REGRESSION OUTLIERS

1. Identification of Outliers

An outlier is an extreme observation. Typically points further than, say, three or four standard deviations from the mean are considered as "outliers". In regression however, the situation is somewhat more complex in the sense that some outlying points will have more influence on the regression than others. In JMPIN there is one diagnostic that can be used to identify possibly influential outliers, known as *Cook's Distance*, or simply *Cook's D*. Given a regression of Y on $(x_1,..,x_k)$ using data set $(y_j, x_{1j},..,x_{kj}), j = 1,..,n$, if

s = estimated *root mean square error*,

$\hat{y}_j$ = regression estimate of the conditional mean $E(Y_j \mid x_{1j},..,x_{kj})$,

$\hat{y}_j(i)$ = regression estimate of the conditional mean $E(Y_j \mid x_{1j},..,x_{kj})$ with the i^{th} data point $(y_i, x_{1i},..,x_{ki})$ removed,

then *Cook's Distance* for point i is given by

$$D_i = \frac{\sum_{j=1}^{n}\left(\hat{y}_j - \hat{y}_j(i)\right)^2}{(k+1)s^2}, \; i = 1,..,n$$

Intuitively, D_i is a normalized measure of the influence of point i on *all* predicted mean values, $\hat{y}_j$, $j = 1,..,n$. Cook's D can be obtained using **Fit Model** in JMPIN as follows:

(i) Right click on the heading of the **Parameter Estimates** table,

(ii) Select the **Save Columns** options, and click on **Cook's D Influence**.

(iii) A new data column will appear, contain the Cook's D Influence values.

To identify potential outliers, one *Rule of Thumb* is to treat point i as an *outlier* when:

$$D_i \geq \frac{4}{n-(k+1)}$$

As with all Rules of Thumb, this provides only a rough guideline (and often tends to identify too many points as potential outliers). The best strategy is to look at the distribution of Cook's D values and see whether there are any conspicuously large values relative to the others. If these values are roughly of the magnitude $4/(n-k-1)$ or larger, then they are worth investigating further.

2. Treatment of Outliers

The key point to stress here is that the above procedure can only serve to identify points that are *suspicious* from a statistical perspective. It does *not* mean that these points should automatically be eliminated! The removal of data points can be dangerous. While this will always improves the "fit" of your regression, it may end up destroying some of the most important information in your data.

Hence the first question that should be asked is whether there exists some *substantive information* about these points that suggests that they should be removed. Do they involve special properties or circumstances not relevant for the situation under investigation? Do they involve possible measurement errors? If no such distinguishing features can be found, then there are no clear grounds for eliminating outliers.

An alternative approach is to perform the regression both *with* and *without* these outliers, and examine their specific influence on the results. If this influence is minor, then it may not matter whether or not they are omitted. On the other hand, if their influence is substantial, then it is probably best to present the results of *both* analyses, and simply alert the reader to the fact that these points may be questionable.