Problem Set 2 (Simple regression)

ECON 441, 2019 Fall

Please solve all the problems below. This is NOT a group exercise, so everyone needs to write up their own solutions. WARNING: the exercises are modified slightly to better fit your needs!

1 Wooldridge, Chapter 2, Exercise 1 (10 pts)

Let *kids* denote the number of children ever born to a woman, and let *educ* denote years of education for the woman. A simple model relating fertility to years of education is

$$kids = \beta_0 + \beta_1 educ + U,$$

where U is the unobserved error.

1. What kinds of factors are contained in U? (Give me 4.)

2. Will a simple regression uncover the *ceteris paribus* (=causal) effect of education on fertility? Why or why not? (1 sentence)

2 Wooldridge, Chapter 2, Exercise 4 (20 pts)

You are interested in the effect of the (tobacco) smoking behavior of the mother on birth weight. You have data about 1,338 births in the US that contain the infant's birth weight in ounces (*bwght*) and the average number of cigarettes the mother smoke during pregnancy (*cigs*).

1. What is the dependent and what is the independent variable?

2. You run a linear regression, and you got from the output that

$$\widehat{bwght} = 112.90 - 0.504cigs$$

   What is the predicted birth weight of the mother did not smoke during pregnancy? What is the predicted birth weight if the mother smoke 20 cigarettes (=1 pack) during pregnancy? Comment on the difference. (Which one is smaller, big difference, small difference in your opinion... etc)

3. Does this simple regression necessarily capture the causal relationship between the child's birth weight and the mother's smoking behavior? Explain and interpret the slope coefficient (depending on your answer).

4. To predict a birth weight of 110 ounces, what would *cigs* have to be? Comment on your result. (Does it make sense? Optional: why do you think this happened?).

3 Wooldridge, Chapter 2, Exercise 5 (10 pts)

This is a macroeconomic application, for those who like macro better! As you know from your macro class, the consumption function gives the consumption of an individual as a function of their income. We collected 100 families annual income (*inc*, in dollars) and consumption (*cons*, in dollars), then we estimated the parameters of a linear consumption function

$$cons = \beta_0 + \beta_1 inc + \epsilon,$$

to get the output

$$\widehat{cons} = -115.4 + 0.803 inc,\ n = 100, R^2 = 0.754$$

1. Here the slope coefficient is also called the marginal propensity to consume in macro. It tells you what share of an additional dollar you spend on consumption (as opposed to savings). Interpret the slope coefficient, comment on its sign and magnitude.

2. What is the fitted value for a family that has $45,000 annual income?

4 Wooldridge, Chapter 2, C1 (30 pts)

The data in the file 401K.dta are a subset of the 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. 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 dollar contribution of the worker. For example, if $mrate = .50$, then a \$1 contribution by the worker is matched by a 50 cents contribution by the firm.

1. What do you think the LHS and RHS variables are here? Define your variables.

2. Write down the linear model.

3. Estimate the simple linear model with OLS. Write down the Stata code you used. Please do not forget to print your Stata log file and submit it with the homework.

4. Interpret the slope coefficient and the intercept.

5. Find the predicted *prate* when *mrate* $= 3.5$. Is this a reasonable prediction? Explain why.

6. How much of the variation of *prate* is explained here by *mrate* in the linear model? (Which measure would you use from the output, interpret that measure.) Is this a lot in your opinion?

5 Wooldridge, Chapter 2, C10 (30 pts)

The data set CATHOLIC includes test score information on over 7,000 students in the US who were in eighth grade in 1988. The variables $math12$ and $read12$ are scores on twelfth grade standardized math and English reading tests, respectively.

1. How many students are in the sample? Find the means and standard deviations of $math12$ and $read12$. Write down what Stata code you used (or where you clicked in detail).

2. How do you think $math12$ and $read12$ are related, why?

3. Run the simple regression of $math12$ on $read12$ to obtain the OLS intercept and slope estimates. Write down what Stata code you used (or where you clicked in detail). Report the results (n and R^2 too).

4. Does the intercept have a meaningful interpretation? Explain why not or interpret.

5. Interpret the $\hat{\beta}_1$ you found (causal or not causal interpretation?). Are you surprised of its sign/magnitude? Interpret the R^2.