QM222 SECTION D1: Modeling Business Decisions Midterm

***BOSTON UNIVERSITY***

***School of Management***

**Fall 2015 ANSWERS**

**SECTION 1 Your Regression: I’ve answered this with these regressions on the Questrom salary data base**

. regress StartingSalary SimulatedGPA

Source | SS df MS Number of obs = 1416

-------------+------------------------------ F( 1, 1414) = 601.78

Model | 6.8165e+10 1 6.8165e+10 Prob > F = 0.0000

Residual | 1.6017e+11 1414 113272510 R-squared = 0.2985

-------------+------------------------------ Adj R-squared = 0.2980

Total | 2.2833e+11 1415 161365689 Root MSE = 10643

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

StartingSa~y | Coef. Std. Err. t P>|t| [95% Conf. Interval]

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

SimulatedGPA | 13689.38 558.0396 24.53 0.000 12594.71 14784.06

\_cons | 13685.26 1539.322 8.89 0.000 10665.66 16704.86

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

. regress StartingSalary SimulatedGPA international

Source | SS df MS Number of obs = 1416

-------------+------------------------------ F( 2, 1413) = 310.46

Model | 6.9706e+10 2 3.4853e+10 Prob > F = 0.0000

Residual | 1.5863e+11 1413 112261996 R-squared = 0.3053

-------------+------------------------------ Adj R-squared = 0.3043

Total | 2.2833e+11 1415 161365689 Root MSE = 10595

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

StartingSal~y | Coef. Std. Err. t P>|t| [95% Conf. Interval]

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

SimulatedGPA | 13968.59 560.6326 24.92 0.000 12868.83 15068.35

international | -2680.008 723.3242 -3.71 0.000 -4098.913 -1261.104

\_cons | 13439.21 1533.878 8.76 0.000 10430.29 16448.14

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

**Answer these questions based on your simpler regression (with the fewest variables):**

1. (4 points) What does each observation in your data set represent? (in a few words at most)

**A Questrom graduate**

1. (5 points) Use the value of the coefficient on your key variable (the variable that appears in both regressions) 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.)

**If the (simulated) GPA increased by 1 point, the Starting Salary would on average be $13,689 higher.**

1. (5 points) What specifically do you learn from the t-statistic on this coefficient in *this* regression? (Do not answer what generally the t-test tells us. Answer about what the t-statistic in THIS case.)

**Since the |t| >2, I am more than 95% confident that the coefficient is not zero (or the opposite sign), which means that I am more than 95% certain that GPA has a positive effect on starting salaries.**

1. (5 points) What specifically do you learn from the p-value on this coefficient in this regression output? (The more specifically you can answer this question, the more points you get.)

**P = 0.000 which means that I am more thant 99.9% certain that that the coefficient is not zero (or the opposite sign), which means that I am more than 99.9% certain that GPA has a positive effect on starting salaries.**

**Alternatively, you could say that the likelihood that the coefficient is zero (or the opposite sign) is less than .0005.**

**Answer these questions based on your multiple regression, the regression that adds a variable that had been omitted in the simpler regression:**

1. (4 points) Use the value of the coefficient on your key variable in the multiple regression (i.e. the key variable that you discussed above) 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.)

**If the (simulated) GPA increased by 1 point (but international status remained the same), the Starting Salary would on average be $13,969 higher.**

1. (4 points) What was the sign of the omitted-variable bias in the simpler regression, positive or negative?

Explain exactly how you know, including the calculations (if any) that led you to this answer.

**Negative: 13689 – 13969 = -280 which means that when the variable international was missing, the coefficient on (Simulated)GPA picked up a negative omitted variable bias of -280.**

1. (4 points) What is the sign of the correlation between your key variable (X1) and the variable omitted in the first regression (X2)? Explain exactly how you know, including the calculations (if any) that led you to this answer.

**Bias = -280 = a1 b2.**

**a1 has the sign of the correlation between international students and GPA.**

**b2 is the coefficient on international in the second regression or -2680.**

**Since the bias is negative and b2 is negative, a1 is positive.**

In everyday language (no statistics terms), what does this sign tell us? In your answer, use the real variable names (not X1, X2 ).

**International students get higher GPAs on average.**

1. (3 points) What would be the coefficient a1 if you ran a regression with X2 as the dependent variable and X1 as the explanatory variable? Show your calculations.

**-280 = a1 (-2680) or a1 =-280 /-2680 = .1045**

**SECTION 2 Voting**

On the last page of this test are regressions based on the GSS for the years 1972-2014. The dependent variable ***grass*** is an indicator (dummy) variable that indicates whether or not the survey respondent believes that marijuana should be legalized.

1. **Use regression 1 to answer these questions:**
2. (3 points) How many people answered this question? **34,147**
3. (4 points) In everyday words (not statistical terms), interpret what the coefficient on year and its t-stat tell us.

**Holding gender, white-race, and age constant, the average likelihood that people believe that marijuana should be legalized increases by .00749. I am more than 95% certain that time has a positive effect on this likelihood because year’s |t| >2.**

1. (4 points) Is the sign on year what you would expect, and why do you say this?

**I do expect that as time passes and people get more used to marijuana as a non-lethal drug (sometimes helping medicinally), support for its legalization increases.**

1. (4 points) In everyday words, interpret what the coefficient on age and its t-stat tells us.

**Holding gender, white-race, and year constant, each year older a person is makes him/her .00439 less likely to believe that marijuana should be legalized. I am more than 95% certain that age has a positive effect on this likelihood because |t| >2.**

1. (4 points) Is this sign on age what you would expect, ***in light of the fact that we are also controlling for year,***  and why do you say that?

**Although we are controlling for the passage of time changing attitudes, as people get older they tend to get more conservative (and as they become parents, they get to fear drugs more). This is as true in 2014 as it was in 1972…..**

1. (4 points) Predict the likelihood that a black woman aged 30 in the year 2014 would believe that marijuana should be legal.

**Grass= -14.50862 + .0799(0) + .0515(0) -.00439\*30 + .007487\*2014= .4385**

1. (5 points) You strongly believe that people who are parents are less likely to support legalizing marijuana (even controlling for gender, age, race and year.) Also, in this dataset, whites are less likely to have had children.

If I were to add in an indicator variable *child* (=1 if the person ever had a child) to Regression 1, what would happen to the coefficient on white? Explain how you know, including any calculations you did to arrive at this answer. (If you can explain without doing any calculations, then don’t include calculations.)

**Right now, the coefficient on white is picking up not just its own effect, but the fact that whites have fewer children which means they are more likely to support legalization (a positive bias). So if you add the variable child, it will make the coefficient on white go down (become less positive).**

**Alternatively, bias = a1 b2. a1 – which has the sign of the correlation between white and child, is negative. b2 is the effect of child on grass, which is also negative. Therefore the bias is positive, and separately controlling for child will make the coefficient on white less positive.**

1. (4 points) A previous study found than males were 10% more likely to support legalizing marijuana than females (controlling for race, age and year). Are the results in Regression 1 ***significantly*** different from that previous study? Explain, including any calculations needed to answer this question.

**The coefficient we measured was .0799, with a 95% confidence interval between .0706 and .0892. Therefore, the results are significantly different.**

1. **Use regression 2 to answer these questions:**
2. (4 points) In regression 2, we have added the variable year-squared. In everyday words, interpret what the coefficient on yearsq and its t-stat tells us.

**Since its t-stat is greater than 2, it is significant, which means that a quadratic relationship between grass and age fits better than a linear one.**

**Optional: Specifically, since year had a negative sign and yearsq has a positive sign, I know that the relationship is shaped like a bowl.**

1. (3 points) How can the coefficient on year be negative in Regression 1 but positive in Regression 2? Explain.

**I gave credit to most answers that realized that you can’t consider year separately from yearsq, since the total effect of year includes that of year squared.**

**I also gave credit to reasonable reasoning about missing variable bias, where the missing variable is the quadratic of year – although I like the answer less.**

1. **Use regression 3 to answer these questions:**
2. (3 points) In regression 3, we have replaced *age* with a set of indicator variables for the decade in which the person was born.

Actually, the GSS does not include the variable “yearborn.” However, you can calculate it from the variables already there as part of regression 1). What Stata command would you use to create yearborn?

**Yearborn = year - age**

1. (4 points) Here is the “sum” of the variable “yearborn”

sum yearborn

Variable | Obs Mean Std. Dev. Min Max

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

yearborn | 59388 1947.183 20.72048 1883 1996

What Stata command(s) would you use to make the variable born60s (used in Regression 3)?

**gen born60s = 0**

**replace born60s =1 if yearborn >= 1960 & yearborn <1970**

**Alternatively, gen born60s = yearborn >= 1960 & yearborn <1970**

1. (5 points) Which regression fits best?

**Regression 2 which has the highest adjusted R-squared.**

1. (4 points) In everyday words, interpret what the coefficient on born60s and its t-stat (in Regression 3) tells us.

**Holding gender, white and year constant, a person who was born in the 60s was .10444 more likely to support legalizing marijuana than a person born before 1950.**

1. (6 points) Based on regression 3, what are the characteristics of the groups least likely to want to legalize marijuana?

**Gender: female**

**Race: nonwhite**

**When born: before 1950**

**SECTION 3 Word Question**

(5 points) The UN’s World Health Organization just announced that eating red meat may cause cancer. They base this on statistical studies that show that ***middle aged people who eat a lot of red meat have higher rates of colon cancer.*** What is an alternative, quite likely, reason that people who eat a lot of red meat have higher colon cancer, one that does NOT suggest that people should stop eating meat? Explain.

**One good answer was: People who eat a lot of red meat also probably have other non-healthy habits like lack of exercise, smoking etc.**

**Note that answers had to give a “reason that people who eat a lot of red meat have higher colon cancer, one that does NOT suggest that people should stop eating meat?”**

**Source: GSS 1972-2014 Definitions:**

grass: an indicator variable =1 if person believes marijuana should be legalized

male: an indicator variable for being male

white : an indicator variable for being Caucasian

age: the age of the respondent

year: the calendar year of the survey

**Regression 1**

Source | SS df MS Number of obs = 34147

-------------+------------------------------ F( 4, 34142) = 653.44

Model | 496.737275 4 124.184319 Prob > F = 0.0000

Residual | 6488.58618 34142 .190047044 R-squared = 0.0711

-------------+------------------------------ Adj R-squared = 0.0710

Total | 6985.32345 34146 .204572233 Root MSE = .43594

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

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

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

male | .0799013 .0047614 16.78 0.000 .0705689 .0892337

white | .0515032 .0062145 8.29 0.000 .0393227 .0636838

age | -.0043914 .0001358 -32.35 0.000 -.0046575 -.0041253

year | .0074874 .0001995 37.53 0.000 .0070964 .0078785

\_cons | -14.50862 .3976841 -36.48 0.000 -15.28809 -13.72914

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

**Regression 2**

Source | SS df MS Number of obs = 34147

-------------+------------------------------ F( 5, 34141) = 642.52

Model | 600.775447 5 120.155089 Prob > F = 0.0000

Residual | 6384.54801 34141 .187005302 R-squared = 0.0860

-------------+------------------------------ Adj R-squared = 0.0859

Total | 6985.32345 34146 .204572233 Root MSE = .43244

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

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

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

male | .0787643 .0047234 16.68 0.000 .0695064 .0880222

white | .0506972 .0061646 8.22 0.000 .0386143 .0627801

age | -.004464 .0001347 -33.14 0.000 -.004728 -.0042

year | -1.58008 .0673077 -23.48 0.000 -1.712006 -1.448155

yearsq | .0003982 .0000169 23.59 0.000 .0003651 .0004313

\_cons | 1567.731 67.08265 23.37 0.000 1436.246 1699.215

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

**Regression 3**

. regr grass male white year born50s born60s born70s born80s born90

Source | SS df MS Number of obs = 34256

-------------+------------------------------ F( 8, 34247) = 301.13

Model | 460.371423 8 57.5464279 Prob > F = 0.0000

Residual | 6544.5772 34247 .191099285 R-squared = 0.0657

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

Total | 7004.94862 34255 .204494194 Root MSE = .43715

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

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

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

male | .0837273 .0047653 17.57 0.000 .0743871 .0930675

white | .0465241 .0062205 7.48 0.000 .0343317 .0587164

year | .00428 .0002297 18.63 0.000 .0038297 .0047302

born50s | .1464544 .006172 23.73 0.000 .1343571 .1585517

born60s | .1044433 .0072604 14.39 0.000 .0902128 .1186738

born70s | .1512566 .0096674 15.65 0.000 .1323081 .1702051

born80s | .2473589 .0132841 18.62 0.000 .2213217 .2733961

born90s | .1882654 .0248165 7.59 0.000 .1396242 .2369066

\_cons | -8.387474 .4569265 -18.36 0.000 -9.283065 -7.491883

Source: GSS 1972-2014

Definitions:

vote is an indicator variable for whether the person voted in the previous presidential election

age is the age in years

educ is the highest level of education achieved in years

. regress vote educ

Source | SS df MS Number of obs = 55420

-------------+------------------------------ F( 1, 55418) = 3034.41

Model | 568.521426 1 568.521426 Prob > F = 0.0000

Residual | 10383.0281 55418 .187358405 R-squared = 0.0519

-------------+------------------------------ Adj R-squared = 0.0519

Total | 10951.5495 55419 .197613626 Root MSE = .43285

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

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

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

educ | .0318844 .0005788 55.09 0.000 .0307499 .0330189

\_cons | .319073 .0076635 41.64 0.000 .3040525 .3340934

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

. regress vote educ age

Source | SS df MS Number of obs = 55249

-------------+------------------------------ F( 2, 55246) = 4016.38

Model | 1386.12245 2 693.061225 Prob > F = 0.0000

Residual | 9533.17372 55246 .172558624 R-squared = 0.1269

-------------+------------------------------ Adj R-squared = 0.1269

Total | 10919.2962 55248 .197641474 Root MSE = .4154

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

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

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

educ | .0405764 .0005706 71.12 0.000 .0394581 .0416947

age | .0072756 .0001057 68.80 0.000 .0070684 .0074829

\_cons | -.1331514 .0098635 -13.50 0.000 -.152484 -.1138188

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