Problem Set 1

EC 204: Empirical Economics 2

Due by: Thursday July 16, 2020 at 10AM (EST)

True/False/Uncertain (20 points)

Instructions: For the each of the following statements, indicate whether it is true, false, or if the assertion is indeterminate. In all cases, defend your answer with a concise explanation. **To earn full credit, you must include an explanation with your answer.**

1. Suppose that X is a normally distributed random variable with pdf $f(X)$. Suppose further that the population mean is μ. Then, the sample median of X is an unbiased estimator of the population mean.

2. Suppose x is a discrete random variable that takes on only values 0 or 1. The random variable y is a continuous random variable. $\mathbb{E}(y|x=0) = 2$, $\mathbb{E}(y|x=1) = 5$, $Pr(x=0) = 0.2$ and $Pr(x=1) = 0.8$ are given. The unconditional expectation $\mathbb{E}(y)$ is equal to 4.4.

3. Suppose that X takes on the values $-1, 0$, and 2 with probabilities $1/8, 1/2$, and $3/8$ respectively. Then $\mathbb{E}(X^2) = 13/8$.

4. The R^2 is bounded between -1 and 1.

Longer Questions (30 Points)

1. (**Wage Regression**) A researcher is using data for a sample of 3,240 female employees 25 years of age and over to investigate the relationship between employee's hourly wage rates, Y_i (measured in dollars per hour) and their age X_i (measured in years). The population regression equation takes the form of equation (1):

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{1}$$

Preliminary analysis of the sample produces the following sample information:

$$N = 3240, \quad \sum_{i=1}^{N}(Y_i - \bar{Y})^2 = 78434.97, \quad \sum_{i=1}^{N}(X_i - \bar{X})^2 = 25526.17, \quad \sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) = 3666.436$$

$$\sum_{i=1}^{N} Y_i = 34379.16, \quad \sum_{i=1}^{N} X_i = 96143.00, \quad \sum_{i=1}^{N} Y_i^2 = 443227.10$$

$$\sum_{i=1}^{N} X_i^2 = 2878451.00, \quad \sum_{i=1}^{N} X_i Y_i = 1023825.00, \quad \sum_{i=1}^{N} \hat{u}_i^2 = 77908.35$$

Use the above information to answer all of the following questions. Show explicitly all formulas and calculations.

(a) Compute OLS estimates of the intercept coefficient, β_0 and the slope coefficient β_1.

(b) Interpret the slope coefficient estimates calculated above (be sure to state your interpretation in terms of age and hourly wages).

(c) Calculate $\widehat{\sigma}^2$ (the estimator of the error variance).

(d) Compute the R^2. Explain what the value of the R^2 means.

2. (**Academic Performance Regression**) Consider the following model in which course grade is a function of hours studied

$$score = \beta_0 + \beta_1 study + u$$

(a) Suppose you get a random sample from the population that consists of $\{score_i, study_i\}$ for $i = 1, 2, \ldots, n$. Also, you confirmed that *study* has variation in this sample. What is an additional assumption needed for OLS to be unbiased?

(b) Using OLS you obtain $\widehat{score} = 50 + 10study$, $n = 100$, $R^2 = 0.20$. From this result, can you claim that $\beta_1 = 10$? Explain. (Hint: think sample vs. population).

(c) The j^{th} observation takes on the values $score_j = 75$ and $study_j = 3$. Based on the estimated result in (b), calculate the residual for this observation.

3. (**Wooldridge, Ch2, Problem 5**) In the linear consumption function

$$\widehat{cons} + \widehat{\beta}_0 + \widehat{\beta}_1 inc$$

the (estimated) marginal propensity to consume (MPC) out of income is simply the slope, $\widehat{\beta}_1$, while the average propensity to consume (APC) is

$$\frac{\widehat{cons}}{inc} = \frac{\widehat{\beta}_0}{inc} + \widehat{\beta}_1$$

Using observations for 100 families on annual income and consumption (both measured in dollars), the following equation is obtained

$$\widehat{cons} = -124.84 + 0.853 inc$$

$$N = 100, \quad R^2 = 0.692$$

(a) Interpret the intercept in this equation, and comment on its sign and magnitude.

(b) What is the predicted consumption when family income is $30,000?

(c) With *inc* on the x-axis, draw a graph of the estimated MPC and APC.

Computational Exercises (30 Points)

Instructions: **To earn full credit, you must include your STATA code along with your answers to each of the following questions.**

1. (**Data Exercise**) Download the data set `401ksubs.dta` and import into STATA

   (a) What is the mean and standard deviation of net total financial assets (`nettfa`)? How do these two summary statistics differ depending on whether the individual has an IRA? (Hint: re-compute mean and standard deviation of `nettfa` for `pira==1` and `pira==0` separately and compare.)

   (b) Report the largest and smallest values that `inc` takes on.

   (c) What is the mean and median of `inc`? Does `inc` appear to be normally distributed? Generate a new variable for log(`inc`). Is this new variable normally distributed? Compare the two graphically by providing a histogram of each.

   (d) What is the correlation of `age` with `nettfa`? What does the correlation coefficient suggest about the relationship between the two variables?

2. (**Wooldridge, Chapter 2 C5**) For the population of firms in the chemical industry, let `rd` denote annual expenditures on research and development, and let `sales` denote annual sales (both are in millions of dollars).

   (a) Write down a model (not an estimated equation) that implies a constant elasticity between `rd` and `sales`. Which parameter is the elasticity?

   (b) Now, estimate the model using the data in RDCHEM.DTA. Write out the estimated equation. What is the estimated elasticity of `rd` with respect to `sales`? Explain in words what this elasticity means.

3. (**More OLS Practice**) Import the dataset, SAVING.DTA and regress `cons` on `inc`.

   (a) Interpret the coefficient in front of `inc`.

   (b) How much of the variation in consumption is explained by income?

   (c) Now generate a variable (call it `conshat`) that is defined to be the fitted values of the regression model. Next, generate a variable (call it `consres`) that is defined to be the

residuals from the regression. What do you expect to happen if you regress `conshat` on `consres`? Why?