

KEY TERMS

Asymptotic Bias	Asymptotic <i>t</i> Statistics	Lagrange Multiplier (<i>LM</i>)
Asymptotic Confidence Interval	Asymptotic Variance	Statistic
Asymptotic Normality	Asymptotically Efficient	Large Sample Properties
Asymptotic Properties	Auxiliary Regression	<i>n</i> -R-Squared Statistic
Asymptotic Standard Error	Consistency	Score Statistic
	Inconsistency	

PROBLEMS

- 5.1** In the simple regression model under MLR.1 through MLR.4, we argued that the slope estimator, $\hat{\beta}_1$, is consistent for β_1 . Using $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1$, show that $\text{plim } \hat{\beta}_0 = \beta_0$. [You need to use the consistency of $\hat{\beta}_1$ and the law of large numbers, along with the fact that $\beta_0 = E(y) - \beta_1 E(x_1)$.]
- 5.2** Suppose that the model

$$\text{pctstck} = \beta_0 + \beta_1 \text{funds} + \beta_2 \text{risktol} + u$$

satisfies the first four Gauss-Markov assumptions, where pctstck is the percentage of a worker's pension invested in the stock market, funds is the number of mutual funds that the worker can choose from, and risktol is some measure of risk tolerance (larger risktol means the person has a higher tolerance for risk). If funds and risktol are positively correlated, what is the inconsistency in $\hat{\beta}_1$, the slope coefficient in the simple regression of pctstck on funds ?

- 5.3** The data set SMOKE.RAW contains information on smoking behavior and other variables for a random sample of single adults from the United States. The variable $cigs$ is the (average) number of cigarettes smoked per day. Do you think $cigs$ has a normal distribution in the U.S. adult population? Explain.
- 5.4** In the simple regression model (5.16), under the first four Gauss-Markov assumptions, we showed that estimators of the form (5.17) are consistent for the slope, β_1 . Given such an estimator, define an estimator of β_0 by $\tilde{\beta}_0 = \bar{y} - \tilde{\beta}_1\bar{x}$. Show that $\text{plim } \tilde{\beta}_0 = \beta_0$.

COMPUTER EXERCISES

- C5.1** Use the data in WAGE1.RAW for this exercise.

- (i) Estimate the equation

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{tenure} + u.$$

Save the residuals and plot a histogram.

- (ii) Repeat part (i), but with $\log(\text{wage})$ as the dependent variable.
 (iii) Would you say that Assumption MLR.6 is closer to being satisfied for the level-level model or the log-level model?

C5.2 Use the data in GPA2.RAW for this exercise.

- (i) Using all 4,137 observations, estimate the equation

$$\text{colgpa} = \beta_0 + \beta_1 \text{hsperc} + \beta_2 \text{sat} + u$$

and report the results in standard form.

- (ii) Reestimate the equation in part (i), using the first 2,070 observations.
- (iii) Find the ratio of the standard errors on hsperc from parts (i) and (ii). Compare this with the result from (5.10).

C5.3 In equation (4.42) of Chapter 4, compute the LM statistic for testing whether mothedu and fathtedu are jointly significant. In obtaining the residuals for the restricted model, be sure that the restricted model is estimated using only those observations for which all variables in the unrestricted model are available (see Example 4.9).

C5.4 Several statistics are commonly used to detect nonnormality in underlying population distributions. Here we will study one that measures the amount of skewness in a distribution. Recall that any normally distributed random variable is symmetric about its mean; therefore, if we standardize a symmetrically distributed random variable, say $z = (y - \mu_y)/\sigma_y$, where $\mu_y = E(y)$ and $\sigma_y = \text{sd}(y)$, then z has mean zero, variance one, and $E(z^3) = 0$. Given a sample of data $\{y_i : i = 1, \dots, n\}$, we can standardize y_i in the sample by using $z_i = (y_i - \bar{\mu}_y)/\hat{\sigma}_y$, where $\bar{\mu}_y$ is the sample mean and $\hat{\sigma}_y$ is the sample standard deviation. (We ignore the fact that these are estimates based on the sample.) A sample statistic that measures skewness is $n^{-1} \sum_{i=1}^n z_i^3$, or where n is replaced with $(n - 1)$ as a degrees-of-freedom adjustment. If y has a normal distribution in the population, the skewness measure in the sample for the standardized values should not differ significantly from zero.

- (i) First use the data set 401KSUBS.RAW, keeping only observations with $fsize = 1$. Find the skewness measure for inc . Do the same for $\log(inc)$. Which variable has more skewness and therefore seems less likely to be normally distributed?
- (ii) Next use BWGHT2.RAW. Find the skewness measures for $bwght$ and $\log(bwght)$. What do you conclude?
- (iii) Evaluate the following statement: "The logarithmic transformation always makes a positive variable look more normally distributed."
- (iv) If we are interested in the normality assumption in the context of regression, should we be evaluating the unconditional distributions of y and $\log(y)$? Explain.

Appendix 5A

Asymptotic Normality of OLS

We sketch a proof of the asymptotic normality of OLS [Theorem 5.2(i)] in the simple regression case. Write the simple regression model as in equation (5.16). Then, by the usual algebra of simple regression, we can write

$$\sqrt{n}(\hat{\beta}_1 - \beta_1) = (1/\hat{\sigma}_x^2) \left[n^{-1/2} \sum_{i=1}^n (x_i - \bar{x})u_i \right].$$

C6.8 Use the data in HPRICE1.RAW for this exercise.

- (i) Estimate the model

$$price = \beta_0 + \beta_1 lotsize + \beta_2 sqft + \beta_3 bdrms + u$$

© CourseSmart

and report the results in the usual form, including the standard error of the regression. Obtain predicted price, when we plug in $lotsize = 10,000$, $sqft = 2,300$, and $bdrms = 4$; round this price to the nearest dollar.

- (ii) Run a regression that allows you to put a 95% confidence interval around the predicted value in part (i). Note that your prediction will differ somewhat due to rounding error.
(iii) Let $price^0$ be the unknown future selling price of the house with the characteristics used in parts (i) and (ii). Find a 95% CI for $price^0$ and comment on the width of this confidence interval.

C6.9 The data set NBASAL.RAW contains salary information and career statistics for 269 players in the National Basketball Association (NBA).

- (i) Estimate a model relating points-per-game (*points*) to years in the league (*exper*), *age*, and years played in college (*coll*). Include a quadratic in *exper*; the other variables should appear in level form. Report the results in the usual way.
(ii) Holding college years and age fixed, at what value of experience does the next year of experience actually reduce points-per-game? Does this make sense?
(iii) Why do you think *coll* has a negative and statistically significant coefficient? (*Hint:* NBA players can be drafted before finishing their college careers and even directly out of high school.)
(iv) Add a quadratic in *age* to the equation. Is it needed? What does this appear to imply about the effects of age, once experience and education are controlled for?
(v) Now regress $\log(wage)$ on *points*, *exper*, $exper^2$, *age*, and *coll*. Report the results in the usual format.
(vi) Test whether *age* and *coll* are jointly significant in the regression from part (v). What does this imply about whether age and education have separate effects on wage, once productivity and seniority are accounted for?

C6.10 Use the data in BWGHT2.RAW for this exercise.

© CourseSmart

- (i) Estimate the equation

$$\log(bwght) = \beta_0 + \beta_1 npvis + \beta_2 npvis^2 + u$$

- by OLS, and report the results in the usual way. Is the quadratic term significant?
(ii) Show that, based on the equation from part (i), the number of prenatal visits that maximizes $\log(bwght)$ is estimated to be about 22. How many women had at least 22 prenatal visits in the sample?
(iii) Does it make sense that birth weight is actually predicted to decline after 22 prenatal visits? Explain.
(iv) Add mother's age to the equation, using a quadratic functional form. Holding *npvis* fixed, at what mother's age is the birth weight of the child maximized? What fraction of women in the sample are older than the "optimal" age?

KEY TERMS

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

PROBLEMS

- 7.1** Using the data in SLEEP75.RAW (see also Problem 3.3), we obtain the estimated equation

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

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

- (i) All other factors being equal, is there evidence that men sleep more than women? How strong is the evidence?
- (ii) Is there a statistically significant tradeoff between working and sleeping? What is the estimated tradeoff?
- (iii) What other regression do you need to run to test the null hypothesis that, holding other factors fixed, age has no effect on sleeping?

- 7.2** The following equations were estimated using the data in BWGHT.RAW:

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

and

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

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

- In the first equation, interpret the coefficient on the variable *cigs*. In particular, what is the effect on birth weight from smoking 10 more cigarettes per day?
- How much more is a white child predicted to weigh than a nonwhite child, holding the other factors in the first equation fixed? Is the difference statistically significant?
- Comment on the estimated effect and statistical significance of *motheduc*.
- From the given information, why are you unable to compute the *F* statistic for joint significance of *motheduc* and *fatheduc*? What would you have to do to compute the *F* statistic?

- 7.3** Using the data in GPA2.RAW, the following equation was estimated:

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

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

- Is there strong evidence that *hsize*² should be included in the model? From this equation, what is the optimal high school size?
- Holding *hsize* fixed, what is the estimated difference in SAT score between nonblack females and nonblack males? How statistically significant is this estimated difference?
- What is the estimated difference in SAT score between nonblack males and black males? Test the null hypothesis that there is no difference between their scores, against the alternative that there is a difference.
- What is the estimated difference in SAT score between black females and nonblack females? What would you need to do to test whether the difference is statistically significant?

- 7.4** An equation explaining chief executive officer salary is

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

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

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

- (ii) Use equation (7.10) to obtain the exact percentage difference in estimated salary between the utility and transportation industries and compare this with the answer obtained in part (i).
- (iii) What is the approximate percentage difference in estimated salary between the consumer products and finance industries? Write an equation that would allow you to test whether the difference is statistically significant.

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

- (i) If $noPC$ is used in place of PC in equation (7.6), what happens to the intercept in the estimated equation? What will be the coefficient on $noPC$? (*Hint:* Write $PC = 1 - noPC$ and plug this into the equation $\widehat{colGPA} = \beta_0 + \delta_0 PC + \beta_1 hsGPA + \beta_2 ACT$.)
- (ii) What will happen to the R -squared if $noPC$ is used in place of PC ?
- (iii) Should PC and $noPC$ both be included as independent variables in the model? Explain.

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

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

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

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

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

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

- (i) Write an equation that would allow you to estimate the effects of marijuana usage on wage, while controlling for other factors. You should be able to make statements such as, "Smoking marijuana five more times per month is estimated to change wage by $x\%$."
- (ii) Write a model that would allow you to test whether drug usage has different effects on wages for men and women. How would you test that there are no differences in the effects of drug usage for men and women?
- (iii) Suppose you think it is better to measure marijuana usage by putting people into one of four categories: nonuser, light user (1 to 5 times per month), moderate user (6 to 10 times per month), and heavy user (more than 10 times per month). Now, write a model that allows you to estimate the effects of marijuana usage on wage.
- (iv) Using the model in part (iii), explain in detail how to test the null hypothesis that marijuana usage has no effect on wage. Be very specific and include a careful listing of degrees of freedom.

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

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

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

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

- (i) Since it changes nothing important, set the error to zero, $u = 0$. Then, when $d = 0$ we can write the relationship between y and z as the function $f_0(z) = \beta_0 + \beta_1 z$. Write the same relationship when $d = 1$, where you should use $f_1(z)$ on the left-hand side to denote the linear function of z .
- (ii) Assuming that $\delta_1 \neq 0$ (which means the two lines are not parallel), show that the value of z^* such that $f_0(z^*) = f_1(z^*)$ is $z^* = -\delta_0/\delta_1$. This is the point at which the two lines intersect [as in Figure 7.2(b)]. Argue that z^* is positive if and only if δ_0 and δ_1 have opposite signs.
- (iii) Using the data in TWOYEAR.RAW, the following equation can be estimated:

© CourseSmart $\widehat{\log(wage)} = 2.289 - .357 \text{ female} + .50 \text{ totcoll} + .030 \text{ female} \cdot \text{totcoll}$

(0.011)	(.015)	(.003)	(.005)
---------	--------	--------	--------

$n = 6,763, R^2 = .202,$

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

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

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

- (i) If you run a simple regression $score_i$ on $voucher_i$ using a random sample of size n , does the OLS estimator provide an unbiased estimator of the effect of the voucher program?
- (ii) Suppose you can collect additional background information, such as family income, family structure (e.g., whether the child lives with both parents), and parents' education levels. Do you need to control for these factors to obtain an unbiased estimator of the effects of the voucher program? Explain.
- (iii) Why should you include the family background variables in the regression? Is there a situation in which you would not include the background variables?

© CourseSmart

COMPUTER EXERCISES

- C7.1** Use the data in GPA1.RAW for this exercise.

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

- (ii) Test for joint significance of *mothcoll* and *fathcoll* in the equation from part (i) and be sure to report the *p*-value.
- (iii) Add *hsGPA*² to the model from part (i) and decide whether this generalization is needed.

C7.2 Use the data in WAGE2.RAW for this exercise.

- (i) Estimate the model

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

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

- (ii) Add the variables *exper*² and *tenure*² to the equation and show that they are jointly insignificant at even the 20% level.
- (iii) Extend the original model to allow the return to education to depend on race and test whether the return to education does depend on race.
- (iv) Again, start with the original model, but now allow wages to differ across four groups of people: married and black, married and nonblack, single and black, and single and nonblack. What is the estimated wage differential between married blacks and married nonblacks?

C7.3 A model that allows major league baseball player salary to differ by position is

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

where outfield is the base group.

- (i) State the null hypothesis that, controlling for other factors, catchers and outfielders earn, on average, the same amount. Test this hypothesis using the data in MLB1.RAW and comment on the size of the estimated salary differential.
- (ii) State and test the null hypothesis that there is no difference in average salary across positions, once other factors have been controlled for.
- (iii) Are the results from parts (i) and (ii) consistent? If not, explain what is happening.

C7.4 Use the data in GPA2.RAW for this exercise.

- (i) Consider the equation

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

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

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

- (iii) Drop *sat* from the model and reestimate the equation. Now, what is the estimated effect of being an athlete? Discuss why the estimate is different than that obtained in part (ii).
- (iv) In the model from part (i), allow the effect of being an athlete to differ by gender and test the null hypothesis that there is no *ceteris paribus* difference between women athletes and women nonathletes.
- (v) Does the effect of *sat* on *colgpa* differ by gender? Justify your answer.

C7.5 In Problem 4.2, we added the return on the firm's stock, *ros*, to a model explaining CEO salary; *ros* turned out to be insignificant. Now, define a dummy variable, *rosneg*, which is equal to one if *ros* < 0 and equal to zero if *ros* ≥ 0. Use CEOSAL1.RAW to estimate the model

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

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

C7.6 Use the data in SLEEP75.RAW for this exercise. The equation of interest is

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

- © CourseSmart
- (i) Estimate this equation separately for men and women and report the results in the usual form. Are there notable differences in the two estimated equations?
 - (ii) Compute the Chow test for equality of the parameters in the sleep equation for men and women. Use the form of the test that adds *male* and the interaction terms *male*·*totwrk*, ..., *male*·*yngkid* and uses the full set of observations. What are the relevant *df* for the test? Should you reject the null at the 5% level?
 - (iii) Now, allow for a different intercept for males and females and determine whether the interaction terms involving *male* are jointly significant.
 - (iv) Given the results from parts (ii) and (iii), what would be your final model?

C7.7 Use the data in WAGE1.RAW for this exercise.

- (i) Use equation (7.18) to estimate the gender differential when *educ* = 12.5. Compare this with the estimated differential when *educ* = 0.
- (ii) Run the regression used to obtain (7.18), but with *female*·(*educ* − 12.5) replacing *female*·*educ*. How do you interpret the coefficient on *female* now?
- (iii) Is the coefficient on *female* in part (ii) statistically significant? Compare this with (7.18) and comment.

C7.8 Use the data in LOANAPP.RAW for this exercise. The binary variable to be explained is *approve*, which is equal to one if a mortgage loan to an individual was approved. The key explanatory variable is *white*, a dummy variable equal to one if the applicant was white. The other applicants in the data set are black and Hispanic.

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

© CourseSmart

$$\text{approve} = \beta_0 + \beta_1 \text{white} + \text{other factors}.$$

- (i) If there is discrimination against minorities, and the appropriate factors have been controlled for, what is the sign of β_1 ?
- (ii) Regress *approve* on *white* and report the results in the usual form. Interpret the coefficient on *white*. Is it statistically significant? Is it practically large?

- (iii) As controls, add the variables *hrat*, *obrat*, *loanprc*, *unem*, *male*, *married*, *dep*, *sch*, *cosign*, *chist*, *pubrec*, *mortal1*, *mortal2*, and *vr*. What happens to the coefficient on *white*? Is there still evidence of discrimination against nonwhites?
- (iv) Now, allow the effect of race to interact with the variable measuring other obligations as a percentage of income (*obrat*). Is the interaction term significant?
- (v) Using the model from part (iv), what is the effect of being white on the probability of approval when *obrat* = 32, which is roughly the mean value in the sample? Obtain a 95% confidence interval for this effect.

C7.9 There has been much interest in whether the presence of 401(k) pension plans, available to many U.S. workers, increases net savings. The data set 401KSUBS.RAW contains information on net financial assets (*nettfa*), family income (*inc*), a binary variable for eligibility in a 401(k) plan (*e401k*), and several other variables.

- (i) What fraction of the families in the sample are eligible for participation in a 401(k) plan?
- (ii) Estimate a linear probability model explaining 401(k) eligibility in terms of income, age, and gender. Include income and age in quadratic form, and report the results in the usual form.
- (iii) Would you say that 401(k) eligibility is independent of income and age? What about gender? Explain.
- (iv) Obtain the fitted values from the linear probability model estimated in part (ii). Are any fitted values negative or greater than one?
- (v) Using the fitted values $\widehat{e401k}_i$ from part (iv), define $\widehat{e401k}_i = 1$ if $\widehat{e401k}_i \geq .5$ and $\widehat{e401k}_i = 0$ if $\widehat{e401k}_i < .5$. Out of 9,275 families, how many are predicted to be eligible for a 401(k) plan?
- (vi) For the 5,638 families not eligible for a 401(k), what percentage of these are predicted not to have a 401(k), using the predictor $\widehat{e401k}_i$? For the 3,637 families eligible for a 401(k) plan, what percentage are predicted to have one? (It is helpful if your econometrics package has a "tabulate" command.)
- (vii) The overall percent correctly predicted is about 64.9%. Do you think this is a complete description of how well the model does, given your answers in part (vi)?
- (viii) Add the variable *pira* as an explanatory variable to the linear probability model. Other things equal, if a family has someone with an individual retirement account, how much higher is the estimated probability that the family is eligible for a 401(k) plan? Is it statistically different from zero at the 10% level?

C7.10 Use the data in NBASAL.RAW for this exercise.

- (i) Estimate a linear regression model relating points per game to experience in the league and position (guard, forward, or center). Include experience in quadratic form and use centers as the base group. Report the results in the usual form.
- (ii) Why do you not include all three position dummy variables in part (i)?
- (iii) Holding experience fixed, does a guard score more than a center? How much more? Is the difference statistically significant?
- (iv) Now, add marital status to the equation. Holding position and experience fixed, are married players more productive (based on points per game)?
- (v) Add interactions of marital status with both experience variables. In this expanded model, is there strong evidence that marital status affects points per game?
- (vi) Estimate the model from part (iv) but use assists per game as the dependent variable. Are there any notable differences from part (iv)? Discuss.

C7.11 Use the data in 401KSUBS.RAW for this exercise.

- (i) Compute the average, standard deviation, minimum, and maximum values of *nettfa* in the sample.
- (ii) Test the hypothesis that average *nettfa* does not differ by 401(k) eligibility status; use a two-sided alternative. What is the dollar amount of the estimated difference?
- (iii) From part (ii) of Computer Exercise C7.9, it is clear that *e401k* is not exogenous in a simple regression model; at a minimum, it changes by income and age. Estimate a multiple linear regression model for *nettfa* that includes income, age, and *e401k* as explanatory variables. The income and age variables should appear as quadratics. Now, what is the estimated dollar effect of 401(k) eligibility?
- (iv) To the model estimated in part (iii), add the interactions $e401k \cdot (\text{age} - 41)$ and $e401k \cdot (\text{age} - 41)^2$. Note that the average age in the sample is about 41, so that in the new model, the coefficient on *e401k* is the estimated effect of 401(k) eligibility at the average age. Which interaction term is significant?
- (v) Comparing the estimates from parts (iii) and (iv), do the estimated effects of 401(k) eligibility at age 41 differ much? Explain.
- (vi) Now, drop the interaction terms from the model, but define five family size dummy variables: *fsize1*, *fsize2*, *fsize3*, *fsize4*, and *fsize5*. The variable *fsize5* is unity for families with five or more members. Include the family size dummies in the model estimated from part (iii); be sure to choose a base group. Are the family dummies significant at the 1% level?
- (vii) Now, do a Chow test for the model

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

across the five family size categories, allowing for intercept differences. The restricted sum of squared residuals, SSR_r , is obtained from part (vi) because that regression assumes all slopes are the same. The unrestricted sum of squared residuals is $\text{SSR}_u = \text{SSR}_1 + \text{SSR}_2 + \dots + \text{SSR}_5$, where SSR_f is the sum of squared residuals for the equation estimated using only family size f . You should convince yourself that there are 30 parameters in the unrestricted model (5 intercepts plus 25 slopes) and 10 parameters in the restricted model (5 intercepts plus 5 slopes). Therefore, the number of restrictions being tested is $q = 20$, and the df for the unrestricted model is $9,275 - 30 = 9,245$.

C7.12 Use the data set in BEAUTY.RAW, which contains a subset of the variables (but more usable observations than in the regressions) reported by Hamermesh and Biddle (1994).

- (i) Find the separate fractions of men and women that are classified as having above average looks. Are more people rated as having above average or below average looks?
- (ii) Test the null hypothesis that the population fractions of above-average-looking women and men are the same. Report the one-sided p -value that the fraction is higher for women. (Hint: Estimating a simple linear probability model is easiest.)
- (iii) Now estimate the model

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

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

© CourseSmart

- (iv) Is there convincing evidence that women with above average looks earn more than women with average looks? Explain.
- (v) For both men and women, add the explanatory variables *educ*, *exper*, *exper*², *union*, *goodlth*, *black*, *married*, *south*, *bigcity*, *smllcity*, and *service*. Do the effects of the "looks" variables change in important ways?

C7.13 Use the data in APPLE.RAW to answer this question.

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

$$\begin{aligned} \text{ecobuy} = & \beta_0 + \beta_1 \text{ecoprc} + \beta_2 \text{regprc} + \beta_3 \text{faminc} \\ & + \beta_4 \text{hhsize} + \beta_5 \text{educ} + \beta_6 \text{age} + u, \end{aligned}$$

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

- (iii) Are the nonprice variables jointly significant in the LPM? (Use the usual *F* statistic, even though it is not valid when there is heteroskedasticity.) Which explanatory variable other than the price variables seems to have the most important effect on the decision to buy ecolabeled apples? Does this make sense to you?
- (iv) In the model from part (ii), replace *faminc* with *log(faminc)*. Which model fits the data better, using *faminc* or *log(faminc)*? Interpret the coefficient on *log(faminc)*.
- (v) In the estimation in part (iv), how many estimated probabilities are negative? How many are bigger than one? Should you be concerned?
- (vi) For the estimation in part (iv), compute the percent correctly predicted for each outcome, *ecobuy* = 0 and *ecobuy* = 1. Which outcome is best predicted by the model?

C7.14 Use the data in CHARITY.RAW to answer this question. The variable *respond* is a dummy variable equal to one if a person responded with a contribution on the most recent mailing sent by a charitable organization. The variable *resplast* is a dummy variable equal to one if the person responded to the previous mailing, *avggift* is the average of past gifts (in Dutch guilders), and *propresp* is the proportion of times the person has responded to past mailings.

- (i) Estimate a linear probability model relating *respond* to *resplast* and *avggift*. Report the results in the usual form, and interpret the coefficient on *resplast*.
- (ii) Does the average value of past gifts seem to affect the probability of responding?
- (iii) Add the variable *propresp* to the model, and interpret its coefficient. (Be careful here: an increase of one in *propresp* is the largest possible change.)
- (iv) What happened to the coefficient on *resplast* when *propresp* was added to the regression? Does this make sense?
- (v) Add *mailyear*, the number of mailings per year, to the model. How big is its estimated effect? Why might this not be a good estimate of the causal effect of mailings on responding?