CSC334/424: Assignment #1
Due: Wednesday, September 13, 2017 (by midnight)
Total: 50 points

Problem 1(5 points – Due Tuesday September 12) Introduce yourself on D2L by posting to the Class Introductions forum on D2L. Include a bit of information about yourself including some of the following. Note, this

- Name
- Undergraduate Degree
- Major/Degree Program(Concentration)/Time in Program (e.g. 3rd quarter, 2nd yr, graduating this quarter)
- Position at Work, if applicable
- What interests you about Advanced Data Analysis
- Field(s) of Interest and/data
- Hobbies

Problem2(15 points) Perform in R, the following calculations from linear algebra. For the following matrices and vectors. Submit both R code and the solution for credit.

$$Z = \begin{bmatrix} 1 & 5 \\ 1 & -3 \\ 1 & 2 \\ 1 & 4 \end{bmatrix}, Y = \begin{bmatrix} 2 \\ 1 \\ -1 \\ 3 \end{bmatrix}, M = \begin{bmatrix} 20 & 15 & 0 \\ 5 & 25 & 10 \\ 0 & 20 & 5 \end{bmatrix}, N = \begin{bmatrix} -20 & 5 & 10 \\ 0 & -10 & 10 \\ 5 & 20 & -5 \end{bmatrix}, v = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}, w = \begin{bmatrix} 2 \\ 1 \\ -1 \end{bmatrix}$$

a. $v \cdot w$ (dot product)
b. $-3 * w$
c. $M * v$
d. $M + N$
e. $M - N$
f. Z^T
g. $Z^T Z$ (Make sure you get the right dimensions on this matrix)
h. $(Z^T Z)^{-1}$ (You do not have to perform Gaussian-Elimination here … i.e. this is a hint on what the dimensions of the matrix should be ☺)
i. $Z^T Y$
j. $\beta = (Z^T Z)^{-1} Z^T Y$
k. $\det(Z^T Z)$

Problem 3(5 points – data exploration, visualization, and interpretation): Every four years, many of the world's greatest athletes gather to participate in the Summer Olympics. In addition to individual (or team) prowess, the Olympics is also a highly-watched pageant of national pride and competition. The data set (Olympics.xls under the course documents for week 2) for this problem concerns the performance of various countries in the 2012 London Summer Olympics. For each included country, the data contains medal counts, number of athletes (by gender), national population figures, and national GDP (gross domestic product).

It is your job to distill an interesting story or insight in this data and present it to the general public. You must choose the message you would like to communicate. Is there an important trend or lesson that you would like the public to understand? For example, are there ways to evaluate a country's "performance" beyond raw medal counts, and if so, do any surprises emerge? Is there any relationship between the success in Olympics game and the wealth of the people in country? How good/bad are they compared to the peers?

In your write-up, be sure to include the graph(s) you are using to see the relationships and clearly indicate the intended message of your graphic

Problem 4: (20 pts – regression analysis) The Housing dataset contains housing values in the suburbs of Boston in the 70s. There are 506 observations and 13 independent variables. (Note that in R, you can load in this file with simply "read.table("housing.dat")". If you try to specify a separator, R will get confused by the multiple spaces between fields.

1. CRIM: per capita crime rate by town
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.
3. INDUS: proportion of non-retail business acres per town
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
5. NOX: nitric oxides concentration (parts per 10 million)
6. RM: average number of rooms per dwelling
7. AGE: proportion of owner-occupied units built prior to 1940
8. DIS: weighted distances to five Boston employment centers
9. RAD: index of accessibility to radial highways
10. TAX: full-value property-tax rate per $10,000
11. PTRATIO: pupil-teacher ratio by town
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of African Americans by town
13. LSTAT: % lower status of the population
14. MEDV: Median value of owner-occupied homes in $1000's (**output dependent variable**)

a. Fit a linear regression model of CRIM based on the other variables by perform a feature selection on this data by using the forward selection method of the regression analysis. Analyze the output in terms of the order in which the variables are included in the regression model.

b. Compare the model selected by forward selection to both backward selection and stepwise selection.

c. Using a criterion mentioned in class, specify which model bests represents the data. Explain why you utilized the specific criterion.

d. For the model selected in part c, report goodness of fit, the utility of the model (F-test), the estimated coefficients, their standard errors, and statistical significance. Interpret your results.

Problem 5 (5 points – regression application): Briefly describe an application for the multiple regression in a field of interest to you. Identify possible independent variables and the dependent variable for your application. If you read about the application from a research paper or news article, please provide a reference to it.