QM222 SECTION D1: Modeling Business Decisions Midterm

***BOSTON UNIVERSITY***

***Question School of Business***

**Fall 2016 WITH ANSWERS**

***NOTE THAT THE OMITTED VARIABLE BIAS QUESTIONS HERE HAVE ANSWERS IN ITALICS***

**SECTION 1 Your Regression**

* 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, I will staple your regression sheet to your test.
* If you use your own regression, make sure all variables are defined (including your Y variable) on the sheet with your regressions or in your answers below.  **ANSWERS ARE FOR THE REGRESSIONS AT THE END OF THIS TEST.**

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

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

**A SURVEYED PERSON.**

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.

**EACH $000 OF FAMILY INCOME DECREASES THE CLINTON SCORE BY 5.57 PERCENTAGE POINTS. NOTE: PERCENTAGE POINTS ARE DIFFERENT FROM PERCENTAGES. IF CLINTON’S AVERAGE SCORE WAS 50, THEN THIS WOULD BE 5.56/50 = 11.1%**

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

**YES** NO

List three ways that you know based on 3 different numbers in the regression output:

i. **| T | > 2**

ii. **P<.05**

iii. **THE 95% CONFIDENCE INTERVAL DOES NOT INCLUDE ZERO.**

1. When a variable does have a statistically significant effect, what does that mean, in everyday non- statistics terms?

**WE ARE AT LEAST 95% CERTAIN THAT THE COEFFICIENT IS NOT ZERO; OR, WE ARE AT LEAST 95% CERTAIN THAT THE VARAIBLE HAS AN IMPACT OF THAT SIGN.**

**Now answer these questions based on both your multiple regression and your simple regression:**

1. *Compare the two coefficients on the key variable that enters both regressions.*

*In which regression is there omitted variable bias, simple or multiple:* ***SIMPLE***

*Exactly how much is this bias? (Be sure to include a negative if it is a negative bias)*  ***-.05571 +.04857= -.00714***

*Exactly what do you learn from these regressions about the correlation between the two explanatory variables in the multiple regression? Is the correlation (CIRCLE ONE):*

*POSITIVE NEGATIVE CAN’T TELL*

*Explain exactly how you know the amount/sign of the bias and the sign of the correlation (or explain why one can’t tell.)*

***THE VARIABLE NEWINT HAS A POSITIVE COEFFICIENT (WHICH WE CALL B2) IN THE FULL MODEL. THE BIAS IS NEGATIVE. SINCE BIAS = - .00714 = A1 B2, THEN A1 MUST BE NEGATIVE. THE SIGN OF A1 IS THE SIGN OF THE CORRELATION, SO THE CORRELATION IS NEGATIVE.***

*5. Name one other possibly confounding factor that you will or should add to this regression to remove some of the omitted variable bias on the coefficient of the “key” explanatory variable (i.e. the one that is in both regressions above.)*

***BLACK***

*Logically, why do you think this is possibly confounding?*

***BLACKS WOULD BE MORE LIKELY TO BE FAVORABLE TO CLINTON (THIS IS B2). ALSO BLACKS HAVE ON AVERAGE LOWER FAMILY INCOME (THIS HAS THE SIGN OF A1).***

1. *If you added this new possibly-confounding variable to the multiple regression, how would the coefficient on the key variable (the one in all 3 regressions) change? CIRCLE ONE*

***BIAS= A1 B2 = NEGATIVE \* POSITIVE = NEGATIVE BIAS.***

***IF WE ADD IN BLACK, THE COEFFICIENT ON INCOME WOULD LOSE THIS NEGATIVE BIAS SO IT WOULD INCREASE (BECOME LESS NEGATIVE) AND COULD EVEN TURN POSITIVE. SO:***

***THE COEFFICIENT WOULD INCREASE (THIS INCLUDES A NEGATIVE COEFFICIENT BECOMING LESS NEGATIVE)***

6. Going back to the two original regressions, which one fits the data better? CIRCLE ONE

**THE MULTIPLE REGRESSION**

List two ways that you know, using different numbers from the regressions:

**1. THE ADJUSTED R-SQUARED IS HIGHER.**

**2. THE ROOT MSE IS LOWER**

**SECTION 2 College Grads**

The next questions use data on a 2013 survey of college graduates.

1. This dataset divides marital status MARSTA into

1. married

2. living in a marriage-like relationship (but not married)

3. widowed

4. separated

5. divorced

6. never married

Also, in this dataset GENDER is a string variable where males are coded M and female F.

I want to study what kind of MEN live in a marriage-type relationship rather than get married. Therefore, I would like to only include those men who are married OR living in a marriage-like relationship, and make a variable “married” that =1 if the guy is married, 0 if the guy is living in a marriage-like relationship but not married, and missing otherwise. What Stata commands would I write to make this variable?

**THIS WILL NOT BE ON EITHER EXAM. HOWEVER, HERE IS AN ANSWER:**

**This puts 1 if married and 0 if not, for men only. Women have missing**

**gen married= MARSTA==1 if GENDER==”M”**

**This replaces married with missing if the person is not married or living in a marriage-like relationship**

**replace married = . if MARSTA > 2**

**I have made this variable “married” and run the following regression of this variable on:**

age: the age of the man (which goes from 20 to 75.)

citizen: an indicator/dummy variable for if the man is a US citizen

Note: I have erased some numbers.

. regress married age citizen

 Source | SS df MS Number of obs = 39102

-------------+------------------------------ F( 2, 39099) = 658.78

 Model | 72.4920576 2 36.2460288 Prob > F = 0.0000

 Residual | 2151.22399 39099 .055019924 R-squared = 0.0326

-------------+------------------------------ Adj R-squared = 0.0326

 Total | 2223.71605 39101 .056871079 Root MSE = .23456

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

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

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

 age | .0031994 36.12 0.000 .0030258 .003373

 citizen | -.0442343 .0041478 -10.66 0.000 -.0523641 -.0361045

 \_cons | .8264514 .0052134 158.52 0.000 .816233 .8366698

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

1. Use the value of the coefficient on age in a sentence that explains what it tells us. In other words, interpret this coefficient. (Use every-day non-statistics words in your answer. )

**EACH YEAR THAT A MAN AGES, IF HE IS LIVING WITH SOMEONE, HE IS 0.3 PERCENTAGE POINTS MORE LIKELY TO BE MARRIED (HOLDING CITIZENSHIP CONSTANT).**

1. I erased the standard error in the row on age. What exactly is the missing standard error? Show calculations.

**SE = COEF / T = 0.0031994/36.12 = .0000858**

1. Someone believes that (college grad) citizens are 5% more likely than noncitizens to *not* get married but just live together (holding age constant.) Based on your regression, can you say with 95% certainty that that person is wrong? CIRCLE ONE:

 **NO**

Explain your answer, showing any calculations:

**SINCE -.05 IS WITHIN THE 95% CONFIDENCE INTERVAL OF THE COEFFICIENT ON CITIZEN, WE CANNOT SAY WITH 95% CERTAINTY THAT THEY ARE WRONG.**

1. I want to know if immigrants and citizens make the same amount as other people, and if it depends on whether they are “permanent” residents allowed to stay in the US indefinitely. Using the same original data set (but with all people), I make 2 dummy/indicator variables for the type of immigrant:

 permanent: born abroad but is a permanent resident (i.e. has a green card)

 temporary: born abroad but is a temporary resident

I then run a regression of salary on age and these two variables:

. reg salary permanent temporary age

 Source | SS df MS Number of obs = 98062

-------------+------------------------------ F( 3, 98058) = 2880.50

 Model | 1.0824e+17 3 3.6079e+16 Prob > F = 0.0000

 Residual | 1.2282e+18 98058 1.2525e+13 R-squared = 0.0810

-------------+------------------------------ Adj R-squared = 0.0810

 Total | 1.3364e+18 98061 1.3629e+13 Root MSE = 3.5e+06

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

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

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

 permanent | 121201.7 53018.67 2.29 0.022 17285.77 225117.7

 temporary | 824318.8 59646.67 13.82 0.000 707412 941225.6

 age | 75157.63 809.202 92.88 0.000 73571.6 76743.65

 \_cons | -1616479 37995.29 -42.54 0.000 -1690949 -1542008

Use the value of the coefficient on **permanent** in a sentence that explains to us exactly what it tells us. In other words, interpret this coefficient. (Use every-day non-statistics words in your answer.)

**PERMANENT RESIDENT NONCITIZENS MAKE ON AVERAGE $121201.7 MORE THAN CITIZENS (THE EXCLUDED CATEGORY), HOLDING AGE CONSTANT.**

1. Jack and Jill are both temporary residents. Jack is 40 years old and Jill is 30 years old. Using this regression, on average how different will their salaries be? Show your calculations.

TO GET MAXIMUM CREDIT, DO THE LEAST POSSIBLE CALCULATIONS YOU WOULD NEED TO GET THIS ANSWER.

**WE CAN IGNORE THE NONCITIZENSHIP DUMMIES AND THE INTERCEPT, SINCE BOTH JACK AND JILL HAVE THE SAME VALUES FOR THESE. SO THE DIFFERENCE IS 10\*75157.6 = 751,576 MORE. (IF YOU RECALL, THESE NUMBERS WERE WACKY)**

1. Someone suggests to me that I create a variable equal to age squared and add it to the regression. I get the following:

. reg salary permanent temporary age agesq

 Source | SS df MS Number of obs = 98062

-------------+------------------------------ F( 4, 98057) = 4522.68

 Model | 2.0816e+17 4 5.2040e+16 Prob > F = 0.0000

 Residual | 1.1283e+18 98057 1.1506e+13 R-squared = 0.1558

-------------+------------------------------ Adj R-squared = 0.1557

 Total | 1.3364e+18 98061 1.3629e+13 Root MSE = 3.4e+06

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

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

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

 permanent | 488036.8 50968.74 9.58 0.000 388138.6 587934.9

 temporary | 589371.5 57224.76 10.30 0.000 477211.6 701531.3

 age | -455168.8 5743.539 -79.25 0.000 -466426.1 -443911.5

 agesq | 5710.62 61.28065 93.19 0.000 5590.511 5830.73

 \_cons | 9501081 124736.9 76.17 0.000 9256598 9745563

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

Is the relationship between salary and age non-linear? CIRCLE ONE:

 **YES** NO CAN’T TELL

Explain how you know:

**BECAUSE AGESQ IS STATISTICALLY SIGNIFICANT**

1. If someone is a permanent citizen, sketch the relationship between salary and age below. The youngest age is 20, so I have started the x-axis at 20. You do not need to draw to scale. However, **DO calculate exactly the salary a 20 year old permanent resident will get, and write the number in as the Y-axis in this graph**

**THE NUMBERS WERE ALL WACKY. THE Y-AXIS NUMBER WAS**

**=488037-455169\*20+5710.6\*400+9501081=3169978**

**TO KNOW THE SLOPE AT EACH AGE, I TAKE THE DERIVATIVE WITH RESPECT TO AGE =**

**-455169+2\*5710\*AGE. AT THE AGE OF 20, THIS SLOPE IS -226745. AT THE AGE OF 60, IT IS +230103. SO THE CURVE IS A SMILE-SHAPED QUADRATIC THAT STARTS OUT NEGATIVE BUT TURNS POSITIVE.**

1. Above, you had to calculate the salary of a 20 year old permanent resident. What is the 95% confidence interval for this prediction?

**751,576 +/- 2\*Root MSE = 751,576 – 6.8 E+06 to 751,576 + 6.8 E+06**

**SECTION 3 Final Question**

Someone did a random survey of people living in big cities. They found that children who live in apartments with cockroaches are more likely to have asthma. Health experts concluded that (at least in big citeis) cockroaches cause asthma. Can you suggest a different **quite likely** reason that would cause this correlation (besides cockroaches causing asthma)? Explain.

***THE ANSWER NEEDS TO BE SOMETHING THAT IS POSITIVELY CORRELATED WITH COCKROACHES AND POSITIVELY AFFECTS ASTHMA. FOR INSTANCE, DIRTY/DUSTY APARTMENTS TEND TO HAVE COCKROACHES AND THE DIRT/DUST CASUESE ASTHMA.***

**Regression for those without their own**

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 would 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 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 Hillary Clinton and relates it to two variables:

* **scoreclinton:** how favorably the person rated Clinton (0 to 100)
* **faminc:** 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.
* **Newsint:** An indicator variable =1 if the person chose “most of the time” as the answer to the question “Would you say you follow what’s going on in government and public affairs?” (Other choices were”only now and then,” “hardly at all” “some of the time” “don’t know”)

. regress scoreclinton faminc

 Source | SS df MS Number of obs = 1053

-------------+------------------------------ F( 1, 1051) = 5.37

 Model | 7054.29698 1 7054.29698 Prob > F = 0.0206

 Residual | 1379918.78 1051 1312.95793 R-squared = 0.0051

-------------+------------------------------ Adj R-squared = 0.0041

 Total | 1386973.08 1052 1318.41547 Root MSE = 36.235

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

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

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

 faminc | -.0557135 .0240358 -2.32 0.021 -.1028772 -.0085498

 \_cons | 46.34476 1.772554 26.15 0.000 42.86661 49.8229

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

. regress scoreclinton faminc newsint

 Source | SS df MS Number of obs = 1053

-------------+------------------------------ F( 2, 1050) = 3.57

 Model | 9365.68343 2 4682.84172 Prob > F = 0.0285

 Residual | 1377607.39 1050 1312.00704 R-squared = 0.0068

-------------+------------------------------ Adj R-squared = 0.0049

 Total | 1386973.08 1052 1318.41547 Root MSE = 36.222

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

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

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

 faminc | -.0485728 .0246221 -1.97 0.049 -.0968868 -.0002587

 newsint | 1.365406 1.028712 1.33 0.185 -.6531583 3.38397

 \_cons | 43.37354 2.854956 15.19 0.000 37.77147 48.9756

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