**Econ 300 Spring 2020**

**Problem set 2**

**Due 2/12**

**Instructions: Please write/type your answers NEATLY and hand in a hard copy before class starts. If you choose to write the solution by hand, please write in legible writing and do not try to fit answers in weird spaces.**

**The total points for this problem set is 100 and a further 10 points are available as a bonus. For STATA questions, paste codes/screenshot (It is very important to show your STATA work).**

**Please do not copy answers from each other or directly from the internet. It is okay to discuss the problems with classmates but answers must be written in your own words.**

**Question 1** (20 points). In order to improve performance, Chicago public schools decide to give bonus salary to teachers whose students achieve top 20 percentile of a standardized test. In order to evaluate if this bonus salary is effective in raising students’ performance, one researcher runs the following regression:

Avg\_student\_score=0+1Receive\_Bonus+

1. Give two examples of omitted variables that may bias

**Teacher quality, Students’ ability**

b. Explain the direction of bias and the plausible stories leading to the bias

**The teachers of better quality and teaching skills tend to receive bonus because their students are more likely to do well in exams. So this would lead to an upward bias of the estimates of Receive-Bonus. The class’s average score would be higher than the other classes due to teacher’s quality, not because the teachers work harder in order to get the bonus.**

**In another case, if there is a class with many talented students, then the average score in this class would be higher than the other classes due to the students’ ability not bonus salary. This would cause upward bias, too.**

**Question 2** (30 points (+10 bonus)). You are studying the causal effect of experience on wages (i.e. return to experience). ttl\_exp is total work experience. You run the following model:

Wage=0+ttl\_exp+

1. Interpret the coefficient for ttl\_exp in this model 1.

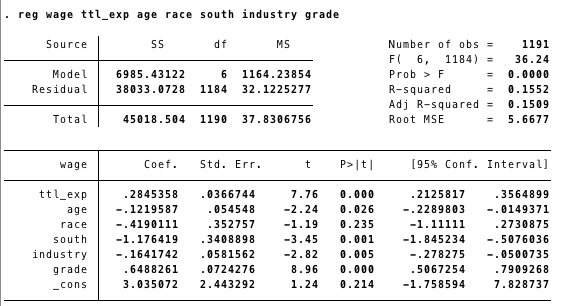
**One unit increase in total work experience is associated with**  **unit increase in average wage, holding the other variables constant.**

Suppose instead that you run the following model 2 now:

Wage=0+ttl\_exp+2age+3race+4south+5industry+6grade+

where race is a dummy variable equal to 1 for Whites, south refers to a dummy variable equals to 1 for individuals living in southern states, industry is a dummy variable for white-collar jobs, and grade picks up the highest year of schooling for the individual. Note: Wage is measured as hourly wage rate, and experience is measured in years.

The Stata output is shown below:



1. Interpret the estimate for ttl\_exp in this model. How is the interpretation here different from model a?

**Estimated coefficient:0.28**

**One unit increase in total work experience is associated with 0.28 unit increase in average wage, holding the other variables constant. You need to hold other variables constant in this case as compared to part a.**

c. Write out the null and alternative hypotheses to test whether there is any relationship between wage and experience.

**Ho: **

**≠**

d. Use any method explain whether you reject the null hypothesis specified in c. (t/2 for =0.01, n=1184 is 2.58)

**P-value= 0.000 < 0.01 => p-value <  reject null hypothesis**

e. What is the null hypothesis for the F-test that all independent variables have no explanatory power in wage? Do you reject or fail to reject the null based on the Stata output? Explain.

**Ho: **

**** ****i**≠**for some i where i=1,2,3,4,5,6.

**Reject the null. Because p-value of F-test=0, which is clearly less than 0.05.**

f. Explain in words what R-square means in this example. Does big R-square necessary mean better model for this research project?

**R-squared is a statistical measure of how close the data are to the fitted regression line. It is the percentage of the dependent variable variation that is explained by the independent variables in the model. 15.5% of the total variation in wages can be explained by this model.**

**R-squared does not indicate whether a regression model is adequate. You can have a low R-squared value for a good model, or a high R-squared value for a model that does not fit the data! Furthermore, if your R-squared value is low but you have statistically significant variables, you can still draw important conclusions about how changes in the predictor values are associated with changes in the response value.**

g. (Bonus worth 10 points) Explain why adding additional variables such as *race*, *south*, *industry* and *grade* in model 2 might be a good idea as compared to model 1?

**All these variables are possibly predictors of wages and are likely to be correlated with total experience of an individual. Therefore, in order to avoid omitted variable bias, it is a good idea to include these variables as additional controls.**

**Question 3** (50 points). (Stata application)

Load the data nlsw\_ps2.dta. Our research question is to examine whether being in the union affects one’s wage. For the following questions that require STATA commands, you can either paste the STATA output or write/type the key results.

1. Before getting to the data, what is your prior expectation? Do you think being in the union has a positive, negative or no effect on one’s wage?

**My prior is that Union is likely to have a positive effect on wages since unions can bargain for higher wages.**

**(However, any answer works as long as it is justified).**

1. Run a simple regression of hourly wage on one’s union status. Write down/paste your codes.

**reg wage union**

c. What is the coefficient and standard error of union status? Interpret the meaning of the coefficient and its standard error.

reg wage union

Source | SS df MS Number of obs = 944

-------------+---------------------------------- F(1, 942) = 19.08

Model | 314.534631 1 314.534631 Prob > F = 0.0000

Residual | 15527.6427 942 16.4836972 R-squared = 0.0199

-------------+---------------------------------- Adj R-squared = 0.0188

Total | 15842.1774 943 16.7997639 Root MSE = 4.06

------------------------------------------------------------------------------

wage | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

union | 1.331179 .30474 4.37 0.000 .7331309 1.929226

\_cons | 7.338289 .1526925 48.06 0.000 7.038632 7.637946

------------------------------------------------------------------------------

**Coefficient: 1.331**

**Standard error: 0.305**

**The coefficient says that being in the union can increase the average hourly wage by 1.331. In other words, on average, the workers in unions earn $ 1.331 more than the workers not being in unions. Note that here we don’t have other variables.**

**The standard error shows how spread-out the sampling distribution of the coefficient for union is.**

d. Look at its t-stat or p-value, do you think there is a relationship between union status and wage? Explain.

**Yes, p-value is 0.000 which is less than alpha of 0.05.**

e. Now based on the confidence interval, do you think there is a relationship between union status and wage? Explain

**95% Confidence interval does not contain zero => There is a statistically significant relationship between union status and wages.**

f. Using the estimated regression, let’s make predictions for wage given each person’s union status. What is the predicted wage for someone in the union?

**Predicted wage= 7.338 + 1.331 = 8.669**

g. Now include two additional variables race and age in your regression. Write down/paste the codes. What is the coefficient and standard error for union?

reg wage union race age

Source | SS df MS Number of obs = 944

-------------+---------------------------------- F(3, 940) = 7.47

Model | 368.872753 3 122.957584 Prob > F = 0.0001

Residual | 15473.3046 940 16.4609623 R-squared = 0.0233

-------------+---------------------------------- Adj R-squared = 0.0202

Total | 15842.1774 943 16.7997639 Root MSE = 4.0572

------------------------------------------------------------------------------

wage | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

union **| 1.360404 .3052671** 4.46 0.000 .7613197 1.959488

race | -.5014688 .2797634 -1.79 0.073 -1.050502 .0475642

age | -.0184276 .0439907 -0.42 0.675 -.1047589 .0679036

\_cons | 8.691264 1.794702 4.84 0.000 5.169178 12.21335

------------------------------------------------------------------------------

h. What values are the ESS and TSS from the STATA output? Explain what they mean in this example.

**ESS= 368.872753 is the part of the variation in wages that is explained by the model with union status, race and age ; TSS= 15842.1774 is the total variation in wages.**

1. Compare the R-square with adjusted R-square, which one do you trust more and why?

**R-square is 0.0233 whereas Adjusted R-square is somewhat smaller at 0.0202. I trust the adjusted R-squared more because it considers the extra control variables we add in the model.**

j. Compare the simple regression with the multiple regression. Which model have more explanatory power in explaining the variation in wage? Why?

**The adjusted R-squared in the first simple regression is .0188, while the adjusted R-squared in the second regression is .0202. The latter is bigger than the former. So the second model has more explanatory power in the variation in wage. This is because race and age are predictors of average wages.**