

## Multivariate Outliers

### Univariate outliers

- For a single variable  $X$ , let the standard score for the  $i^{th}$  observation be:

$$z_i = \frac{X_i - \bar{X}}{s}$$

- Outliers are cases with extreme standard scores, for example,

$$|z_i| > 4 \text{ or more}$$

## Multivariate Outliers

- For a multivariate vector  $\mathbf{X}$ , the squared Mahalanobis distance between that observation and the sample mean vector is:

$$D_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}})' \mathbf{S}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}})$$

- Analogously, we can define a multivariate outlier as a case with a large Mahalanobis distance.

## Multivariate Outliers

- On a multivariate normal quantile plot, these cases would appear as points in the upper right that are substantially above the line for the expected  $\chi^2$  quantiles.
- However, the quantile plot itself is not resistant to the effects of outliers.
- A few discrepant observations can not only affect the mean vector, but also inflate the entries in the covariance matrix.

## Multivariate Outliers

- Thus, the effect of a few outlying observations is spread through all the  $D^2$  values, making it harder to use this tool to detect extreme values.
- A reasonable solution would be to use a multivariate trimming procedure to calculate squared distances that are not affected by potential outliers.

## Iterative Multivariate Trimming

- On each iteration, some proportion of the observations with the largest  $D^2$  values are temporarily set aside.
- From the remaining observations, compute

a trimmed mean:  $\bar{X}_{(-)}$

a trimmed covariance matrix:  $S_{(-)}$

## Robust Mahalanobis Distances

- Compute Mahalanobis distances for all observations using the trimmed mean and the trimmed covariance matrix, which are (more) robust to the effect of extreme observations:

$$D_i^2 = (\mathbf{X}_i - \bar{\mathbf{X}}_{(-)})' \mathbf{S}_{(-)}^{-1} (\mathbf{X}_i - \bar{\mathbf{X}}_{(-)})$$

## Mahalanobis Distance

With  $p$  variables,  $D^2 \sim \chi_p^2$

Therefore, a Q-Q plot of the ordered distance values

$$D_{(i)}^2$$

against the corresponding quantiles of  $\chi_p^2$  should yield a straight line.

## Detecting Multivariate Outliers

- Using the trimmed mean vector and the trimmed covariance matrix, outliers will appear as points in the upper right that are substantially above the line for the expected  $\chi^2$  quantiles.
- The SAS Macro outlier (contained in the file outlier.sas) can be used to produce a Q-Q plot to detect outliers.