**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** | ***=average(A2:A2201)*** | ***78376.55*** |
| **Median** | ***=median(A2:A2201)*** | ***60984*** |
| **Variance** | ***=var(A2:A2201) or***  ***=var.s(A2:A2201)*** | ***4577060948*** |
| **Standard deviation** | ***=stdev(A2:A2201)*** | ***67653.98*** |
| **Range** | ***=min(A2:A2201)***  ***=max(A2:A2201)*** | ***0***  ***382000*** |
| **25th percentile** | ***=percentile(a2:a2201,.25) or***  ***=percentile.inc(a2:a2201,.25)*** | ***40000*** |
| **75th percentile** | ***=percentile(a2:a2201,.75)*** | ***93631.75*** |

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

***Possible answers:***

***These data cover all business graduates. Do you think BU Questrom students earn more or less than the average graduate? (My guess is that they earn more.)***

***Other than that, look for other reasons that you would be different than typical business graduates…. Your special abilities, your networks, etc.***

**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?

***Not necessarily. It is possible that children and teenagers who are inherently more prone to violence are also more likely to watch more violent TV shows. If that is the case, banning violent TV shows would be akin to "treating the symptoms" instead of the underlying problem.***

1. 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?

***The positive correlation between CEO pay and profits is probably because companies which hire the best CEO's end up being the most profitable. To hire the best CEOs, they need to pay a large salary. Paying a mediocre CEO more money is not going to make the firm more profitable. In fact, profits will probably go down!***

***Alternatively, companies that are profitable might increase their CEO’s pay because they have the money to do so. Again, it is not the profits causing the CEO to do better.***

1. 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?

***One possibility is that schools with smaller class sizes are more likely to be located in more affluent neighborhoods and cater to affluent families. If this were true, children who go to such schools are also likely to have other advantages such as proper nutrition, stable family environment, better supervision at home, etc. Thus we cannot conclude that it is small class size -- and not one of these other factors -- which contribute to success at the college level.***

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

***We should add protection to the places that don’t show very much damage. The planes that were damaged in these areas must not have survived to return home for analysis!***

# 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?

***minutes per mile***

What is the explanatory variable?

***Age***

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

***For each year older, the person took .0178 more minutes to run each mile***

How many observations are there?

***2,949***

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

***A run. (or, a day he rain)***

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

***Very little, since the colleague never ran when he was 0 years old. You need an intercept to fit the line.***

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?

***minutes per mile***

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

***For each additional mile he ran (in a run), it took him on average .04530 more minutes per mile. In other words, the farther he ran, the slower he ran on average. We all slow down after a while!***

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) ?

***Holding distance constant, for each year older, the person took .01957 more minutes to run each mile. This is greater than in the first regression, because you are holding distance constant. So age and distance must be correlated. Consequently, in the first regression, the coefficient on age was partially picking up the impact of distance. Later in the semester, we will explain exactly why the coefficient rose in the second regression.***

# 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.

**[0.00287 – 2\*0.0006, 0.00287 + 2\*0.0006] = [0.0017, 0.0041]**

1. 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.

**Null Hypothesis: slope coefficient = 0.004**

**Alternative Hypothesis: slope coefficient ≠ 0.004**

**The psychology expert’s guess falls within the 95% confidence interval around the estimated slope, so we do not reject that the true effect is actually .004.**

**Alternatively you can construct the t-stat, and see if its absolute value is greater or equal to 2 or not, under a 95% confidence level.**

**t = .00287 - .004 = -1.88 The absolute value is less than 2, we cannot reject**

**.0006 that the coefficient = .004 with 95% certainty**

1. What is the standard error of the intercept? Show your calculations.

***Since t = coefficient/se se=coefficient/t***

***5.3677/132.456 = 0.0405***

1. What is the average change in happiness for every $50,000 in income?

**50\*0.00287 = 0.144**

1. Additional Challenge Bonus 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?)*

***Multiply the coefficient .00287 by 10 ( = .0287) . For instance, when income is in $1,000s, for an income of $100,000, b X is:***

***.00287 income = .00287 \* 100 = 0.287***

***When income is in $10,000, for an income of $100,000, b X is:***

***.0287 income = .0287 \* 10 = 0.287***

**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.

***Players that were injured in the previous season are expected to have an average WS48 that’s .0325 lower relative to players that were not injured.***

***Note: When the dummy variable INJURED increases by 1, it goes from 0 to 1: i.e. we are comparing those injured with those not injured***

1. 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.

***Holding age constant, players that were injured in the previous season are expected to have an average WS48 that’s .0224 lower relative to players that were not injured.***

***The coefficients are different because, in the simple first regression, INJURED picked up the effect of the confounding factor Age.***

***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?

***We are comparing two people of the same Age, and asking about the impact of an injury (INJURED). Therefore, we need to use the second regression that holds Age constant.***

***The one who is injured will on average have a - .0274 lower WS48***

1. 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?

***Note:***

***Older people (Age) have lower WS48.***

***In both equations, INJURED decreases WS48.***

***Without controlling for Age, the impact of INJURED was a larger negative.***

***So it must be that the people who were injured were older, causing the impact of INJURED to pick up this confounding age factor.***

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

***Centers*** is the excluded (reference) category!

***Forwards*** have .01194 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 -.01194 – (-.02485) = + .01291 HIGHER WS48 than Guards***

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

**WS48 = .2152 - .0289\*0 - .00284\*30 - .01194\*0 -.02485\*0  
 = .2152 - .00284\*30**

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

***WS48* = .2152 - .0289\*0 - .00284\*30 - .01194\*0 -.02485\*1  
 = .01291**

## 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?

***= 61.809 – 5.68 = 56.129***

* 1. 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?

***t-stat on Pay Program = -3.19 |t|>2 so YES***

* 1. How much of the variation in SCORE is explained by this pay program?

***Adj R2= .0175***

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?

***Because in Regression 2, OldScore is being held constant. This tells us that OldScore was a confounding factor in Regression 1. (In fact, it must be that Pay Program had a negative sign in Regression 1 because it must have had a worse OLDSCORE.)***

* 1. In words, what is the interpretation of the coefficient on PAY\_Program in Regression 1?

***Those with the PayProgram scored 5.68 less than other schools.***

* 1. In words, what is the interpretation of the coefficient on PAY\_Program in Regression 2?

***For two schools with the same OldScore, those with the PayProgram scored 3.73 more.***

* 1. In words, what is the interpretation of the coefficient on OLD\_SCORE in Regression 2?

***If someone had a 1 point higher OldScore, they had a .826 higher (new) Score on average, holding Pay\_Program constant. Said a different way, if there were two schools with Pay Programs but one school had a 1 point higher OldScore, their (new) score was .826 higher on average. Similarly, if there were two schools without Pay Programs but one school had a 1 point higher OldScore, their (new) score was .826 higher on average.***

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?

***For two schools with the same OldScore and the same Poverty Rate, those with the PayProgram scored 5.88 more.***

* 1. In words, what is the interpretation of the coefficient on OLD\_SCORE in Regression 3?

***If school had a 1 point higher OldScore, they had a .797 higher (new) Score on average, holding Pay\_Program AND Poverty Rate constant.***

1. Which regression gives us the best estimate of causal effect of PAY\_PROGRAM. Why?

***Regression 3, since it holds more confounding factor constant, therefore isolating the impact of Pay Program for 2 otherwise identical schools (in terms of Oldscore and Poverty Rate)***

***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.

***gen time=\_n***

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

***gen weekend = Day == "Fri" | Day=="Sat" | Day == "Sun"***

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):

**. regress Gross time weekend**

**Source | SS df MS Number of obs = 133**

**-------------+---------------------------------- F(2, 130) = 42.72**

**Model | 1.4297e+15 2 7.1483e+14 Prob > F = 0.0000**

**Residual | 2.1754e+15 130 1.6733e+13 R-squared = 0.3966**

**-------------+---------------------------------- Adj R-squared = 0.3873**

**Total | 3.6050e+15 132 2.7311e+13 Root MSE = 4.1e+06**

**------------------------------------------------------------------------------**

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

**-------------+----------------------------------------------------------------**

**time | -81379.98 9248.284 -8.80 0.000 -99676.61 -63083.36**

**weekend | 1743349 717492.4 2.43 0.016 323875.7 3162822**

**\_cons | 6983532 788358.6 8.86 0.000 5423858 8543205**

**------------------------------------------------------------------------------**

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

***Every day, Gross decreases by -81380 (holding weekend constant).***

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

***On weekend days, Gross is 1,743,349 more than non-weekend days (controlling for time, i.e. excluding any time trend)***

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):

**. xi:regress Gross time i.Day**

**i.Day \_IDay\_1-7 (\_IDay\_1 for Day==Fri omitted)**

**Source | SS df MS Number of obs = 133**

**-------------+---------------------------------- F(7, 125) = 11.87**

**Model | 1.4395e+15 7 2.0564e+14 Prob > F = 0.0000**

**Residual | 2.1655e+15 125 1.7324e+13 R-squared = 0.3993**

**-------------+---------------------------------- Adj R-squared = 0.3657**

**Total | 3.6050e+15 132 2.7311e+13 Root MSE = 4.2e+06**

**------------------------------------------------------------------------------**

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

**-------------+----------------------------------------------------------------**

**time | -81323 9413.275 -8.64 0.000 -99953.04 -62692.96**

**\_IDay\_2 | -2021385 1350698 -1.50 0.137 -4694585 651814.9**

**\_IDay\_3 | 278613 1350436 0.21 0.837 -2394067 2951293**

**\_IDay\_4 | -620881.3 1350534 -0.46 0.647 -3293756 2051994**

**\_IDay\_5 | -1899400 1351584 -1.41 0.162 -4574352 775551.8**

**\_IDay\_6 | -1606004 1350928 -1.19 0.237 -4279658 1067650**

**\_IDay\_7 | -1903762 1351223 -1.41 0.161 -4578000 770476.7**

**\_cons | 8837266 1129044 7.83 0.000 6602748 1.11e+07**

**------------------------------------------------------------------------------**

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

***Looking at the labels of the daily dummies (which you see in the Stata variable list), you can see that the excluded day is Friday (or Day 1). Stata chooses to exclude some category, usually the first it sees.***

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

***On Mondays (Day 2), on average Gross is 2,021,385 less than on Fridays -- the excluded reference category -- controlling for time constant, i.e. excluding any time trend)***

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

***Saturday here is Day 3, Tuesday is day 5, so the difference is 278,613 - (-1,899,400)= 2,178,013***

**Note: We did not do part 2.**

## 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)?

**In regression 1**

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

**Negative (because this coefficient is more negative than the one in regression2.**

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

**-5.68 – 3.73 = - 9.41**

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

**In words: It must be that OldScore is correlated with who chooses the Pay Program, and particularly that schools with bad (old) scores chose the pay program.**

**See next page.**



# 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 Resources→Datasets→Drinking GPA.) 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 = 3.173 – 0.170 \* Drinks - 0.94 \* Male***

***( 0.090 ) ( 0.005) (0.063)***

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 = 3.238 - 0.243 \* Male - 0.013 \* Drinks For Men - 0.021\* Drinks For Women***

***(0.118) (0.188) (0.007) (0.007)***

1. Is there a significant relationship between drinking and GPA on average? Explain.

***The first regression suggests that after controlling for average differences in GPA between males and females, there is a significant and negative relationship between drinking and GPA (|t-stat| >2). The second regression suggests that this negative relationship is strongly significant for females and barely so for males at a 95% significance level.***

1. What is the predicted effect of an increase in 10 drinks/mo. on the drinking scale on GPA…
   1. For men? ***The predicted change in GPA is: ‑0.013 \* 10 = -0.13.***
   2. For women? ***The predicted change in GPA is: ‑0.021 \* 10 = -0.21.***
2. What is the difference in predicted GPAs for men versus women …
   1. who don’t drink at all (drinks = 0)?

***This is just the coefficient on “Male” in the second coefficient: males who don’t drink at all have predicted GPAs that are on average 0.243 lower than females who don’t drink at all.***

***(Reminder: We can ignore the intercept in our calculations since it is the same for men and women.)***

* 1. who drink the mean amount (drinks = 18.8)?

***Ignoring the intercept, the latter terms for men are ‑0.243 ‑ 0.013 \* 18.8 = -0.4874 while for women they are - 0.021 \* 18.8 = -0.3948***

***The difference? -0.4874 - (-0.3948) = ‑0.0926***

1. 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.

***The coefficient on drinking is not causal. There are omitted variables correlated with drinking and GPA at the same time: student motivation, health status (e.g. whether the student is suffering from depression) and etc.***

1. 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.

***We might consider targeting the advertising campaign at women since the negative effect of drinking on GPA is larger and more significant for them. However, as explained in 4. above, the evidence may not be causal; we cannot be sure that there aren’t other underlying factors driving the relationship between alcohol consumption and poor academic performance, in general or between male***

# 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? \_\_\_\_**A**\_\_\_\_\_\_\_

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

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.)

**Combined effect, c1 = -0.0480**

**Direct effect, b1 = -0.0551**

**Difference: Bias = c1 – b1 = -0.0480 – (-0.0551) = 0.1031**

**Omitted variable’s effect, b2 = 7.800**

**Given that the b1 < c1, there is a positive bias term (a1\*b2). If b2 is also positive, then a1 must be positive.**

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

**Since**

**bias = b2 \* a1**

**= 0.1031 = 7.800 \*a1**

**Thus,**

**a1 = (0.1031/7.800) = 0.013218**

1. Explain in words what we learn from the sign of this correlation:

**On average, under-represented minorities have higher annual family incomes than non-minorities.**

1. Which coefficient on Income is biased, Regression A or B? \_\_\_\_\_\_\_\_\_\_\_

In words, can you explain WHY it is biased?

**Regression A: since there is a missing variable that is being partially picked up by the coefficient on income.**

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

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

**Positive … as we calculated above, the bias is= c1 – b1 = -0.0480 – (-0.0551) = 0.1031**

**It has this bias because the income variable picks up that fact that Trump votes are less likely to be URM’s and URM’s on average make less income.**

1. Which regression is the better fit, and how do you know?

**Regression B has the higher adjusted Rsq.**

1. Overall, what do you conclude about the relationship between family income and feelings towards Trump then?

**The true impact of income on feelings toward Trump in Jan 2016 was probably negative, but its t-stat on income in regression B was only -1.49% so we are not 95% it is negative.**

1. Overall, what do you conclude about the relationship between race and feelings towards Trump then?

**The true impact of being a URM on feelings toward Trump in Jan 2016 was negative with approximately 95% certainty (t=-1.97).**

1. If average income was 57,275 in this sample in January 2016, how warm were people’s feelings towards Trump then?

**43.27 - .0480\*57.275 = 40.52**

**(using regression A, since we are talking about the overall average, not a specific race)**

1. http://www.sportingcharts.com/dictionary/nba/win-shares-per-48-minutes.aspx [↑](#footnote-ref-1)