Multivariate outlier detection with Mahalanobis' distance.

Dr. Jon Starkweather, Research and Statistical Support consultant.

This month's article deals with a procedure for evaluating the presence of multivariate outliers. One of the core assumptions of most parametric multivariate techniques is multivariate normality; which implies the absence of multivariate outliers. It is important to realize, cases which are multivariate outliers may not necessarily be univariate outliers. Stated another way; being an outlier on one of the variables under consideration is not a necessary condition of being a multivariate outlier. One way to check for multivariate outliers is with Mahalanobis' distance (Mahalanobis, 1927; 1936). Mahalanobis' distance can be thought of as a metric for estimating how *far* each case is from the center of all the variables' distributions (i.e. the centroid in multivariate space). Mahalanobis' distance accounts for the different scale and variance of each of the variables in a set in a probabilistic way; in other words, if one considers the probability of a case being a member of the multivariate distribution, then one must account for the density function, or standard deviation, of each variable in the multivariate set (see: Wicklin, 2012; Wikipedia, 2013). When using R there are multiple ways of calculating the Mahalanobis distance of a given data set. One way is using the chemometrics package (Filzmoser & Varmuza, 2013). The chemometrics package contains a function (Moutlier) for calculating and plotting both the Mahalanobis' distance and a robust version of the Mahalanobis' distance. The robust Mahalanobis' distance is based on the minimum covariance determinant (MCD) estimate. Below are illustrative examples for discovering multivariate outliers among two data sets; one which adheres to multivariate normality and one which contains multivariate outliers. Keep in mind, the chemometrics package has more than 10 dependent packages; therefore, as always, it is recommended that all available CRAN repository packages be downloaded directly after one installs R.

Examples

First, we create a multivariate normal data set (n = 1000) using the MASS package (Venables & Ripley, 2002). The set.seed function below simply allows us to replicate the results produced.

```
set.seed(201307)
n <- 1000
library(MASS)
Sigma <- matrix(c(1.0, .80, .50, .20,
                  .80, 1.0, .05, .05,
                  .50, .05, 1.0, .05,
                  .20, .05, .05, 1.0), ncol = 4)
x <- mvrnorm(n, Sigma, mu = c(100,100,100,100), empirical = TRUE)
detach("package:MASS"); rm(Sigma)

df.1 <- data.frame(x); rm(x)
names(df.1) <- c("y", "x1", "x2", "x3")
summary(df.1)
       y                x1               x2               x3
 Min.   : 96.37   Min.   : 96.55   Min.   : 97.22   Min.   : 96.67
 1st Qu.: 99.34   1st Qu.: 99.31   1st Qu.: 99.34   1st Qu.: 99.28
```

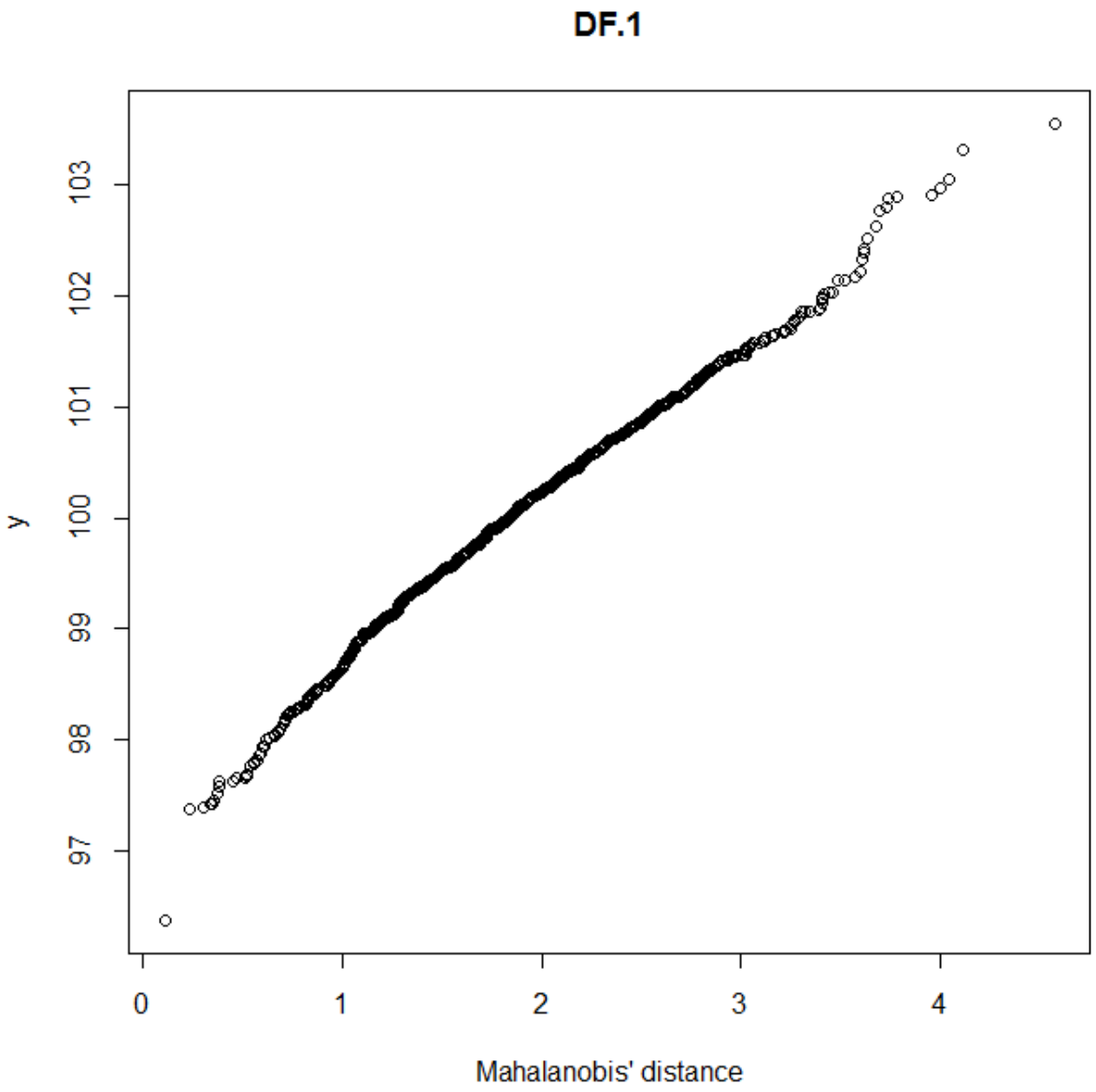
```
 Median :100.00   Median :100.00   Median :100.00   Median :100.00
 Mean   :100.00   Mean   :100.00   Mean   :100.00   Mean   :100.00
 3rd Qu.:100.65   3rd Qu.:100.61   3rd Qu.:100.67   3rd Qu.:100.73
 Max.   :103.56   Max.   :103.50   Max.   :103.31   Max.   :103.02
```

Next, we replace some values with some severe outlier values (here using cases 50, 150, 250, 500, 750, 850, 950 as the outliers) by adding the existing values to five standard deviations for each variable (i.e. column).

```
out <- c(50,150,250,500,750,850,950)
df.2 <- df.1
df.2[out,] <- df.2[out,] + c(sd(df.1[,1])*5, sd(df.1[,2])*5,
                             sd(df.1[,3])*5, sd(df.1[,4])*5)
```

Next, we calculate the Mahalanobis' distances using the Moutlier function of the chemometrics package. When using the Moutlier function, you simply supply the function with the numeric data frame (or matrix), the quantile cutoff point beyond which you want to identify points as outliers, and whether or not you want a plot. The Moutlier function returns several elements, including the Mahalanobis' distance ($md) and the robust [Mahalanobis'] distance ($rd). Below we are using the [standard] Mahalanobis' distance (md.1$md). Note; if 'plot = TRUE' the function will open a graphics window showing both the Mahalanobis' distance ($md) and the robust [Mahalanobis'] distance ($rd).

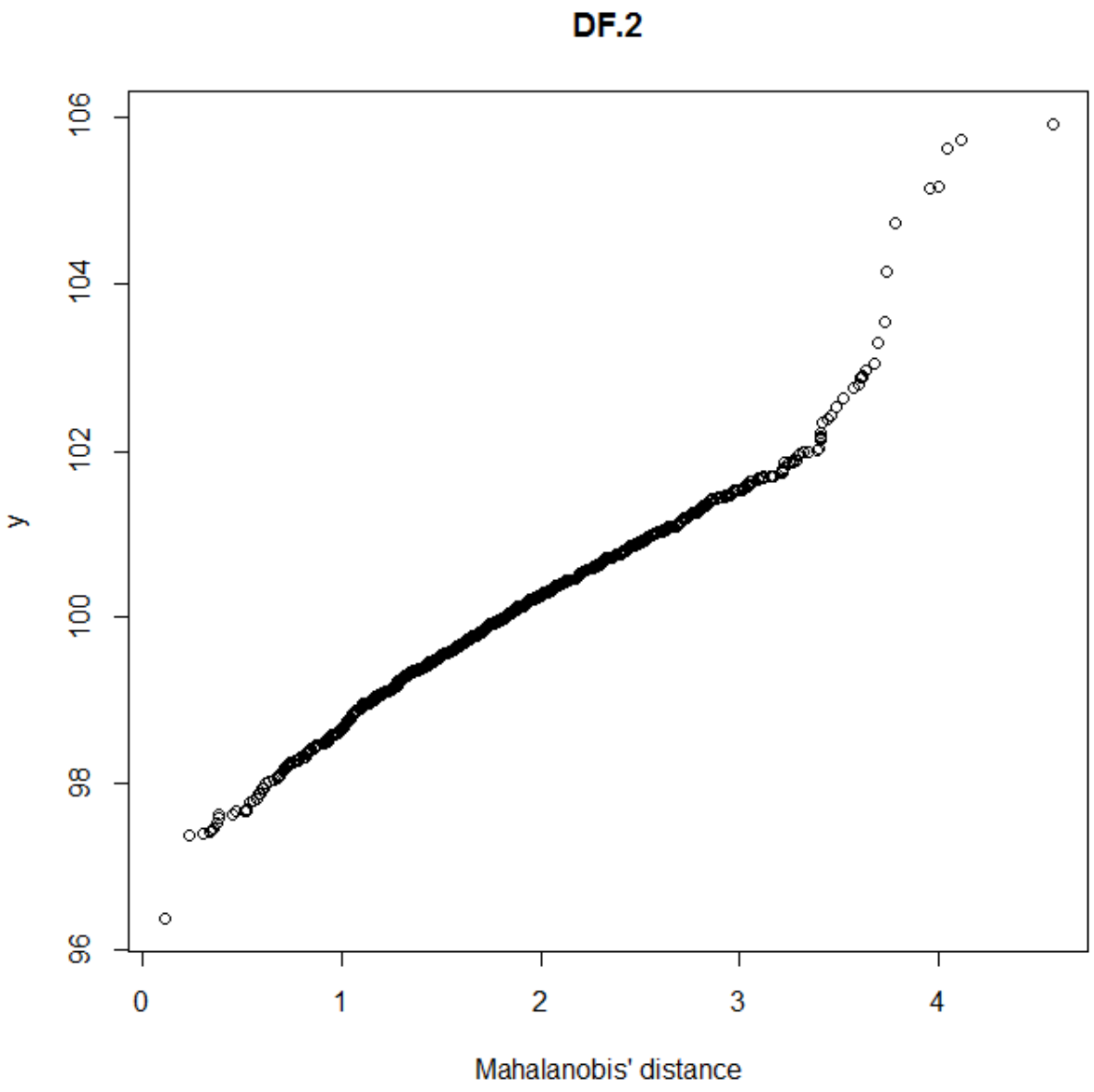
```
library(chemometrics)
md.1 <- Moutlier(df.1, quantile = 0.99, plot = FALSE)
md.1$cutoff
[1] 3.643721
summary(md.1$md)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1115  1.3670  1.8310  1.8740  2.3000  4.5740
qqplot(md.1$md, df.1$y, plot.it = TRUE, xlab = "Mahalanobis' distance",
       ylab = "y", main = "DF.1")
```



Notice, in the plot above; cases with extreme Mahalanobis distances (upper right) are likely to be true multivariate outliers.

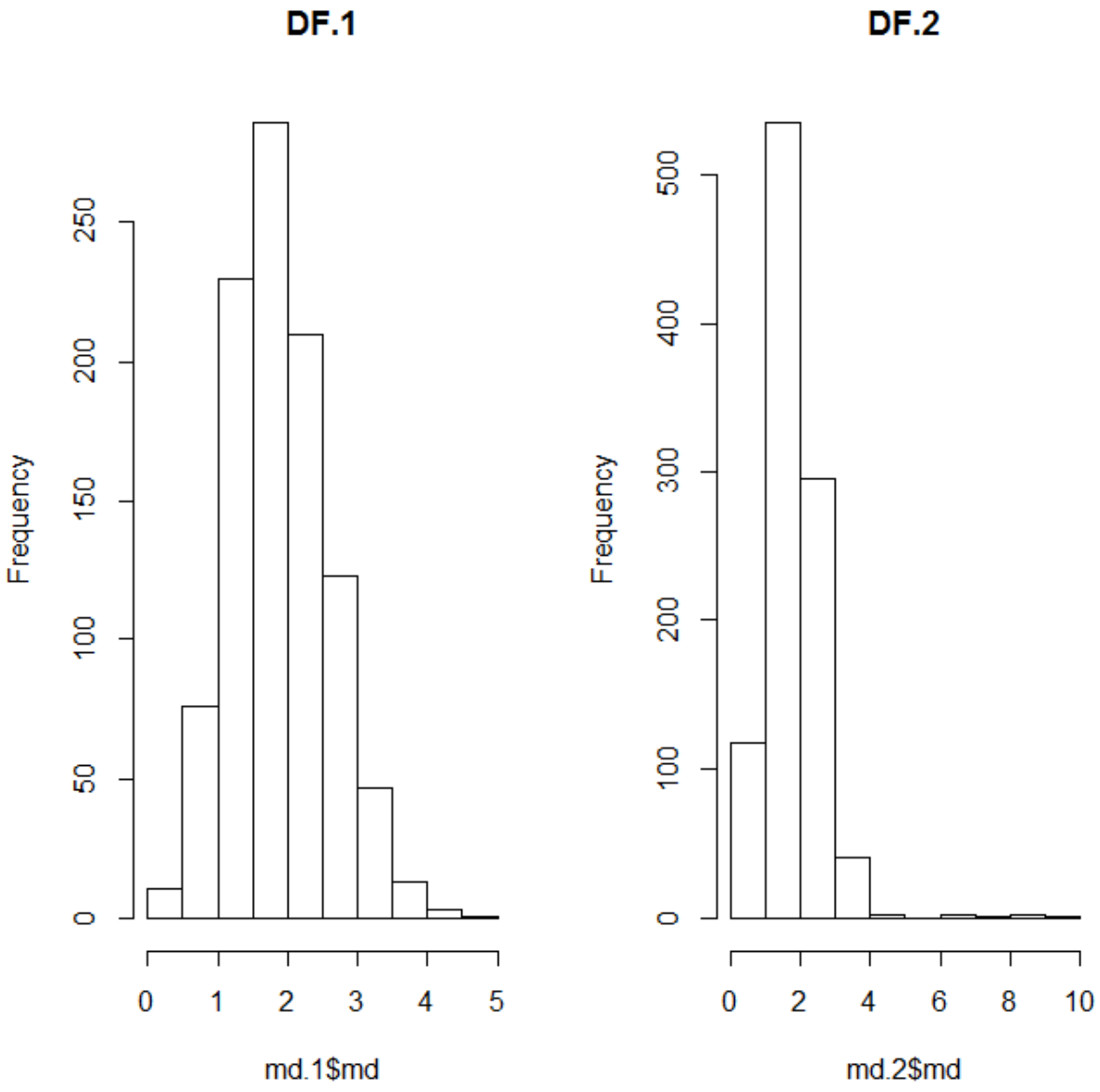
Next, we do the same as above with the second data frame – which contains severe outliers.

```
md.2 <- Moutlier(df.2, quantile = 0.99, plot = FALSE)
md.2$cutoff
[1] 3.643721
summary(md.2$md)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1368  1.2940  1.7250  1.8160  2.2050  9.1300
qqplot(md.1$md, df.2$y, plot.it = TRUE, xlab = "Mahalanobis' distance",
       ylab = "y", main = "DF.2")
```



In the plot above, cases with extreme Mahalanobis distances (upper right) are likely to be true multivariate outliers. We can also compare the Mahalanobis' distances of each data file with simple histograms.

```
par(mfrow = c(1,2))
hist(md.1$md, main = "DF.1")
hist(md.2$md, main = "DF.2")
```



However, just looking at the top 6 most extreme Mahalanobis' distance reveals the presence of outliers (regardless of 'cutoff') in the second data set.

```
head(sort(md.1$md, decreasing = TRUE))
[1] 4.573500 4.115382 4.044138 4.003354 3.954523 3.781593
head(sort(md.2$md, decreasing = TRUE))
[1] 9.130241 8.763086 8.395173 7.352192 6.955474 6.897083
```

Simply use the 'which' function to identify which cases are outliers according to the 'cutoff' values provided by the Moutlier function.

```
which(md.1$md > md.1$cutoff)
 [1]  26  59 219 300 509 584 648 668 689 944
which(md.2$md > md.2$cutoff)
```

```
[1]  50 150 250 500 509 584 668 689 750 850 944 950
```

An R script file with the same information as contained in this article is available at the Research and Statistical Support Do-It-Yourself Introduction to R course website.

Until next time, *happy computing*…

References / Resources

Filzmoser, P., & Varmuza, K. (2013). Package Chemometrics. Documentation available at: http://cran.r-project.org/web/packages/chemometrics/index.html

Mahalanobis, P. C. (1927). Analysis of race mixture in Bengal. *Journal and Proceedings of the Asiatic Society of Bengal, 23*, 301 – 333. Available at: http://www.unt.edu/rss/class/Jon/MiscDocs/1927_Mahalanobis.pdf

Mahalanobis, P. C. (1936). *On the generalised distance in statistics*. Proceedings of the National Institute of Sciences of India 2 (1): 49 – 55. Available at: http://www.unt.edu/rss/class/Jon/MiscDocs/1936_Mahalanobis.pdf

Wicklin, R. (2012). *What is Mahalanobis' distance?* The DO loop (SAS programing blog). Retrieved on 2013-07-11 from: http://blogs.sas.com/content/iml/2012/02/15/what-is-mahalanobis-distance/

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer.

Wikipedia. (2013). *Mahanobis' distance*. Retrieved on 2013-07-11 from: http://en.wikipedia.org/wiki/Mahalanobis_distance