**In-Class Exercise: September 8**

**Describing Data: What Salary Can You Expect to Get?**

Use the dataset “***ACS Business Major Earnings 2015***” (available on <http://sites.bu.edu/qm222projectcourse> *→Other Materials→*Datasets for In-Class Exercises). This dataset gives the 2015 salaries of a Census sample of US 25 year olds who had been business majors in college.

1. Fill in the descriptive statistics for earnings in this table:

|  |  |  |
| --- | --- | --- |
|  | Excel Formula | Value |
| **Mean** |  |  |
| **Median** |  |  |
| **Standard deviation** |  |  |
| **Range** |  |  |
| **25th percentile**  |  |  |
| **75th percentile**  |  |  |

1. Suppose you wanted to predict *your* future earnings. Are the statistics you calculated above a good way to do that? Why or why not?

**In-Class Exercise: September 14**

**Introduction to Stata**

1. Open Stata.
2. Open the stata dataset 2015busmajors25earnings (sites.bu.edu/qm222projectcourse→Other Materials→DataSetsforIn-Class Exercises
3. Get descriptive statistics by typing **sum** in the command window. (**sum** is short for **summarize** … you can shorten command or variable names as long as there aren’t more than one that start that way.)

Note that there is a variable listed as us2015a\_f~1p. The ~ means that the variable name was too long to show. (You can read its name in the variables window).

Why are there no observations for us2015a\_schl?

1. Learn what values this variable takes by typing **tab** **us2015a\_schl**.

An easy way to do this is by typing **tab** and then clicking on the variable name in the variables window. (Note: we need a codebook to interpret what these values mean.

These sure look like numbers! So make this numeric **destring us2015a\_schl, replace**

1. See what values educd has by typing **tab educd**
2. These seem like they are strings, but Stata gave numerical descriptive statistics. Why? Because in Stata, it is possible to give labels to specific values of specific variables. To see the underlying numbers, type **tab educd, nol** (The nol is short for nolabel)

Note: If you had click on the command tab educd in the Review window, it will be copied to the command window so you just need to add **,nol**

1. You decide that you want to save some of this output. So you open a log file by clicking on the notebook icon towards the left. NEVER choose the default type (.scml) … Instead Choose **Log (.log)**

Why? Only Stata can read the default type by Word can read a .log file

Also, put the file where you want it by choosing the folder on the top where it says Save in. In the computer lab, you probably just want to put it on the desktop.

Then retype **sum,**  then retype **tab educd** (or just click those command on the top left).

If that’s all you want, type log close. Look for the file on your desk top and click it to open.

1. Then clear the dataset from Stata by typing **clear**
2. What if your data is not a Stata data set? Let’s use the Excel dataset ACS Business Major Earnings 2015 (on the same page). Download this, open it and save it somewhere (e.g. the desktop).
3. Back in Stata, click File, then choose Import, and then Excel spreadsheet. Browse the desktop and find your file. Change the worksheet to the data worksheet. Click “Import first row as variable names”. Click OK. Easy peasy.
4. **Clear** this data set. What if your data was in a text file? The easiest is a text file ending in **.csv** A lot of sources can give you a .csv file. This stands for **c**omma **s**eparated **v**ariables. Go back to our website and download the file NCVS PERSONAL VICTIMIZATION 1993-2013, and save it to your desktop if you can – If you click on it, it will open in Excel but you can save it to the desktop as csv using save as. Keep it as csv.
5. Back in Stata, click File, then choose Import, then “text data(delimited, \*.csv …). Browse the desktop and find your file. Change the worksheet to the data worksheet. Click “Import first row as variable names”. Click OK. The defaults seem good so we can just click OK.
6. If we have time, we can then:

**browse** (This shows you the data set)

**edit** (This shows you the data set so you can edit it)

**sum** and check out the data, noticing the variables with different numbers of observations.

**generate** a new variable

**generate** a new variable with an if statement. On logical statements:

Many commands in Stata can be applied to a subset of the data only with an if statement. “If” statements are also called logical statements.

In Stata logical statements (only), you can use these “operators”:

**==** ( double equal signs) equals (use in logical statements only)

**&**  and

**|**  or

**!=** not equal to

**< > <= >=** obvious

Example: sum wagp **if** agep>=25

# Class 6 Examples of Right (or Wrong) Statistics

***Your team name:***

***Your team members:***

***Answer these questions as a team 1 sentence per answer. Only write correct answer, since if you cross out I won’t know if you did that after our discussion. 1 point per correct answer.***

1. Children and teenagers who watch violent TV shows and/or play violent video games are more likely to exhibit violent behavior (Y). One possible reason could be that violent TV shows/games increase teens’ and children’ violent behavior, suggesting we should limit childrens’ access to these. Can you suggest a very different reason that would be quite likely that could also lead to this correlation that would not lead to this policy suggestion?
2. High CEO pay is positively correlated with a company being more profitable. One possible reason is that more motivated CEOs make their firms more profitable, suggesting that companies should increase CEO pay. Can you suggest a very different reason for this correlation that would be quite likely, that would not lead us to the conclusion that companies should increase CEO pay to make them more profitable?
3. In the US, people who go to schools with smaller classes in grades K-6 are more likely to eventually graduate college . The experts conclude that smaller classes are important for high-quality education. Can you suggest a different quite likely reason that would cause this correlation?

d. In the WW2 fighter plane example, where should we add protection to the plane? Explain.

# Class 7: Regression

A colleague of mine gave me his running statistics for the past 20 year period. Variables were:

Minutes per mile average: 7.64

 Age average: 45.0

Distance (miles ran): average: 4.8

I used it to estimate the following regression (copied here from the results window using copy past and changing font to Courier New 9 point.)

. regress minutespermile age

 Source | SS df MS Number of obs = 2,949

-------------+---------------------------------- F(1, 2947) = 219.91

 Model | 27.135562 1 27.135562 Prob > F = 0.0000

 Residual | 363.635601 2,947 .123391789 R-squared = 0.0694

-------------+---------------------------------- Adj R-squared = 0.0691

 Total | 390.771163 2,948 .132554669 Root MSE = .35127

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

minutesper~e | Coef. Std. Err. t P>|t| [95% Conf. Interval]

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

 age | .0178476 .0012035 14.83 0.000 .0154877 .0202074

 \_cons | 6.84023 .0540412 126.57 0.000 6.734268 6.946193

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

What is the dependent variable?

What is the explanatory variable?

In words, what does the coefficient on age .0178 tell us? Is its sign what you expected?

How many observations are there?

What do you think an observation in this data set is?

What do we learn from the intercept (\_cons) 6.840?

Running a multiple regression:

**. regress minutesper age distance**

 Source | SS df MS Number of obs = 2,949

-------------+---------------------------------- F(2, 2946) = 169.87

 Model | 40.4043817 2 20.2021909 Prob > F = 0.0000

 Residual | 350.366781 2,946 .118929661 R-squared = 0.1034

-------------+---------------------------------- Adj R-squared = 0.1028

 Total | 390.771163 2,948 .132554669 Root MSE = .34486

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

minutesper~e | Coef. Std. Err. t P>|t| [95% Conf. Interval]

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

 age | .0195712 .0011928 16.41 0.000 .0172325 .02191

 distance | .0453009 .0042888 10.56 0.000 .0368916 .0537103

 \_cons | 6.547485 .059858 109.38 0.000 6.430117 6.664853

What is the dependent variable?

In words, what does the coefficient on distance .04530 tell us? Is its sign what you expected?

In words, what does the coefficient on age .01957 tell us? Why is it different than the coefficient on age in the simple regression (.0178) ?

# Class 9: Standard Errors on Regression Coefficients: Does Money Buy Happiness?

The General Social Survey is an annual survey of Americans. We are using data from the 2012 survey wave. GSS respondents rate their happiness from 1 to 7, with 7 being completely happy. It also asks about family income (measured in $1000s)

happy – respondent’s happiness on a 1-7 scale

income – total family income in $1000s

Use this regression of happiness on income taken to answer the questions below. (Note: we’ve erased the Lower 95% and Upper 95% and other values.)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | ***Coefficients*** | ***Standard Error*** | ***t Stat*** | ***P-value*** | ***Lower 95%*** | ***Upper 95%*** |
| **Intercept** | **5.36774257** |  | **132.456** | **0** | **5.2882321** | **5.4472531** |
| **income** | **0.00286912** | **0.0006** | **4.76911** | **2.1E-06** |  |  |

1. Write out the regression equation (Y = b0+b1X) using the actual variable names and coefficient values.
2. What is the predicted happiness of someone who earns $30,000 per year?
3. What is the 95% confidence interval for the coefficient on income? Show your calculations.
4. A psychology expert believes that the true effect of $1000 on happiness is actually 0.004 points (on the 7-point scale). Do you think he is wrong? (Hint: Start by setting up a Hypothesis Test.) Show your calculations.
5. What is the standard error of the intercept? Show your calculations.
6. What is the average change in happiness for every $50,000 in income?
7. Additional Question in case you are finished early: If instead of measuring income in $1000s, the regression was measured income in $10,000s what would the value of the coefficient be on the new income$10000 variable? *(Hint: To check if you’ve done it right, what’s your intuition about whether the number should be bigger or smaller than .00287?)*

# Class 10: Stata for Assignment 3

Follow along with me with your data or open in stata the file class10acs2014data.dta on sites.bu.edu/qm222projectcourse >> Other Materials >> Data Sets for In-Class Exercises

1. If you have your own computer, tell stata where to put and find data by going

cd “put location here” for instance cd "c:/0000QM222Fall2017/"

1. Open a file that will save everything you do: e.g. **log using jsmithassignment3, text**  This will open a file for the results that ends in .log. DO NOT make the default type.
2. Summarize to see which variables are string variables, which are numeric. This also tells you how many (non-missing) observations you have for each variable.

 **sum** (this will summarize all variables)

1. For each variable, we want to figure out what the “missing” code is and make it into a stata missing value (either . or “” ).

 Let’s do this by going **tab variablename, missing**

 For those using my data, **tab metro, missing**

But this is a numeric variable, and I am looking at “value labels.” To see how it really is coded, go

 **tab metro, missing nolabel**

I see that the value 4 is really missing, so I replace it with a .

 **replace metro=. if metro== 5**

5. I want to make a dummy variable for being Asian from the variable race.

 I do the **tab** and **tab, nolabel** again and see that they are Asian if race is 4, 5 or 6. I make the

 variable. DO IT using if.

 There doesn’t seem to be any truly missing race, although I worry about two or more races.

 I want to make them missing. DO IT!

6. Maybe I want to put all of my variable changes in a file that I can rerun (e.g. if I collect more data.) To do this, I make a do-file. Follow me as I do it. Note that I am cutting and pasting the commands.

Then, erase all of the data ( clear ) and run this do-file.

7. I want more information on usual hours worked each week. I can go

 **sum uhourswk, detail**

Note: Is 99 a code for missing? We can’t know from this evidence. Instead, we need to look at the codebook or other documentation.

Other things we could do:

tab marst, summarize(incwage) tab marst gender, column

**In-Class Exercise – Class 12 Basketball Injuries**

One area that uses data analyses a lot is sports. This example is based on a project by a past QM222 student Jonathan Wong, who asked: “How do players do the season after they have an injury that keeps them from playing part of the previous season?” The source of the data was [www.basketball-reference.com](http://www.basketball-reference.com) and data is from 1977-2014.

The dependent variable is Win Shares per 48 minutes (***WS48***) which is a basketball statistic that measures how much a player contributes to winning on average during a 48 minute game. It “takes into account the various things a basketball player does to win or lose a game.” [[1]](#footnote-1) The average WS48 in this data is .116.

A simple regression of WS48 on a dummy variable for having been injured in the previous season (***INJURED***) leads to this regression:

(Note: When writing regressions, be sure to put either the coefficient’s standard error or the coefficient’s t-statistic in parentheses underneath the coefficient, note what is in parentheses.)

***WS48*** = .1203 - .03251 ***INJURED***

 (66.37) (-6.34)

t-statistics in parentheses

However, the age of the basketball player (***Age***) might affect both the injury rate and performance in the game. Therefore, Jonathan ran the following regression:

***WS48*** = .1486 - .0224 ***INJURED*** - .00279 ***Age***

 (18.38) (-5.41) (-7.37)

t-statistics in parentheses

1. Interpret the coefficient on ***INJURED*** in the first regression in a sentence.
2. Interpret the coefficient on ***INJURED*** in the second regression in a sentence. (The meaning is different in the two regressions). Explain **why** the ***INJURED*** coefficient is different than in the first regression.

***WS48*** = .1203 - .03254 ***INJURED***

 (66.37) (-6.34)

***WS48*** = .1992 - .0274 ***INJURED*** - .00279 ***Age***

 (18.38) (-5.41) (-7.37)

t-statistics in parentheses

1. There are two 25 year old basketball players with similar abilities. One was injured last season, one wasn’t. On average, how different will their WS48’s be next year?
2. Hard, needs intuition and thought (or luck): Based on the difference between the coefficient on ***INJURED*** in the simple regression and the multivariate regression, do you think that injured players are older or younger than non-injured players, on average? CIRCLE ONE:

 OLDER YOUNGER CAN’T TELL

There are **three** basketball positions – forward, guard and center. I therefore make two new dummy variables, ***Forward*** and ***Guard***. I run this regression:

***WS48*** = .2152 - .0289 ***INJURED*** - .00284 ***Age*** - .01194 ***Forward***  -.02485 ***Guard***

 (18.80) (-5.79) (-7.58) (-2.69) (-5.58)

t-statistics in parentheses

e. Holding ***Age*** and ***INJURED*** constant, what is the difference in ***WS48*** between ***Forwards*** and ***Centers***? CIRCLE AND FILL IN BLANK

***Forwards*** have \_\_\_\_\_\_\_\_\_\_\_\_\_\_ (CIRCLE ONE) HIGHER or LOWER ***WS48*** than ***Centers.***

f. Holding Age and INJURED constant, what is the difference in WS48 between forwards and guards? CIRCLE AND FILL IN BLANK

***Forwards*** have \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ (CIRCLE ONE) HIGHER or LOWER ***WS48*** than ***Guards***

g. On average, what ***WS48*** do I predict for a 30 year old ***Center*** who was not injured?

h. On average, what ***WS48*** do I predict for a 30 year old ***Guard*** who was not injured?

## In-class Exercise Class 13 and 14: Interpreting Multiple Regression

Some people think that we should pay money to high school students who perform well on a test, a program called “Pay for Performance”. Supporters think this gives students an incentive to learn and try hard. However, some people oppose paying students to learn, saying it is costly and that it crowds out “intrinsic motivation” (that is, it takes away love of learning).

Boloxia is a large city that has a metropolitan-area-wide school district. There are already some schools in the district that implemented the “Pay for Performance” program in 2006 and have been using it for several years. The other schools do not offer the pay for performance program.

You have a dataset of all the schools in the region, with the following variables:

* SCORE: Score on the Math Test in 2012
* OLD\_SCORE: Score on the Math Test in 2000
* PAY\_PROGRAM= 1 if the school offered the “Pay for Performance” program from 2008 through 2012, 0 otherwise
* POVERTY RATE : (0 to 100) = the poverty rate in the school district

I have run several regressions on these data. You can find the regressions as PayforPerformance log on our website (Other materials – Data and other materials used in class)

Your objective is to evaluate whether the Pay for Performance Program is successful.**(Regression 1)**

1. The first regression is a regression of SCORE on PAY\_PROGRAM. (t-stats in parentheses)

***Regression 1:***

***Score = 61.809 – 5.68 Pay\_Program adjR2=.0175***

 ***(93.5) (-3.19)***

* 1. What is the average SCORE of a school that offered PAY\_PROGRAM?
	2. Is there a statistically significant difference (at the 95% level) between the average SCORE at schools that offer the PAY\_PROGRAM and average SCORE at schools that do not?
	3. How much of the variation in SCORE is explained by this pay program?
1. We then ran a regression of SCORE on PAY\_PROGRAM and OLDSCORE: (t-stats in parens)

***Regression 2:***

***Score = 10.80 + 3.73 Pay\_Program + 0.826 OldScore adjR2=.6687***

 ***(6.52) (3.46) (31.68)***

* 1. Why is the coefficient on PAY\_PROGRAM different in Regression 1 v. 2?
	2. In words, what is the interpretation of the coefficient on PAY\_Program in Regression 1?
	3. In words, what is the interpretation of the coefficient on PAY\_Program in Regression 2?
	4. In words, what is the interpretation of the coefficient on OLD\_SCORE in Regression 2?
1. We then ran a regression of SCORE on PAY\_PROGRAM and OLD\_SCORE and POVERTY\_RATE. (t-stats in parentheses)

***Regression 3: adjR2=.6727***

***Score = 14.55 + 5.88 Pay\_Program + 0.797 OldScore – 0.213 Poverty***

 ***(7.10) (4.59) (28.97) (-3.05)***

* 1. In words, what is the interpretation of the coefficient on PAY\_Program in Regression 3?
	2. In words, what is the interpretation of the coefficient on OLD\_SCORE in Regression 3?
1. Which regression gives us the best estimate of causal effect of PAY\_PROGRAM. Why?

***Team In-Class Exercise Class 15***

1. Open hobbit data (Materials used in class – Data Sets for In-Class exercise.)

2. Browse the data to see what it looks like.

3. Make a variable Time.

4. Make a dummy variable weekend that is 1 on weekend days.

5. Run a regression of Gross on time and the weekend dummy

 a) Report the regression here (with t-stats in parentheses below each coefficient):

b) In words, interpret the coefficient on time…. Exactly what does it tell us?

c) In words, interpret the coefficient on weekend ... Exactly what does it tell us?

6. Run a regression of Gross on time and dummies for day of the week

[Hint use the xi: regress y i.x]

a) Report the regression here (with t-stats in parentheses below each coefficient):

b) What is the reference category, the excluded day of the week?

c) In words, interpret the coefficient on one of the day of the week variables ... Exactly what does it tell us?

d) What is the difference in gross between Saturday and Tuesday? Show how you calculated it.

## In-class Exercise Class 18: Omitted Variable Bias in the Pay Program

If you recall from an earlier exercise, Boloxia is a large city that has a metropolitan-area-wide school district that had some schools who implemented the “Pay for Performance” program in 2006 and have been using it for several years. The other schools do not offer the pay for performance program.

You have a dataset of all the schools in the region, with the following variables:

* SCORE: Score on the Math Test in 2012
* OLD\_SCORE: Score on the Math Test in 2000
* PAY\_PROGRAM= 1 if the school offered the “Pay for Performance” program from 2008 through 2012, 0 otherwise

We have run the following regressions (t-stats in parentheses)

***Regression 1:***

***Score = 61.809 – 5.68 Pay\_Program adjR2=.0175***

 ***(93.5) (-3.19)***

***Regression 2:***

***Score = 10.80 + 3.73 Pay\_Program + 0.826 OldScore adjR2=.6687***

 ***(6.52) (3.46) (31.68)***

1. Which of the coefficients on Pay Program have an omitted variable bias (the one in regression 1 or in regression 2)?

2. What is the sign of this omitted variable bias?

3. What is the value of the omitted variable bias? Show your calculations.

4. Draw a graph to explain why this bias occurs, and then explain why it occurs in words.

# Class 20. Multiple Slopes for a Single Variable

***GPA and Drinking***

Many people are concerned about the consequences of drinking on college campuses, for both health and academic performance. At BU, there are mandatory alcohol education courses and new alcohol enforcement policies.

The BU Dean of Students has asked you to think about whether to target a campus-wide advertising campaign that highlights the danger of drinking for academic performance at men or at women.

You have access to a survey about drinking linked to students GPAs, conducted at an anonymous university (in Other Materials→Datasets for In-Class Exercises→drinkingdata) Warning: this is not BU data! Use the dataset to answer to following questions. Drinking is measured as drinks per month.

First, run a regression of GPA on drinking and gender. Fill in the coefficients, and put standard errors below each coefficient in the parentheses.

 GPA = \_\_\_\_\_\_\_\_\_\_\_\_ + \_\_\_\_\_\_\_\_\_\_\_\* Drinks + \_\_\_\_\_\_\_\_\_\_\_\_\* Male

 ( ) ( ) ( )

Next, run a regression of GPA on drinking and gender, but let the slopes on the drinks coefficient vary between men and women. Fill in the coefficients and standard errors below.

GPA = \_\_\_\_\_\_\_\_ + \_\_\_\_\_\_\_\_\_\_ Male + \_\_\_\_\_\_\_\_\_\_\* Drinks For Men + \_\_\_\_\_\_\_\_\_\* Drinks For Women

 ( ) ( ) ( ) ( )

1. Is there a significant relationship between drinking and GPA on average? Explain.
2. What is the predicted effect of an increase in 10 drinks/mo. on the drinking scale on GPA …
	1. For men?
	2. For women?
3. What is the difference in predicted GPAs for men versus women …
	1. who don’t drink at all (drinks =0)?
	2. who drink the mean amount (drinks = 18.8)?
4. Do these regressions give causal evidence of drinking on GPA? Explain. To answer this, think about what important factors might be left out of the regression.
5. Should you target the advertising campaign at men or at women? Answer the question using the regressions above and any other relevant analyses of these data to make your case.

# Class 20 Omitted variable bias: Measuring & Predicting Bias, Inferring Correlation

In January 2016, *American National Election Studies* did a survey of attitudes towards presidential candidates of 1200 respondents randomly chosen from a “large and diverse set of over a million respondents who have volunteered to complete surveys online” and get paid a small amount for each survey they fill out.

While January 2016 was long before the primaries were finished, it is still interesting to see who tended to support the two eventual nominees. One survey question asked people to rate their feelings for the candidates, from 0 (“Very cold or unfavorable feeling”) to 100 (“very warm or favorable feeling”). The regressions below use the responses to this question for Donald Trump, and relate it to two variables:

* URM: A dummy variable for under-represented minorities
* Income: Family income in $000, “topcoded” at $320(000). This means that anyone whose income was greater than $320,000 had a value of $320,000. Only 1,053 respondents wrote their income.
* Educ: Education, measured in terms of degrees (from 1=not high school to 5=postBAdegree)

You start by running a simple regression of attitudes toward Trump on Income:

 Regression A :

Trump = 43.27 - .0480 Income

 (26.0) (-1.30)

(t-statistics in parentheses) Adjusted R2= 0.0007

You are concerned that Regression A didn’t control for race, so you add URM:

Regression B:

Trump = 45.26 - .0551 Income - 7.800 URM

 (21.6) (-1.49) (-1.97)

(t-statistics in parentheses) Adjusted R2= .0029

1. Of these two regressions, which regression is the limited model? \_\_\_\_\_\_\_\_\_\_\_

Of these two regressions, which regression is the full model? \_\_\_\_\_\_\_\_\_\_\_\_

1. Use one of the methods learned in class to figure out the sign of the correlation between Income and URM. (This is the sign of the coefficient a1 in the background model, a regression with URM the left hand side and Income on the right hand side.)

Challenge question if you finish the exercise early: What is the ***value*** of a1 ?

1. Explain in words what we learn from the sign of this correlation:
2. Which coefficient on Income is biased, Regression A or B? \_\_\_\_\_\_\_\_\_\_\_

In words, can you explain WHY it is biased?

1. What is the sign of the bias, positive or negative? \_\_\_\_\_\_\_\_\_\_

In words, can you explain WHY the bias has this sign?

1. Which regression is the better fit, and how do you know?
2. Overall, what do you conclude about the relationship between family income and feelings towards Trump then?

1. Overall, what do you conclude about the relationship between race and feelings towards Trump then?
2. If average income was 57,275 in this sample in January 2016, how warm were people’s feelings towards Trump then?
1. http://www.sportingcharts.com/dictionary/nba/win-shares-per-48-minutes.aspx [↑](#footnote-ref-1)