Homework Set 6

Due on TBA

All questions are from the end-of-chapter exercises in Chapters 8 and 9. The question numbers refer to those in the book. I highly recommend you reading the textbook and lecture notes before completing the homework questions. When reading the textbook, please pay attention to the sections on how to interpret the estimated coefficients. Note that Exercise 9.12 is an optional question.

Exercises

8.3 After reading this chapter's analysis of test scores and class size, an educator comments, "In my experience, student performance depends on class size, but not in the way your regressions say. Rather, students do well when class size is less than 20 students and do very poorly when class size is greater than 25. There are no gains from reducing class size below 20 students, the relationship is constant in the intermediate region between 20 and 25 students, and there is no loss to increasing class size when it is already greater than 25." The educator is describing a "threshold effect" in which performance is constant for class sizes less than 20, then jumps and is constant for class sizes between 20 and 25, and then jumps again for class sizes greater than 25. To model these threshold effects, define the binary variables

$$STRsmall = 1 \text{ if } STR < 20, \text{ and } STRsmall = 0 \text{ otherwise;}$$
$$STRmoderate = 1 \text{ if } 20 \leq STR \leq 25, \text{ and } STRmoderate = 0 \text{ otherwise; and}$$
$$STRlarge = 1 \text{ if } STR > 25, \text{ and } STRlarge = 0 \text{ otherwise}$$

a. Consider the regression $TestScore_i = \beta_0 + \beta_1 STRsmall_i + \beta_2 STRlarge_i + u_i$. Sketch the regression function relating $TestScore$ to STR for hypothetical values of the regression coefficients that are consistent with the educator's statement.

b. A researcher tries to estimate the regression $TestScore_i = \beta_0 + \beta_1 STRsmall_i + \beta_2 STRmoderate_i + \beta_3 STRlarge + u_i$ and finds that her computer crashes. Why?

8.7 This problem is inspired by a study of the "gender gap" in earnings in top corporate jobs [Bertrand and Hallock (2001)]. The study compares total compensation among top executives in a large set of U.S. public corporations in the 1990s. (Each year these publicly traded corporations must report total compensation levels for their top five executives.)

a. Let *Female* be an indicator variable that is equal to 1 for females and 0 for males. A regression of the logarithm of earnings onto *Female* yields

$$\widehat{\ln(Earnings)} = \underset{(0.01)}{6.48} - \underset{(0.05)}{-0.44} Female, SER = 2.65$$

i. The estimated coefficient on *Female* is -0.44. Explain what this value means.

ii. The *SER* is 2.65. Explain what this value means.

iii. Does this regression suggest that female top executives earn less than top male executives? Explain.

iv. Does this regression suggest that there is gender discrimination? Explain.

b. Two new variables, the market value of the firm (a measure of firm size, in millions of dollars) and stock return (a measure of firm performance, in percentage points), are added to the regression:

$$\widehat{\ln(Earnings)} = \underset{(0.03)}{3.86} - \underset{(0.04)}{0.28}\, Female + \underset{(0.004)}{0.37}\, \ln(MarketValue) + \underset{(0.003)}{0.004}\, Return$$
$$n = 46,670, \bar{R}^2 = 0.345$$

i. The coefficient on $\ln(MarketValue)$ is 0.37. Explain what this value means.

ii. The coefficient on *Female* is now -0.28. Explain why it has changed from the regression in (a).

c. Are large firms more likely to have female top executives than small firms? Explain.

9.5 The demand for a commodity is given by $Q = \beta_0 + \beta_1 P + u$, where Q denotes quantity, P denotes price, and u denotes factors other than price that determine demand. Supply for the commodity is given by $Q = \gamma_0 + \gamma_1 P + v$, where v denotes factors other than price that determine supply. Suppose that u and v both have a mean of zero, have variances σ_u^2 and σ_v^2, and are mutually uncorrelated.

a. Solve the two simultaneous equations to show how Q and P depend on u and v.

b. Derive the means of P and Q.

c. Derive the variance of P, the variance of Q, and the covariance between Q and P.

d. A random sample of observations of (Q_i, P_i) is collected, and Q_i is regressed on P_i. (That is, Q_i is the regressand and P_i is the regressor.) Suppose that the sample is very large.

i. Use your answers to (b) and (c) to derive values of the regression coefficient (*Hint*: Use Equation (4.7) and (4.8).)

ii. A researcher uses the slope of this regression as an estimate of the slope of the demand function (β_1). Is the estimated slope too large or too small? (*Hint*: Remember that demand curves slope down and supply curves slope up.)

9.12 (Optional) Consider the one-variable regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ and suppose that it satisfies the least squares assumptions in Key Concept 4.3. The regressor X_i is missing, but data on a related variable Z_i are available, and the value of X_i is estimated using $\tilde{X}_i = E(X_i|Z_i)$. Let $w_i = \tilde{X}_i - X_i$.

a. Show that $\tilde{X}_i$ is the minimum mean square error estimator of X_i using Z_i. That is, let $\hat{X}_i = g(Z_i)$ be some other guess of X_i based on Z_i and show that $\text{var}(\hat{X}_i - X_i) \geq \text{var}(\tilde{X}_i - X_i)$. (*Hint*: Review Exercise 2.27.)

b. Show that $E(w_i|\tilde{X}_i) = 0$.

c. Suppose that $E(u_i|Z_i) = 0$ and that $\tilde{X}_i$ is used as the regressor in place of X_i. Show that $\hat{\beta}_1$ is consistent. Is $\hat{\beta}_0$ consistent?

9.13 Assume that the regression model $Y_i = \beta_0 + \beta_1 X_i + u_i$ satisfies the least squares assumptions in Key Concept 4.3. You and a friend collect a random sample of 300 observations on Y and X.

a. Your friend reports that he inadvertently scrambled the X observations for 20% of the sample. For these scrambled observations, the value of X does not correspond to X_i for the i$^{\text{th}}$ observation, but rather to the value of X for some other observation. In the notation of Section 9.2, the measured value of the regressor, $\tilde{X}_i$, is equal to X_i, for 80% of the observations, but is equal to a randomly selected X_j for the remaining 20% of the observations. You regress Y_i on $\tilde{X}_i$. Show that $E(\hat{\beta}_1) = 0.8\beta_1$.

b. Explain how you could construct an unbiased estimate of β_1 using the OLS estimator in (a).

c. Suppose now that your friend tells you that the X's were scrambled for the first 60 observations, but that the remaining 240 observations are correct. You estimate β_1 by regressing Y on X using only the correctly measured 240 observations. Is this estimator of β_1 better than the estimator you proposed in (b)? Explain.

Empirical Exercise

E8.1 Use the data set **CPS08** described in Empirical Exercise 4.1 to answer the following questions.

a. Run a regression of average hourly earnings (AHE) on age (Age), gender ($Female$), and education ($Bachelor$). If Age increases from 25 to 26, how are earnings expected to change? If Age increases from 33 to 34, how are earnings expected to change?

b. Run a regression of the logarithm of average hourly earnings, $\ln(AHE)$, on Age, $Female$, and $Bachelor$. If Age increases from 25 to 26, how are earnings expected to change? If Age increases from 33 to 34, how are earnings expected to change?

c. Run a regression of the logarithm of average hourly earnings, $\ln(AHE)$, on $\ln(Age)$, $Female$, and $Bachelor$. If Age increases from 25 to 26, how are earnings expected to change? If Age increases from 33 to 34, how are earnings expected to change?

d. Run a regression of the logarithm of average hourly earnings, $\ln(AHE)$, on *Age*, Age^2, *Female*, and *Bachelor*. If *Age* increases from 25 to 26, how are earnings expected to change? If *Age* increases from 33 to 34, how are earnings expected to change?

e. Do you prefer the regression in (c) to the regression in (b)? Explain.

f. Do you prefer the regression in (d) to the regression in (b)? Explain.

g. Do you prefer the regression in (d) to the regression in (c)? Explain.

h. Pot the regression relation between *Age* and $\ln(AHE)$ from (b), (c), and (d) for males with a high school diploma. Describe the similarities and differences between the estimated regression functions. Would your answer change if you plotted the regression function for females with college degree?

i. Run a regression of $\ln(AHE)$ on *Age*, Age^2, *Female*, *Bachelor* and the interaction term $Female \times Bachelor$. What does the coefficient on the interaction term measure? Alexis is a 30-year-old female with a bachelor's degree. What does the regression predict for her value of $\ln(AHE)$? Jane is a 30-year-old female with a high school degree. What does the regression predict for her value of $\ln(AHE)$? What is the predicted difference between Alexis's and Jane's earnings? Bob is a 30-year-old male with a bachelor's degree. What does the regression predict for his value of $\ln(AHE)$? Jim is a 30-year-old male with a high school degree. What does the regression predict for his value of $\ln(AHE)$? What is the predicted difference between Bob's and Jim's earnings?

j. Is the effect of *Age* on earnings different for men than for women? Specify and estimate a regression that you can use to answer this question?

k. Is the effect of *Age* on earnings different for high school graduates than for college graduates? Specify and estimate a regression that you can use to answer this question.

l. After running all these regressions (and any others that you want to run), summarize the effect of age on earnings for young workers.