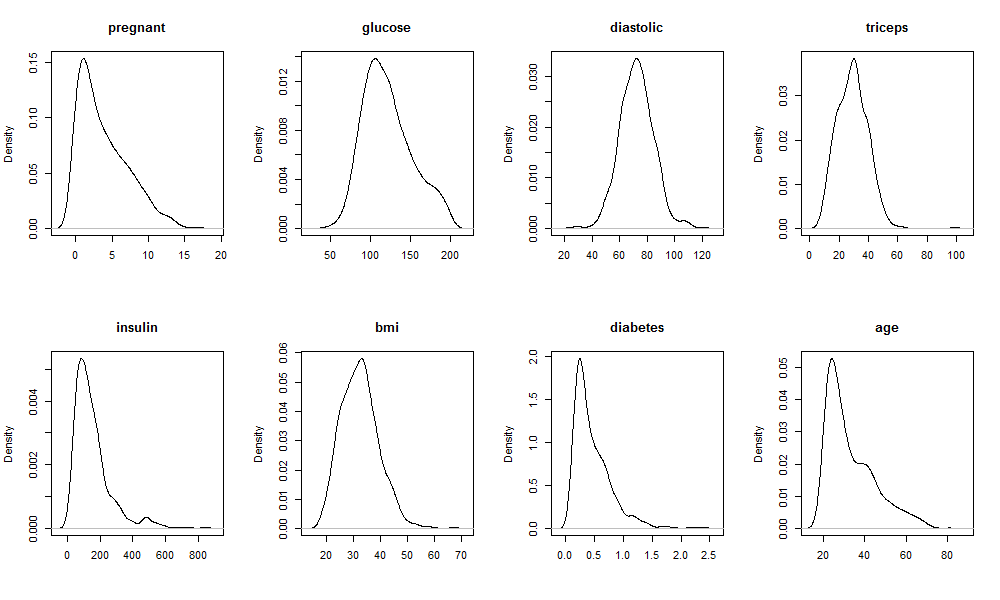
Alyssa Semerdjian

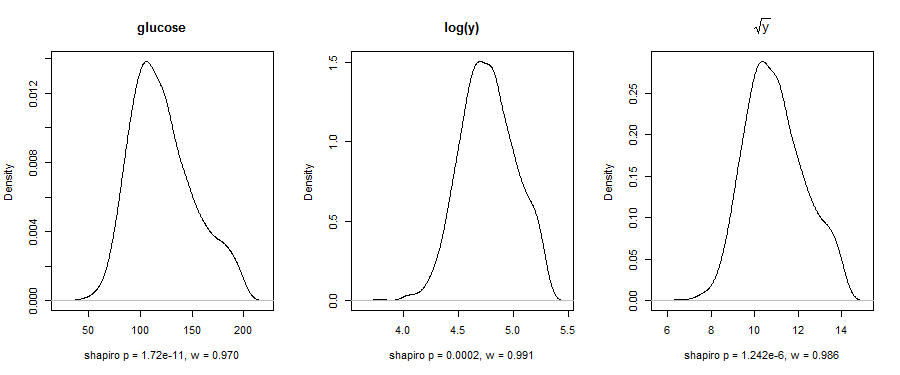
STAT 510

22 Jan 2018

Homework 1

1. a)The dataset has 9 rows and 768 columns

b)The diastolic variable appears to be the most normally distributed. The peak of the curve is at the center of the data and the curve appears very symmetrical. The number of times each person has been pregnant appears to be the least normally distributed. There are a couple of other variables with peaks that are not centered, but most of those are fairly symmetrical with only a few bumps in the graph. The pregnancy variable is very skewed to the right.

c) Below are graphs of the density of glucose variables, log(glucose) and sqrt(glucose). It is difficult to visually discern which is the most normally distributed. The one that looks the most symmetrical is log(glucose, but they are all slightly off center and not quite symmetrical, but close. Shapiro tests conducted on the data all returned p-values less than 0.05, suggesting you must reject the null hypothesis that the data is normally distributed. However, the W-value for log(glucose) was the closest to 1, and the p-value is the largest indicating that it is the most normally distributed out of the three.

d) Pima contains 768 rows, many of which have NA values for at least one variable. Data2 has 392 rows with no NA values. na.omit told R to throw out all rows with missing data.

2. a)Insulin is the most significant variable with a p-value of <2x10-16.

b) The estimate value for insulin in this model is 0.13321. This means that a one unit increase in insulin will result in a 0.13321 unit increase in glucose.

c) The model uses 384 degrees of freedom.

d) mean(pima$glucose,na.rm=T) gives you a value of 121.6868. mean(data2$glucose) gives you a value of 122.6276. The value for data2 is higher because that dataset does not include any rows with NA values while pima$glucose with na’s removed includes rows that have na’s for other variables, but not na values for glucose. So mean(pima$glucose,na.rm=T)uses more data points than mean(pima$glucose,na.rm=T) resulting in a different value. The model used mean(data2$glucose) because there cannot be missing values for any of the variables used in the model.

e) The residual for glucose for the first data point is -15.906. According to this model, glucose of this individual is lower than what’s expected.