11.8 Consider the linear probability model $Y_i = \beta_0 + \beta_1 X_i + u_i$, where $\Pr(Y_i = 1|X_i) = \beta_0 + \beta_1 X_i$.

a. Show that $E(u_i|X_i) = 0$.

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 white and one black, apply for a mortgage. They have the same values for all the regressors other than race. How much more likely is the black 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 that 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) Which model would you use for:

a. A study explaining the number of minutes that a person spends talking on a cell phone during the month?

b. A study explaining grades (A through F) in a large Principles of Economics class?

c. A study of consumers' choices for Coke, Pepsi, or generic cola?

d. A study of the number of cell phones owned by a family?

Empirical Exercises

E11.1 It has been conjectured that workplace smoking bans induce 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 textbook Web site **www.pearsonhighered.com/stock_watson** 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 Web site.

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 a reason, based on the substance of this regression, explaining the change in the estimated effect of a smoking ban between (b) and (c).

d. Test the hypothesis that the coefficient on *smkban* is zero 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. Based on the regression in (c), is there a nonlinear relationship between *age* and the probability of smoking? Plot the relationship between the probability of smoking and *age* for $18 \leq age \leq 65$ for a white, non-Hispanic male college graduate with no workplace smoking ban.

E11.2 This exercise uses the same data as Empirical Exercise 11.1.

a. Estimate a probit model using the same regressors as in Empirical Exercise 11.1(c).

b. Test the hypothesis that the coefficient on *smkban* is zero in the population version of this probit regression against the alternative that it is

[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.

nonzero, at the 5% significance level. Compare your t-statistic and your conclusion with those of Empirical Exercise 11.1(d) based on the linear probability model.

c. Test the hypothesis that the probability of smoking does not depend on the level of education in this probit model. Compare your results with those in Empirical Exercise 11.1(e) using the linear probability model.

d. Mr. A is white, non-Hispanic, 20 years old, and a high school dropout. Using the probit regression from (a) 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?

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

f. Repeat (d) and (e) using the linear probability model from Empirical Exercise 11.1(c).

g. Based on the answers to (d) through (f), do the probit and linear probability model results differ? If they do, which results make more sense? Are the estimated effects large in a real-world sense?

h. Are there important remaining threats to internal validity?

E11.3 In this exercise you will study health insurance, health status, and employment using a random sample of more than 8000 workers in the United States surveyed in 1996. The data are available on the textbook Web site **www.pearsonhighered.com/stock_watson** in the file **Insurance**.[8] A detailed description is given in **Insurance_Description**, available on the Web site.

a. Are the self-employed less likely to have health insurance than wage earners? If so, is the difference large in a real-world sense? Is the difference statistically significant?

b. The self-employed might systematically differ from wage earners in their age, education, and so forth. After you control for these other factors, are the self-employed less likely to have health insurance?

c. How does health insurance status vary with age? Are older workers more likely to have health insurance? Less likely?

[8]These data were provided by Professor Harvey Rosen of Princeton University and were used in his paper with Craig Perry, "The Self-Employed Are Less Likely Than Wage-Earners to Have Health Insurance. So What?" in Douglas Holtz-Eakin and Harvey S. Rosen, eds., *Entrepeneurship and Public Policy* (Cambridge, MA: MIT Press, 2004).

d. Is the effect of self-employment on insurance status different for older workers than it is for younger workers?

e. It has been argued that the self-employed are less likely to be insured, but despite this, they are just as healthy as wage-earners. Is this right? Does the argument hold up for young workers? For older workers? Are there potential two-way causality problems that might undermine the internal validity of this kind of statistical analysis?

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 the article by Alicia H. Munnell, Geoffrey M. B. Tootell, Lynne E. Browne, and James McEneaney, "Mortgage Lending in Boston: Interpreting HMDA Data," *American Economic Review*, 1996, pp. 25–53.

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.