Econometrics (ECON 21020) – Winter 2020
Problem Set #3
Date Due: February 5, 2020

1. Stock and Watson, Exercises 4.3, 4.4, 4,7, 4.11(a), 4.12, 5.1, 5.5, 5.6

2. Suppose that Col_GPA $= \beta_0 + \beta_1 \text{PC} + U$, where Col_GPA denotes the student's college GPA and PC is a binary variable with $PC = 1$ if the student owns a PC and $PC = 0$ otherwise. Define noPC as a dummy variable for whether the student does not own a PC, with noPC=1 if the student does not own a PC and noPC=0 if the student does own a PC. Suppose that, in the data, there are 34 students who do not own a PC and 53 students who own a PC. The sample average of Col_GPA for those students without a PC is 2.5, and the sample average of Col_GPA for those students with a PC is 3.5. Suppose further that, in the data, the sample standard deviation of Col_GPA for those students without a PC is .62, and the sample standard deviation of Col_GPA for those students with a PC is .47.

   (a) What are the OLS estimates of β_0 and β_1? Why can you compute them in this fashion?

   (b) Test $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 > 0$ at the 5% significance level. What is the p-value for this hypothesis test? What do you conclude?

   (c) Suppose that the researcher instead wishes to do an OLS regression of Col_GPA on noPC. What would be the resulting OLS estimates for that regression?

3. Suppose

$$Y = \beta_0 + \beta_1 X + U \ ,$$

   where Y is a binary random variable. Suppose further that $E[U|X] = 0$ and $0 < \text{Var}[X] < \infty$.

   (a) What is $E[Y|X]$? What is $P\{Y = 1|X\}$?

   (b) What is $\text{Var}[Y|X]$?

   (c) What is $\text{Var}[U|X]$? Is the model homoskedastic or heteroskedastic?

   (d) Suppose X can take any real-number value. What is a potential issue with the linear model in this case?

4. The following question involves the California Test Score data set. The data set is described in Appendix 4.1 of Stock and Watson, and can be downloaded from the course website.

(a) Load the California Test Score data set into `R`. How many observations do you have in the data set?

(b) The variable *avginc* is average district income measured in 1000*s* of dollars. Define a new variable, *income*, which is the variable *avginc* multiplied by 1000.

   i. What does the variable *income* measure?

   ii. What is the mean and standard deviation of *avginc*?

   iii. What is the mean and standard deviation of *income*? Given your result to part (ii), are the mean and standard deviation for *income* what you expected? why?

(c) i. What is the mean math score across all districts?

   ii. What fraction of districts have an average class size of 20 or fewer students? What is the mean math score in districts with average class size of 20 or fewer students?

   iii. What fraction of districts have an average class size of more than 20 students? What is the mean math score in districts with average class size greater than 20?

   iv. What is the connection between your answer in (i) and your answers in (ii) and (iii)?

   v. Calculate a test at the 10% level of whether the mean math score in districts with average class size of 20 or fewer students is equal to the mean math score in districts with average class size greater than 20. Formally state your null hypothesis in terms of population level conditional expectations. Describe your testing procedure. Can you reject the null hypothesis?

   vi. What is the covariance between *avginc* and mean math score? What is covariance between *income* and mean math score? Are the two covariances the same or different? Explain.

   vii. What is the correlation between *avginc* and mean math score? What is correlation between *income* and mean math score? Are the two correlations the same or different? Explain.