Problem Set 2: Basic Statistics

Introduction

In this Problem Set, we will practice doing basic statistical analyses and visualizations in R with a health dataset. Download the dataset associated with this assignment, "Session5PimaDiabetes.csv" from the Problem Set 2 folder on GitHub. Diabetes incidence and risk is extremely high among those with Pima Indian heritage from the Southwestern US. This dataset contains data for 768 women over the age of 21 with 8 continuous variables related to Diabetes risk along with a binary response variable called Diabetes. You will submit a short written report and accompanying R script file for this problem set. It is good practice to document your R code as you run it, so be sure to save code that you run into your R script file as you go along. Someone else running your R script file should be able to reproduce your results exactly.

Instructions

Follow the instructions below. Be sure to include the answers to any questions posed in your report. For the purpose of reproducibility, start your R analysis off by setting your seed to the number 1389 with the command: `set.seed(1389)`.

Data Quality Control

First, assess the missing data for all samples and variables. **Discard (and document) any variables with greater than 40% missing data.** Use multivariate imputation using the "mice" package to impute the remaining missing data points. During imputation, do you include the response variable or not?

With this new imputed data, assess each individual independent (predictor) variable for skewness and kurtosis of each independent variable using the `e1071` package. Note that the Kurtosis function in `e1017` has a normality expectation of 0 instead of 3. **Report any variable with a skewness less than -1 or greater than 1.**

For any variable with excess skewness (in this case any value between -1 & 1 is acceptable), anchor the variable to 1 and transform with the BoxCox. Use the `boxcox` function from the `MASS` packages and the `bcPower` function from the the `car` package. **Report the "optimal" lambda for each variable that needs transformation.**

Use the `uni.plot` function (with it's default setting) in the `mvoutlier` package to assess multivariate outliers. Report how many outliers it suggests there are. Report which ones (use the row numbers as ID's for each sample). Report which

sample is the most extreme outlier (hint: look at the mahalanobis distances in the output object)?

Data Source

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C. and Johannes, R. S. (1988) Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the Symposium on Computer Applications in Medical Care (Washington, 1988), ed. R. A. Greenes, pp. 261–265. Los Alamitos, CA: IEEE Computer Society Press.

The report

Develop a report (I recommend a Word (or other text editor) document) for your problem set that includes answers to all of the questions posed above, showing plots where appropriate.

Save your report as a PDF file and submit your report through the course 2GW site. Clean up your code and submit it as a supplementary file, along with your main report.

Due date

Day 7, Week 5