**PP249: Problem Set 2**

**Due September 30th, 2022**

You will be asked to upload both the PDF file and code for this problem set. Please organize and neatly comment your code!

# Question 1 [Lecture 3]

You are given cleaned data from the [Tennessee Star Experiment](https://www.aft.org/sites/default/files/STARSummer99.pdf), which aims to address the impact of class size on student outcomes. This question walks through the following tasks, either in R or Stata (please be sure to share your code!):

1. Load in the dataset, which is stored on bCourses as “data\_tennessee\_star\_q1.csv”
2. Conduct summary statistics as requested in Question 1.1
3. Conduct a regression with *robust* standard errors, where “math\_score” is the outcome and our only predictor variable is “treat”
4. Answer the questions below!

# Question 1.1

Before we get into the regression, we want to run some summary statistics. Fill in the table below for the outcome variable, math\_score.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Treatment Group** | ***n*** | **Minimum** | **Mean** | **Standard Deviation** | **Maximum** |
| Control (treat=0) |  |  |  |  |  |
| Treatment (treat=1) |  |  |  |  |  |

# Question 1.2

Conduct the regression with robust standard errors (using vote as the outcome and treat as the predictor). Please answer the following:

* What is control group mean?
* What is the treatment group mean?
* What is the estimated treatment effect?
* What is the estimated standard error on the treatment effect?
* What is the “t statistic”?
* Is the impact of treat on math\_scores statistically significant?

# Question 1.3

Conduct a two-sample t-test for the difference in means between the treatment group and the control group. What is the t statistic? Does this differ from your answer to Question 1.2? Did you conduct this testing assuming equal or unequal variances?

# Question 1.4

Consider the case where we double the sample size. Holding all else equal, we would see the following:

* Standard Error: Increase, Decrease, or Stay the Same?
* *t*-statistic: Increase, Decrease, or Stay the Same?
* *p*-value: Increase, Decrease, or Stay the Same?

# Question 2 [Lecture 3]

You are given cleaned data from the [Gerber and Green experiment](https://isps.yale.edu/sites/default/files/publication/2012/12/ISPS00-001.pdf) on the effects of canvassing, direct mail, and phone calls on voter turnout. This question walks through the following tasks, either in R or Stata (please be sure to share your code!):

1. Load in the dataset, which is stored on bCourses as “data\_gerber\_q2.csv”
2. Conduct a regression with *robust* standard errors, where “vote” is the outcome and our only predictor variable is “treat”
3. Answer the questions below!

# Question 2.1

Conduct the regression with robust standard errors (using vote as the outcome and treat as the predictor). Please answer the following:

* What is control group mean?
* What is the treatment group mean?
* What is the estimated treatment effect?
* What is the estimated standard error on the treatment effect?
* What is the “t statistic”?
* Is the impact of treat on vote statistically significant?

# Question 2.2

Conduct a two-sample t-test for the difference in means between the treatment group and the control group. What is the t statistic? Does this differ from your answer to Question 2.1? Did you conduct this testing assuming equal or unequal variances?

# Question 2.3

Consider the case where we double the sample size. Holding all else equal, we would see the following:

* Standard Error: Increase, Decrease, or Stay the Same?
* *t*-statistic: Increase, Decrease, or Stay the Same?
* *p*-value: Increase, Decrease, or Stay the Same?

# Question 3: replicating measures of Black-white inequality as in Bayer and Charles (2018) [Lecture 4]

The goal of this exercise is to replicate measures of racial gap in log median income (as defined in Bayer and Charles (2018)) for three samples: (i) for working men only, (ii) for all men, and (iii) for all men or women

Please read below:

1. Instructions for requesting US Census data from IPUMS.org
2. Instructions for running analysis and plotting figures
3. Please run “measuring\_bw\_inequality\_cleaning.do” to help you clean the dataset before you start running quantile regressions to measure the racial income level gaps at the median.

**(1) Instructions for requesting US Census data from IPUMS.org**

1. Visit [www.ipums.org](http://www.ipums.org).
2. Select IPUMS USA for ACS/Census data.
3. Select Get Data.
4. Scroll through the different categories of variables per Person to select the following variables (or use the Search function):
   1. race
   2. hispan
   3. age
   4. sex
   5. ind1950
   6. incwage
   7. incbusfm
   8. incfarm
   9. incbus
   10. incbus00
5. Select View Cart.
6. Select Add More Samples.
7. Check “default sample from each year.”
   1. Deselect the years before 1950.
8. Submit Sample Selections.
9. Click Create Data Extract.
10. Select Customize Sample Sizes and enter “0.05” as the density for all samples.
11. Describe your extract with “Bayer and Charles (2018) Replication – PP 249” or a similar title.
12. Click “Submit Extract.”
13. You will be redirected to a page to Sign In or Create an Account.
    1. Click create an account in the bottom right and fill out the necessary information.
    2. Submit your account creation form and then log in.
14. You should be redirected back to the “Extract Request” page where you should click “Submit Extract” again.
    1. If you are not for some reason, you can repeat steps 1-11 now that you are logged in, but IPUMS should have saved your progress.
15. You will then be redirected to a page where you can download and revise extracts.
    1. This extract will be “processing.”
    2. You will receive an email and link when your extract is ready for download.
    3. [FYI: I’ve called my raw dataset usa\_00053.dta – this is the name that’ll appear in the solutions]

**(2) Instructions for running analysis and plotting figures**

1. Import ACS and Census data from 1950-2019 from the Census.
2. Data Preparation
   1. Inflation Adjustment
      1. Use the dataset and methodology [here](https://www.census.gov/topics/income-poverty/income/guidance/current-vs-constant-dollars.html) to adjust for inflation.
      2. Bring in the CPI dataset by merging on year.
      3. [FYI: I’ve called my raw dataset “cpi\_acs” – this is the name that’ll appear in the solutions]
      4. Generate a new real income variable using the formula linked above.
      5. *(The above steps are done for you in the cleaning code)*
   2. Samples: Create a flag for each of the samples plotted on the graph
      1. Sample 1: Limit to non-Hispanic white and Black men who work using sex, race, hispan, and empstat.
      2. Sample 2: Limit to all non-Hispanic white and Black men using sex, race, and hispan.
      3. Sample 3: Limit to all non-Hispanic white and Black men and women using race and hispan.
      4. *(The above steps are done for you in the cleaning code)*
   3. Consolidating Years: Years 2007, 2014 and 2019 as displayed on the graph will contain multiple years of data.
      1. Assign years 2005-2006 a value of 2007.
      2. Assign year 2013 a value of 2014.
      3. Assign year 2018 a value of 2019.
      4. *(You need to add this to the cleaning code)*
   4. Log Adjustment
      1. Use the real earnings variable from 2a to generate a log earnings variable at the individual level by taking the log of that variable.
      2. *(This is done for you in the cleaning code)*
   5. Age Controls
      1. Generate dummy variable to control for each of the six 5-year age categories to account for cohort size and life-cycle effects
      2. *(This is done for you in the cleaning code)*
3. Plotting Figure
   1. Perform Quantile Regression on Each Sample
      1. A quantile regression allows us to calculate the earnings gaps between white and Black individuals at different spots (quantiles) in the earnings distribution.
      2. Identify the log point gap at the median.
         1. This gap is represented by the coefficient on the black dummy variable in the quantile regression.
      3. Control for six 5-year age categories mentioned above to account for cohort size and life-cycle effects.
      4. Control for “other” races.
         1. This allows us to hone in on just the Black-White Gap.
      5. Perform this quantile regression for each year in the data, and for each sample definition.
   2. Store Quantile Regression Coefficients per Year per Sample as a Dataset
      1. For each sample, these datasets show the log point difference between: Log earnings for Black individuals at the median of the Black earnings distribution vs. white individuals at the median of the white earnings distribution
      2. You can use the parmest command in Stata to save parameters. You will initially save one dataset for each year, and you will ultimately want to combine this into a single dataset for plotting.
   3. Use These 3 Datasets to Produce 3 Figures
      1. Figures to be created for each sample
         1. Racial gap in log median income for working men only
         2. Racial gap in log median income for all men
         3. Racial gap in log median income for all men and women
      2. Basic Parameters of Each Figure
         1. X-axis : Year
         2. Y-axis : Black-white gap (in log points)