ECO 480 Econometrics 1

Take-Home Exam 3

Due Wednesday, 3/24/2021

**Statement of Academic Honesty*:*** *On the first page of your Word-file before you answer the assigned questions, you must type in the following passage, then put your name, sign, and date:*

**For this exam, I make the following truthful statements:**

**1. I have not received any non-instructor approved assistance, I have not given any non-instructor approved assistance to another student taking this exam, including discussing the exam with students in another section of the course.**

**2. I did not plagiarize someone else’s work and turn it in as my own.**

**3. I understand that acts of academic dishonesty may be penalized to the full extent allowed by the University at Buffalo Student Conduct Code, including receiving a failing grade for the course with a transcript notation and being expelled from the university. I recognize that I am responsible for understanding the provisions of the University at Buffalo Student Conduct Code as they relate to this academic exercise.**

**Name: \_\_Hao Wu\_\_\_\_ Sign: \_\_Hao Wu\_\_\_\_\_ Date: \_\_\_\_\_03/20/2021\_\_\_\_**

***Important Note: If you want to receive a grade for your take-home exam, you must have this statement. If you do not have this statement, I will not accept your work and your grade would be zero.***

Instruction: You have one week to complete this exam. This exam consists of data analysis. You may NOT discuss the exam questions with your classmates. No late work will be accepted and all the files must be submitted via UBLearns. Make sure you upload it to the correct submission slot because no credit will be given for incorrect submissions.

Important: It is extremely important to write a clean well-commented program for transparency and replication purposes in *all* empirical work. You should always be able to reproduce your result from raw data to support your claim.

There are 3 items to hand in: (1) Typed write-up (i.e., word-file) answering the assigned questions, reporting your results, and interpreting your findings; if the question asks for graphs or tables, these must be in the word-file in an organized manner with your interpretation, (2) do-file (i.e., program file), and (3) log-file (i.e., output file that shows the results). You MUST use Stata. For questions involving data analysis, you will NOT get any credit if you do not provide a program code and the output. You may not use Excel. Do not submit any undigested log-file that contains errors. Put all your answers in the word file and do NOT say “please see log-file (or do-file) for answers. You will not receive any credit for answers that is not stated in the word file.

1. The following table describes variables from heightwage.dta we will use in this question, which you have seen already. Persico, Postlewaite, and Silverman (2004) analyzed data from the National Longitudinal Survey of Youth 1979 cohort to assess the relationship between height and wages for white men who were between 14 and 22 years old in 1979. This data set consists of answers from individuals who were asked questions in various years between 1979 and 1996
2. Estimate two regress models: one in which adult wages is regressed on adult height for all respondents, and another in which adult wages is regressed on adult height and adolescent height for all respondents. Discuss differences across the two models. Explain why the coefficient on adult height changed. [Hint: Think in terms of omitted variable bias.]

|  |  |
| --- | --- |
|  | (1) |
|  | wage96 |
| height85 | 0.315\*\*\* |
|  | (0.000) |
|  |  |
| \_cons | -6.977 |
|  | (0.210) |
| *N* | 6713 |

*p*-values in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

Wage96 = -6.977 + 0.315\*height85

|  |  |
| --- | --- |
|  | (1) |
|  | wage96 |
| height85 | -0.107 |
|  | (0.660) |
|  |  |
| height81 | 0.457 |
|  | (0.063) |
|  |  |
| \_cons | -9.248 |
|  | (0.109) |
| *N* | 6594 |

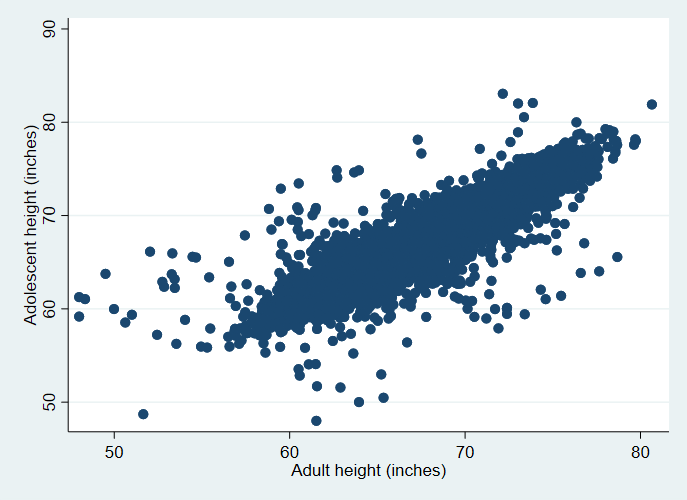
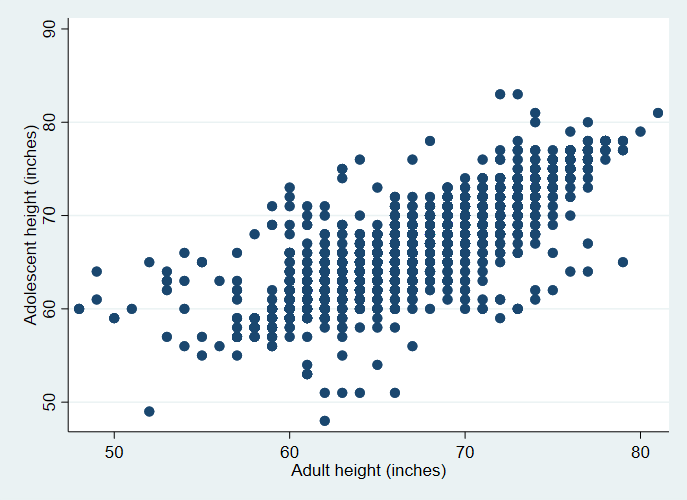
*p*-values in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

Wage96 = -9.248 +( -0.107) \* height85 + 0.457\* height81

We add more factor, so that the coefficient will be changed.

1. Assess the multicollinearity of the two height variables using (1) a plot, and (2) an auxiliary regression (i.e., a regression that is not directly the one of interest but yields information helpful in analyzing the equation we really care about). Run the plot once without a jitter subcommand and once with it, and choose the more informative of the two plots.



After we use jitter subcommand, the plot is useful. We could know the approach confidence interval of the data, and we also can except some outlier. If we run a regress between height81 and height85 we could get a conclusion which is make sense, because our height usually growth when we are teenager.

1. Notice that IQ is omitted from the model. Is this a problem? Why or why not?

This is a problem. Most people with high IQ are more likely to master more skills or proficient in a certain field, which also means that high-paying occupations are more likely to be obtained by them.

1. Notice that eye color is omitted from the model. Is this a problem? Why or why not?

This is not a problem. Eye color does not affect the level of salary. No job position will limit your eye color, but another important factor is skin color. In fact, most companies have prejudices.

1. Estimate a regression model with adult wages as the dependent variable and adult height, adolescent height, and a dummy variable for males as the independent variable. Does controlling for gender affect the results?

|  |  |
| --- | --- |
|  | (1) |
|  | wage96 |
| male | -1.000 |
|  | (0.312) |
|  |  |
| height81 | 0.487\* |
|  | (0.049) |
|  |  |
| height85 | -0.0476 |
|  | (0.849) |
|  |  |
| \_cons | -14.72 |
|  | (0.063) |
| *N* | 6594 |

*p*-values in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

We can easily conclude that gender control will affect the outcome.

1. Generate a female dummy variable. Estimate a model with both a male dummy variable and a female dummy variable. What happens? Why?

They conflict, because the dummy variable will only be 0 or 1. If dummy is 1, it means yes, and if dummy is 0, it means no. If we create a new variable to represent female, then the dummy variable will be confused, because 0 in the male variable represents female.

1. Re-estimate the model from part (e) separately for males and females. Do these results differ from the model in which male was included a dummy variable? Why or why not?

|  |  |
| --- | --- |
|  | (1) |
|  | wage96 |
| height81 | -0.00603 |
|  | (0.991) |
|  |  |
| height85 | 0.321 |
|  | (0.520) |
|  |  |
| \_cons | -6.728 |
|  | (0.663) |
| *N* | 3149 |

*p*-values in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

|  |  |
| --- | --- |
|  | (1) |
|  | wage96 |
| height81 | 0.783\*\*\* |
|  | (0.000) |
|  |  |
| height85 | -0.269 |
|  | (0.207) |
|  |  |
| \_cons | -20.82\*\* |
|  | (0.005) |
| *N* | 3445 |

*p*-values in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

The top matrix is for female and the bottom matrix is for male. Yes, the result is different from including the dummy variable. Because dummy variables can only express yes or no, we should compare them separately instead of bringing them directly into regression.

1. Every observation is categorized into one of four regions based on where the subjects lived in 1996. The four regions are Northeast (norest96), Midwest (norcen96), South (south96), and West (west96). Add dummy variables for regions to a control for regional effect. What are the regional variables you would include? Explain.

Make norest96 as 0, norcen96 as 1, south96 as 2, and west96 as 3.

1. Estimate a regression model (e) with male and regional dummy variables. First exclude West, then interpret the coefficient of each regional dummy variables.

|  |  |
| --- | --- |
|  | (1) |
|  | wage96 |
| male | 1.371 |
|  | (0.081) |
|  |  |
| regions | -2.262\*\*\* |
|  | (0.000) |
|  |  |
| \_cons | 16.10\*\*\* |
|  | (0.000) |
| *N* | 5527 |

*p*-values in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

1. Re-estimate the model in part (i) except exclude South instead. Interpret the coefficient of each regional dummy variables.

|  |  |
| --- | --- |
|  | (1) |
|  | wage96 |
| male | 0.984 |
|  | (0.320) |
|  |  |
| regions | -0.352 |
|  | (0.392) |
|  |  |
| \_cons | 15.32\*\*\* |
|  | (0.000) |
| *N* | 4084 |

*p*-values in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

1. You are the boss! Use the data in this file to estimate a model that you think sheds light on an interesting relationship. The specification decisions include whether to limit the sample and what variables to include. Report only a single additional specification. Describe in no more than two paragraphs why this is an interesting way to assess the data. [Hint: Decide on which control variable you would like to include in your model and explain.]

Study whether male becoming athletes will have an impact on wages.

First, we have to filter out all the male data, Then we perform regression on the athlete variables and wage variables.

|  |  |
| --- | --- |
|  | (1) |
|  | wage96 |
| athletes | 3.061\*\*\* |
|  | (0.000) |
|  |  |
| \_cons | 13.54\*\*\* |
|  | (0.000) |
| *N* | 3569 |

*p*-values in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

We saw that it has a statically significant., because p is less than 0.001, the regression should be:

Wage = 13.54 + 3.061\*athletes

Because athletes is a dummy variable only include 0 and 1, we get a hypothesis that become athletes will make wage increase 3.061.

1. Do cell phones distract drivers and cause accidents? Worried that this is happening, many states have passed legislation to reduce distracted driving. Fourteen states now have laws making handheld cell phone use while driving illegal, and 44 states have banned texting while driving. This problem looks more closely at the relationship between cell phones and traffic fatalities. The following table describes the variables in the dataset Cellphone\_2012.dta.
2. While we don’t know how many people are using their phones while driving, we can find the number of cell phone subscriptions in a state (in thousands). Estimate a simple regression model with traffic deaths as the dependent variable and number of cell phone subscriptions as the independent variable. Briefly discuss the results. Do you suspect that the estimate on cell phone subscription suffers from omitted variable bias? If so, what would be the direction of bias? Discuss

|  |  |
| --- | --- |
|  | (1) |
|  | numberofdeaths |
| cell\_subscription | 0.0911\*\*\* |
|  | (0.000) |
|  |  |
| \_cons | 124.0\* |
|  | (0.028) |
| *N* | 50 |

*p*-values in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

This set of data only records the sales volume of mobile phones and does not consider the possibility of one person owning multiple mobile phones. In fact, most people keep their old phones while buying new phones. But this will not have an impact on our conclusions, because we are discussing the impact of mobile phone sales on mortality, not whether one person owns multiple mobile phones will increase the accident probability.

1. Add population to the model. What happens to the coefficient on cell phone subscriptions? Why? [Hint: Discuss in terms of omitted variable bias and the direction of bias.]

|  |  |
| --- | --- |
|  | (1) |
|  | numberofdeaths |
| cell\_subscription | -0.211\* |
|  | (0.027) |
|  |  |
| population | 0.000291\*\* |
|  | (0.002) |
|  |  |
| \_cons | 113.9\* |
|  | (0.027) |
| *N* | 50 |

*p*-values in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

Coefficient on cell phone subscriptions becomes negative, because population is real unit. However, the cell phone subscriptions are displayed after dividing by 1,000. The huge unit deviation has affected our regression results.

1. Add total miles driven to the model in addition to population. What happens to the coefficient on cell phone subscription? Why? [Hint: Discuss in terms of omitted variable bias and the direction of bias.]

|  |  |
| --- | --- |
|  | (1) |
|  | numberofdeaths |
| cell\_subscription | 0.00246 |
|  | (0.975) |
|  |  |
| population | -0.0000742 |
|  | (0.404) |
|  |  |
| total\_miles\_driven | 0.0188\*\*\* |
|  | (0.000) |
|  |  |
| \_cons | 4.346 |
|  | (0.916) |
| *N* | 50 |

*p*-values in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

Too little data will make our regression model underfit, but too much data will make our regression model overfit.

1. [Required for ECO 580, Extra credit for ECO 480] Using the data set Growth again, but excluding the data for Malta.
2. Construct a table that shows the sample mean, standard deviation, and minimum and maximum values for the series Growth, TradeShare, YearsSchool, Oil, Rev\_Coups, Assassinations, and RGDP60. Include the appropriate units for all entries.

**Summary of growth**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Obs | Mean | Std. Dev. | Min | Max | Unit |
| growth | 64 | 1.869 | 1.816 | -2.812 | 7.157 | Percentage |
| oil | 64 | 0 | 0 | 0 | 0 | Dummy Variable |
| rgdp60 | 64 | 3130.813 | 2522.979 | 367 | 9895.004 | Increase by 1980 |
| tradeshare | 64 | .542 | .228 | .141 | 1.128 | Null |
| yearsschool | 64 | 3.959 | 2.553 | .2 | 10.07 | Years |
| rev coups | 64 | .17 | .225 | 0 | .97 | Coups per year |
| assasinations | 64 | .282 | .494 | 0 | 2.467 | Assasina per year |

1. Run a regression of Growth on TradeShare, YearsSchool, Rev\_Coups, Assassinations, and RGDP60. What is the value of the coefficient on Rev\_Coups? Interpret the value of this coefficient. Is it large or small in a real-world sense?

|  |  |
| --- | --- |
|  | (1) |
|  | growth |
| tradeshare | 1.341 |
|  | (0.168) |
|  |  |
| yearsschool | 0.564\*\*\* |
|  | (0.000) |
|  |  |
| rev\_coups | -2.150 |
|  | (0.059) |
|  |  |
| assasinations | 0.323 |
|  | (0.511) |
|  |  |
| rgdp60 | -0.000461\*\* |
|  | (0.003) |
|  |  |
| \_cons | 0.627 |
|  | (0.427) |
| *N* | 64 |

*p*-values in parentheses

\* *p* < 0.05, \*\* *p* < 0.01, \*\*\* *p* < 0.001

The coefficient of rev\_coups is -2.15, if we use 2.15 divide by 5 we got 0.43. It is large in a real-world sense.

1. Use the regression to predict the average annual growth rate for a country that has average values for all regressors.

The average annual growth rate for a country that has average values for all regressors is 0.874.

1. Why is Oil omitted from the regression? What would happen if it were included? Explain

Oil is a dummy variable. The value of all countries is set to 0, so it belongs to a constant and will not affect the regression results.