MGT6203 business Analytics

Sample questions, Summer 2018

**(The actual exam is much longer with more questions covering more topics)**

**Part I**

- This part only has multiple choice questions

- This section will be closed book. No references allowed.

- We may even give outputs of analysis and ask for interpretations. We may ask conceptual questions. Questions will be similar to the format asked in HW.

1. If you run a regression without any explanatory variables (Right Hand Side variables), then what will the parameter estimate represent? i.e. your model is “y = intercept + error”. What does the intercept represent?

\_\_\_ A) Mean of y

B) Median y

C) Variance y

D) Standard deviation y

Answer – A

2. In logistic regression, you model for the success and failure of the outcome variable. We define Odds Ratio as a function of the probability ‘p’ of success of the outcome variable. Which is the correct definition of odds ratio?

\_\_\_

1. p
2. (1-p) / p
3. 1-p
4. p/(1-p)

Answer – D

3. Taking (natural) log transformation of a variable that has high variance does what?

\_\_\_

1. Does not change the variance at all
2. Increases the variance
3. Increases the standard deviation
4. Decreases the variation

Answer – D

4. What is the Cook’s (C) distance represent in regression diagnostics??

\_\_\_

1. It measures the fit of the regression
2. It measures the accuracy of the regression
3. It measures the influence for each observation
4. It measures the normality of the residuals

Answer – C

5. Which of the following will happen if you use linear regression to estimate a model that has a binary dependent variable?

\_\_\_ A) The regression will produce an error and then stop

B) The regression will run and produce some output

C) The R software automatically recognizes that the dependent variable is binary and use a logistic regression to fit the data.

D) None of the above

Answer – B

6. You run a log-log model and get the following results:

Log (demand) = 10 + 5.3 log(advertising)

How do you interpret the coefficient for the variable ‘advertising’?

\_\_\_ A) A 1% increase in advertising leads to a +5.3 % change in demand

B) A 1% increase in advertising leads to a +5.3 units change in demand

C) A 1-unit increase in advertising leads to a +5.3 % change in demand

D) A 1-unit increase in advertising leads to a +5.3 units change in demand

Answer – A

7. This is based on the Galton’s dataset on father and son’s heights. We are trying to predict son’s height (**sheight**) based on father’s height (**fheight)**. We have provided outputs for summary statistics, and regression output. Please study this and answer the questions that follow.

|  |  |
| --- | --- |
| **2 Variables:** | fheight sheight |

| **Simple Statistics** | | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **Mean** | **Std Dev** | **Sum** | **Minimum** | **Maximum** |
| **fheight** | 1078 | 67.68710 | 2.74487 | 72967 | 59.00800 | 75.43393 |
| **sheight** | 1078 | 68.68407 | 2.81470 | 74041 | 58.50708 | 78.36479 |

Call:

lm(formula = sheight ~ fheight, data = father.son)

Residuals:

Min 1Q Median 3Q Max

-8.8772 -1.5144 -0.0079 1.6285 8.9685

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 33.88660 1.83235 18.49 <2e-16 \*\*\*

fheight 0.51409 0.02705 19.01 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.437 on 1076 degrees of freedom

Multiple R-squared: 0.2513, Adjusted R-squared: 0.2506

F-statistic: 361.2 on 1 and 1076 DF, p-value: < 2.2e-16

7 A) What is the interpretation of coefficient for **fheight**?

A) a 1 unit increase in father’s height is associated with a 33.88660 unit increase in son’s height

B) a 1 unit increase in father’s height is associated with a 0.51409 unit decrease in son’s height

C) a 1 unit increase in father’s height is associated with a 0.51409 unit increase in son’s height

D) a 1 unit increase in father’s height has no significant impact on son’s height.

Answer: C

7 B) What is the predicted value of son’s height for the average value of father’s height?

A) 0.904169

B) 34.79117

C) 67.68710

D) 68.68407

Answer: D

sheight = 33.887 + 0.514 \* 67.68710 = 68.68407

This happens to be the average value of sheight, which is a proof that regression line passes through the mean of X and Y variables!

**Part II**

- This covers *application of R* for problem solving and *conceptual* questions.

- Questions could be multiple-choice **or** open ended. We are providing sample questions only for the open-ended type. Multiple choice format will be very similar to part 1. The key difference is that in this part is open book and you must submit your R code.

- Open book and open everything.

- Students need to submit their R code in addition to their answers and also cite any references used (websites, resources etc.) to answer questions. Instructors will evaluate these files and the answers submitted to assign grades.

1. How does logistic regression differ from linear regression? *Refer to slides for answers.*
2. What are the disadvantages of using linear regression for a binary classification problem? *Refer to slides for answers.*
3. What are the key assumptions of linear regression? Please explain **briefly.** *Refer to slides for answers.*
4. Vegasinsider.com shows that the odds of the Cleveland Cavaliers becoming the next NBA Finals Champions is 13 to 4 (which is odds against). What is the probability of the Cleveland Cavaliers becoming the next NBA Finals Champions?

**Odds for = 4/13 = p/(1-p), therefore 4 – 4p = 13p,**

**therefore 4 = 17p, hence p = 4/17.**

1. Variable X is the answer to the question “During the last 12 months, was the respondent a patient in a hospital overnight?” with one denoting the answer “Yes” and zero denoting the answer “No”. Variable Y is the answer to the question “Would you say your health in general is excellent, very good, good, fair, or poor?” with 5 being “Excellent”, 4 being “Very Good”, 3 being “Good”, 2 being “Fair”, and 1 being “Poor”. When we estimate the following model: Y=a + b1X + ε, we obtain a = 3.93 and   
   b1= − 0.72, with both coefficients being highly significant. Since b1 < 0, should we conclude that “going to hospital makes people become less healthy”? Why or why not?

**We should not conclude that “going to hospital makes people to be less healthy”.**

**Variable X is a choice variable chosen by people—people who are initially less healthy are more likely to choose to go to hospital. Without controlling for prior health of people, this variable becomes embedded in the error term ε. Because Variable X that measures “going to hospital” is correlated with error term “prior health”, the orthogonality assumption of this linear regression model (cov(X, ε)=0) is violated. In other words, there is a strong “selection bias”.**

**The coefficient we obtained is the sum of the “selection bias” discussed above and the real effect of hospital stay on people’s health. Therefore, we should not conclude that “going to hospital makes people to be less healthy”.**

6. We have estimated a logistic regression model to predict whether any of the 1,000 consumers in a specific sample will purchase a single-family house. Using 0.80 as the cutoff, we have obtained the predicted value of purchasing a single-family house to compare with the true value of purchasing a single-family house. We have the following frequency table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **Predicted Value** | |  |
| 0 | 1 | Total |
| **True Value** | 0 | 650 | 110 | 760 |
| 1 | 50 | 190 | 240 |
|  | Total | 700 | 300 | 1,000 |

A. What is the sensitivity of our model at this current cutoff? What is the specificity of our model at this current cutoff?

**Sensitivity: True Positive / (True Positive + False Negative) = 190/(190 + 50) = 0.79**

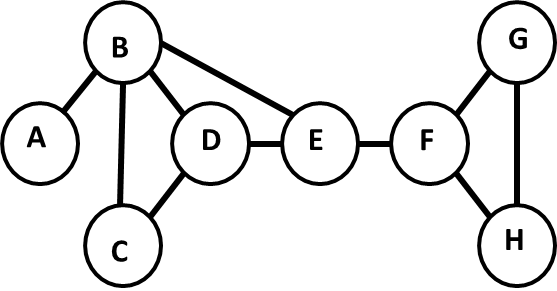
**Specificity: True Negative / (True Negative + False Positive) = 650/(650 + 110)=0.86**

B. If we reduce the cutoff in the model in Part (a) to 0.60, will we see a decrease of sensitivity? Will we see a decrease of specificity?

**No. Sensitivity will increase.**

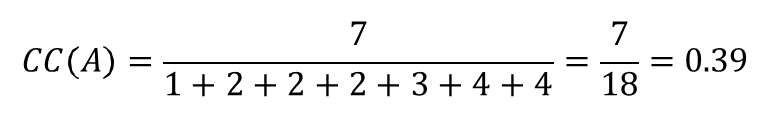
**Yes. Specificity will decrease.**

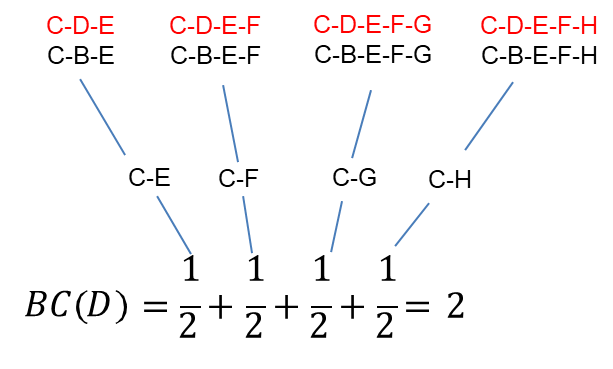
Questions 7 and 8. Given the following social network

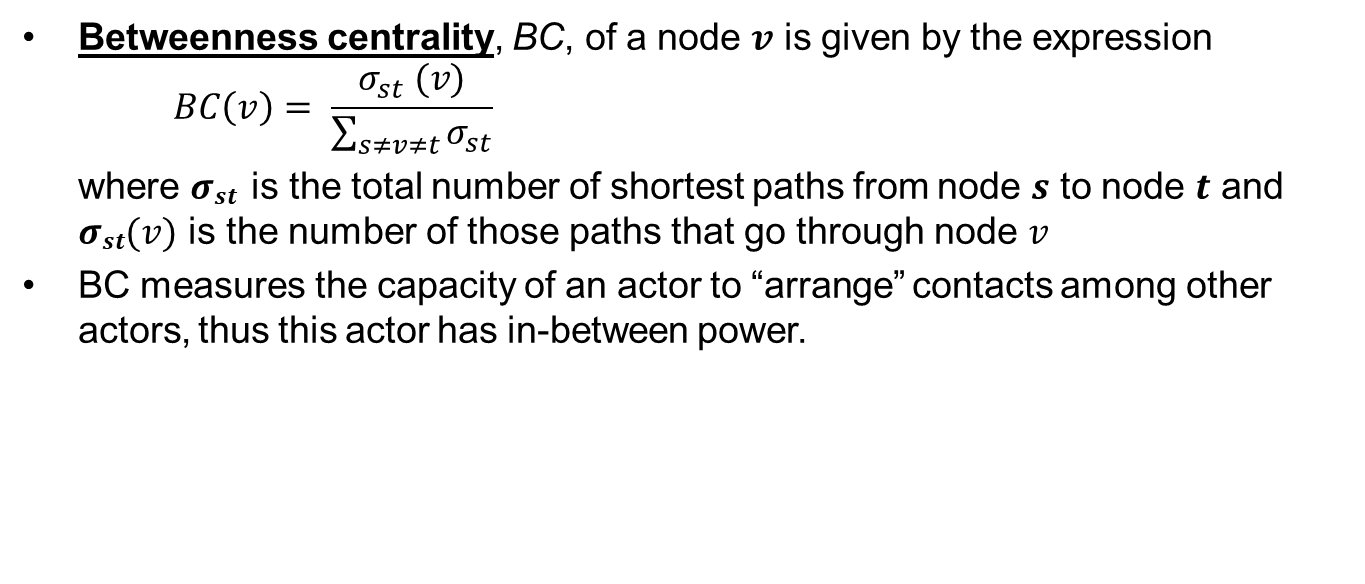


7. Compute the closeness centrality of node A

CC(A) = (N-1)/d(y,A) where d(y,A) is the length of shortest path between another node y and A .



8. Compute the betweenness centrality of node D



9. This question uses the dataset “FlightData.csv”. Please download the data and save it to your hard drive.

**Background Information**

We would like to investigate which factors can be used to predict the **airfare (y variable) of flights**. The data set comes from Department of Transportation for the fourth quarter of 2012. It contains the airfare data for the top 1,000 domestic flight routes (city pairs). “city1” records the name of the first city for each flight route; “city2” records the name of the second city for each flight route; “distance” records the distance (in miles) for each flight route; “passengers” records the average number of passengers per day for each flight route; “airfare” records the average airfare for each flight route; “largestcarrier” records the code of the airline with the largest market share for each flight route; “mktshlargest” records the market share of the airline with the largest market share for each flight route; “lowestfarecarrier” records the code of the airline with the lowest airfare for each flight route; “mktshlowestfare” records the market share of the airline with the lowest airfare for each flight route.

There are also two variables “farelargestcarrier” and “farelowestfarecarrrier” that we will not use in this assignment. There are some observations with missing data in this data set, but we will NOT filter them out for this assignment because they will NOT affect your analyses in this assignment.

**Data Preparation:**

Please create a dummy variable “many” that equals 1 when “passengers” is greater than 500, and equals 0 otherwise.

“WN” is the code for Southwest Airlines. <http://en.wikipedia.org/wiki/List_of_airline_codes>

Please create a dummy variable “southwestdom” that equals 1 when “largestcarrier” is equal to “WN”, and equals 0 otherwise.

1. Use dplyr to group by ‘many’ and ‘southwestdom’. Find the average airfare for each group. There will be four groups.
2. Run linear regression with airfare as the dependent variable, ‘many’ and ‘southwestdom’ as the independent variables. Now, can you use the below table to derive the individual group means obtained in #1 from the regression model? Show your work as to how you got your answers

|  |  |  |
| --- | --- | --- |
| Regression model | many | southwestdom |
|  | 0 | 0 |
|  | 0 | 1 |
|  | 1 | 0 |
|  | 1 | 1 |

1. Create a scatterplot with airfare as x axis and distance as the y axis with the regression line shown. Interpret the plot.
2. Please run a linear regression to understand how “airfare” can be predicted by “distance”. Please write down your interpretation of the two coefficients in such a linear regression.
3. Please run a linear regression to understand how “airfare” can be predicted by “distance” for a subsample in the whole sample with “many” being 1. Please run a linear regression to understand how “airfare” can be predicted by “distance” for a subsample in the whole sample with “many” being 0. Please write down your interpretation of the two coefficients in such a linear regression.
4. Please run ONE linear regression to generate all the results in question 5. Please write down your interpretation of all the coefficients in such a linear regression.
5. Now, can you use ggplot with grouping on ‘many’ to show the two regression lines in the graph? Please interpret the plot explaining how it relates to questions 5, and 6.
6. Now can you create a scatter plot (with regression lines) with facets for two factor variables – many and southwestdom ? Please interpret the plot explaining how intercept and slope is for each facet.

END OF SAMPLE QUESTIONS