PROBLEMS

6.1 The following equation was estimated using the data in CEOSAL1.RAW:

$$\widehat{\log(salary)} = 4.322 + .276\log(sales) + .0215\,roe - .00008\,roe^2$$
$$(.324)\ \ (.033)\qquad\qquad (.0129)\qquad (.00026)$$
$$n = 209,\ R^2 = .282.$$

This equation allows *roe* to have a diminishing effect on log(*salary*). Is this generality necessary? Explain why or why not.

6.2 Let $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ be the OLS estimates from the regression of y_i on $x_{i1}, \ldots, x_{ik}$, $i=1, 2, \ldots, n$. For nonzero constants $c_1, \ldots, c_k$, argue that the OLS intercept and slopes from the regression of $c_0 y_i$ on $c_1 x_{i1}, \ldots, c_k x_{ik}$, $i = 1, 2, \ldots, n$, are given by $\tilde{\beta}_0 = c_0\hat{\beta}_0$, $\tilde{\beta}_1 = (c_0/c_1)\hat{\beta}_1, \ldots, \tilde{\beta}_k = (c_0/c_k)\hat{\beta}_k$. [*Hint*: Use the fact that the $\hat{\beta}_j$ solve the first order conditions in (3.13), and the $\tilde{\beta}_j$ must solve the first order conditions involving the rescaled dependent and independent variables.]

6.3 Using the data in RDCHEM.RAW, the following equation was obtained by OLS:

$$\widehat{rdintens} = 2.613 + .00030\,sales - .0000000070\,sales^2$$
$$(.429)\ \ (.00014)\qquad\qquad (.0000000037)$$
$$n = 32,\ R^2 = .1484.$$

(i) At what point does the marginal effect of *sales* on *rdintens* become negative?

(ii) Would you keep the quadratic term in the model? Explain.

(iii) Define *salesbil* as sales measured in billions of dollars: *salesbil* = *sales*/1,000. Rewrite the estimated equation with *salesbil* and *salesbil*^{2} as the independent variables. Be sure to report standard errors and the *R*-squared. [*Hint*: Note that *salesbil*^{2} = *sales*^{2}/(1,000)^2.]

(iv) For the purpose of reporting the results, which equation do you prefer?

6.4 The following model allows the return to education to depend upon the total amount of both parents' education, called *pareduc*:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 educ\cdot pareduc + \beta_3 exper + \beta_4 tenure + u.$$

(i) Show that, in decimal form, the return to another year of education in this model is

$$\Delta\log(wage)/\Delta educ = \beta_1 + \beta_2 pareduc.$$

What sign do you expect for β_2? Why?

(ii) Using the data in WAGE2.RAW, the estimated equation is

$$\widehat{\log(wage)} = \underset{(.13)}{5.65} + \underset{(.010)}{.047}\, educ + \underset{(.00021)}{.00078}\, educ{\cdot}pareduc + \underset{(.004)}{.019}\, exper + \underset{(.003)}{.010}\, tenure$$

$$n = 722,\ R^2 = .169.$$

(Only 722 observations contain full information on parents' education.) Interpret the coefficient on the interaction term. It might help to choose two specific values for *pareduc*—for example, *pareduc* = 32 if both parents have a college education, or *pareduc* = 24 if both parents have a high school education—and to compare the estimated return to *educ*.

(iii) When *pareduc* is added as a separate variable to the equation, we get:

$$\widehat{\log(wage)} = \underset{(.38)}{4.94} + \underset{(.027)}{.097}\, educ + \underset{(.017)}{.033}\, pareduc - \underset{(.0012)}{.0016}\, educ{\cdot}pareduc + \underset{(.004)}{.020}\, exper + \underset{(.003)}{.010}\, tenure$$

$$n = 722,\ R^2 = .174.$$

Does the estimated return to education now depend positively on parent education? Test the null hypothesis that the return to education does not depend on parent education.

6.5 In Example 4.2, where the percentage of students receiving a passing score on a tenth-grade math exam (*math*10) is the dependent variable, does it make sense to include *sci11*—the percentage of eleventh graders passing a science exam—as an additional explanatory variable?

6.6 When $atndrte^2$ and $ACT{\cdot}atndrte$ are added to the equation estimated in (6.19), the *R*-squared becomes .232. Are these additional terms jointly significant at the 10% level? Would you include them in the model?

6.7 The following three equations were estimated using the 1,534 observations in 401K.RAW:

$$\widehat{prate} = \underset{(.78)}{80.29} + \underset{(.52)}{5.44}\, mrate + \underset{(.045)}{.269}\, age - \underset{(.00004)}{.00013}\, totemp$$

$$R^2 = .100,\ \bar{R}^2 = .098.$$

$$\widehat{prate} = \underset{(1.95)}{97.32} + \underset{(0.51)}{5.02}\, mrate + \underset{(.044)}{.314}\, age - \underset{(.28)}{2.66}\, \log(totemp)$$

$$R^2 = .144,\ \bar{R}^2 = .142.$$

$$\widehat{prate} = \underset{(.78)}{80.62} + \underset{(.52)}{5.34}\, mrate + \underset{(.045)}{.290}\, age - \underset{(.00009)}{.00043}\, totemp + \underset{(.0000000010)}{.0000000039}\, totemp^2$$

$$R^2 = .108,\ \bar{R}^2 = .106.$$

Which of these three models do you prefer? Why?

6.8 Suppose we want to estimate the effects of alcohol consumption (*alcohol*) on college grade point average (*colGPA*). In addition to collecting information on grade point averages and alcohol usage, we also obtain attendance information (say, percentage of lectures attended, called *attend*). A standardized test score (say, *SAT*) and high school GPA (*hsGPA*) are also available.

(i) Should we include *attend* along with *alcohol* as explanatory variables in a multiple regression model? (Think about how you would interpret $\beta_{alcohol}$.)

(ii) Should *SAT* and *hsGPA* be included as explanatory variables? Explain.

6.9 If we start with (6.38) under the CLM assumptions, assume large n, and ignore the estimation error in the $\hat{\beta}_j$, a 95% prediction interval for y^0 is $[\exp(-1.96\hat{\sigma})\exp(\widehat{logy}^0), \exp(1.96\hat{\sigma})\exp(\widehat{logy}^0)]$. The point prediction for y^0 is $\hat{y}^0 = \exp(\hat{\sigma}^2/2)\exp(\widehat{logy}^0)$.

(i) For what values of $\hat{\sigma}$ will the point prediction be in the 95% prediction interval? Does this condition seem likely to hold in most applications?

(ii) Verify that the condition from part (i) is satisfied in the CEO salary example.

COMPUTER EXERCISES

C6.1 Use the data in KIELMC.RAW, only for the year 1981, to answer the following questions. The data are for houses that sold during 1981 in North Andover, Massachusetts; 1981 was the year construction began on a local garbage incinerator.

(i) To study the effects of the incinerator location on housing price, consider the simple regression model

$$\log(price) = \beta_0 + \beta_1 \log(dist) + u,$$

where *price* is housing price in dollars and *dist* is distance from the house to the incinerator measured in feet. Interpreting this equation causally, what sign do you expect for β_1 if the presence of the incinerator depresses housing prices? Estimate this equation and interpret the results.

(ii) To the simple regression model in part (i), add the variables log(*intst*), log(*area*), log(*land*), *rooms*, *baths*, and *age*, where *intst* is distance from the home to the interstate, *area* is square footage of the house, *land* is the lot size in square feet, *rooms* is total number of rooms, *baths* is number of bathrooms, and *age* is age of the house in years. Now, what do you conclude about the effects of the incinerator? Explain why (i) and (ii) give conflicting results.

(iii) Add $[\log(intst)]^2$ to the model from part (ii). Now what happens? What do you conclude about the importance of functional form?

(iv) Is the square of log(*dist*) significant when you add it to the model from part (iii)?

C6.2 Use the data in WAGE1.RAW for this exercise.

(i) Use OLS to estimate the equation

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u$$

and report the results using the usual format.

(ii) Is $exper^2$ statistically significant at the 1% level?

(iii) Using the approximation

$$\%\Delta\widehat{wage} \approx 100(\hat{\beta}_2 + 2\hat{\beta}_3 exper)\Delta exper,$$

find the approximate return to the fifth year of experience. What is the approximate return to the twentieth year of experience?

(iv) At what value of *exper* does additional experience actually lower predicted log(*wage*)? How many people have more experience in this sample?

C6.3 Consider a model where the return to education depends upon the amount of work experience (and vice versa):

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 educ \cdot exper + u.$$

(i) Show that the return to another year of education (in decimal form), holding *exper* fixed, is $\beta_1 + \beta_3 exper$.

(ii) State the null hypothesis that the return to education does not depend on the level of *exper*. What do you think is the appropriate alternative?

(iii) Use the data in WAGE2.RAW to test the null hypothesis in (ii) against your stated alternative.

(iv) Let θ_1 denote the return to education (in decimal form), when $exper = 10$: $\theta_1 = \beta_1 + 10\beta_3$. Obtain $\hat{\theta}_1$ and a 95% confidence interval for θ_1. (*Hint*: Write $\beta_1 = \theta_1 - 10\beta_3$ and plug this into the equation; then rearrange. This gives the regression for obtaining the confidence interval for θ_1.)

C6.4 Use the data in GPA2.RAW for this exercise.

(i) Estimate the model

$$sat = \beta_0 + \beta_1 hsize + \beta_2 hsize^2 + u,$$

where *hsize* is the size of the graduating class (in hundreds), and write the results in the usual form. Is the quadratic term statistically significant?

(ii) Using the estimated equation from part (i), what is the "optimal" high school size? Justify your answer.

(iii) Is this analysis representative of the academic performance of *all* high school seniors? Explain.

(iv) Find the estimated optimal high school size, using log(*sat*) as the dependent variable. Is it much different from what you obtained in part (ii)?

C6.5 Use the housing price data in HPRICE1.RAW for this exercise.

(i) Estimate the model

$$\log(price) = \beta_0 + \beta_1 \log(lotsize) + \beta_2 \log(sqrft) + \beta_3 bdrms + u$$

and report the results in the usual OLS format.

(ii) Find the predicted value of log(*price*), when *lotsize* = 20,000, *sqrft* = 2,500, and *bdrms* = 4. Using the methods in Section 6.4, find the predicted value of *price* at the same values of the explanatory variables.

(iii) For explaining variation in *price*, decide whether you prefer the model from part (i) or the model

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

C6.6 Use the data in VOTE1.RAW for this exercise.

(i) Consider a model with an interaction between expenditures:

$$voteA = \beta_0 + \beta_1 prtystrA + \beta_2 expendA + \beta_3 expendB + \beta_4 expendA \cdot expendB + u.$$

What is the partial effect of *expendB* on *voteA*, holding *prtystrA* and *expendA* fixed? What is the partial effect of *expendA* on *voteA*? Is the expected sign for β_4 obvious?

(ii) Estimate the equation in part (i) and report the results in the usual form. Is the interaction term statistically significant?

(iii) Find the average of *expendA* in the sample. Fix *expendA* at 300 (for $300,000). What is the estimated effect of another $100,000 spent by Candidate B on *voteA*? Is this a large effect?

(iv) Now fix *expendB* at 100. What is the estimated effect of $\Delta expendA = 100$ on *voteA*? Does this make sense?

(v) Now, estimate a model that replaces the interaction with *shareA*, Candidate A's percentage share of total campaign expenditures. Does it make sense to hold both *expendA* and *expendB* fixed, while changing *shareA*?

(vi) (Requires calculus) In the model from part (v), find the partial effect of *expendB* on *voteA*, holding *prtystrA* and *expendA* fixed. Evaluate this at *expendA* = 300 and *expendB* = 0 and comment on the results.

C6.7 Use the data in ATTEND.RAW for this exercise.

(i) In the model of Example 6.3, argue that

$$\Delta stndfnl/\Delta priGPA \approx \beta_2 + 2\beta_4 priGPA + \beta_6 atndrte.$$

Use equation (6.19) to estimate the partial effect when *priGPA* = 2.59 and *atndrte* = 82. Interpret your estimate.

(ii) Show that the equation can be written as

$$stndfnl = \theta_0 + \beta_1 atndrte + \theta_2 priGPA + \beta_3 ACT + \beta_4 (priGPA - 2.59)^2 + \beta_5 ACT^2 + \beta_6 priGPA(atndrte - 82) + u,$$

where $\theta_2 = \beta_2 + 2\beta_4(2.59) + \beta_6(82)$. (Note that the intercept has changed, but this is unimportant.) Use this to obtain the standard error of $\hat{\theta}_2$ from part (i).

(iii) Suppose that, in place of *priGPA*(*atndrte* − 82), you put (*priGPA* − 2.59)·(*atndrte* − 82). Now how do you interpret the coefficients on *atndrte* and *priGPA*?

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

(i) Estimate the model

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

and report the results in the usual form, including the standard error of the regression. Obtain predicted price, when we plug in $lotsize = 10{,}000$, $sqrft = 2{,}300$, 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.

(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?

(v) Would you say that mother's age and number of prenatal visits explain a lot of the variation in log(*bwght*)?

(vi) Using quadratics for both *npvis* and *age*, decide whether using the natural log or the level of *bwght* is better for predicting *bwght*.

C6.11 Use APPLE.RAW to verify some of the claims made in Section 6.3.

(i) Run the regression *ecolbs* on *ecoprc*, *regprc* and report the results in the usual form, including the *R*-squared and adjusted *R*-squared. Interpret the coefficients on the price variables and comment on their signs and magnitudes.

(ii) Are the price variables statistically significant? Report the *p*-values for the individual *t* tests.

(iii) What is the range of fitted values for *ecolbs*? What fraction of the sample reports *ecolbs* = 0? Comment.

(iv) Do you think the price variables together do a good job of explaining variation in *ecolbs*? Explain.

(v) Add the variables *faminc*, *hhsize* (household size), *educ*, and *age* to the regression from part (i). Find the *p*-value for their joint significance. What do you conclude?

C6.12 Use the subset of 401KSUBS.RAW with *fsize* = 1; this restricts the analysis to single person households; see also Computer Exercise C4.8.

(i) What is the youngest age of people in this sample? How many people are at that age?

(ii) In the model

$$nettfa = \beta_0 + \beta_1 inc + \beta_2 age + \beta_3 age^2 + u,$$

what is the literal interpretation of β_2? By itself, is it of much interest?

(iii) Estimate the model from part (ii) and report the results in standard form. Are you concerned that the coefficient on *age* is negative? Explain.

(iv) Because the youngest people in the sample are 25, it makes sense to think that, for a given level of income, the lowest average amount of net total financial assets is at age 25. Recall that the partial effect of *age* on *nettfa* is $\beta_2 + 2\beta_3 age$, so the partial effect at age 25 is $\beta_2 + 2\beta_3(25) = \beta_2 + 50\beta_3$; call this θ_2. Find $\hat{\theta}_2$ and obtain the two-sided *p*-value for testing H_0: $\theta_2 = 0$. You should conclude that $\hat{\theta}_2$ is small and very statistically insignificant. [*Hint:* One way to do this is to estimate the model $nettfa = \alpha_0 + \beta_1 inc + \theta_2 age + \beta_3(age - 25)^2 + u$, where the intercept, α_0, is different from β_0. There are other ways, too.]

(v) Because the evidence against H_0: $\theta_2 = 0$ is very weak, set it to zero and estimate the model

$$nettfa = \alpha_0 + \beta_1 inc + \beta_3(age - 25)^2 + u.$$

In terms of goodness-of-fit, does this model fit better than that in part (ii)?

(vi) For the estimated equation in part (v), set *inc* = 30 (roughly, the average value) and graph the relationship between *nettfa* and *age*, but only for $age \geq 25$. Describe what you see.

(vii) Check to see whether including a quadratic in *inc* is necessary.

C6.13 Use the data in MEAP00_01 to answer this question.

(i) Estimate the model

$$math4 = \beta_0 + \beta_1 lexppp + \beta_2 lenroll + \beta_3 lunch + u$$

by OLS, and report the results in the usual form. Is each explanatory variable statistically significant at the 5% level?

(ii) Obtain the fitted values from the regression in part (i). What is the range of fitted values? How does it compare with the range of the actual data on *math4*?

(iii) Obtain the residuals from the regression in part (i). What is the building code of the school that has the largest (positive) residual? Provide an interpretation of this residual.

(iv) Add quadratics of all explanatory variables to the equation, and test them for joint significance. Would you leave them in the model?

(v) Returning to the model in part (i), divide the dependent variable and each explanatory variable by its sample standard deviation, and rerun the regression. (Include an intercept unless you also first subtract the mean from each variable.) In terms of standard deviation units, which explanatory variable has the largest effect on the math pass rate?

Appendix 6A

6A. A Brief Introduction to Bootstrapping

In many cases where formulas for standard errors are hard to obtain mathematically, or where they are thought not to be very good approximations to the true sampling variation of an estimator, we can rely on a **resampling method**. The general idea is to treat the observed data as a population that we can draw samples from. The most common resampling method is the **bootstrap**. (There are actually several versions of the bootstrap, but the most general, and most easily applied, is called the *nonparametric bootstrap*, and that is what we describe here.)

Suppose we have an estimate, $\hat{\theta}$, of a population parameter, θ. We obtained this estimate, which could be a function of OLS estimates (or estimates that we cover in later chapters), from a random sample of size n. We would like to obtain a standard error for $\hat{\theta}$ that can be used for constructing t statistics or confidence intervals. Remarkably, we can obtain a valid standard error by computing the estimate from different random samples drawn from the original data.

Implementation is easy. If we list our observations from 1 through n, we draw n numbers randomly, with replacement, from this list. This produces a new data set (of size n) that consists of the original data, but with many observations appearing multiple times (except in the rather unusual case that we resample the original data). Each time we randomly sample from the original data, we can estimate θ using the same procedure that we used on the original data. Let $\hat{\theta}^{(b)}$ denote the estimate from bootstrap sample b. Now, if we repeat the resampling and estimation m times, we have m new estimates, $\{\hat{\theta}^{(b)}: b = 1, 2, \ldots, m\}$. The **bootstrap standard error** of $\hat{\theta}$ is just the sample standard deviation of the $\hat{\theta}^{(b)}$, namely,

KEY TERMS

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

PROBLEMS

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

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

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:

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

and

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

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

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

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

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

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

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

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

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

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

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

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

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

7.4 An equation explaining chief executive officer salary is

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

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.

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

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

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

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 $\widehat{colGPA} = \hat{\beta}_0 + \hat{\delta}_0 PC + \hat{\beta}_1 hsGPA + \hat{\beta}_2 ACT$.)

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

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

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 $inlf = \beta_0 + \beta_1 nwifeinc + \beta_2 educ + \ldots$ and rearrange.)

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

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

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:

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

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 a 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?

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 $ros \geq 0$. Use CEOSAL1.RAW to estimate the model

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

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

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

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

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

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

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

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

C7.7 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:

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

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

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

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

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

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

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 $\widehat{2e401k} \geq .5$ and $\widehat{401k} = 0$ if $\widehat{2e401k} < .5$. 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 $e401k \cdot (age - 41)$ and $e401k \cdot (age - 41)^2$. Note that the average age in the sample is about 41, so that in the new model, the coefficient on *e401k* is the estimated effect of 401(k) eligibility at the average age. Which interaction term is significant?

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

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

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

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

across the five family size categories, allowing for intercept differences. The restricted sum of squared residuals, SSR_r, is obtained from part (vi) because that regression assumes all slopes are the same. The unrestricted sum of squared residuals is $SSR_{ur} = SSR_1 + SSR_2 + ... + SSR_5$, 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.

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

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

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

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

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

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

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

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

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

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