Worksheet 6 Outliers

Load the Fish1.txt file from blackboard into R.

The data was recorded from an investigation into the levels of lead found in fish. Perform exploratory data analysis on the data set (summary statistics, histograms and scatter plots).

To produce all scatterplots try:

pairs(Fish1)

Using the pairs function, it is possible to include histograms of each variable along the diagonal panel.

?pairs

Identify any possible outliers graphically. We can test these points using the outliers package:

install.packages(“outliers”)

library(outliers)

Check for univariate outliers using grubbs.test

?grubbs.test

**Be careful how you interpret the results of grubbs test, the alternative hypothesis is always stated in the output. The p-value tells you whether the observation is significantly different from the rest.**

Use the rm.outlier function to replace the outlier with the mean. You will need to look up the help for rm.outlier to do this.

To check for multivariate outliers we can use the mahalanobis function which takes the form:

mahalanobis(x, center, cov, inverted = FALSE, ...)

where x is the matrix containing the data

center is a vector of data containing the variable means

cov is the covariance matrix calculated from the data

inverted = FALSE specifies whether the covariance matrix or the inverse of the covariance matrix has been entered.

Create a vector containing the variance of each variable from the Fish1 data set:

cov\_matrix <- cov(Fish1)

Create a vector containing the variable means:

center<- colMeans(Fish1)

We can now calculate the mahalanobis distance for the data set Fish1:

mahalanobis(Fish1, center, cov\_matrix)

We get the following output:

4.291035 3.678726 1.591098 1.717256 3.376488 9.190784 1.699392 1.244229 3.899651 3.228352 1.680254 10.144276 2.973726 3.284734

Each value represents the square of the mahalanobis distance (a k-dimensional z-score). We had four different variables so each value represents a 4 dimensional z-score for each observation (there are 14 observations). To check whether these distances are significant from the centroid of the data set we compare them to the critical value obtained from the chi-squared distribution with significance α and k degrees of freedom. We will compare the distances to the critical value from the chi-squared distribution with α=0.05 and k=4.

qchisq(0.95,df = 4)

Are there any multivariate outliers?

Load the Fish2.txt file from blackboard into R and repeat the process above.