**­­API-202I – Spring 2018**

**Assignment #1: Causality, Validity and Bivariate Regression (and Stata)**

**Part I – Counterfactuals**

In 2014, Colorado legalized recreational marijuana use for its residents. Suppose a policymaker is worried that such legalization would have an adverse effect on student performance among high school students. To this end, the policymaker would like to acquire evidence and estimate the causal impact of this reform on SAT scores of high school students in Colorado.

1. Explain what the ideal counterfactual would be here. Why can’t this counterfactual be observed?

Legalization of marijuana has no impact on the SAT performance of the same high school students, holding all other factors equal including time. There is a temporal element in this requirement which cannot be satisfied. The students cannot exist in a reality where marijuana is legalized and not legalized.

1. Suppose a researcher decided use data to estimate impact by using two different strategies: 1) Comparing average SAT scores among seniors in Colorado in 2015 to average SAT scores among seniors in Colorado in 2013. 2) Comparing average SAT scores among seniors in Colorado in 2015 to average SAT scores among seniors in Utah in 2015. For each strategy above:
   1. Describe which data that is designed to mimic the counterfactual

In (1) the location, schooling-level and state schooling system is held constant, which mimics the location angle; and in (2) the year is held constant, which mimics the temporal factors.

* 1. Provide an explanation for why using that data as a counterfactual is problematic

In (1) the seniors in 2015 are different to those in 2013, which is problematic as it introduces a source of error/ variation. Similarly, in (2) the students in Utah are different to those in Colorado, which have two different education systems, which also runs against the counterfactual.

* 1. Of the two strategies, which one do you think is *most* likely to estimate a causal effect? How confident are you that this strategy identifies a causal effect? Explain

I would assess that the first strategy would be most likely to demonstrate a causal effect as it holds more factors constant (location, schools, demographics etc). You could try to compare the average scores of 2013 and 2015 grades in prior years to try and control for those variations. For example, if the senior grade from 2013, consistently had similar grades to the seniors from 2015 in prior years, then it would be plausible that the change in policy may have impacted SAT scores.

**Part II – Does Globalization Hurt the Poor in Developing Countries?**

Read the article published in Vox about some micro-evidence of the effects manufacturing jobs have on the short and long run wellbeing of low-wage workers in Ethiopia: <http://www.vox.com/2016/9/29/13096580/globalization-poverty-experiment-blattman-dercon>. Pay particular attention to the design of the study and the results the authors obtained.

1. In a short paragraph, describe the study. Make sure you include in your description
   1. What is the key question the authors are trying to answer?

What impact do low-wage manufacturing jobs have on ordinary workers in developing countries?

* 1. What are the outcomes (dependent variables) of interest?

Relative income levels, the health perceptions, and risk of disability in different trial groups

* 1. What intervention (treatment) did the authors examine?

The researchers set up three groups: – factory workers, entrepreneurs, and control group in Ethiopia. The intervention is factory employment, and entrepreneurial training.

1. Explain clearly what internal validity implies in the context of this study. As far as you can tell, does this study have high internal validity?

Internal validity in this context attempts to capture the causal effect of interest for the population represented by our sample. In this case, the study appears to have high internal validity. They conducted a randomized control trial sampling from a group that applied for factory jobs. They randomly assigned them into groups of factory, entrepreneur, and control group across 5 factories with significant number of participants.

1. For what populations or contexts would this study have high external validity? Are there populations for which you think this study has low external validity? Explain clearly.

The external validity would likely be high for countries similar to Ethiopia with a similarly developed entrepreneurial opportunities. Presumably, factory conditions and the commercial environment would differ significantly across countries. For example, it may not easily apply in similarly poor countries like Cambodia, which have the ILO’s better factories program to monitors factory conditions. Factories are better unionized and able to push for slightly better conditions, not to say that conditions are acceptable by any means. Conversely, some areas may have factories but in remote areas with few other commercial opportunities, even for entrepreneurs, which would limit the applicability of this research.

**Part III – Stata**

Using results from the Stata tutorial, the India\_agric\_yields.dta dataset, and any additional commands you need, create a “do-file” to give you the information necessary to answer the following questions. You need to show your work in answering questions (1)-(2), i.e. write down the formulas you are using, and use the numbers from the Stata output to compute the relevant confidence interval or test statistic.

Important: For this and all future assignments, you will need to submit your Stata output (your do and your log files) with your answers. Directions for creating “do” and “log” files is included in the Stata tutorial.

|  |
| --- |
| Note: The answer you submit for this part should contain:   * + Your “ps1.do” Do-file   + A log file where the results (i.e. output) produced by “ps1.do”are displayed.   + Your complete write-up of answers to questions 1-3 |

1. Test the null hypothesis that the average yield (output/plot area) is 5000 Kg/Hectare. Use a 5% significance level. In doing so, indicate:
2. The null hypothesis

H0: Mean yeild = 5000kg/hectare. See log file for commands used.

1. The sampling distribution and test statistic used

t = 0.44213, df = 4654. Because this is sampled, it will be normally distributed as per the central limit theorem with a sample mean of 5225.494.

1. The p-value of this test and your interpretation of the p-value

p-value = 0.6584; The p-value is greater than 0.05, therefore we fail to reject the null hypothesis that the average yield is 5000kg/hectares with 95% confidence.

1. The 95% confidence interval for the agricultural yield  
   95 percent confidence interval: 4225.621kg/hectares & 6225.367kg/hectares, which 5000kg/hectares falls within.
2. Are there statistically significant differences in the yields of the plantation when a woman is the head of the household, relative to when a man is the head of the household? Conduct a hypothesis test and indicate:
   1. The null hypothesis

Mean yield of female headed households = Mean yield of male headed households. See log file for commands used.

* 1. The sampling distribution and t-statistic used

t = 5.5162. Because this is sampled, it will be normally distributed as per the central limit theorem with a sample mean of 5225.494. To reject, the T-statistic would need to be more than 1.96.

* 1. The value of the t-statistic and your interpretation of the p-value associated

p-value = 3.951e-08. As the | t | > 1.96 and the p-value is greater than 0.05, we can reject the null hypothesis with 95% confidence.

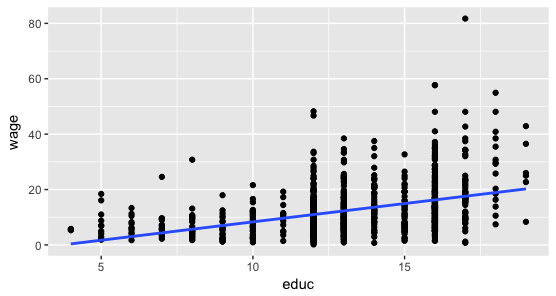
1. Interpret your results in the context of the study. Do you think this is conclusive evidence that women are less productive in farming than men are? Explain briefly.

This is not conclusive evidence on the comparative productivity of women farmers. For one, we only have data by head of household, which is not the same as women farmers. Even when it comes to female head of households (1660), there were just over half as many data points compared to male head of households (2995), suggesting these aren’t in proportion to women in the population. Also, the mean for households with women heads was substantially higher than for men.

**Part IV – Education and Wages**

Use the **gender.dta** dataset for this problem.

1. Type “**scatter wage educ**” to get a visual representation of the relationship between hourly wages and years of education. Show this figure.



1. Based on this figure, what can you conclude about the relationship between wages and education?

Positive correlation. As education increases, so does wage and vice-a-versa.

1. Type “**regress wage educ**” and record the output from that command
2. Write down the PRF and the SRF being estimated here

PRF: Wage = B0 + B1(educ) + e

SRF: Wage = B^0 + B^1(educ)

1. Interpret the coefficient associated to education in the regression  
   B^1 = 1.32329, which shows a positive correlation between wage and education. The slope of the regression is 1.323.
2. Is the coefficient for education statistically significant? Explain

t-value is 13.538 and the p-value is less than 2e-16 \*\*\*. Therefore we can reject the null hypothesis that there is no correlation with 95% confidence as the t-statistic> 1.96 and p-value< 0.05.

1. Does this imply that a policy designed to increase educational attainment will increase wages? Explain

No, this only shows a correlation. Regressions do not show causation. We would need to vary education levels while holding all other variables constant to explore causation.

1. What is the interpretation for the intercept?

B^0 = -4.91523 is the intercept for wages for zero years of education. However, the sample includes no data- points for anything below around 2.5 years of education. -4.9 on wages seems technically impossible but it is where the line of best fit intercepts the Y-axis.

1. What is the predicted wage for a person with 10 years of education? Show your calculations

Wage = -4.91523 + 1.32329\*10 = 8.31767 hourly wage.

1. Calculate the predicted values from the regression using **“predict yhat, xb”**

Wage = -4.92 + 1.32329 \* educ

* 1. a. Type **“sum yhat”**. Explain what you see.

Residuals:

Min 1Q Median 3Q Max

-16.811 -5.195 -1.496 3.221 64.150

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -4.91523 1.30692 -3.761 0.00018 \*\*\*

educ 1.32329 0.09774 13.538 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.157 on 948 degrees of freedom

Multiple R-squared: 0.162, Adjusted R-squared: 0.1611

F-statistic: 183.3 on 1 and 948 DF, p-value: < 2.2e-16

* 1. This is a summary of the sample regression.

b. Type **“sum yhat if educ ==10”**. Have you seen this before? Explain  
Residuals:

Min 1Q Median 3Q Max

-12.402 -5.928 -2.100 3.717 69.153

Coefficients:

Estimate Std. Error t value Pr(>|t|)

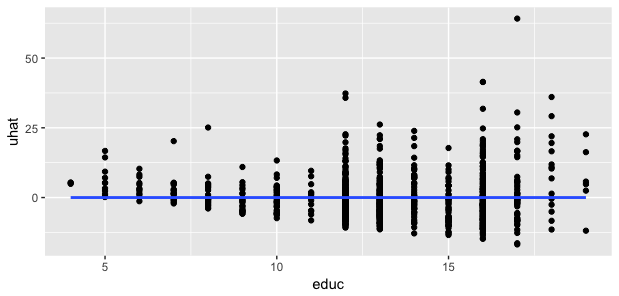
(Intercept) 12.5772 0.2932 42.889 < 2e-16 \*\*\*

educ == 10TRUE -4.4986 1.5278 -2.944 0.00331 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

This gives a value for wages as 12.5772-4.4986 = 8.0786, which is close to the predicted value for wage at education = 10 from in Q3 (f)

1. Calculate the residuals from the regression using “predict uhat, resid”
   * 1. a. Type **“sum uhat”.** Are the residuals centered at zero? Interpret what you observe
     2. > summary(uhattest)
     3. Residuals:
     4. Min 1Q Median 3Q Max
     5. -9.0937 -1.0937 -0.0937 2.9063 5.9063
     6. Coefficients:
     7. Estimate Std. Error t value Pr(>|t|)
     8. (Intercept) 1.309e+01 8.794e-02 148.9 <2e-16 \*\*\*
     9. uhat -3.536e-17 1.079e-02 0.0 1
   1. It appears as though the residuals are centered at near-zero suggesting that the errors from the regression are not skewed to any particular side.
      1. b. Type **“regress uhat educ”**. Do you find something that makes sense?
      2. Call:
      3. lm(formula = educ ~ uhat, data = gender.data)
      4. Residuals:
      5. Min 1Q Median 3Q Max
      6. -9.0937 -1.0937 -0.0937 2.9063 5.9063
      7. Coefficients:
      8. Estimate Std. Error t value Pr(>|t|)
      9. (Intercept) 1.309e+01 8.794e-02 148.9 <2e-16 \*\*\*
      10. uhat -3.536e-17 1.079e-02 0.0 1
   2. /// given the t-statistic is 0, we cannot reject the null hypothesis, therefore there is no correlation between education and uhat with 95% confidence.
      1. c. Type **“scatter uhat educ”**. Does the graph help you to understand what is going on?
      2. 
      3. At lower education levels there appears to be less variation in wages compared to higher levels of education.