CSC334/424: Assignment #1
Due: Friday, January 22, 2016 (by 10pm)
Total: 50 points

Problem 1(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.csv, in the assignments topic on D2L) 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 2: (**10 points – regression analysis**) In a study of genetic variation in sugar maple, seeds were collected from native trees in the eastern United States and Canada and planted in a nursery in Wooster, Ohio. The time of leafing out of these seedlings can be related to the latitude and mean July temperature of the place of origin of the seed. The variables are X1 = latitude, X2 = July mean temperature, and Y = weighted mean index of leafing out time. (Y is a measure of the degree to which the leafing out process has occurred. A high value is indicative that the leafing out process is well advanced.) *The data is in the file maple.txt in the D2L assignments topic.*
(a) (2 points) Find the regression of LeafIndex on Latitude. Is latitude a useful predictor of leaf index?
(b) (2 points) Repeat part (a) for the regression of LeafIndex on JulyTemp.
(c) (6 points) Find the regression of LeafIndex on Latitude and JulyTemp. Compare the results of this analysis with your results from (a) and (b). How different are the slope coefficients in each case? What best explains the differences in their values?

Problem 3: (**30 points – regression analysis**) The data in the file *chicinsur.txt* are collected from 47 zip-code areas in the Illinois area. There are 8 columns in the data file but not all are relevant here. The response variable of interest is the number of new home insurance policies (NEWPOL) (minus canceled policies) per 100 housing units. The predictor variables are the percent minority population living in the area (PCTMINOR), the number of fires per 1000 housing units (FIRES), the number of thefts per 1000 in population (THEFTS), the percent of housing units built before 1940 (PCTOLD), and the median income (INCOME). We are interested in which predictors are significant predictors of insurance policies issued.
(a) (5 points) Before running any regressions make a prediction as to what the sign of the coefficient of each predictor should be expected to be. (5 points) Obtain the correlation matrix for the variables PCT-MINOR FIRES THEFTS PCTOLD INCOME NEWPOL. Do the simple correlations support your predictions about the signs?
(b) Run a multiple regression of NEWPOL on the variables listed above.

i. (5 points) Comment on the overall significance of the regression fit.
ii. (5 points) Which predictors have coefficients that are significantly different from zero at the .05 level?
iii. (5 points) Do any of the predictors have signs that are different than suggested by their simple correlations? If so, explain what may be happening. If not, explain how such a thing can happen.
iv. (5 points) Examine a plot of residuals versus predicted values. Do you see any problems?

Problem 4 (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.