CHAPTER 11

Regression with a Binary Dependent Variable

Two people, identical but for their race, walk into a bank and apply for a mortgage, a large loan so that each can buy an identical house. Does the bank treat them the same way? Are they both equally likely to have their mortgage application accepted? By law, they must receive identical treatment. But whether they actually do is a matter of great concern among bank regulators.

Loans are made and denied for many legitimate reasons. For example, if the proposed loan payments take up most or all of the applicant's monthly income, a loan officer might justifiably deny the loan. Also, even loan officers are human and they can make honest mistakes, so the denial of a single minority applicant does not prove anything about discrimination. Many studies of discrimination thus look for statistical evidence of discrimination, that is, evidence contained in large data sets showing that whites and minorities are treated differently.

But how, precisely, should one check for statistical evidence of discrimination in the mortgage market? A start is to compare the fraction of minority and white applicants who were denied a mortgage. In the data examined in this chapter, gathered from mortgage applications in 1990 in the Boston, Massachusetts, area, 28% of black applicants were denied mortgages but only 9% of white applicants were denied. But this comparison does not really answer the question that opened this chapter because the black applicants and the white applicants were not necessarily "identical but for their race." Instead, we need a method for comparing rates of denial, *holding other applicant characteristics constant*.

This sounds like a job for multiple regression analysis—and it is, but with a twist. The twist is that the dependent variable—whether the applicant is denied—is binary. In Part II, we regularly used binary variables as regressors, and they caused no particular problems. But when the dependent variable is binary, things are more difficult: What does it mean to fit a line to a dependent variable that can take on only two values, 0 and 1?

The answer to this question is to interpret the regression function as a conditional probability. This interpretation is discussed in Section 11.1, and it allows us to apply the multiple regression models from Part II to binary dependent variables. Section 11.1 goes over this "linear probability model." But the predicted probability interpretation also suggests that alternative, nonlinear regression models can do a better job modeling these probabilities. These methods, called "probit" and "logit" regression, are discussed in Section 11.2. Section 11.3, which is optional, discusses the method used to estimate the coefficients of the probit and logit regressions, the method of

maximum likelihood estimation. In Section 11.4, we apply these methods to the Boston mortgage application data set to see whether there is evidence of racial bias in mortgage lending.

The binary dependent variable considered in this chapter is an example of a dependent variable with a limited range; in other words, it is a **limited dependent variable**. Models for other types of limited dependent variables—for example, dependent variables that take on multiple discrete values—are surveyed in Appendix 11.3.

11.1 Binary Dependent Variables and the Linear Probability Model

Whether a mortgage application is accepted or denied is one example of a binary variable. Many other important questions also concern binary outcomes. What is the effect of a tuition subsidy on an individual's decision to go to college? What determines whether a teenager takes up smoking? What determines whether a country receives foreign aid? What determines whether a job applicant is successful? In all these examples, the outcome of interest is binary: The student does or does not go to college, the teenager does or does not take up smoking, a country does or does not receive foreign aid, the applicant does or does not get a job.

This section discusses what distinguishes regression with a binary dependent variable from regression with a continuous dependent variable and then turns to the simplest model to use with binary dependent variables, the linear probability model.

Binary Dependent Variables

The application examined in this chapter is whether race is a factor in denying a mortgage application; the binary dependent variable is whether a mortgage application is denied. The data are a subset of a larger data set compiled by researchers at the Federal Reserve Bank of Boston under the Home Mortgage Disclosure Act (HMDA) and relate to mortgage applications filed in the Boston, Massachusetts, area in 1990. The Boston HMDA data are described in Appendix 11.1.

Mortgage applications are complicated. During the period covered by these data, the decision to approve a loan application typically was made by a bank loan officer. The loan officer must assess whether the applicant will make his or her loan payments. One important piece of information is the size of the required loan payments relative to the applicant's income. As anyone who has borrowed money knows, it is much easier to make payments that are 10% of your income than 50%! We therefore begin by looking at the relationship between two variables: the binary dependent variable *deny*, which equals 1 if the mortgage application was denied and equals 0 if it was accepted, and the continuous variable *P/I ratio*, which is the ratio of the applicant's anticipated total monthly loan payments to his or her monthly income.

FIGURE 11.1 Scatterplot of Mortgage Application Denial and the Payment-to-Income Ratio

Mortgage applicants with a high ratio of debt payments to income (*P/I ratio*) are more likely to have their application denied (*deny* = *1* if denied; *deny* = 0 if approved). The linear probability model uses a straight line to model the probability of denial, conditional on the *P/I ratio*.

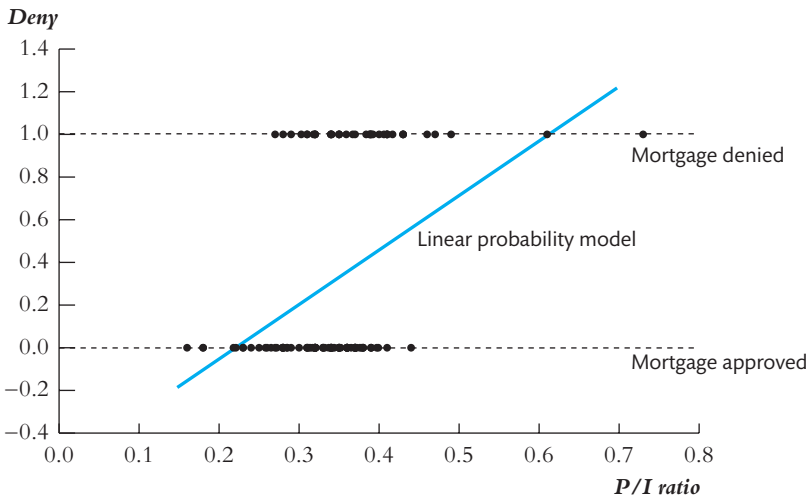


Figure 11.1 presents a scatterplot of *deny* versus *P/I ratio* for 127 of the 2380 observations in the data set. (The scatterplot is easier to read using this subset of the data.) This scatterplot looks different from the scatterplots of Part II because the variable *deny* is binary. Still, it seems to show a relationship between *deny* and *P/I ratio*: Few applicants with a payment-to-income ratio less than 0.3 have their application denied, but most applicants with a payment-to-income ratio exceeding 0.4 are denied.

This positive relationship between *P/I ratio* and *deny* (the higher the *P/I ratio*, the greater the fraction of denials) is summarized in Figure 11.1 by the OLS regression line estimated using these 127 observations. As usual, this line plots the predicted value of *deny* as a function of the regressor, the payment-to-income ratio. For example, when *P/I ratio* = 0.3, the predicted value of deny is 0.20. But what, precisely, does it mean for the predicted value of the binary variable *deny* to be 0.20?

The key to answering this question—and more generally to understanding regression with a binary dependent variable—is to interpret the regression as modeling the *probability* that the dependent variable equals 1. Thus the predicted value of 0.20 is interpreted as meaning that, when *P/I ratio* is 0.3, the probability of denial is estimated to be 20%. Said differently, if there were many applications with *P/I ratio* = 0.3, then 20% of them would be denied.

This interpretation follows from two facts. First, from Part II, the population regression function is the expected value of Y given the regressors, $E(Y|X_1, \ldots, X_k)$. Second, from Section 2.2, if Y is a 0–1 binary variable, its expected value (or mean) is the probability that $Y = 1$; that is, $E(Y) = 0 \times Pr(Y = 0) + 1 \times Pr(Y = 1) = Pr(Y = 1)$. In the regression context, the expected value is conditional on the value of the regressors, so the probability is conditional on X. Thus for a binary variable,

$E(Y|X_1, \ldots, X_k) = \Pr(Y = 1|X_1, \ldots, X_k)$. In short, for a binary dependent variable, the predicted value from the population regression is the probability that $Y = 1$ given X.

The linear multiple regression model applied to a binary dependent variable is called the linear probability model: *linear* because it is a straight line and *probability model* because it models the probability that the dependent variable equals 1 (in our example, the probability of loan denial).

The Linear Probability Model

The **linear probability model** is the name for the multiple regression model of Part II when the dependent variable is binary rather than continuous. Because the dependent variable Y is binary, the population regression function corresponds to the probability that the dependent variable equals 1 given X. The population coefficient β_1 on a regressor X is the *change in the probability* that $Y = 1$ associated with a *unit change* in X. Similarly, the OLS predicted value, $\hat{Y}_i$, computed using the estimated regression function, is the predicted probability that the dependent variable equals 1, and the OLS estimator $\hat{\beta}_1$ estimates the change in the probability that $Y = 1$ associated with a unit change in X.

Almost all of the tools of Part II carry over to the linear probability model. The coefficients can be estimated by OLS. Ninety-five percent confidence intervals can be formed as ± 1.96 standard errors, hypotheses concerning several coefficients can be tested using the F-statistic discussed in Chapter 7, and interactions between variables can be modeled using the methods of Section 8.3. Because the errors of the linear probability model are always heteroskedastic (Exercise 11.8), it is essential that heteroskedasticity-robust standard errors be used for inference.

One tool that does not carry over is the R^2. When the dependent variable is continuous, it is possible to imagine a situation in which the R^2 equals 1: All the data lie exactly on the regression line. This is impossible when the dependent variable is binary unless the regressors are also binary. Accordingly, the R^2 is not a particularly useful statistic here. We return to measures of fit in the next section.

The linear probability model is summarized in Key Concept 11.1.

Application to the Boston HMDA data. The OLS regression of the binary dependent variable, *deny*, against the payment-to-income ratio, *P/I ratio*, estimated using all 2380 observations in our data set is

$$\widehat{deny} = \underset{(0.032)}{-0.080} + \underset{(0.098)}{0.604}\, P/I\ ratio. \tag{11.1}$$

The estimated coefficient on *P/I ratio* is positive, and the population coefficient is statistically significantly different from 0 at the 1% level (the t-statistic is 6.13). Thus applicants with higher debt payments as a fraction of income are more likely to have their application denied. This coefficient can be used to compute the predicted

KEY CONCEPT 11.1

The Linear Probability Model

The linear probability model is the linear multiple regression model,

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, \tag{11.2}$$

applied to a binary dependent variable Y_i. Because Y is binary, $E(Y|X_1, X_2, \ldots, X_k) = \Pr(Y = 1|X_1, X_2, \ldots, X_k)$, so for the linear probability model,

$$\Pr(Y = 1|X_1, X_2, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k.$$

The regression coefficient β_1 is the difference in the probability that $Y = 1$ associated with a unit difference in X_1, holding constant the other regressors, and so forth for $\beta_2, \ldots, \beta_k$. The regression coefficients can be estimated by OLS, and the usual (heteroskedasticity-robust) OLS standard errors can be used for confidence intervals and hypothesis tests.

change in the probability of denial given a change in the regressor. For example, according to Equation (11.1), if *P/I ratio* increases by 0.1, the probability of denial increases by $0.604 \times 0.1 \cong 0.060$—that is, by 6.0 percentage points.

The estimated linear probability model in Equation (11.1) can be used to compute predicted denial probabilities as a function of *P/I ratio*. For example, if projected debt payments are 30% of an applicant's income, *P/I ratio* is 0.3, and the predicted value from Equation (11.1) is $-0.080 + 0.604 \times 0.3 = 0.101$. That is, according to this linear probability model, an applicant whose projected debt payments are 30% of income has a probability of 10.1% that his or her application will be denied. [This is different from the probability of 20% based on the regression line in Figure 11.1 because that line was estimated using only 127 of the 2380 observations used to estimate Equation (11.1).]

What is the effect of race on the probability of denial, holding constant the *P/I ratio*? To keep things simple, we focus on differences between black applicants and white applicants. To estimate the effect of race, holding constant *P/I ratio*, we augment Equation (11.1) with a binary regressor that equals 1 if the applicant is black and equals 0 if the applicant is white. The estimated linear probability model is

$$\widehat{deny} = \underset{(0.029)}{-0.091} + \underset{(0.089)}{0.559}\, P/I\ ratio + \underset{(0.025)}{0.177}\, black. \tag{11.3}$$

The coefficient on *black*, 0.177, indicates that an African American applicant has a 17.7% higher probability of having a mortgage application denied than a white applicant, holding constant their payment-to-income ratio. This coefficient is significant at the 1% level (the *t*-statistic is 7.11).

Taken literally, this estimate suggests that there might be racial bias in mortgage decisions, but such a conclusion would be premature. Although the payment-to-income ratio plays a role in the loan officer's decision, so do many other factors, such as the applicant's earning potential and his or her credit history. If any of these variables is correlated with the regressors *black* given the *P/I ratio*, its omission from Equation (11.3) will cause omitted variable bias. Thus we must defer any conclusions about discrimination in mortgage lending until we complete the more thorough analysis in Section 11.3.

Shortcomings of the linear probability model. The linearity that makes the linear probability model easy to use is also its major flaw. Because probabilities cannot exceed 1, the effect on the probability that $Y = 1$ of a given change in X must be nonlinear: Although a change in *P/I ratio* from 0.3 to 0.4 might have a large effect on the probability of denial, once *P/I ratio* is so large that the loan is very likely to be denied, increasing *P/I ratio* further will have little effect. In contrast, in the linear probability model, the effect of a given change in *P/I ratio* is constant, which leads to predicted probabilities in Figure 11.1 that drop below 0 for very low values of *P/I ratio* and exceed 1 for high values! But this is nonsense: A probability cannot be less than 0 or greater than 1. This nonsensical feature is an inevitable consequence of the linear regression. To address this problem, we introduce new nonlinear models specifically designed for binary dependent variables, the probit and logit regression models.

11.2 Probit and Logit Regression

Probit and **logit**[1] regression are nonlinear regression models specifically designed for binary dependent variables. Because a regression with a binary dependent variable Y models the probability that $Y = 1$, it makes sense to adopt a nonlinear formulation that forces the predicted values to be between 0 and 1. Because cumulative probability distribution functions (c.d.f.'s) produce probabilities between 0 and 1 (Section 2.1), they are used in logit and probit regressions. Probit regression uses the standard normal c.d.f. Logit regression, also called **logistic regression**, uses the logistic c.d.f.

Probit Regression

Probit regression with a single regressor. The probit regression model with a single regressor X is

$$\Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X), \tag{11.4}$$

where Φ is the cumulative standard normal distribution function (tabulated in Appendix Table 1).

[1]Pronounced prō-bit and lō-jit.

For example, suppose that Y is the binary mortgage denial variable (*deny*), X is the payment-to-income ratio (*P/I ratio*), $\beta_0 = -2$, and $\beta_1 = 3$. What then is the probability of denial if *P/I ratio* $= 0.4$? According to Equation (11.4), this probability is $\Phi(\beta_0 + \beta_1 P/I\ ratio) = \Phi(-2 + 3P/I\ ratio) = \Phi(-2 + 3 \times 0.4) = \Phi(-0.8)$. According to the cumulative normal distribution table (Appendix Table 1), $\Phi(-0.8) = \Pr(Z \leq -0.8) = 21.2\%$. That is, when *P/I ratio* is 0.4, the predicted probability that the application will be denied is 21.2%, computed using the probit model with the coefficients $\beta_0 = -2$ and $\beta_1 = 3$.

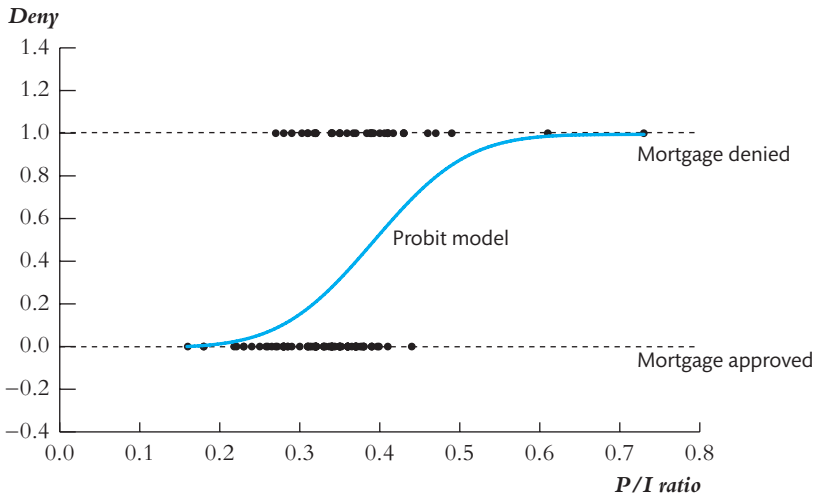
In the probit model, the term $\beta_0 + \beta_1 X$ plays the role of "z" in the cumulative standard normal distribution table in Appendix Table 1. Thus the calculation in the previous paragraph can, equivalently, be done by first computing the "z-value," $z = \beta_0 + \beta_1 X = -2 + 3 \times 0.4 = -0.8$, and then looking up the probability in the tail of the normal distribution to the left of $z = -0.8$, which is 21.2%.

The probit coefficient β_1 in Equation (11.4) is the difference in the z-value associated with a unit difference in X. If β_1 is positive, a greater value for X increases the z-value and thus increases the probability that $Y = 1$; if β_1 is negative, a greater value for X decreases the probability that $Y = 1$. Although the effect of X on the z-value is linear, its effect on the probability is nonlinear. Thus in practice the easiest way to interpret the coefficients of a probit model is to compute the predicted probability, or the change in the predicted probability, for one or more values of the regressors. When there is just one regressor, the predicted probability can be plotted as a function of X.

Figure 11.2 plots the estimated regression function produced by the probit regression of *deny* on *P/I ratio* for the 127 observations in the scatterplot. The

FIGURE 11.2 Probit Model of the Probability of Denial Given *P/I Ratio*

The probit model uses the cumulative normal distribution function to model the probability of denial given the payment-to-income ratio or, more generally, to model $Pr(Y = 1|X)$. Unlike the linear probability model, the probit conditional probabilities are always between 0 and 1.



estimated probit regression function has a stretched "S" shape: It is nearly 0 and flat for small values of *P/I ratio*, it turns and increases for intermediate values, and it flattens out again and is nearly 1 for large values. For small values of the payment-to-income ratio, the probability of denial is small. For example, for $P/I\ ratio = 0.2$, the estimated probability of denial based on the estimated probit function in Figure 11.2 is $\Pr(deny = 1|P/I\ ratio = 0.2) = 2.1\%$. When $P/I\ ratio = 0.3$, the estimated probability of denial is 16.1%. When $P/I\ ratio = 0.4$, the probability of denial increases sharply to 51.9%, and when $P/I\ ratio = 0.6$, the denial probability is 98.3%. According to this estimated probit model, for applicants with high payment-to-income ratios, the probability of denial is nearly 1.

Probit regression with multiple regressors. In all the regression problems we have studied so far, leaving out a determinant of Y that is correlated with the included regressors results in omitted variable bias. Probit regression is no exception. In linear regression, the solution is to include the additional variable as a regressor. This is also the solution to omitted variable bias in probit regression.

The probit model with multiple regressors extends the single-regressor probit model by adding regressors to compute the z-value. Accordingly, the probit population regression model with two regressors, X_1 and X_2, is

$$Pr(Y = 1|X_1, X_2) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2). \tag{11.5}$$

For example, suppose that $\beta_0 = -1.6$, $\beta_1 = 2$, and $\beta_2 = 0.5$. If $X_1 = 0.4$ and $X_2 = 1$, the z-value is $z = -1.6 + 2 \times 0.4 + 0.5 \times 1 = -0.3$. So the probability that $Y = 1$ given $X_1 = 0.4$ and $X_2 = 1$ is $\Pr(Y = 1|X_1 = 0.4, X_2 = 1) = \Phi(-0.3) = 38\%$.

Effect of a change in X. In general, the regression model can be used to determine the expected change in Y arising from a change in X. When Y is binary, its conditional expectation is the conditional probability that it equals 1, so the expected change in Y arising from a change in X is the change in the probability that $Y = 1$.

Recall from Section 8.1 that, when the population regression function is a nonlinear function of X, this expected change is estimated in three steps: First, compute the predicted value at the original value of X using the estimated regression function; next, compute the predicted value at the changed value of X, $X + \Delta X$; finally, compute the difference between the two predicted values. This procedure is summarized in Key Concept 8.1. As emphasized in Section 8.1, this method *always* works for computing predicted effects of a change in X, no matter how complicated the nonlinear model. When applied to the probit model, the method of Key Concept 8.1 yields the estimated effect on the probability that $Y = 1$ of a change in X.

The probit regression model, predicted probabilities, and estimated effects are summarized in Key Concept 11.2.

The Probit Model, Predicted Probabilities, and Estimated Effects

KEY CONCEPT 11.2

The population probit model with multiple regressors is

$$\Pr(Y = 1 | X_1, X_2, \ldots, X_k) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k), \quad (11.6)$$

where the dependent variable Y is binary, Φ is the cumulative standard normal distribution function, and X_1, X_2, and so on are regressors. The model is best interpreted by computing predicted probabilities and the effect of a change in a regressor.

The predicted probability that $Y = 1$, given values of $X_1, X_2, \ldots, X_k$, is calculated by computing the z-value, $z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$, and then looking up this z-value in the normal distribution table (Appendix Table 1).

The coefficient β_1 is the difference in the z-value arising from a unit difference in X_1, holding constant $X_2, \ldots, X_k$.

The effect on the predicted probability of a change in a regressor is computed by (1) computing the predicted probability for the initial value of the regressor, (2) computing the predicted probability for the new or changed value of the regressor, and (3) taking their difference.

Application to the mortgage data. As an illustration, we fit a probit model to the 2380 observations in our data set on mortgage denial (*deny*) and the payment-to-income ratio (*P/I ratio*):

$$\widehat{\Pr(deny = 1 | P/I\,ratio)} = \underset{(0.16)}{\Phi(-2.19} + \underset{(0.47)}{2.97}\, P/I\,ratio). \quad (11.7)$$

The estimated coefficients of −2.19 and 2.97 are difficult to interpret because they affect the probability of denial via the z-value. Indeed, the only things that can be readily concluded from the estimated probit regression in Equation (11.7) are that the payment-to-income ratio is positively related to probability of denial (the coefficient on *P/I ratio* is positive) and that this relationship is statistically significant ($t = 2.97/0.47 = 6.32$).

What is the change in the predicted probability that an application will be denied when the payment-to-income ratio increases from 0.3 to 0.4? To answer this question, we follow the procedure in Key Concept 8.1: Compute the probability of denial for $P/I\,ratio = 0.3$ and for $P/I\,ratio = 0.4$, and then compute the difference. The probability of denial when $P/I\,ratio = 0.3$ is $\Phi(-2.19 + 2.97 \times 0.3) = \Phi(-1.30) = 0.097$. The probability of denial when $P/I\,ratio = 0.4$ is $\Phi(-2.19 + 2.97 \times 0.4) = \Phi(-1.00) = 0.159$. The estimated change in the probability of denial is $0.159 - 0.097 = 0.062$. That is, an increase in the payment-to-income ratio from 0.3 to 0.4 is associated with an increase in the probability of denial of 6.2 percentage points, from 9.7% to 15.9%.

Because the probit regression function is nonlinear, the effect of a change in X depends on the starting value of X. For example, if $P/I\ ratio = 0.5$, the estimated denial probability based on Equation (11.7) is $\Phi(-2.19 + 2.97 \times 0.5) = \Phi(-0.71) = 0.239$. Thus the change in the predicted probability when *P/I ratio* increases from 0.4 to 0.5 is 0.239 − 0.159, or 8.0 percentage points, larger than the increase of 6.2 percentage points when *P/I ratio* increases from 0.3 to 0.4.

What is the effect of race on the probability of mortgage denial, holding constant the payment-to-income ratio? To estimate this effect, we estimate a probit regression with both *P/I ratio* and *black* as regressors:

$$\widehat{\Pr(deny = 1 \mid P/I\ ratio, black)} = \underset{(0.16)}{\Phi(-2.26} + \underset{(0.44)}{2.74\, P/I\ ratio} + \underset{(0.083)}{0.71\, black}). \quad (11.8)$$

Again, the values of the coefficients are difficult to interpret, but the sign and statistical significance are not. The coefficient on *black* is positive, indicating that an African American applicant has a higher probability of denial than a white applicant, holding constant their payment-to-income ratio. This coefficient is statistically significant at the 1% level (the *t*-statistic on the coefficient multiplying *black* is 8.55). For a white applicant with $P/I\ ratio = 0.3$, the predicted denial probability is 7.5%, while for a black applicant with $P/I\ ratio = 0.3$, it is 23.3%; the difference in denial probabilities between these two hypothetical applicants is 15.8 percentage points.

Estimation of the probit coefficients. The probit coefficients reported here were estimated using the method of maximum likelihood, which produces efficient (minimum variance) estimators in a wide variety of applications, including regression with a binary dependent variable. The maximum likelihood estimator is consistent and normally distributed in large samples, so *t*-statistics and confidence intervals for the coefficients can be constructed in the usual way.

Regression software for estimating probit models typically uses maximum likelihood estimation, so this is a simple method to apply in practice. Standard errors produced by such software can be used in the same way as the standard errors of regression coefficients; for example, a 95% confidence interval for the true probit coefficient can be constructed as the estimated coefficient ± 1.96 standard errors. Similarly, *F*-statistics computed using maximum likelihood estimators can be used to test joint hypotheses. Maximum likelihood estimation is discussed further in Section 11.3, with additional details given in Appendix 11.2.

Logit Regression

The logit regression model. The logit regression model is similar to the probit regression model except that the cumulative standard normal distribution function Φ in Equation (11.6) is replaced by the cumulative standard logistic distribution function, which we denote by *F*. Logit regression is summarized in Key Concept 11.3. The logistic

KEY CONCEPT 11.3

Logit Regression

The population logit model of the binary dependent variable Y with multiple regressors is

$$\begin{aligned}\Pr(Y = 1 | X_1, X_2, \ldots, X_k) &= F(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k) \\ &= \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)}}. \end{aligned} \tag{11.9}$$

Logit regression is similar to probit regression except that the cumulative distribution function is different.

cumulative distribution function has a specific functional form, defined in terms of the exponential function, which is given as the final expression in Equation (11.9).

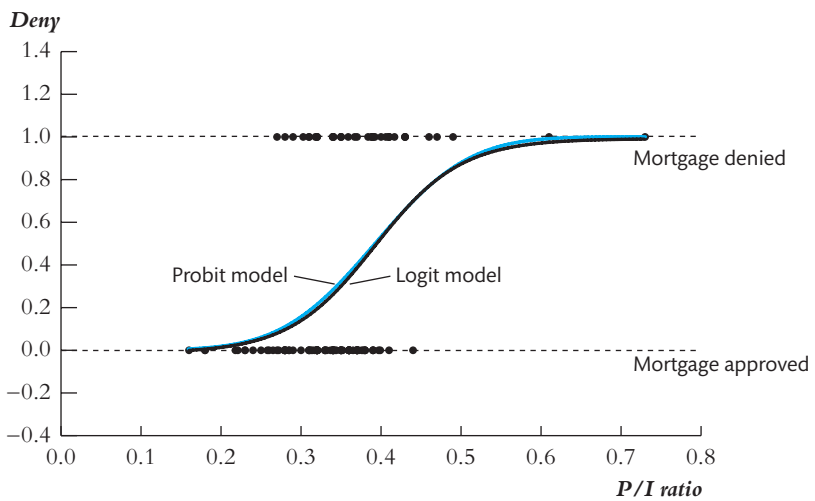
As with probit, the logit coefficients are best interpreted by computing predicted probabilities and differences in predicted probabilities.

The coefficients of the logit model can be estimated by maximum likelihood. The maximum likelihood estimator is consistent and normally distributed in large samples, so t-statistics and confidence intervals for the coefficients can be constructed in the usual way.

The logit and probit regression functions are similar. This is illustrated in Figure 11.3, which graphs the probit and logit regression functions for the dependent variable *deny* and the single regressor *P/I ratio*, estimated by maximum likelihood using the same 127 observations as in Figures 11.1 and 11.2. The differences between the two functions are small.

FIGURE 11.3 Probit and Logit Models of the Probability of Denial Given *P/I Ratio*

These logit and probit models produce nearly identical estimates of the probability that a mortgage application will be denied, given the payment-to-income ratio.



Historically, the main motivation for logit regression was that the logistic cumulative distribution function could be computed faster than the normal cumulative distribution function. With the advent of more powerful computers, this distinction is no longer important.

Application to the Boston HMDA data. A logit regression of *deny* against *P/I ratio* and *black*, using the 2380 observations in the data set, yields the estimated regression function

$$\widehat{\Pr(deny = 1 \mid P/I\,ratio, black)} = F(\underset{(0.35)}{-4.13} + \underset{(0.96)}{5.37}\,P/I\,ratio + \underset{(0.15)}{1.27}\,black). \quad (11.10)$$

The coefficient on *black* is positive and statistically significant at the 1% level (the t-statistic is 8.47). The predicted denial probability of a white applicant with $P/I\,ratio = 0.3$ is $1/[1 + e^{-(-4.13+5.37\times0.3+1.27\times0)}] = 1/[1 + e^{2.52}] = 0.074$, or 7.4%. The predicted denial probability of an African American applicant with $P/I\,ratio = 0.3$ is $1/[1 + e^{1.25}] = 0.222$, or 22.2%, so the difference between the two probabilities is 14.8 percentage points.

Comparing the Linear Probability, Probit, and Logit Models

All three models—linear probability, probit, and logit—are just approximations to the unknown population regression function $E(Y|X) = \Pr(Y = 1|X)$. The linear probability model is easiest to use and to interpret, but it cannot capture the nonlinear nature of the true population regression function. Probit and logit regressions model this nonlinearity in the probabilities, but their regression coefficients are more difficult to interpret. So which should you use in practice?

There is no one right answer, and different researchers use different models. Probit and logit regressions frequently produce similar results. For example, according to the estimated probit model in Equation (11.8), the difference in denial probabilities between a black applicant and a white applicant with $P/I\,ratio = 0.3$ was estimated to be 15.8 percentage points, whereas the logit estimate of this gap, based on Equation (11.10), was 14.9 percentage points. For practical purposes, the two estimates are very similar. One way to choose between logit and probit is to pick the method that is easier to use in your statistical software.

The linear probability model provides the least sensible approximation to the nonlinear population regression function. Even so, in some data sets there may be few extreme values of the regressors, in which case the linear probability model still can provide an adequate approximation. In the denial probability regression in Equation (11.3), the estimated black/white gap from the linear probability model is 17.7 percentage points, larger than the probit and logit estimates but still qualitatively similar. The only way to know this, however, is to estimate both a linear and a nonlinear model and to compare their predicted probabilities.

11.3 Estimation and Inference in the Logit and Probit Models[2]

The nonlinear models studied in Sections 8.2 and 8.3 are nonlinear functions of the independent variables but are linear functions of the unknown coefficients (parameters). Consequently, the unknown coefficients of those nonlinear regression functions can be estimated by OLS. In contrast, the probit and logit regression functions are nonlinear functions of the coefficients. That is, the probit coefficients $\beta_0, \beta_1, \ldots, \beta_k$ in Equation (11.6) appear *inside* the cumulative standard normal distribution function Φ, and the logit coefficients in Equation (11.9) appear *inside* the cumulative standard logistic distribution function F. Because the population regression function is a nonlinear function of the coefficients $\beta_0, \beta_1, \ldots, \beta_k$, those coefficients cannot be estimated by OLS.

This section provides an introduction to the standard method for estimation of probit and logit coefficients, maximum likelihood; additional mathematical details are given in Appendix 11.2. Because it is built into modern statistical software, maximum likelihood estimation of the probit and logit coefficients is easy in practice. The theory of maximum likelihood estimation, however, is more complicated than the theory of least squares. We therefore first discuss another estimation method, nonlinear least squares, before turning to maximum likelihood.

Nonlinear Least Squares Estimation

Nonlinear least squares is a general method for estimating the unknown parameters of a regression function when, like the probit coefficients, those parameters enter the population regression function nonlinearly. The nonlinear least squares estimator, which was introduced in Appendix 8.1, extends the OLS estimator to regression functions that are nonlinear functions of the parameters. Like OLS, nonlinear least squares finds the values of the parameters that minimize the sum of squared prediction mistakes produced by the model.

To be concrete, consider the nonlinear least squares estimator of the parameters of the probit model. The conditional expectation of Y given the X's is $E(Y|X_1, \ldots, X_k) = \Pr(Y = 1|X_1, \ldots, X_k) = \Phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)$. Estimation by nonlinear least squares fits this conditional expectation function, which is a nonlinear function of the parameters, to the dependent variable. That is, the nonlinear least squares estimator of the probit coefficients is the values of $b_0, \ldots, b_k$ that minimize the sum of squared prediction mistakes:

$$\sum_{i=1}^{n} [\, Y_i - \Phi(b_0 + b_1 X_{1i} + \cdots + b_k X_{ki}) \,]^2. \tag{11.11}$$

The nonlinear least squares estimator shares two key properties with the OLS estimator in linear regression: It is consistent (the probability that it is close to the true

[2]This section contains more advanced material that can be skipped without loss of continuity.

value approaches 1 as the sample size gets large), and it is normally distributed in large samples. There are, however, estimators that have a smaller variance than the nonlinear least squares estimator; that is, the nonlinear least squares estimator is inefficient. For this reason, the nonlinear least squares estimator of the probit coefficients is rarely used in practice, and instead the parameters are estimated by maximum likelihood.

Maximum Likelihood Estimation

The **likelihood function** is the joint probability distribution of the data, treated as a function of the unknown coefficients. The **maximum likelihood estimator (MLE)** of the unknown coefficients consists of the values of the coefficients that maximize the likelihood function. Because the MLE chooses the unknown coefficients to maximize the likelihood function, which is in turn the joint probability distribution, in effect the MLE chooses the values of the parameters to maximize the probability of drawing the data that are actually observed. In this sense, the MLEs are the parameter values "most likely" to have produced the data.

To illustrate maximum likelihood estimation, consider two i.i.d. observations, Y_1 and Y_2, on a binary dependent variable with no regressors. Thus Y is a Bernoulli random variable, and the only unknown parameter to estimate is the probability p that $Y = 1$, which is also the mean of Y.

To obtain the maximum likelihood estimator, we need an expression for the likelihood function, which in turn requires an expression for the joint probability distribution of the data. The joint probability distribution of the two observations Y_1 and Y_2 is $\Pr(Y_1 = y_1, Y_2 = y_2)$. Because Y_1 and Y_2 are independently distributed, the joint distribution is the product of the individual distributions [Equation (2.24)], so $\Pr(Y_1 = y_1, Y_2 = y_2) = \Pr(Y_1 = y_1)\Pr(Y_2 = y_2)$. The Bernoulli distribution can be summarized in the formula $\Pr(Y = y) = p^y(1 - p)^{1-y}$: When $y = 1$, $\Pr(Y = 1) = p^1(1 - p)^0 = p$, and when $y = 0$, $\Pr(Y = 0) = p^0(1 - p)^1 = 1 - p$. Thus the joint probability distribution of Y_1 and Y_2 is $\Pr(Y_1 = y_1, Y_2 = y_2) = [p^{y_1}(1-p)^{1-y_1}] \times [p^{y_2}(1-p)^{1-y_2}] = p^{(y_1+y_2)}(1-p)^{2-(y_1+y_2)}$.

The likelihood function is the joint probability distribution, treated as a function of the unknown coefficients. For $n = 2$ i.i.d. observations on Bernoulli random variables, the likelihood function is

$$f(p; Y_1, Y_2) = p^{(Y_1+Y_2)}(1 - p)^{2-(Y_1+Y_2)}. \tag{11.12}$$

The maximum likelihood estimator of p is the value of p that maximizes the likelihood function in Equation (11.12). As with all maximization or minimization problems, this can be done by trial and error; that is, you can try different values of p and compute the likelihood $f(p; Y_1, Y_2)$ until you are satisfied that you have maximized this function. In this example, however, maximizing the likelihood function using calculus produces a simple formula for the MLE: The MLE is $\hat{p} = \frac{1}{2}(Y_1 + Y_2)$.

In other words, the MLE of p is just the sample average! In fact, for general n, the MLE $\hat{p}$ of the Bernoulli probability p is the sample average; that is, $\hat{p} = \overline{Y}$ (this is shown in Appendix 11.2). In this example, the MLE is the usual estimator of p, the fraction of times $Y_i = 1$ in the sample.

This example is similar to the problem of estimating the unknown coefficients of the probit and logit regression models. In those models, the success probability p is not constant but rather depends on X; that is, it is the success probability conditional on X, which is given in Equation (11.6) for the probit model and Equation (11.9) for the logit model. Thus the probit and logit likelihood functions are similar to the likelihood function in Equation (11.12) except that the success probability varies from one observation to the next (because it depends on X_i). Expressions for the probit and logit likelihood functions are given in Appendix 11.2.

Like the nonlinear least squares estimator, the MLE is consistent and normally distributed in large samples. Because regression software commonly computes the MLE of the probit coefficients, this estimator is easy to use in practice. All the estimated probit and logit coefficients reported in this chapter are MLEs.

Statistical inference based on the MLE. Because the MLE is normally distributed in large samples, statistical inference about the probit and logit coefficients based on the MLE proceeds in the same way as inference about the linear regression function coefficients based on the OLS estimator. That is, hypothesis tests are performed using the t-statistic, and 95% confidence intervals are formed as ± 1.96 standard errors. Tests of joint hypotheses on multiple coefficients use the F-statistic in a way similar to that discussed in Chapter 7 for the linear regression model. All of this is completely analogous to statistical inference in the linear regression model.

An important practical point is that some statistical software reports tests of joint hypotheses using the F-statistic, while other software uses the chi-squared statistic. The chi-squared statistic is $q \times F$, where q is the number of restrictions being tested. Because the F-statistic is, under the null hypothesis, distributed as χ^2_q/q in large samples, $q \times F$ is distributed as χ^2_q in large samples. Because the two approaches differ only in whether they divide by q, they produce identical inferences, but you need to know which approach is implemented in your software so that you use the correct critical values.

Measures of Fit

In Section 11.1, it was mentioned that the R^2 is a poor measure of fit for the linear probability model. This is also true for probit and logit regression. Two measures of fit for models with binary dependent variables are the fraction correctly predicted and the pseudo-R^2. The **fraction correctly predicted** uses the following rule: If $Y_i = 1$ and the predicted probability exceeds 50% or if $Y_i = 0$ and the predicted probability is less than 50%, then Y_i is said to be correctly predicted. Otherwise, Y_i is said to be incorrectly predicted. The fraction correctly predicted is the fraction of the n observations $Y_1, \ldots, Y_n$ that are correctly predicted.

An advantage of this measure of fit is that it is easy to understand. A disadvantage is that it does not reflect the quality of the prediction: If $Y_i = 1$, the observation is treated as correctly predicted whether the predicted probability is 51% or 90%.

The **pseudo-R^2** measures the fit of the model using the likelihood function. Because the MLE maximizes the likelihood function, adding another regressor to a probit or logit model increases the value of the maximized likelihood, just like adding a regressor necessarily reduces the sum of squared residuals in linear regression by OLS. This suggests measuring the quality of fit of a probit model by comparing values of the maximized likelihood function with all the regressors to the value of the likelihood with none. This is, in fact, what the pseudo-R^2 does. A formula for the pseudo-R^2 is given in Appendix 11.2.

11.4 Application to the Boston HMDA Data

The regressions of the previous two sections indicated that denial rates were higher for black than white applicants, holding constant their payment-to-income ratio. Loan officers, however, legitimately weigh many factors when deciding on a mortgage application, and if any of those other factors differ systematically by race, the estimators considered so far have omitted variable bias.

In this section, we take a closer look at whether there is statistical evidence of discrimination in the Boston HMDA data. Specifically, our objective is to estimate the effect of race on the probability of denial, holding constant those applicant characteristics that a loan officer might legally consider when deciding on a mortgage application.

The most important variables available to loan officers through the mortgage applications in the Boston HMDA data set are listed in Table 11.1; these are the variables we will focus on in our empirical models of loan decisions. The first two variables are direct measures of the financial burden the proposed loan would place on the applicant, measured in terms of his or her income. The first of these is the *P/I ratio*; the second is the ratio of housing-related expenses to income. The next variable is the size of the loan, relative to the assessed value of the home; if the loan-to-value ratio is nearly 1, the bank might have trouble recouping the full amount of the loan if the applicant defaults on the loan and the bank forecloses. The final three financial variables summarize the applicant's credit history. If an applicant has been unreliable paying off debts in the past, the loan officer legitimately might worry about the applicant's ability or desire to make mortgage payments in the future. The three variables measure different types of credit histories, which the loan officer might weigh differently. The first concerns consumer credit, such as credit card debt; the second is previous mortgage payment history; and the third measures credit problems so severe that they appeared in a public legal record, such as filing for bankruptcy.

TABLE 11.1 Variables Included in Regression Models of Mortgage Decisions

Variable	Definition	Sample Average
Financial Variables		
P/I ratio	Ratio of total monthly debt payments to total monthly income	0.331
housing expense-to-income ratio	Ratio of monthly housing expenses to total monthly income	0.255
loan-to-value ratio	Ratio of size of loan to assessed value of property	0.738
consumer credit score	1 if no "slow" payments or delinquencies 2 if one or two slow payments or delinquencies 3 if more than two slow payments 4 if insufficient credit history for determination 5 if delinquent credit history with payments 60 days overdue 6 if delinquent credit history with payments 90 days overdue	2.1
mortgage credit score	1 if no late mortgage payments 2 if no mortgage payment history 3 if one or two late mortgage payments 4 if more than two late mortgage payments	1.7
public bad credit record	1 if any public record of credit problems (bankruptcy, charge-offs, collection actions) 0 otherwise	0.074
Additional Applicant Characteristics		
denied mortgage insurance	1 if applicant applied for mortgage insurance and was denied, 0 otherwise	0.020
self-employed	1 if self-employed, 0 otherwise	0.116
single	1 if applicant reported being single, 0 otherwise	0.393
high school diploma	1 if applicant graduated from high school, 0 otherwise	0.984
unemployment rate	1989 Massachusetts unemployment rate in the applicant's industry	3.8
condominium	1 if unit is a condominium, 0 otherwise	0.288
black	1 if applicant is black, 0 if white	0.142
deny	1 if mortgage application denied, 0 otherwise	0.120

Table 11.1 also lists some other variables relevant to the loan officer's decision. Sometimes the applicant must apply for private mortgage insurance.[3] The loan officer knows whether that application was denied, and that denial would weigh negatively with the loan officer. The next four variables, which concern the applicant's employment status, marital status, and educational attainment, as well as the unemployment rate in the applicant's industry, relate to the prospective ability of the applicant to repay. In the event of foreclosure, characteristics of the property are relevant as well, and the next variable indicates whether the property is a condominium. The final two variables in Table 11.1 are whether the applicant is black or white and

[3]Mortgage insurance is an insurance policy under which the insurance company makes the monthly payment to the bank if the borrower defaults. During the period of this study, if the loan-to-value ratio exceeds 80%, the applicant typically was required to buy mortgage insurance.

whether the application was denied or accepted. In these data, 14.2% of applicants are black, and 12.0% of applications are denied.

Table 11.2 presents regression results based on these variables. The base specifications, reported in columns (1) through (3), include the financial variables in Table 11.1 plus the variables indicating whether private mortgage insurance was denied and whether the applicant is self-employed. In the 1990s, loan officers commonly used thresholds, or cutoff values, for the loan-to-value ratio, so the base specification for that variable uses binary variables for whether the loan-to-value ratio is high (≥ 0.95), medium (between 0.8 and 0.95), or low (< 0.8; this case is omitted to avoid perfect multicollinearity). The regressors in the first three columns are similar to those in the base specification considered by the Federal Reserve Bank of Boston researchers in their original analysis of these data.[4] The regressions in columns (1) through (3) differ only in how the denial probability is modeled, using a linear probability model, a logit model, and a probit model, respectively.

Because the coefficients of the logit and probit models in columns (2)–(6) are not directly interpretable, the table reports standard errors but not confidence intervals. In addition, because the aim of these regressions is to approximate the loan officers' decision rule, it is of interest to know whether individual variables—especially the applicant's race—enter that decision rule. Thus the table reports, through asterisks, whether the test that the coefficient is 0 rejects at the 5% or 1% significance level.

Because the regression in column (1) is a linear probability model, its coefficients are estimated changes in predicted probabilities arising from a unit change in the independent variable. Accordingly, an increase in *P/I ratio* of 0.1 is estimated to increase the probability of denial by 4.5 percentage points (the coefficient on *P/Iratio* in column (1) is 0.449, and $0.449 \times 0.1 \cong 0.045$). Similarly, having a high loan-to-value ratio increases the probability of denial: A loan-to-value ratio exceeding 95% is associated with an 18.9 percentage point increase (the coefficient is 0.189) in the denial probability, relative to the omitted case of a loan-to-value ratio less than 80%, holding the other variables in column (1) constant. Applicants with a poor credit rating also have a more difficult time getting a loan, all else being constant, although interestingly the coefficient on consumer credit is statistically significant but the coefficient on mortgage credit is not. Applicants with a public record of credit problems, such as filing for bankruptcy, have much greater difficulty obtaining a loan: All else equal, a public bad credit record is estimated to increase the probability of denial by 0.197, or 19.7 percentage points. Being denied private mortgage insurance is estimated to be virtually decisive: The estimated coefficient of 0.702 means that being denied mortgage insurance increases your chance of being denied a mortgage by 70.2 percentage points, all else

[4]The difference between the regressors in columns (1) through (3) and those in Munnell et al. (1996), table 2 (1), is that Munnell et al. include additional indicators for the location of the home and the identity of the lender, data that are not publicly available; an indicator for a multifamily home, which is irrelevant here because our subset focuses on single-family homes; and net wealth, which we omit because this variable has a few very large positive and negative values and thus risks making the results sensitive to a few specific outlier observations.

TABLE 11.2 Mortgage Denial Regressions Using the Boston HMDA Data

Dependent variable: *deny* = 1 if mortgage application is denied, = 0 if accepted; 2380 observations.

Regression Model	*LPM*	*Logit*	*Probit*	*Probit*	*Probit*	*Probit*
Regressor	**(1)**	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
black	0.084** (0.023)	0.688** (0.182)	0.389** (0.098)	0.371** (0.099)	0.363** (0.100)	0.246 (0.448)
P/I ratio	0.449** (0.114)	4.76** (1.33)	2.44** (0.61)	2.46** (0.60)	2.62** (0.61)	2.57** (0.66)
housing expense-to-income ratio	−0.048 (0.110)	−0.11 (1.29)	−0.18 (0.68)	−0.30 (0.68)	−0.50 (0.70)	−0.54 (0.74)
medium loan-to-value ratio (0.80 ≤ *loan-value ratio* ≤ 0.95)	0.031* (0.013)	0.46** (0.16)	0.21** (0.08)	0.22** (0.08)	0.22** (0.08)	0.22** (0.08)
high loan-to-value ratio (*loan-value ratio* > 0.95)	0.189** (0.050)	1.49** (0.32)	0.79** (0.18)	0.79** (0.18)	0.84** (0.18)	0.79** (0.18)
consumer credit score	0.031** (0.005)	0.29** (0.04)	0.15** (0.02)	0.16** (0.02)	0.34** (0.11)	0.16** (0.02)
mortgage credit score	0.021 (0.011)	0.28* (0.14)	0.15* (0.07)	0.11 (0.08)	0.16 (0.10)	0.11 (0.08)
public bad credit record	0.197** (0.035)	1.23** (0.20)	0.70** (0.12)	0.70** (0.12)	0.72** (0.12)	0.70** (0.12)
denied mortgage insurance	0.702** (0.045)	4.55** (0.57)	2.56** (0.30)	2.59** (0.29)	2.59** (0.30)	2.59** (0.29)
self-employed	0.060** (0.021)	0.67** (0.21)	0.36** (0.11)	0.35** (0.11)	0.34** (0.11)	0.35** (0.11)
single				0.23** (0.08)	0.23** (0.08)	0.23** (0.08)
high school diploma				−0.61** (0.23)	−0.60* (0.24)	−0.62** (0.23)
unemployment rate				0.03 (0.02)	0.03 (0.02)	0.03 (0.02)
condominium					−0.05 (0.09)	
black × *P/I ratio*						−0.58 (1.47)
black × *housing expense-to-income ratio*						1.23 (1.69)
additional credit rating indicator variables	no	no	no	no	yes	no
constant	−0.183** (0.028)	−5.71** (0.48)	−3.04** (0.23)	−2.57** (0.34)	−2.90** (0.39)	−2.54** (0.35)

(continued)

(Table 11.2 continued)

F-Statistics and p-Values Testing Exclusion of Groups of Variables

	(1)	(2)	(3)	(4)	(5)	(6)
applicant single; high school diploma; industry unemployment rate				5.85 (< 0.001)	5.22 (0.001)	5.79 (< 0.001)
additional credit rating indicator variables					1.22 (0.291)	
race interactions and black						4.96 (0.002)
race interactions only						0.27 (0.766)
difference in predicted probability of denial, white vs. black (percentage points)	8.4%	6.0%	7.1%	6.6%	6.3%	6.5%

These regressions were estimated using the $n = 2380$ observations in the Boston HMDA data set described in Appendix 11.1. The linear probability model was estimated by OLS, and probit and logit regressions were estimated by maximum likelihood. Standard errors are given in parentheses under the coefficients, and *p*-values are given in parentheses under the *F*-statistics. The change in predicted probability in the final row was computed for a hypothetical applicant whose values of the regressors, other than race, equal the sample mean. Individual coefficients are statistically significant at the *5% or **1% level.

equal. Of the nine variables (other than race) in the regression, the coefficients on all but two are statistically significant at the 5% level, which is consistent with loan officers' considering many factors when they make their decisions.

The coefficient on *black* in regression (1) is 0.084, indicating that the difference in denial probabilities for black and white applicants is 8.4 percentage points, holding constant the other variables in the regression. This is statistically significant at the 1% significance level $(t = 3.65)$.

The logit and probit estimates reported in columns (2) and (3) yield similar conclusions. In the logit and probit regressions, eight of the nine coefficients on variables other than race are individually statistically significantly different from 0 at the 5% level, and the coefficient on *black* is statistically significant at the 1% level. As discussed in Section 11.2, because these models are nonlinear, specific values of all the regressors must be chosen to compute the difference in predicted probabilities for white applicants and black applicants. A conventional way to make this choice is to consider an "average" applicant who has the sample average values of all the regressors other than race. The final row in Table 11.2 reports this estimated difference in probabilities, evaluated for this average applicant. The estimated racial differentials are similar to each other: 8.4 percentage points for the linear probability model [column (1)], 6.0 percentage points for the logit model [column (2)], and 7.1 percentage points for the probit model [column (3)]. These estimated race effects and the coefficients on *black* are less than in the regressions of the previous sections, in which the only regressors were *P/I ratio* and *black*, indicating that those earlier estimates had omitted variable bias.

The regressions in columns (4) through (6) investigate the sensitivity of the results in column (3) to changes in the regression specification. Column (4) modifies

column (3) by including additional applicant characteristics. These characteristics help to predict whether the loan is denied; for example, having at least a high school diploma reduces the probability of denial (the estimate is negative, and the coefficient is statistically significant at the 1% level). However, controlling for these personal characteristics does not change the estimated coefficient on *black* or the estimated difference in denial probabilities (6.6%) in an important way.

Column (5) breaks out the six consumer credit categories and four mortgage credit categories to test the null hypothesis that these two variables enter linearly; this regression also adds a variable indicating whether the property is a condominium. The null hypothesis that the credit rating variables enter the expression for the z-value linearly is not rejected, nor is the condominium indicator significant, at the 5% level. Most importantly, the estimated racial difference in denial probabilities (6.3%) is essentially the same as in columns (3) and (4).

Column (6) examines whether there are interactions. Are different standards applied to evaluating the payment-to-income and housing expense-to-income ratios for black applicants versus white applicants? The answer appears to be no: The interaction terms are not jointly statistically significant at the 5% level. However, race continues to have a significant effect, because the race indicator and the interaction terms are jointly statistically significant at the 1% level. Again, the estimated racial difference in denial probabilities (6.5%) is essentially the same as in the other probit regressions.

In all six specifications, the effect of race on the denial probability, holding other applicant characteristics constant, is statistically significant at the 1% level. The estimated difference in denial probabilities between black applicants and white applicants ranges from 6.0 percentage points to 8.4 percentage points.

One way to assess whether this differential is large or small is to return to a variation on the question posed at the beginning of this chapter. Suppose two individuals apply for a mortgage, one white and one black, but otherwise having the same values of the other independent variables in regression (3); specifically, aside from race, the values of the other variables in regression (3) are the sample average values in the HMDA data set. The white applicant faces a 7.4% chance of denial, but the black applicant faces a 14.5% chance of denial. The estimated racial difference in denial probabilities, 7.1 percentage points, means that the black applicant is nearly twice as likely to be denied as the white applicant.

The results in Table 11.2 (and in the original Boston Fed study) provide statistical evidence of racial patterns in mortgage denial that, by law, ought not be there. This evidence played an important role in spurring policy changes by bank regulators.[5] But economists love a good argument, and not surprisingly these results have also stimulated a vigorous debate.

Because the suggestion that there is (or was) racial discrimination in lending is charged, we briefly review some points of this debate. In so doing, it is useful to adopt the framework of Chapter 9—that is, to consider the internal and external validity of

[5]These policy shifts include changes in the way that fair lending examinations were done by federal bank regulators, changes in inquiries made by the U.S. Department of Justice, and enhanced education programs for banks and other home loan origination companies.

the results in Table 11.2, which are representative of previous analyses of the Boston HMDA data. A number of the criticisms made of the original Federal Reserve Bank of Boston study concern internal validity: possible errors in the data, alternative nonlinear functional forms, additional interactions, and so forth. The original data were subjected to a careful audit, some errors were found, and the results reported here (and in the final published Boston Fed study) are based on the "cleaned" data set. Estimation of other specifications—different functional forms and/or additional regressors—also produces estimates of racial differentials comparable to those in Table 11.2. A potentially more difficult issue of internal validity is whether there is relevant nonracial financial information obtained during in-person loan interviews, but not recorded on the loan application itself, that is correlated with race; if so, there still might be omitted variable bias in the Table 11.2 regressions. Finally, some have questioned external validity: Even if there was racial discrimination in Boston in 1990, it is wrong to implicate lenders elsewhere today. Moreover, racial discrimination might be less likely using modern online applications because the mortgage can be approved or denied without a face-to-face meeting. The only way to resolve the question of external validity is to consider data from other locations and years.[6]

11.5 Conclusion

When the dependent variable Y is binary, the population regression function is the probability that $Y = 1$, conditional on the regressors. Estimation of this population regression function entails finding a functional form that does justice to its probability interpretation, estimating the unknown parameters of that function, and interpreting the results. The resulting predicted values are predicted probabilities, and the estimated effect of a change in a regressor X is the estimated change in the probability that $Y = 1$ arising from the change in X.

A natural way to model the probability that $Y = 1$ given the regressors is to use a cumulative distribution function, where the argument of the c.d.f. depends on the regressors. Probit regression uses a normal c.d.f. as the regression function, and logit regression uses a logistic c.d.f. Because these models are nonlinear functions of the unknown parameters, those parameters are more complicated to estimate than linear regression coefficients. The standard estimation method is maximum likelihood. In practice, statistical inference using the maximum likelihood estimates proceeds the same way as it does in linear multiple regression; for example, 95% confidence intervals for a coefficient are constructed as the estimated coefficient ± 1.96 standard errors.

[6]If you are interested in further reading on this topic, a good place to start is the symposium on racial discrimination and economics in the Spring 1998 issue of the *Journal of Economic Perspectives*. The article in that symposium by Helen Ladd (1998) surveys the evidence and debate on racial discrimination in mortgage lending. A more detailed treatment is given in Goering and Wienk (1996). The U.S. mortgage market has changed dramatically since the Boston Fed study, including a relaxation of lending standards, a bubble in housing prices, the financial crisis of 2008–2009, and a return to tighter lending standards. For an introduction to changes in mortgage markets, see Green and Wachter (2008).

James Heckman and Daniel McFadden, Nobel Laureates

The 2000 Nobel Prize in Economics was awarded jointly to two econometricians, James J. Heckman of the University of Chicago and Daniel L. McFadden of the University of California at Berkeley, for fundamental contributions to the analysis of data on individuals and firms. Much of their work addressed difficulties that arise with limited dependent variables.

Heckman was awarded the prize for developing tools for handling sample selection. As discussed in Section 9.2, sample selection bias occurs when the availability of data is influenced by a selection process related to the value of the dependent variable. For example, suppose you want to estimate the relationship between earnings and some regressor, X, using a random sample from the population. If you estimate the regression using the subsample of employed workers—that is, those reporting positive earnings—the OLS estimate could be subject to selection bias. Heckman's solution was to specify a preliminary equation with a binary dependent variable indicating whether the worker is in or out of the labor force (in or out of the subsample) and to treat this equation and the earnings equation as a system of simultaneous equations. This general strategy has been extended to selection problems that arise in many fields, ranging from labor economics to industrial organization to finance.

McFadden was awarded the prize for developing models for analyzing discrete choice data (does a high school graduate join the military, go to college, or get a job?). He started by considering the problem of an individual maximizing the expected utility of each possible choice, which could depend on observable variables (such as wages, job characteristics, and family background). He then derived models for the individual choice probabilities with unknown coefficients, which in turn could be estimated by maximum likelihood. These models and their extensions have proven widely useful in analyzing discrete choice data in many fields, including labor economics, health economics, and transportation economics.

For more information on these and other Nobel laureates in economics, visit the Nobel Foundation website, http://www.nobel.se/economics.

Henrik Montgomery/ Pressens Bild/AP Images

James J. Heckman

Paul Sakuma/AP Images

Daniel L. McFadden

Despite its intrinsic nonlinearity, sometimes the population regression function can be adequately approximated by a linear probability model—that is, by the straight line produced by linear multiple regression. The linear probability model, probit regression, and logit regression all give similar bottom-line answers when they are applied to the Boston HMDA data: All three methods estimate substantial differences in mortgage denial rates for otherwise similar black applicants and white applicants.

Binary dependent variables are the most common example of limited dependent variables, which are dependent variables with a limited range. The final quarter of the 20th century saw important advances in econometric methods for analyzing other limited dependent variables (see the box "James Heckman and Daniel McFadden, Nobel Laureates"). Some of these methods are reviewed in Appendix 11.3.

Summary

1. When Y is a binary variable, the population regression function shows the probability that $Y = 1$ given the value of the regressors, $X_1, X_2, \ldots, X_k$.
2. The linear multiple regression model is called the linear probability model when Y is a binary variable because the probability that $Y = 1$ is a linear function of the regressors.
3. Probit and logit regression models are nonlinear regression models used when Y is a binary variable. Unlike the linear probability model, probit and logit regressions ensure that the predicted probability that $Y = 1$ is between 0 and 1 for all values of X.
4. Probit regression uses the standard normal cumulative distribution function. Logit regression uses the logistic cumulative distribution function. Logit and probit coefficients are estimated by maximum likelihood.
5. The values of coefficients in probit and logit regressions are not easy to interpret. Changes in the probability that $Y = 1$ associated with changes in one or more of the X's can be calculated using the general procedure for nonlinear models outlined in Key Concept 8.1.
6. Hypothesis tests on coefficients in the linear probability, logit, and probit models are performed using the usual t- and F-statistics.

Key Terms

limited dependent variable (393)
linear probability model (395)
probit (397)
logit (397)
logistic regression (397)
likelihood function (405)
maximum likelihood estimator (MLE) (405)
fraction correctly predicted (406)
pseudo-R^2 (407)

Review the Concepts

11.1 Suppose a linear probability model yields a predicted value of Y that is equal to 1.3. Explain why this is nonsensical.

11.2 In Table 11.2, the estimated coefficient on *black* is 0.084 in column (1), 0.688 in column (2), and 0.389 in column (3). In spite of these large differences, all three models yield similar estimates of the marginal effect of race on the probability of mortgage denial. How can this be?

11.3 What is maximum likelihood estimation? What are the advantages of using maximum likelihood estimators such as the probit and the logit, instead of the linear probability model? How would you choose between the probit and the logit?

11.4 What measures of fit are typically used to assess binary dependent variable regression models?

Exercises

Exercises 11.1 through 11.5 are based on the following scenario: Seven hundred income-earning individuals from a district were randomly selected and asked whether they are government employees ($Gov_i = 1$) or not ($Gov_i = 0$); data were also collected on their gender ($Male_i = 1$ if male and $= 0$ if female) and their years of schooling ($Schooling_i$, in years). Note, *Schooling* refers to the number of years of education received by people ages 25 and older. The following table summarizes several estimated models.

11.1 Using the results in column (1):

a. Does the probability of working for the government depend on *Schooling*? Explain.

b. Friedrich Fürnrohr has 16 years of schooling. What is the probability that he will be employed by the government?

c. Hans Schneider never went to college (12 years of schooling). What is the probability that Hans will get a government job?

d. The sample included values of *Schooling* between 0 and 18 years, and only five people in the sample had more than 15 years of schooling. Günter Mayer has completed his PhD and has been a student for

Dependent Variable: *Gov*

	Probit (1)	Logit (2)	Linear Probability (3)	Probit (4)	Logit (5)	Linear Probability (6)	Probit (7)
Schooling	0.272 (0.029)	0.551 (0.062)	0.035 (0.003)				0.548 (0.091)
Male				−0.242 (0.125)	−0.455 (0.234)	−0.050 (0.025)	4.352 (1.291)
Male × Schooling							−0.344 (0.096)
Constant	−4.107 (0.358)	−8.146 (0.800)	−0.172 (0.027)	−1.027 (0.098)	−1.717 (0.179)	0.152 (0.021)	−7.702 (1.238)

24 years. What is the model's prediction for the probability that Günter will be employed by the government? Do you think that this prediction is reliable? Why or why not?

11.2 **a.** Answer (a) through (c) from Exercise 11.1 using the results in column (2).

b. Sketch the predicted probabilities from the probit and logit in columns (1) and (2) for values of *Schooling* between 0 and 18. Are the probit and logit models similar?

11.3 **a.** Answer (a) through (c) from Exercise 11.1 using the results in column (3).

b. Sketch the predicted probabilities from the probit and linear probability in columns (1) and (3) as a function of *Schooling* for values of *Schooling* between 0 and 18. Do you think that the linear probability is appropriate here? Why or why not?

11.4 Using the results in columns (4) through (6):

a. Compute the estimated probability of being employed by the government for men and for women.

b. Are the models in (4) through (6) different? Why or why not?

11.5 Using the results in column (7):

a. Liam Johansson is a man with 10 years of schooling. What is the probability that government will employ him?

b. Anneli Karlsson is a woman with 12 years of schooling. What is the probability that government will employ her?

c. Does the effect of schooling on government employment depend on gender? Explain.

11.6 Use the estimated probit model in Equation (11.8) to answer the following questions:

a. A black mortgage applicant has a *P/I ratio* of 0.35. What is the probability that his application will be denied?

b. Suppose the applicant reduced this ratio to 0.30. What effect would this have on his probability of being denied a mortgage?

c. Repeat (a) and (b) for a white applicant.

d. Does the marginal effect of the *P/I ratio* on the probability of mortgage denial depend on race? Explain.

11.7 Repeat Exercise 11.6 using the logit model in Equation (11.10). Are the logit and probit results similar? Explain.

11.8 Consider the linear probability model $Y_i = \beta_0 + \beta_1 X_i + u_i$, and assume that $E(u_i|X_i) = 0$.

a. Show that $Pr(Y_i = 1|X_i) = \beta_0 + \beta_1 X_i$.

b. Show that $\text{var}(u_i|X_i) = (\beta_0 + \beta_1 X_i)[1 - (\beta_0 + \beta_1 X_i)]$. [*Hint:* Review Equation (2.7).]

c. Is u_i heteroskedastic? Explain.

d. (Requires Section 11.3) Derive the likelihood function.

11.9 Use the estimated linear probability model shown in column (1) of Table 11.2 to answer the following:

a. Two applicants, one self-employed and one in salaried employment, apply for a mortgage. They have the same values for all the regressors other than employment status. How much more likely is the self-employed applicant to be denied a mortgage?

b. Construct a 95% confidence interval for your answer to (a).

c. Think of an important omitted variable that might bias the answer in (a). What is it, and how would it bias the results?

11.10 (Requires Section 11.3 and calculus) Suppose a random variable Y has the following probability distribution: $\Pr(Y = 1) = p$, $\Pr(Y = 2) = q$, and $\Pr(Y = 3) = 1 - p - q$. A random sample of size n is drawn from this distribution, and the random variables are denoted $Y_1, Y_2, \ldots, Y_n$.

a. Derive the likelihood function for the parameters p and q.

b. Derive formulas for the MLE of p and q.

11.11 (Requires Appendix 11.3) State which model you would use for:

a. A study explaining the number of hours a person spends working in a factory during one week.

b. A study explaining the level of satisfaction (0 through 5) a person gains from their job.

c. A study of consumers' choices for mode of transport—bus, car, or bicycle.

d. A study of the number of rainy days in a week.

Empirical Exercises

E11.1 In April 2008, the unemployment rate in the United States stood at 5.0%. By April 2009, it had increased to 9.0%, and it had increased further, to 10.0%, by October 2009. Were some groups of workers more likely to lose their jobs than others during the Great Recession? For example, were young workers more likely to lose their jobs than middle-aged workers? What about workers with a college degree versus those without a degree or women versus men? On the text website, **http://www.pearsonglobaleditions.com**, you will find the data file **Employment_08_09**, which contains a random sample of 5440 workers who were surveyed in April 2008 and reported that they were employed full-time. A detailed description is given in **Employment_08_09_Description**,

available on the website. These workers were surveyed one year later, in April 2009, and asked about their employment status (employed, unemployed, or out of the labor force). The data set also includes various demographic measures for each individual. Use these data to answer the following questions.

a. What fraction of workers in the sample were employed in April 2009? Use your answer to compute a 95% confidence interval for the probability that a worker was employed in April 2009, conditional on being employed in April 2008.

b. Regress *Employed* on *Age* and Age^2, using a linear probability model.

i. Based on this regression, was age a statistically significant determinant of employment in April 2009?

ii. Is there evidence of a nonlinear effect of age on the probability of being employed?

iii. Compute the predicted probability of employment for a 20-year-old worker, a 40-year-old worker, and a 60-year-old worker.

c. Repeat (b) using a probit regression.

d. Repeat (b) using a logit regression.

e. Are there important differences in your answers to (b)–(d)? Explain.

f. The data set includes variables measuring the workers' educational attainment, sex, race, marital status, region of the country, and weekly earnings in April 2008.

i. Construct a table like Table 11.2 to investigate whether the conclusions on the effect of age on employment from (b)–(d) are affected by omitted variable bias.

ii. Use the regressions in your table to discuss the characteristics of workers who were hurt most by the Great Recession.

g. The results in (a)–(f) were based on the probability of employment. Workers who are not employed can either be (i) unemployed or (ii) out the labor force. Do the conclusions you reached in (a)–(f) also hold for workers who became unemployed? (*Hint:* Use the binary variable *Unemployed* instead of *Employed*.)

h. These results have covered employment transitions during the Great Recession, but what about transitions during normal times? On the text website, you will find the data file **Employment_06_07**, which measures the same variables but for the years 2006–2007. Analyze these data and comment on the differences in employment transitions during recessions and normal times.

E11.2 Believe it or not, workers used to be able to smoke inside office buildings. Smoking bans were introduced in several areas during the 1990s. Supporters of these bans argued that in addition to eliminating the externality of secondhand

smoke, they would encourage smokers to quit by reducing their opportunities to smoke. In this assignment, you will estimate the effect of workplace smoking bans on smoking, using data on a sample of 10,000 U.S. indoor workers from 1991 to 1993, available on the text website, **http://www.pearsonglobaleditions.com**, in the file **Smoking**. The data set contains information on whether individuals were or were not subject to a workplace smoking ban, whether the individuals smoked, and other individual characteristics.[7] A detailed description is given in **Smoking_Description**, available on the website.

a. Estimate the probability of smoking for (i) all workers, (ii) workers affected by workplace smoking bans, and (iii) workers not affected by workplace smoking bans.

b. What is the difference in the probability of smoking between workers affected by a workplace smoking ban and workers not affected by a workplace smoking ban? Use a linear probability model to determine whether this difference is statistically significant.

c. Estimate a linear probability model with *smoker* as the dependent variable and the following regressors: *smkban*, *female*, *age*, age^2, *hsdrop*, *hsgrad*, *colsome*, *colgrad*, *black*, and *hispanic*. Compare the estimated effect of a smoking ban from this regression with your answer from (b). Suggest an explanation, based on the substance of this regression, for the change in the estimated effect of a smoking ban between (b) and (c).

d. Test the hypothesis that the coefficient on *smkban* is 0 in the population version of the regression in (c) against the alternative that it is nonzero, at the 5% significance level.

e. Test the hypothesis that the probability of smoking does not depend on the level of education in the regression in (c). Does the probability of smoking increase or decrease with the level of education?

f. Repeat (c)–(e) using a probit model.

g. Repeat (c)–(e) using a logit model.

h. i. Mr. A is white, non-Hispanic, 20 years old, and a high school dropout. Using the probit regression and assuming that Mr. A is not subject to a workplace smoking ban, calculate the probability that Mr. A smokes. Carry out the calculation again, assuming that he is subject to a workplace smoking ban. What is the effect of the smoking ban on the probability of smoking?

ii. Repeat (i) for Ms. B, a female, black, 40-year-old college graduate.

iii. Repeat (i)–(ii) using the linear probability model.

[7]These data were provided by Professor William Evans of the University of Maryland and were used in his paper with Matthew Farrelly and Edward Montgomery, "Do Workplace Smoking Bans Reduce Smoking?" *American Economic Review*, 1999, 89(4): 728–747.

iv. Repeat (i)–(ii) using the logit model.

v. Based on your answers to (i)–(iv), do the logit, probit, and linear probability models differ? If they do, which results make most sense? Are the estimated effects large in a real-world sense?

APPENDIX

11.1 The Boston HMDA Data Set

The Boston HMDA data set was collected by researchers at the Federal Reserve Bank of Boston. The data set combines information from mortgage applications and a follow-up survey of the banks and other lending institutions that received these mortgage applications. The data pertain to mortgage applications made in 1990 in the greater Boston metropolitan area. The full data set has 2925 observations, consisting of all mortgage applications by blacks and Hispanics plus a random sample of mortgage applications by whites.

To narrow the scope of the analysis in this chapter, we use a subset of the data for single-family residences only (thereby excluding data on multifamily homes) and for black applicants and white applicants only (thereby excluding data on applicants from other minority groups). This leaves 2380 observations. Definitions of the variables used in this chapter are given in Table 11.1.

These data were graciously provided to us by Geoffrey Tootell of the Research Department of the Federal Reserve Bank of Boston. More information about this data set, along with the conclusions reached by the Federal Reserve Bank of Boston researchers, is available in Munnell et al. (1996).

APPENDIX

11.2 Maximum Likelihood Estimation

This appendix provides a brief introduction to maximum likelihood estimation in the context of the binary response models discussed in this chapter. We start by deriving the MLE of the success probability p for n i.i.d. observations of a Bernoulli random variable. We then turn to the probit and logit models and discuss the pseudo-R^2. We conclude with a discussion of standard errors for predicted probabilities. This appendix uses calculus at two points.

MLE for n i.i.d. Bernoulli Random Variables

The first step in computing the MLE is to derive the joint probability distribution. For n i.i.d. observations on a Bernoulli random variable, this joint probability distribution is the extension of the $n = 2$ case in Section 11.3 to general n:

$$\begin{aligned}\Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_n = y_n) &= [p^{y_1}(1-p)^{(1-y_1)}] \times [p^{y_2}(1-p)^{(1-y_2)}] \times \cdots \times [p^{y_n}(1-p)^{(1-y_n)}] \\ &= p^{(y_1+\cdots+y_n)}(1-p)^{n-(y_1+\cdots+y_n)}. \end{aligned} \tag{11.13}$$

The likelihood function is the joint probability distribution, treated as a function of the unknown coefficients. Let $S = \sum_{i=1}^{n} Y_i$; then the likelihood function is

$$f_{Bernoulli}(p; Y_1, \ldots, Y_n) = p^S(1-p)^{n-S}. \tag{11.14}$$

The MLE of p is the value of p that maximizes the likelihood in Equation (11.14). The likelihood function can be maximized using calculus. It is convenient to maximize not the likelihood but rather its logarithm (because the logarithm is a strictly increasing function, maximizing the likelihood or its logarithm gives the same estimator). The log likelihood is $S\ln(p) + (n-S)\ln(1-p)$, and the derivative of the log likelihood with respect to p is

$$\frac{d}{dp}\ln\left[f_{Bernoulli}(p; Y_1, \ldots, Y_n)\right] = \frac{S}{p} - \frac{n-S}{1-p}. \tag{11.15}$$

Setting the derivative in Equation (11.15) to 0 and solving for p yields the MLE $\hat{p} = S/n = \overline{Y}$.

MLE for the Probit Model

For the probit model, the probability that $Y_i = 1$, conditional on $X_{1i}, \ldots, X_{ki}$, is $p_i = \Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$. The conditional probability distribution for the i^{th} observation is $\Pr[Y_i = y_i | X_{1i}, \ldots, X_{ki}] = p_i^{y_i}(1-p_i)^{1-y_i}$. Assuming that $(X_{1i}, \ldots, X_{ki}, Y_i)$ are i.i.d., $i = 1, \ldots, n$, the joint probability distribution of $Y_1, \ldots, Y_n$, conditional on the X's, is

$$\begin{aligned}
&\Pr(Y_1 = y_1, \ldots, Y_n = y_n | X_{1i}, \ldots, X_{ki}, i = 1, \ldots, n) \\
&\quad = Pr(Y_1 = y_1 | X_{11}, \ldots, X_{k1}) \times \cdots \times Pr(Y_n = y_n | X_{1n}, \ldots, X_{kn}) \\
&\quad = p_1^{y_1}(1-p_1)^{1-y_1} \times \cdots \times p_n^{y_n}(1-p_n)^{1-y_n}.
\end{aligned} \tag{11.16}$$

The likelihood function is the joint probability distribution, treated as a function of the unknown coefficients. It is conventional to consider the logarithm of the likelihood. Accordingly, the log likelihood function is

$$\begin{aligned}
&\ln[f_{probit}(\beta_0, \ldots, \beta_k; Y_1, \ldots, Y_n X_{1i}, \ldots, X_{ki}, i = 1, \ldots, n)] \\
&\quad = \sum_{i=1}^{n} Y_i \ln[\Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})] \\
&\qquad + \sum_{i=1}^{n} (1 - Y_i)\ln[1 - \Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})],
\end{aligned} \tag{11.17}$$

where this expression incorporates the probit formula for the conditional probability, $p_i = \Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$.

The MLE for the probit model maximizes the likelihood function or, equivalently, the logarithm of the likelihood function given in Equation (11.17). Because there is no simple formula for the MLE, the probit likelihood function must be maximized using a numerical algorithm on the computer.

Under general conditions, maximum likelihood estimators are consistent and have a normal sampling distribution in large samples.

MLE for the Logit Model

The likelihood for the logit model is derived in the same way as the likelihood for the probit model. The only difference is that the conditional success probability p_i for the logit model is given by Equation (11.9). Accordingly, the log likelihood of the logit model is given by Equation (11.17), with $\Phi(\beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki})$ replaced by $[1 + e^{-(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots+\beta_k X_{ki})}]^{-1}$. As with the probit model, there is no simple formula for the MLE of the logit coefficients, so the log likelihood must be maximized numerically.

Pseudo-R^2

The pseudo-R^2 compares the value of the likelihood of the estimated model to the value of the likelihood when none of the X's are included as regressors. Specifically, the pseudo-R^2 for the probit model is

$$\text{pseudo} - R^2 = 1 - \frac{\ln(f^{\max}_{probit})}{\ln(f^{\max}_{Bernoulli})}, \tag{11.18}$$

where $f^{\max}_{probit}$ is the value of the maximized probit likelihood (which includes the X's) and $f^{\max}_{Bernoulli}$ is the value of the maximized Bernoulli likelihood (the probit model excluding all the X's).

Standard Errors for Predicted Probabilities

For simplicity, consider the case of a single regressor in the probit model. Then the predicted probability at a fixed value of that regressor, x, is $\hat{p}(x) = \Phi(\hat{\beta}_0^{MLE} + \hat{\beta}_1^{MLE} x)$, where $\hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{MLE}$ are the MLEs of the two probit coefficients. Because this predicted probability depends on the estimators $\hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{MLE}$, and because those estimators have a sampling distribution, the predicted probability will also have a sampling distribution.

The variance of the sampling distribution of $\hat{p}(x)$ is calculated by approximating the function $\Phi(\hat{\beta}_0^{MLE} + \hat{\beta}_1^{MLE} x)$, a nonlinear function of $\hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{MLE}$, by a linear function of $\hat{\beta}_0^{MLE}$ and $\hat{\beta}_1^{MLE}$. Specifically, let

$$\hat{p}(x) = \Phi(\hat{\beta}_0^{MLE} + \hat{\beta}_1^{MLE} x) \cong c + a_0(\hat{\beta}_0^{MLE} - \beta_0) + a_1(\hat{\beta}_1^{MLE} - \beta_1), \tag{11.19}$$

where the constant c and factors a_0 and a_1 depend on x and are obtained from calculus. [Equation (11.19) is a first-order Taylor series expansion; $c = \Phi(\beta_0 + \beta_1 x)$; and a_0 and a_1 are the partial derivatives, $a_0 = \partial\Phi(\beta_0 + \beta_1 x)/\partial\beta_0\big|_{\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}}$ and $a_1 = \partial\Phi(\beta_0 + \beta_1 x)/\partial\beta_1\big|_{\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}}$.] The variance of $\hat{p}(x)$ now can be calculated using the approximation in Equation (11.19) and the expression for the variance of the sum of two random variables in Equation (2.32):

$$\begin{aligned} \text{var}[\hat{p}(x)] &\cong \text{var}[c + a_0(\hat{\beta}_0^{MLE} - \beta_0) + a_1(\hat{\beta}_1^{MLE} - \beta_1)] \\ &= a_0^2 \text{var}(\hat{\beta}_0^{MLE}) + a_1^2 \text{var}(\hat{\beta}_1^{MLE}) + 2a_0 a_1 \text{cov}(\hat{\beta}_0^{MLE}, \hat{\beta}_1^{MLE}). \end{aligned} \tag{11.20}$$

Using Equation (11.20), the standard error of $\hat{p}(x)$ can be calculated using estimates of the variances and covariance of the MLEs.

APPENDIX

11.3 Other Limited Dependent Variable Models

This appendix surveys some models for limited dependent variables, other than binary variables, found in econometric applications. In most cases, the OLS estimators of the parameters of limited dependent variable models are inconsistent, and estimation is routinely done using maximum likelihood. There are several advanced references available to the reader interested in further details; see, for example, Greene (2018), Ruud (2000), and Wooldridge (2010).

Censored and Truncated Regression Models

Suppose you have cross-sectional data on car purchases by individuals in a given year. Car buyers have positive expenditures, which can reasonably be treated as continuous random variables, but nonbuyers spend \$0. Thus the distribution of car expenditures is a combination of a discrete distribution (at 0) and a continuous distribution.

Nobel laureate James Tobin developed a useful model for a dependent variable with a partly continuous and partly discrete distribution (Tobin, 1958). Tobin suggested modeling the i^{th} individual in the sample as having a desired level of spending, Y_i^*, that is related to the regressors (for example, family size) according to a linear regression model. That is, when there is a single regressor, the desired level of spending is

$$Y_i^* = \beta_0 + \beta_1 X_i + u_i, i = 1, \ldots, n. \tag{11.21}$$

If Y_i^* (what the consumer wants to spend) exceeds some cutoff, such as the minimum price of a car, the consumer buys the car and spends $Y_i = Y_i^*$, which is observed. However, if Y_i^* is less than the cutoff, spending of $Y_i = 0$ is observed instead of Y_i^*.

When Equation (11.21) is estimated using observed expenditures Y_i in place of Y_i^*, the OLS estimator is inconsistent. Tobin solved this problem by deriving the likelihood function using the additional assumption that u_i has a normal distribution, and the resulting MLE has been used by applied econometricians to analyze many problems in economics. In Tobin's honor, Equation (11.21), combined with the assumption of normal errors, is called the *tobit* regression model. The tobit model is an example of a *censored regression model*, so called because the dependent variable has been "censored" above or below a certain cutoff.

Sample Selection Models

In the censored regression model, there are data on buyers and nonbuyers, as there would be if the data were obtained via simple random sampling of the adult population. If, however, the data are collected from sales tax records, then the data would include only buyers: There would

be no data at all for nonbuyers. Data in which observations are unavailable above or below a threshold (data for buyers only) are called truncated data. The *truncated regression model* is a regression model applied to data in which observations are simply unavailable when the dependent variable is above or below a certain cutoff.

The truncated regression model is an example of a sample selection model, in which the selection mechanism (an individual is in the sample by virtue of buying a car) is related to the value of the dependent variable (expenditure on a car). As discussed in the box "James Heckman and Daniel McFadden, Nobel Laureates" in Section 11.5, one approach to estimation of sample selection models is to develop two equations, one for Y_i^* and one for whether Y_i^* is observed. The parameters of the model can then be estimated by maximum likelihood, or, in a stepwise procedure, estimating the selection equation first and then estimating the equation for Y_i^*. For additional discussion, see Ruud (2000, Chapter 28), Greene (2018, Chapter 19), or Wooldridge (2010, Chapter 17).

Count Data

Count data arise when the dependent variable is a counting number—for example, the number of restaurant meals eaten by a consumer in a week. When these numbers are large, the variable can be treated as approximately continuous, but when they are small, the continuous approximation is a poor one. The linear regression model, estimated by OLS, can be used for count data, even if the number of counts is small. Predicted values from the regression are interpreted as the expected value of the dependent variable, conditional on the regressors. So when the dependent variable is the number of restaurant meals eaten, a predicted value of 1.7 means, on average, 1.7 restaurant meals per week. As in the binary regression model, however, OLS does not take advantage of the special structure of count data and can yield nonsense predictions: for example, −0.2 restaurant meals per week. Just as probit and logit eliminate nonsense predictions when the dependent variable is binary, special models do so for count data. The two most widely used models are the Poisson and negative binomial regression models.

Ordered Responses

Ordered response data arise when mutually exclusive qualitative categories have a natural ordering, such as obtaining a high school diploma, obtaining some college education (but not graduating), or graduating from college. Like count data, ordered response data have a natural ordering, but unlike count data, they do not have natural numerical values.

Because there are no natural numerical values for ordered response data, OLS is inappropriate. Instead, ordered data are often analyzed using a generalization of probit called the *ordered probit model*, in which the probability of each outcome (e.g., a college education), conditional on the independent variables (such as parents' income), is modeled using the cumulative normal distribution.

Discrete Choice Data

A *discrete choice* or *multiple choice* variable can take on multiple unordered qualitative values. One example in economics is the mode of transport chosen by a commuter: She might take the subway, ride the bus, drive, or make her way under her own power (walk, bicycle). If we were to analyze these choices, the dependent variable would have four possible outcomes (subway, bus, car, and human-powered). These outcomes are not ordered in any natural way. Instead, the outcomes are a choice among distinct qualitative alternatives.

The econometric task is to model the probability of choosing the various options given various regressors such as individual characteristics (how far the commuter's house is from the subway station) and the characteristics of each option (the price of the subway). As discussed in the box in Section 11.5, models for analysis of discrete choice data can be developed from principles of utility maximization. Individual choice probabilities can be expressed in probit or logit form, and those models are called *multinomial probit* and *multinomial logit* regression models.