Outlier Analysis for CAMP Exploratory Analysis

Haineycf

December 2, 2016

definition: "In statistics, an outlier is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error; the latter are sometimes excluded from the data set." - Wikipedia

There are essential aspects to this definition that need to be addressed: it needs to be a single point and needs to be distant from other points. It can be a true value, for example a mutation in a gene causing a gene to overexpress. On the other side, it can be incorrect information such as in a probe for the gene is not unique and can be mapped to two or more different parts of the genome experimental causing the signal to be higher than the other probes in the gene.

For the CAMP data, an outlier can be someone entering data incorrectly, or the test subject could have some serious other health effects unknown to the researchers.

An essential part of exploratory analysis is to determine outliers and the never ending question of removing them or include them in further analysis. Removing outliers can improve performance of downstream analysis, especially machine learning algorithms that are sensitive to extremes. If the outlier is a true value and not merely an input mistake it warrants further discussion.

Detecting Outliers

The one of the statistical method utilized in this report is Gubb's Test. This test will detect the point with the largest deviation from the sample mean and perform a t-test to determine if it should be labeled as an outlier. This test needs to be iterated until there are no more outliers.

Testing outliers one by one is rather tedious, especially if there are a large number of elements. One way to get around this is to an inital Grubb's test to identify if there are outliers, than use other methods to remove the outlier. This is the path this analysis will take.

Removing Outliers

There are a number of ways to remove outliers. One could look at the distribution of data and simply move it by eye, but this is neither scientific nor precise. There are a number of scripts and computational methods to look at these values; I will review a couple of them. This essay will focus on removing outliers before performing machine learning algorithms. The car package for R has a script to find outliers after constructing a linear model. This will not be reviewed here.

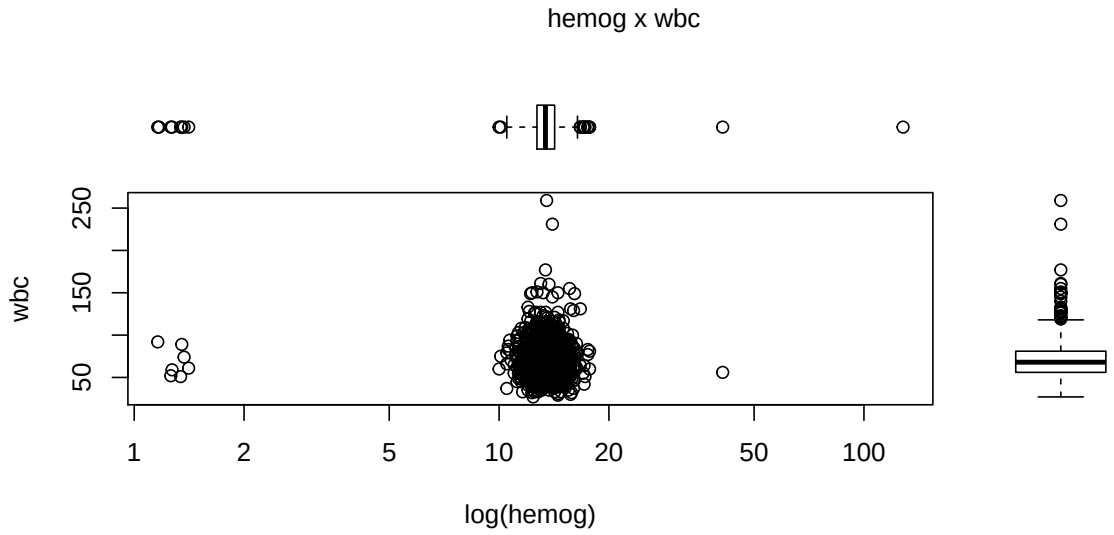
The first method utilizes Chebyshev's Theorem: the probability of a random variable differing from its mean by more than k standard deviations is less than or equal to 1/kˆ2. I have looked at removing observations that are outside of 2 standard deviation in one round. This is the first method utilized with the hemog and wbc data. Only one round of outlier removal is recommended.

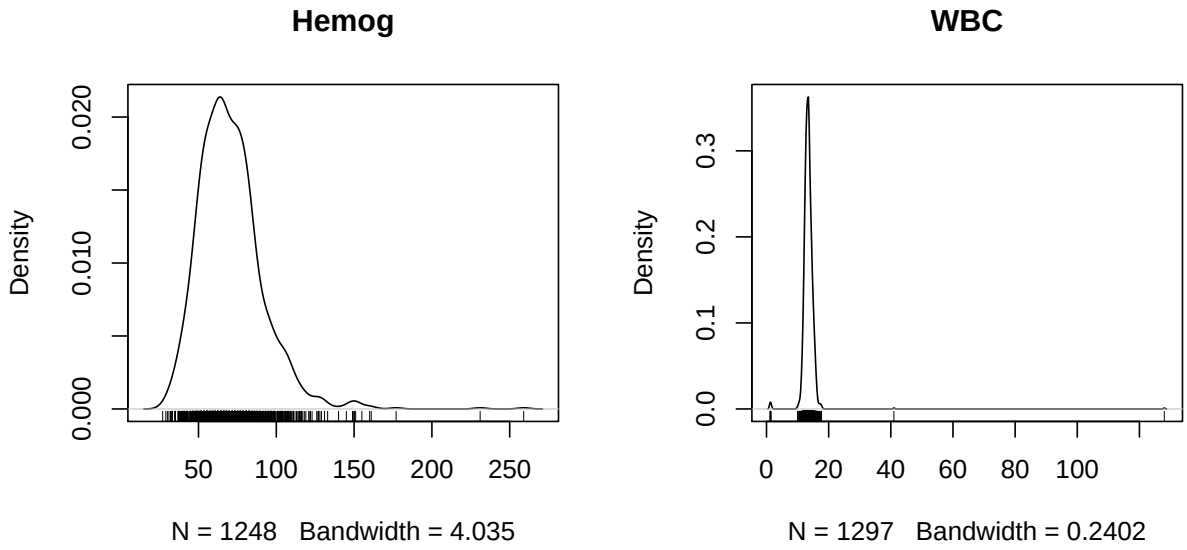
A second method uses interquartile analysis to evaluate the outliers. This method tags values that are tagged outside a specific interquartile, 25 and 75, but this can be altered if the tails of the distribution are wide, i.e. there may be meaningful values in there.

A method similar to interquartile analysis involves taking the labeling of outliers from boxplot and removing those.

Dataset before outliers removed.

The data for wbc and hemog were obtained from the CAMP study data. It is a longevity study. The purpose of graphing hemog x wbc is a graphical representation purposes of outliers only and not a study of interactions of the two.





hemog

A Grubb's test was performed on hemog data. The significance was p = 0e+00. This indicates there are outliers, and we'll move to the next step, removing outliers for hemog data.

From the graphs, hemog has a high point around 123. Mayo Clinic reports Hemoglobin is considered high when it's above 17.5 for men and 15.5 for females (Mayo Clinic). The conclusion is 123 hemog is probably a

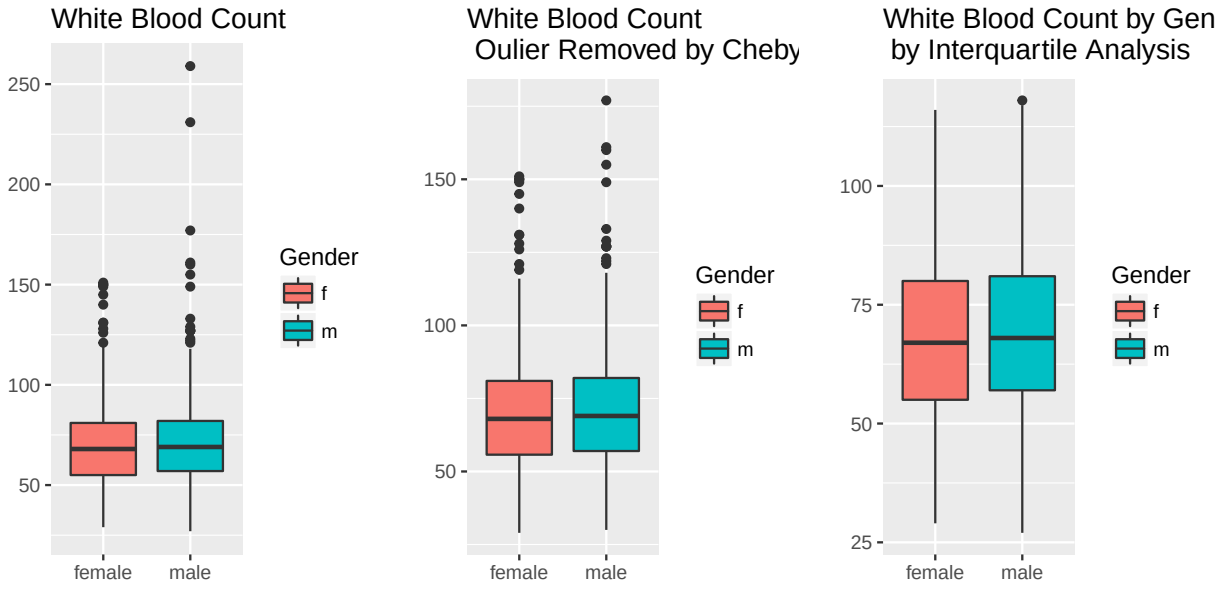
typing error. The other extreme is the low hemoglobin. The graph shows numerous values under 2. Mayo Clinic reports a low hemog value would be between 12 and 13.5 for adults. The values below 2 are most likely entry errors. I would recommend changing the values to NAs instead of fixing them.

I have used two methods to eliminate the outliers: Chebyshev's and interquartile. The Chebyshev's theorem boundaries were two standard deviation. Interquartile boundaries were 25 and 75.

The original data has two females with values above 20 and no males above 20 for hemog. From the graphs it is easy to see the interquartile analysis removed more points than the Chebyshev's. Chebyshev's largest value is 17.7, while interquartile is 16.4. Chebyshev removed two values, one around 41 and the other around 121. Interquartile also removed more, not reaching what Mayo Clinic calls "high hemoglobin's". Checking the Grubbs test again, Chebyshev removal has a outlier probability of 0.1086723 and interquartile has a significance of 1.

White Blood Count

Mayo Clinic reports the normal wbc is 45 to 110. Our data shows values in access of 100, but they are not single outliers. From the Grubbs test the p-value is 0 this suggest some of them maybe outliers. The maximum wbc is 259.



Removing outliers for wbc was trickier than hemog. There were obvious outliers for hemog. Wbc on the other had is skewed, there are more data shifted to the right, complicating outlier analysis. When Chebyshev's points were removed, the outlier probability is 6.452493e-05, interquartile has a p-value of 1e+00. Both Chebyshev's and Interquartile make an improvement over the original 0e+00, but it is hard to say exactly what to do. Best bet is two obtain a second option about removing the individual because they are truly sicker than the study intended or change just the wbc data.