UNM Statistics Qualifying Exam Take-Home
August 2017

Due 3:00pm August 16 Wednesday, 2017. Return to Ana Parra Lombard in the Math/Stat Dept Office SMLC 395.

Directions: The answer to each problem should be presented as a summary. It should be word processed and double spaced, and be identified by your "Code Word" (do not include your UNM ID and name). **A suggested length of the report to each problem is no longer than 3 pages**. Create brief, well-organized appendixes for each problem.

In your data analysis, RAW AND UNINTERPRETED COMPUTER OUTPUT IS UNACCEPTABLE. You should have a caption by every figure and table that describes it and tells the reader briefly what you see. Organize the sections to tell the story you uncovered, not the circuitous path you may have taken to get there. Remember that even that best data analysis is worthless if your reader cannot understand it.

You may **not** consult any other person when working on this exam or discuss your exam with anyone else regardless of whether or not the person is taking the exam. You may use your course notes as well as any available books or web resources for the exam. Questions pertaining to clarification about these questions can be directed to Guoyi Zhang, gzhang12@math.unm.edu.

1. The data sets ("train.csv","test.csv" and variable description) are available at:
`www.math.unm.edu/~gzhang12/data.html`

   Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return has become automatic. Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. The duration of travel, departure, and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. For example, monitoring these data may detect significant events in the city. The bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, the day of the week, season, the hour of the day, etc. can affect the rental behaviors. The provided data sets "train.csv" and "test.csv" are subsets of the two-year daily log corresponding to years 2011 and 2012 from Capital Bike share system, Washington D.C., USA. File "description.csv" describes the variables of interest.

   (a) Plot or describe the training dataset ("train.csv") in a way that helps you understand the variables in the dataset and effects of potential explanatory variables on the response variable *count of total rental bikes*. (5pts).

   (b) Determine a regression model to predict *count of total rental bikes* using the training dataset. Treat categorical variables carefully. Consider transformations of variables and interactions if necessary. If you use model selection criteria to determine your final model, check model assumptions before and after applying the technique. Try to address the deviations from those model assumptions. Summarize the evidence for the decisions made to arrive at your final model and move other model fit details to the appendix (30 pts).

   (c) Write out the statistical model (in notation) with all parameters retained in the final model. Interpret each covariate's effect. Identify outliers and influential data points.(5pts)

   (d) Use your final model to predict the *count of total rental bikes* for the observations in the test dataset ("test.csv"). Compute the mean squared error of the predictions.(5pts)

   (e) Summarize main findings of your analysis.(5pts)

2. The data set (oncology2.dat) is available at: `www.math.unm.edu/~gzhang12/data.html`

read in the data using the following R code:
ex.data< −read.table(file="C:/qualifyingexam/2017august/oncology2.dat",head=T)

The goal of this analysis is to determine the reliability of size measurements of tumors in cancer patients depending on two variables: oncologist and shape. Researchers randomly recruited 26 oncologists from a city to measure simulated tumors. Shapes are in three types: "small", "oblong" and "large".

The simulated tumors were made of one of two materials chosen to physically resemble the texture of size of tumors which are found in cancer patients and they were made in one of three shapes: "small", "oblong" and "large". Two copies of each simulated tumor were made, all were placed randomly in rows on a folded blanket and then covered with a sheet of half-inch foam. The oncologists then independently measured each tumor with their usual equipment (ruler and calipers) and recorded the size obtained. "Size" is cross-sectional area, which they define as the product of the longest dimension and the shortest dimension of a tumor. Think about the following questions. Is there large oncologist to oncologist variability in measurement of simulated tumor sizes? Does this vary by tumor shape? Find a good statistical model for the size measurement. Make sure that your analysis includes appropriate displays, that you mention any unusual features of the data, and that you provide conclusions about your analysis.

Be sure to check the model assumptions such as constant variance, independence, normality and also check for outliers. Write a succinct, coherent, and complete summary of your analysis.