**QTM 3610: Applied Multivariate Statistics**

Fall 2019, Problem Set #2

**Due: Tuesday 2/11 on Canvas before class. Make sure that all work is your own! Start a single .R script file in RStudio and save it as “LastName\_PSET2.R”. Then, clearly label (using comments) each problem and subproblem (i.e., Problem 1E)**

**The first two problems have a manual component. Either: Submit it in hard copy at the beginning of class on 2/11, or submit it electronically via Canvas. If you choose to submit it electronically, you may take a clear picture of your work, or do the work on OneNote and export it as PDF.**

**Problem 1. Basic Matrix Computations**

Matrix Computations.

A = B = C =

First, do each of the following manually, or explain why it is impossible. If you would like to do so, you can either show the calculations in Word, write out the calculations in OneNote and export to PDF before submitting, or writing it out (nicely) and scanning your work::

a.) A+BT

b.) AC

c.) BC

d.) BBT

Then, do each of the following in your R script (Name your R script “LastName\_Matrices.R”), or explain why it is impossible:

e.) A+BT

f.) AC

g.) BC

h.) BBT

i.) (AC)-1

j.) Find the determinant of Matrix A

k.) Find the determinant of Matrix C

**Problem 2. Linear Regression using Matrices**

In class, we solved for the Linear Regression questions using matrices. On Canvas, there are four data sets. For your last name, use the appropriate dataset. Using matrices, manually calculate (and show all work) the regression coefficients.

1. Last Name: A-E Group1Data.csv
2. Last Name: F-J Group2Data.csv
3. Last Name: K-Q Group3Data.csv
4. Last Name: R-Z Group4Data.csv

Then, do the same in an R script to double check your work. Make sure to create the appropriate matrices in R and then do the calculations to find the matrix for the regression coefficients.

**Problem 3. Checking Assumptions of Multivariate Analysis**

The data file CarSales.csv shows the car sales from major dealers across the country in the old days when cars were much cheaper. It includes the number of cards sold (“sold”), the average price (“price”), the number of seats (“Seats”), and some indicators of the car brand (a set of dummies and a categorical variable, “make”). It also includes a few survey variables that resulted from surveys conducted by dealers one year after selling the car to a buyer. It includes “Disregarding the price, how happy are you with the car?,” “How happy are you relative to the price?”, and “How safe do you feel in your car?”. These variables are average customer ratings ranging from 1 to 10, with 10 being the best.

In your R Script, run the following analyses and then answer the questions on the Word Document:

1. Run the Shapiro-Wilk Test. Is there a violation of multivariate normality? How do you know?
2. For each variable, plot the density plot and the Q-Q Plots. Which variables “look” normal? Which variables do not?
3. Run the Shapiro-Wilk test for each variable. Which ones does it say violate multivariate normality?
4. Would you recommend that multivariate analysis continue with this data as is? Why or why not?
5. Run a Box M test to examine the equality of covariance matrices between brands. Is there equality of covariance matrices? How do you know?

**Problem 4. Checking Assumptions of Multiple Regression Analysis**

Using the CarSales.csv data again, do the following in your R script and answer the following:

1. Run a multiple regression with Sold as the dependent variable and the rest of the non-categorical variables as your independent variables. Also, don’t include the dummy variables. What is your multiple regression model?
2. Now, run a multiple regression model that includes the dummy variables. What happens when you include all the dummy variables? How did you fix this?

For the full multiple regression model (that includes all the dummies):

1. Run a Bonferonni outlier test and a Q-Q plot for the residuals. Are there any outliers? How do you know?
2. Properly interpret each non-categorical variable (one sentence for each).
3. Properly interpret each categorical variable (one sentence for each).
4. What is the r-squared here? In one sentence, how do you interpret the r-squared?
5. Check for normality of residuals. Are residuals normal?
6. Evaluate Homoscedasticity. Is there equal variance?
7. Evaluate Collinearity. Is there multicollinearity?
8. Evaluate Linearity. Is there a violation?
9. Evaluate Independence. Is there a violation?
10. Would you recommend that we continue with regression analysis on this data as is?