Problem Set 1 -- BUAN/MIS 6356

Turn in your code as ps1.R in eLearning. Add your answers to the interpretations as comments within your code. Your code should be able to be run. All the data files can be found in wooldridge.db

This problem set is due Friday Aug. 31th at 11:59 pm.

1.1 Use the data in WAGE1.RAW for this exercise.

(i) Find the average education level in the sample. What are the lowest and highest years of education?

(ii) Find the average hourly wage in the sample. Does it seem high or low?

(iii) The wage data are reported in 1976 dollars. Using the *Economic Report of the President* (2011 or later), obtain and report the Consumer Price Index (CPI) for the years 1976 and 2010.

(iv) Use the CPI values from part (iii) to find the average hourly wage in 2010 dollars. Now does the average hourly wage seem reasonable?

(v) How many women are in the sample? How many men?

1.2 The data in MEAP01.RAW are for the state of Michigan in the year 2001. Use these data to answer the following questions.

(i) Find the largest and smallest values of *math4*. Does the range make sense? Explain.

(ii) How many schools have a perfect pass rate on the math test? What percentage is this of the total sample?

(iii) How many schools have math pass rates of exactly 50%?

(iv) Compare the average pass rates for the math and reading scores. Which test is harder to pass?

(v) Find the correlation between *math4* and *read4*. What do you conclude?

(vi) The variable *exppp* is expenditure per pupil. Find the average of *exppp* along with its standard deviation. Would you say there is wide variation in per pupil spending?

(vii) Suppose School A spends $6,000 per student and School B spends $5,500 per student. By what percentage does School A's spending exceed School B's? Compare this to 100 · [log(6,000) – log(5,500)], which is the approximation percentage difference based on the difference in the natural logs. (See Section A.4 in Appendix A.)

1.3 The data in 401K.RAW are a subset of data analyzed by Papke (1995) to study the relationship between participation in a 401(k) pension plan and the generosity of the plan. The variable *prate* is the percentage of eligible workers with an active account; this is the variable we would like to explain. The measure of generosity is the plan match rate, *mrate*. This variable gives the average amount the firm contributes to each worker's plan for each \$1 contribution by the worker. For example, if $mrate = 0.50$, then a \$1 contribution by the worker is matched by a 50¢ contribution by the firm.

(i) Find the average participation rate and the average match rate in the sample of plans.

(ii) Now, estimate the simple regression equation

$$\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 mrate,$$

and report the results along with the sample size and *R*-squared.

(iii) Interpret the intercept in your equation. Interpret the coefficient on *mrate*.

(iv) Find the predicted *prate* when $mrate = 3.5$. Is this a reasonable prediction? Explain what is happening here.

(v) How much of the variation in *prate* is explained by *mrate*? Is this a lot in your opinion?

1.4 The data set in CEOSAL2.RAW contains information on chief executive officers for U.S. corporations. The variable *salary* is annual compensation, in thousands of dollars, and *ceoten* is prior number of years as company CEO.

(i) Find the average salary and the average tenure in the sample.

(ii) How many CEOs are in their first year as CEO (that is, $ceoten = 0$)? What is the longest tenure as a CEO?

(iii) Estimate the simple regression model

$$\log(salary) = \beta_0 + \beta_1 ceoten + u,$$

and report your results in the usual form. What is the (approximate) predicted percentage increase in salary given one more year as a CEO?

1.5 Use the data in WAGE2.RAW to estimate a simple regression explaining monthly salary (*wage*) in terms of IQ score (*IQ*).

(i) Find the average salary and average IQ in the sample. What is the sample standard deviation of IQ? (IQ scores are standardized so that the average in the population is 100 with a standard deviation equal to 15.)

(ii) Estimate a simple regression model where a one-point increase in *IQ* changes *wage* by a constant dollar amount. Use this model to find the predicted increase in wage for an increase in *IQ* of 15 points. Does *IQ* explain most of the variation in *wage*?

(iii) Now, estimate a model where each one-point increase in *IQ* has the same percentage effect on *wage*. If *IQ* increases by 15 points, what is the approximate percentage increase in predicted *wage*?

1.6 We used the data in MEAP93.RAW for Example 2.12. Now we want to explore the relationship between the math pass rate (*math10*) and spending per student (*expend*).

(i) Do you think each additional dollar spent has the same effect on the pass rate, or does a diminishing effect seem more appropriate? Explain.

(ii) In the population model

$$math10 = \beta_0 + \beta_1 \log(expend) + u,$$

argue that $\beta_1/10$ is the percentage point change in *math10* given a 10% increase in *expend*.

(iii) Use the data in MEAP93.RAW to estimate the model from part (ii). Report the estimated equation in the usual way, including the sample size and *R*-squared.

(iv) How big is the estimated spending effect? Namely, if spending increases by 10%, what is the estimated percentage point increase in *math10*?

(v) One might worry that regression analysis can produce fitted values for *math10* that are greater than 100. Why is this not much of a worry in this data set?

1.7 Use the data in HPRICE1.RAW to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

where *price* is the house price measured in thousands of dollars.

(i) Write out the results in equation form.

(ii) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

(iii) What is the estimated increase in price for a house with an additional bedroom that is 140 square feet in size? Compare this to your answer in part (ii).

(iv) What percentage of the variation in price is explained by square footage and number of bedrooms?

(v) The first house in the sample has $sqrft = 2{,}438$ and $bdrms = 4$. Find the predicted selling price for this house from the OLS regression line.

(vi) The actual selling price of the first house in the sample was $300,000 (so $price = 300$). Find the residual for this house. Does it suggest that the buyer underpaid or overpaid for the house?

1.8 The file CEOSAL2.RAW contains data on 177 chief executive officers and can be used to examine the effects of firm performance on CEO salary.

(i) Estimate a model relating annual salary to firm sales and market value. Make the model of the constant elasticity variety for both independent variables. Write the results out in equation form.

(ii) Add *profits* to the model from part (i). Why can this variable not be included in logarithmic form? Would you say that these firm performance variables explain most of the variation in CEO salaries?

(iii) Add the variable *ceoten* to the model in part (ii). What is the estimated percentage return for another year of CEO tenure, holding other factors fixed?

(iv) Find the sample correlation coefficient between the variables log(*mktval*) and *profits*. Are these variables highly correlated? What does this say about the OLS estimators?

1.9 Use the data in ATTEND.RAW for this exercise.

(i) Obtain the minimum, maximum, and average values for the variables *atndrte*, *priGPA*, and *ACT*.

(ii) Estimate the model

$$atndrte = \beta_0 + \beta_1 priGPA + \beta_2 ACT + u,$$

and write the results in equation form. Interpret the intercept. Does it have a useful meaning?

(iii) Discuss the estimated slope coefficients. Are there any surprises?

(iv) What is the predicted *atndrte* if $priGPA = 3.65$ and $ACT = 20$? What do you make of this result? Are there any students in the sample with these values of the explanatory variables?

(v) If Student A has $priGPA = 3.1$ and $ACT = 21$ and Student B has $priGPA = 2.1$ and $ACT = 26$, what is the predicted difference in their attendance rates?

1.10 Use the data in HTV.RAW to answer this question. The data set includes information on wages, education, parents' education, and several other variables for 1,230 working men in 1991.

(i) What is the range of the *educ* variable in the sample? What percentage of men completed 12th grade but no higher grade? Do the men or their parents have, on average, higher levels of education?

(ii) Estimate the regression model

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + u$$

by OLS and report the results in the usual form. How much sample variation in *educ* is explained by parents' education? Interpret the coefficient on *motheduc*.

(iii) Add the variable *abil* (a measure of cognitive ability) to the regression from part (ii), and report the results in equation form. Does "ability" help to explain variations in education, even after controlling for parents' education? Explain.

(iv) (Requires calculus) Now estimate an equation where *abil* appears in quadratic form:

$$educ = \beta_0 + \beta_1 motheduc + \beta_2 fatheduc + \beta_3 abil + \beta_4 abil^2 + u.$$

Using the estimates $\hat{\beta}_3$ and $\hat{\beta}_4$, use calculus to find the value of *abil*, call it $abil^*$, where *educ* is minimized. (The other coefficients and values of parents' education variables have no effect; we are holding parents' education fixed.) Notice that *abil* is measured so that negative values are permissible. You might also verify that the second derivative is positive so that you do indeed have a minimum.

(v) Argue that only a small fraction of men in the sample have "ability" less than the value calculated in part (iv). Why is this important?

(vi) If you have access to a statistical program that includes graphing capabilities, use the estimates in part (iv) to graph the relationship beween the predicted education and *abil*. Let *motheduc* and *fatheduc* have their average values in the sample, 12.18 and 12.45, respectively.