. The binomial logit is an estimation technique for equations with dummy dependent variables that avoids the unboundedness problem of the linear probability model by using a variant of the cumulative logistic function:

$$\ln\left(\frac{D_i}{[1 - D_i]}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

5. The binomial logit is best estimated using the maximum likelihood technique and a large sample. A slope coefficient from a logit measures the impact of a one-unit increase of the independent variable in

10. As with the binomial logit, the maximum likelihood computer package doesn't estimate these precise equations. Instead, it estimates versions of Equations 13.20 and 13.21 that are similar to Equation 13.12.

11. For an interesting and yet accessible example of the estimation of a multinomial logit model, see Kang H. Park and Peter M. Kerr, "Determinants of Academic Performance: A Multinomial Logit Approach," *The Journal of Economic Education*, 1990, pp. 101–111.

question (holding the other explanatory variables constant) on the log of the odds of a given choice.

6. The binomial probit model is an estimation technique for equations with dummy dependent variables that uses the cumulative normal distribution function. The binomial probit has properties quite similar to the binomial logit except that it takes more computer time to estimate than a logit and is based on the normal distribution.

7. The multinomial logit model is an extension of the binomial logit that allows more than two discrete alternatives to be considered simultaneously. One alternative is chosen as a base alternative, and then each other possible choice is compared to that base alternative with a logit equation.

Exercises

(Answers to even-numbered exercises are in Appendix A.)

1. Write the meaning of each of the following terms without referring to the book (or your notes), and compare your definition with the version in the text for each:
   a. linear probability model
   b. R_p^2
   c. binomial logit model
   d. log of the odds
   e. binomial probit model
   f. sequential binary model
   g. multinomial logit model

2. On graph paper, plot each of the following linear probability models. For what range of X_1 is $1 < \hat{D}_i$? How about $\hat{D}_i < 0$?
   a. $\hat{D}_i = 0.3 + 0.1X_i$
   b. $\hat{D}_i = 3.0 - 0.2X_i$
   c. $\hat{D}_i = -1.0 + 0.3X_i$

3. Bond ratings are letter ratings (Aaa = best) assigned to firms that issue debt. These ratings measure the quality of the firm from the point of view of the likelihood of repayment of the bond. Suppose you've been hired by an arbitrage house that wants to predict *Moody's Bond Ratings* before they're published in order to buy bonds whose ratings are going to improve. In particular, suppose your firm wants to distin-

guish between A-rated bonds (high quality) and B-rated bonds (medium quality) and has collected a data set of 200 bonds with which to estimate a model. As you arrive on the job, your boss is about to buy bonds based on the results of the following model (standard errors in parentheses):

$$\hat{Y}_i = 0.70 + \underset{(0.05)}{0.05P_i} + \underset{(0.02)}{0.05PV_i} - \underset{(0.002)}{0.020D_i}$$

$$\bar{R}^2 = .69 \qquad DW = 0.50 \qquad n = 200$$

where: Y_i = 1 if the rating of the *i*th bond = A, 0 otherwise
P_i = the profit rate of the firm that issues the *i*th bond
PV_i = the standard deviation of P_i over the last 5 years
D_i = the ratio of debt to total capitalization of the firm that issued the *i*th bond

a. What econometric problems, if any, exist in this equation?
b. What suggestions would you have for a rerun of this equation with a different specification?
c. Suppose that your boss rejects your suggestions, saying, "This is the real world, and I'm sure that my model will forecast bond ratings just as well as yours will." How would you respond? (*Hint:* Saying, "Okay, boss, you win," is sure to keep your job for you, but it won't get much credit on this question.)

4. Show that the logistic function, $D = 1/(1 + e^{-Z})$, is indeed equivalent to the binomial logit model, $\ln[D/(1 - D)] = Z$, where $Z = \beta_0 + \beta_1X_1 + \beta_2X_2 + \epsilon$.

5. Plot each of the following binomial logit models. For what range of X_i is $1 < D_i$? How about $D_i < 0$? (*Hint:* When you finish, compare your answers to those for Exercise 2 above.)
   a. $\ln[D_i/(1 - D_i)] = 0.3 + 0.1X_i$
   b. $\ln[D_i/(1 - D_i)] = 3.0 - 0.2X_i$
   c. $\ln[D_i/(1 - D_i)] = -1.0 + 0.3X_i$

6. R. Amatya[12] estimated the following logit model of birth control for 1,145 continuously married women aged 35 to 44 in Nepal:

12. Ramesh Amatya, "Supply-Demand Analysis of Differences in Contraceptive Use in Seven Asian Nations, Late 1970s" (paper presented at the Annual Meetings of the Western Economic Association, 1988, Los Angeles).

$$\widehat{\ln\left(\frac{D_i}{[1 - D_i]}\right)} = -4.47 + 2.03WN_i + 1.45ME_i$$
$$\qquad\qquad\qquad\qquad (0.36) \qquad\quad (0.14)$$
$$\qquad\qquad\qquad t = 5.64 \qquad\quad 10.36$$

where: D_i = 1 if the *i*th woman has ever used a recognized form of birth control, 0 otherwise

WN_i = 1 if the *i*th woman wants no more children, 0 otherwise

ME_i = number of methods of birth control known to the *i*th woman

a. Explain the theoretical meaning of the coefficients for WN and ME. How would your answer differ if this were a linear probability model?

b. Do the signs, sizes, and significance of the estimated slope coefficients meet your expectations? Why or why not?

c. What is the theoretical significance of the constant term in this equation?

d. If you could make one change in the specification of this equation, what would it be? Explain your reasoning.

7. What happens if we define a dummy dependent variable over a range other than zero to one? For example, suppose that in the research cited above, Amatya had defined D_i as being equal to 2 if the *i*th woman had ever used birth control and zero otherwise.
   a. What would happen to the size and theoretical meaning of the estimated logit coefficients? Would they stay the same? Would they change? (If so, how?)
   b. How would your answers to part a above change if Amatya had estimated a linear probability model instead of a binomial logit?

8. Return to our data on women's labor force participation and consider the possibility of adding A_i, the age of the *i*th woman, to the equation. Be careful when you develop your expected sign and functional form because the expected impact of age on labor force participation is difficult to pin down. For instance, some women drop out of the labor force when they get married, but others continue working even while they're raising their children. Still others work until they get married, stay at home to have children, and then return to the work force once the children reach school age. Malcolm Cohen et al., for example, found the age of a woman to be relatively unimportant in

determining labor force participation, except for women who were 65 and older and were likely to have retired.[13] The net result for our model is that age appears to be a theoretically irrelevant variable. A possible exception, however, is a dummy variable equal to one if the *i*th woman is 65 or over and zero otherwise.

a. Look over the data set in Table 13.1. What problems do you see with adding an independent variable equal to one if the *i*th woman is 65 or older and zero otherwise?
b. If you go ahead and add the dummy implied above to Equation 13.15 and reestimate the model, you obtain the equation below. Which equation do you prefer, Equation 13.15 or the one below? Explain your answer.

$$\widehat{\ln\left(\frac{D_i}{[1 - D_i]}\right)} = -5.89 - \underset{\substack{(1.18) \\ t = -2.19}}{2.59M_i} + \underset{\substack{(0.31) \\ 2.19}}{0.69S_i} - \underset{\substack{(0.30) \\ -0.01}}{0.03AD_i}$$

$$n = 30 \qquad R_p^2 = .80 \qquad \text{iterations} = 5$$

where: $AD_i = 1$ if the age of the *i*th woman is > 65, 0 otherwise

9. To get practice in actually estimating your own linear probability, logit, and probit equations, test the possibility that age (A_i) is actually a relevant variable in our women's labor force participation model. That is, take the data from Table 13.1 and estimate each of the following equations. Then use our specification criteria to compare your equation with the parallel version in the text (without A_i). Explain why you do or do not think that age is a relevant variable.
   a. the linear probability model $D = f(M,A,S)$
   b. the logit $D = f(M,A,S)$
   c. the probit $D=f(M,A,S)$

10. An article published in a book edited by A. Kouskoulaf and B. Lytle[14] presents coefficients from an estimated logit model of the choice between the car and public transportation for the trip to work in Boston. All three public transportation modes in Boston (bus, subway, and train, of which train is the most preferred) were lumped together as a

13. Malcolm Cohen, Samuel A. Rea, Jr. and Robert I. Lerman, *A Micro Model of Labor Supply* (Washington, D.C.: U.S. Bureau of Labor Statistics, 1970), p. 212.

14. "The Use of the Multinomial Logit in Transportation Analysis," in A. Kouskoulaf and B. Lytle, eds. *Urban Housing and Transportation* (Detroit: Wayne State University, 1975), pp. 87–90.

single alternative to the car in a binomial logit model. The dependent variable was the log of the odds of taking public transportation for the trip to work, so the first coefficient implies that as income rises, the log of the odds of taking public transportation falls, and so on.

Independent Variable	**Coefficient**
Family income (9 categories with 1 = low and 9 = high)	−0.12
Number employed in the family	−1.09
Out-of-pocket costs (cents)	−3.16
Wait time (tenths of minutes)	0.18
Walk time (tenths of minutes)	−0.03
In-vehicle travel time (tenths of minutes)	−0.01

The last four variables are defined as the difference between the value of the variable for taking public transportation and its value for taking the car.

a. Do the signs of the estimated coefficients agree with your prior expectations? Which one(s) differ?
b. The transportation literature hypothesizes that people would rather spend time traveling in a vehicle than waiting for or walking to that vehicle. Do the sizes of the estimated coefficients of time support this hypothesis?
c. Since trains run relatively infrequently, the researchers set wait time for train riders fairly high. Most trains run on known schedules, however, so the average commuter learns that schedule and attempts to hold down wait time. Does this fact explain any of the unusual results indicated in your answers to parts a and b above?

11. Suppose that you want to build a multinomial logit model of how students choose which college to attend. For the sake of simplicity, let's assume that there are only four colleges to choose from: your college (c), the state university (u), the local junior college (j), and the nearby private liberal arts college (a). Further assume that everyone agrees that the important variables in such a model are the family income (Y) of each student, the average SAT scores of each college (SAT), and the tuition (T) of each college.
    a. How many equations should there be in such a multinomial logit system?
    b. If your college is the base, write out the definition of the dependent variable for each equation.