SM222 SECTION B6: Modeling Business Decisions Midterm

***BOSTON UNIVERSITY***

***School of Management***

**Fall 2014**

**SECTION 1 Your Regression I AM ANSWERING THIS FOR THE REGRESSIONS ATTACHED AT THE END OF THIS TEST.**

Answer the following questions regarding the regressions that you have brought with you or the regressions that Professor Kahn gives to you.

Be sure to put your name on the page with your regression. When you complete the test, staple your regression sheet to your test.

Make sure all variables are defined (including your Y variable)

**Answer these questions based on your simple (1 variable) regression:**

1. What does each observation in your data set represent? (in a few words at most)

**A male (person) i.e. a guy.**

1. Use the value of the coefficient on your variable in a sentence that explains what it tells us. In other words, interpret this coefficient. (Do not use statistics terms in your answer. Be specific but concise.) Note: If your “simple” regression includes two (or more) X-variables that are different categories of the same categorical variable, answer this question and the next only about the first of these variables.

**If the person is a veteran, he would be 2.86% less likely to be working than non-veterans.**

1. Does this variable have a statistically significant effect on your dependent variable? Circle one:

**YES**

List three ways that you know based on the regression output:

1. **The |t-stat | > 2**
2. **The P-value is less than .05**
3. **The coefficient’s 95% confidence interval doesn’t include 0.**

**Note that the R-squared does not indicate whether the coefficient is significantly different from zero. For instance, the R-squared (and adj R-squared) here is only .0006 because veteran status does not explain *very much* of the variation in working, but it is still very significant.**

When a variable does not have a statistically significant effect, what does that mean, in everyday non- statistics terms?

**It means that we are not 95% certain that the variable affects Y (here, that veterans are less likely to be working.)**

**Now answer these questions based on your multiple regression:**

1. Use the value of the coefficient on your key variable in the multiple regression (i.e. the key variable that was also in the simple regression) in a sentence that explains what it tells us. In other words, interpret this coefficient. (Do not use statistics terms in your answer. Be specific but concise.)

**Holding age, citizenship, highschool and college education and marital status constant, veterans are 1.92% less likely than non-veterans to be working.**

1. Compare the two coefficients on the key variable that enters both regressions. Explain as specifically as possible why in the multiple regression, the coefficient it fell, rose, or stayed the same. This question will be graded based on whether it identified precisely why we see this direction of change.

**The coefficient has increased, i.e. it has become a less negative number. Therefore, in the simple 1-variable regression, it was holding a negative bias.**

**Why? It must be that the variables added in the multiple regression were confounding factors. Specifically, it must be that veterans are also people who have factors that make them less likely to be working – e.g. they are less likely to be married and/or less educated on average. (Not that the excluded education category is having less than a high school degree.)**

1. For two other variables in your multiple regression, explain specifically what we learn from the coefficient. (If you have multiple dummy variables for a categorical variable, this counts here as one variable. If you have two variables in total in your multiple regression, you obviously can only explain 1.)

**Those who have a highschool education are 1.58% more likely to be working than those who have less than a high school degree, while those with at least a college degree are 10.3% more likely to be working than those who have less than a high school degree (holding all other variables in the regression constant.)**

**A person who is 1 year older than another (but otherwise similar in terms of the other variables in the regression) is 0.35% less likely to be working.**

1. Why is it important to include all of these other variables in your regression, if they are not the focus of your research question? Explain fully.

**Because, as we saw in Q5 above, the effect of the variable that you are focusing on might be obscured (hidden) by the missing confounding factors biasing it. In other words, your key variable (here – veteran) can be picking up the effect of important omitted variables.**

**SECTION 2 Other Questions**

1. The ACS codebook says the following about its education variable:

**schl Highest Educational attainment**

01 No schooling completed

02 Nursery school, preschool

03 Kindergarten

04 Grade 1

05 Grade 2

06 Grade 3

07 Grade 4

08 Grade 5

09 Grade 6

10 Grade 7

11 Grade 8

12 Grade 9

13 Grade 10

14 Grade 11

15 12th grade - no diploma

16 Regular high school diploma

17 GED or alternative credential

18 Some college, but less than 1 year

19 1 or more years of college credit, no degree

20 Associate's degree

21 Bachelor's degree

22 Master's degree

23 Professional degree beyond a bachelor's degree

24 Doctorate degree

Instead of these 24 categories, you would like to have only 4 categories of educational attainment (1) less than high school (2) high school diploma or GED but not college diploma (3) college diploma (4) higher diploma. Then you would like to run a regression of the variable **earnings** on the different educational categories. What Stata commands would you write to create education variables and run the regression.

**Here is the quickest but not the only way:**

**gen lessthanhigh= schl < 16**

**gen highscl= schl >=16 & schl <21**

**gen college= schl == 21**

**gen morethancol= schl> 21**

**regress earnings highscl college morethancol**

**You could instead have used “ | “ i.e. “or”, such as**

**gen highscl= schl==16 | schl==17 | schl==18 | schl==19 | schl==20**

**You could also make each variable in two steps such as:**

**gen college= 0**

**replace college=1 if schl == 21**

**For full credit, you needed to use correct if statements, and you needed to realize that you could only put 3 of the 4 education variables in to the regression. The fourth one is the excluded category.**

1. Here is a regression of the quantity of fish sold in a fish market daily on the daily precipitation (inches of rain) and on the day of the week: **Some numbers have been erased.**

Source | SS df MS Number of obs = 97

-------------+------------------------------ F( 5, 91) = 5.48

Model | 152589134 5 30517826.8 Prob > F = 0.0002

Residual | 506735892 91 5568526.29 R-squared = 0.2314

-------------+------------------------------ Adj R-squared = 0.1892

Total | 659325026 96 6867969.03 Root MSE = 2359.8

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

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

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

mon | -1010.766 766.6924 -1.32 0.191 -2533.706 512.1747

tues | -2305.435 756.0636 -3.05 0.003 -3807.262 -803.6073

wed | -1836.111 747.2473 -3320.426 -351.7958

thurs | -4.749902 747.3892 -0.01 0.995 -1489.347 1479.847

precipitation | -2544.961 721.489 -3.53 0.001

\_cons | 7415.935 820.3249 9.04 0.000 5786.46 9045.409

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

1. What is the average difference in the quantity of fish sold on Monday and on Tuesday? (Show the calculations that you used to derive this answer.) If you cannot calculate the answer, explain why not.

**AVERAGE DIFFERENCE: -1010.8 -( -2305.4)= + 1294.6 more bought on Monday than Tuessday**

1. What is the amount of fish sold on Fridays? ? (Show the calculations that you used to derive this answer.) If you cannot calculate the answer, explain why not.

**To know this, we would need to know what the precipitation is. For instance, if there was zero precipitation, on average the quantity of fish sold would be 7415.9.**

**Note that Friday is the reference, excluded category. It is what it is when it’s not Mon, Tues,Wed, or Thurs.**

1. We have erased the t-stat on Wednesday. What is it? (Show the calculations that you used to derive this answer.)

**T-STAT: coefficient/se = -1836.11/747.247 = -2.457**

1. What common sense fact do we learn from this t-stat? (NO statistics terms…. The more non-statistics-sounding your answer, the more points you get.)

**We are more than 95% certain that less fish is sold on Wednesdays than on Fridays.**

**Note that Friday is the reference, excluded category. It is what it is when it’s not Mon, Tues,Wed, or Thurs.**

**You would also have gotten the whole points if you said:**

**We are more than 95% certain that more fish is sold on Wednesdays than on the average of Fridays, Saturdays and Sundays. (We did not clarify that the fish market is only open on weekdays.)**

1. On a Tuesday with .3 inches of rain, how much fish is sold on average? Within what range am I 95% certain that the sales of fish will be? Show your calculations.

**QUANTITY: = -2305.4 -2545.0\*0.3 + 7415.9 = 4347.0**

**RANGE: = 4347+/- 2\*2359.8 = -372.6 to 9066.6 (or could use 1.96 instead of 2)**

1. I am 95% confident that the coefficient on precipitation falls in what range? (Show your calculations)

**= -2544.961 +/- 2\*721.489 = -3987.94 to -1101.98**

1. Below are two regressions of quarterly sales of JCrew on quarter dummy/indicator variables, on the variable time, and on the variable time squared.

**Regression 1:**

Source | SS df MS Number of obs = 31

-------------+------------------------------ F( 4, 26) = 234.26

Model | 2.4594e+11 4 6.1486e+10 Prob > F = 0.0000

Residual | 6.8241e+09 26 262465918 R-squared = 0.9730

-------------+------------------------------ Adj R-squared = 0.9688

Total | 2.5277e+11 30 8.4256e+09 Root MSE = 16201

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

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

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

Q2 | 3482.262 8106.972 0.43 0.671 -13181.86 20146.38

Q3 | 10820.9 8126.657 1.33 0.195 -5883.684 27525.48

Q4 | 60838.98 8391.06 7.25 0.000 43590.91 78087.05

time | 9562.238 326.3744 29.30 0.000 8891.366 10233.11

\_cons | 126056.9 7534.938 16.73 0.000 110568.6 141545.2

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

**Regression 2:**

Source | SS df MS Number of obs = 31

-------------+------------------------------ F( 5, 25) = 180.36

Model | 2.4595e+11 5 4.9190e+10 Prob > F = 0.0000

Residual | 6.8183e+09 25 272732372 R-squared = 0.9730

-------------+------------------------------ Adj R-squared = 0.9676

Total | 2.5277e+11 30 8.4256e+09 Root MSE = 16515

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

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

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

Q2 | 3476.16 8264.111 0.42 0.678 -13544.09 20496.42

Q3 | 10820.9 8284.071 1.31 0.203 -6240.465 27882.26

Q4 | 60710.84 8598.574 7.06 0.000 43001.75 78419.93

time | 9757.495 1379.142 7.08 0.000 6917.099 12597.89

timesquared | -6.101781 41.82536 -0.15 0.885 -92.24273 80.03917

\_cons | 125013.5 10495.2 11.91 0.000 103398.3 146628.8

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

a) Which of these two regressions fits best? CIRCLE ONE:  **Regression 1**

How do you know? List two ways.

**(i) highest adjusted R-squared**

**(ii) lowest Room MSE (SEE)**

**(Also, because the only difference in the included variables is that Regression 2 has timesquared and this**

**variable had a |t-stat| <1.)**

b) Is the relationship between revenues and time linear or not? CIRCLE ONE:  **NO**

How do you know? **Timesquared in Regression 2 had a |t-stat| <1.**

c) What common sense fact do we learn from the coefficient on time in **Regression 1**? (NO statistics terms…. The more non-statistics-sounding your answer, the more points you get.)

**Each quarter, JCrews revenues increase on average 9562.238.**

d) What does each observation in this data set represent? (in a few words at most)

**A quarter.**

1. We ran an equation with a dummy variable for being an ***entrepreneur*** as the dependent variable and a dummy variable for whether or not the person was ***native-born*** as the explanatory variable. The sample was about 20,000 people whose BA major had been in STEM (science, technology, engineering and math, including social sciences).

***Entrepreneur*** = .0748 - .0204 ***native-born***

(.0051) (.0025)

(standard errors in parentheses)

1. In words, say exactly what the coefficient on native-born tells us.

**On average, when an individual is native born, there is a 2.04 percentage point decrease in the likelihood of being an entrepreneur.**

1. Foreign-born people are more likely to be engineers than native-born people. Engineers are also more likely to be entrepreneurs. If a dummy for engineering were added to the regression, would the coefficient on native-born become more negative or less negative? Why? Explain fully. (You might choose to use algebra or a graph, or just reasoning.)

**Here is a simple way to answer this in words:**

**Native born are less likely to be engineers and engineers are more likely to be entrepreneurs.**

**So without the variable “engineer” in the regression, the variable native will pick up the fact that native born are less likely to be engineers and therefore less likely to be entrepreneurs. This will “bias” the coefficient on native born to be more negative.**

**If you take add in a control variable for engineer and therefore take off this negative bias, the coefficient on native-born will be less negative.**

**Or, you could do this arithmetic:**

**Full model: Entrepreneur = b0 + b1\*native-born + b2\*Engineer**

**Substitute in the background relationship: Engineer = a0 + a1\*native-born**

**Entrepreneur = b0 + b1\*native-born + b2\*(a0 + a1\*native-born)**

**Entrepreneur= (b0+b2\*a0) + (b1+b2\*a1)\*native-born**

**Note from this that the bias = b2\*a1**

**The mis-specified limited model is:**

**Entrepreneur = c0 + c1\*native-born OR Enrepreneur = .0748 - .0204 *native-born***

**However, c1 ( -.0204) includes native-born’s direct effect PLUS the bias b2\*a1 where:**

**b2 is how being an engineer impacts being an entrepreneur, We were told this is positive.**

**a1 is how being native-born impacts being an engineer. We know this is negative (since foreign born are more likely to be entrepreneurs).**

**So the bias= b2\*a1 = a positive number times a negative number, which is negative.**

**If we included Engineer in the regression, we’d get rid of this negative bias on the coefficient on native-born.**

**Taking a negative amount *away from* the negative coefficient *would make it less negative* (more positive).**

**Thus, the fact that native-born are less likely to be engineers is likely a major reason that native born are also less likely to be entrepreneurs.**

These are regressions of ACS (the American Community Survey) survey data from 2008 to 2012 for males not currently in the military only. The dependent variable is whether or not they are currently working , with working=1.

The other variables are:

veteran =1 if the person is a veteran (This is the key variable)

age is the person’s age

citizen = 1 if the person is a citizen

highschool=1 if the person has graduated high school but not college

collegeplus=1 if the person has at least a college education

married=1 if the person is married

Source | SS df MS Number of obs = 3791981

-------------+------------------------------ F( 1,3791979) = 2268.91

Model | 312.761737 1 312.761737 Prob > F = 0.0000

Residual | 522712.223791979 .137846813 R-squared = 0.0006

-------------+------------------------------ Adj R-squared = 0.0006

Total | 523024.9823791980 .137929256 Root MSE = .37128

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

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

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

veteran | -.0285964 .0006003 -47.63 0.000 -.0297731 -.0274197

\_cons | .8380244 .0002025 4137.65 0.000 .8376275 .8384214

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

Source | SS df MS Number of obs = 3791981

-------------+------------------------------ F( 6,3791974) =44263.60

Model | 34233.8619 6 5705.64364 Prob > F = 0.0000

Residual | 488791.123791974 .128901496 R-squared = 0.0655

-------------+------------------------------ Adj R-squared = 0.0655

Total | 523024.9823791980 .137929256 Root MSE = .35903

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

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

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

veteran | -.0192179 .0005961 -32.24 0.000 -.0203862 -.0180496

age | -.0035962 .0000176 -203.77 0.000 -.0036308 -.0035616

citizen | -.0141807 .0006666 -21.27 0.000 -.0154872 -.0128742

highschool | .0158482 .0004135 38.32 0.000 .0150377 .0166587

collegeplus | .1032049 .0006725 153.47 0.000 .1018869 .104523

married | .1831442 .000393 466.04 0.000 .182374 .1839145

\_cons | .8748984 .0009198 951.18 0.000 .8730957 .8767012

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

Didn’t use:

These are regressions of ACS (the American Community Survey) survey data from 2008 to 2012. The dependent variable is annual earnings.

The other variables are:

veteran =1 if the person is a veteran

age is the person’s age

citizen = 1 if the person is a citizen

collegeplus=1 if the person has at least a college education

female=1 if the person is female

married=1 if the person is married

. regress earnings veteran if working==1

Source | SS df MS Number of obs = 6113163

-------------+------------------------------ F( 1,6113161) = 7783.34

Model | 2.4212e+13 1 2.4212e+13 Prob > F = 0.0000

Residual | 1.9016e+166113161 3.1107e+09 R-squared = 0.0013

-------------+------------------------------ Adj R-squared = 0.0013

Total | 1.9040e+166113162 3.1147e+09 Root MSE = 55774

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

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

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

veteran | 8107.189 91.89402 88.22 0.000 7927.08 8287.298

\_cons | 47965.5 23.32128 2056.73 0.000 47919.79 48011.21

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

Source | SS df MS Number of obs = 6113163

-------------+------------------------------ F( 5,6113157) = .

Model | 2.8519e+15 5 5.7037e+14 Prob > F = 0.0000

Residual | 1.6189e+166113157 2.6481e+09 R-squared = 0.1498

-------------+------------------------------ Adj R-squared = 0.1498

Total | 1.9040e+166113162 3.1147e+09 Root MSE = 51460

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

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

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

veteran | -3617.944 87.25648 -41.46 0.000 -3788.964 -3446.924

age | 744.1531 1.925493 386.47 0.000 740.3792 747.927

citizen | 13527.51 78.36973 172.61 0.000 13373.91 13681.11

collegeplus | 50297.63 63.90603 787.06 0.000 50172.37 50422.88

female | -21346.22 42.56528 -501.49 0.000 -21429.64 -21262.79

\_cons | 9246.092 106.4477 86.86 0.000 9037.459 9454.726

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