

1 Outliers and Influential Cases (Soc 504)

Throughout the semester we have assumed that all observations are clustered around the mean of x and y , and that variation in y is attributable to variation in x . Occasionally, we find that some cases don't fit the general pattern we observe in the data, but rather that some cases appear 'strange' relative to the majority of cases. We call these cases 'outliers.' We can think of outliers in (at least) two dimensions: outliers on x and outliers on y .

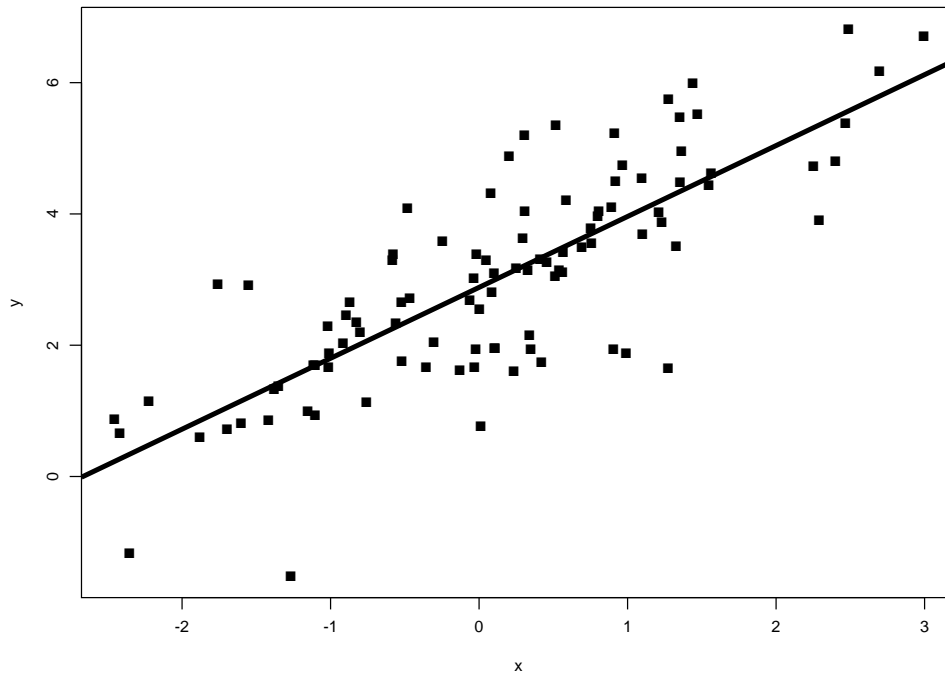


Figure 1. Y regressed on X, no outliers.

In the above plot, there are few if any outliers present, in either X or Y dimensions. The regression coefficients for this model are: $b_0 = 2.88$, $b_1 = 1.08$. In the following plot, there is one Y outlier. The regression coefficients for this model are: $b_0 = 2.93$, $b_1 = 1.08$.

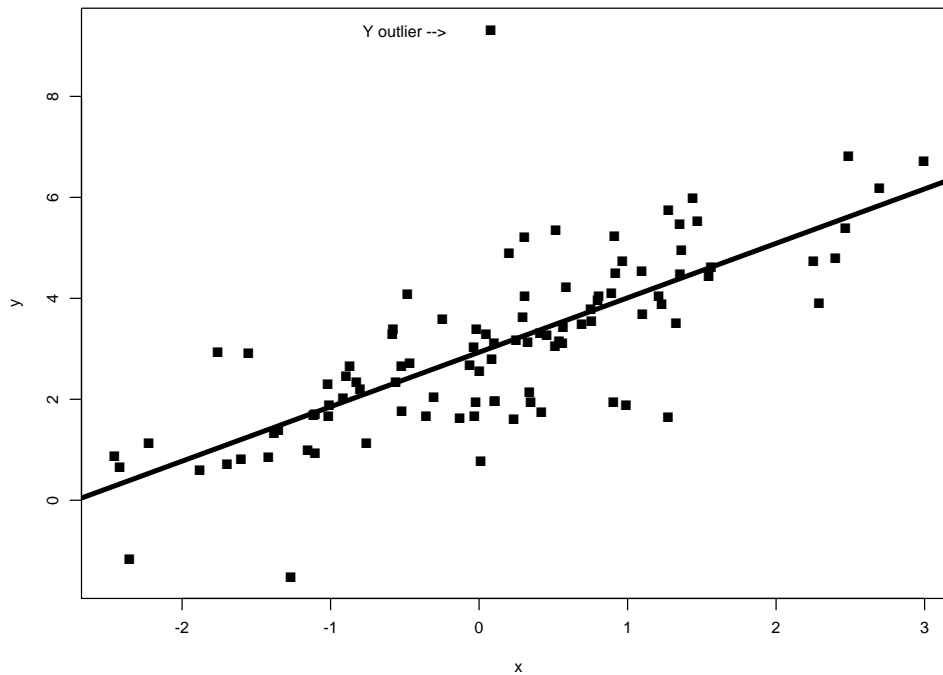


Figure 2. Y regressed on X, One Outlier on Y.

Notice that the regression line is only slightly affected by this outlier. The same would be true if we had an outlier that was only an outlier in the X dimension. However, if a variable is an outlier in both X and Y dimensions, we have a problem, as indicated by the next figure.

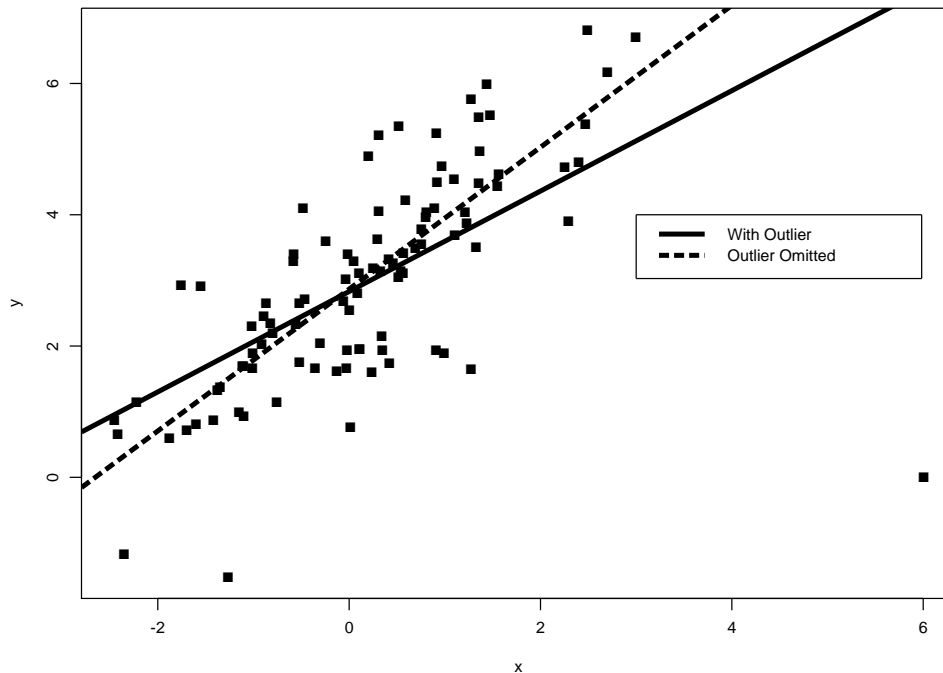


Figure 3. Y regressed on X, One Outliers on Y.

In this figure, it is apparent that the outlier at the far right edge of the figure is pulling the regression line away from where it ‘should’ be. Outliers, therefore, have the potential to cause serious problems in regression analyses.

1.1 Detecting Outliers

There are numerous ways to detect outliers in data. The simplest method is to construct plots like the ones above to examine outliers. This procedure works well in two dimensions, but may not work well in higher dimensions. Additionally, what may appear as an outlier in one dimension may in fact not be an outlier when all variables are jointly modeled in a regression.

Numerically, there are several statistics that you can compute to detect outliers, but we will concentrate on only two: studentized residuals via dummy regression, and DFBetas via regression. These statistics are produced by most software packages on request.

Studentized residuals can be obtained by constructing a dummy variable representing an observation that is suspected to be an outlier (or do all of them, one at a time) and

including the dummy variable in the regression model. If the dummy variable coefficient for a particular case is significant, it indicates that the observation is an outlier. The t test on the dummy variable coefficient is the studentized residual, which can also be obtained in other ways.

Many packages will produce studentized residuals if you ask for them. However, it is important to realize that these residuals will follow a t distribution, implying that (in a large sample) approximately 5% of them will appear ‘extreme.’ We correct for this problem by adjusting our t test critical value, using a ‘Bonferroni correction.’ The correction is to replace the usual critical t with $t_{\frac{\alpha}{2n}}$. This moves the critical t further out into the tails of the distribution, making it more difficult to obtain a ‘significant’ residual.

The above approach finds outliers, but it doesn’t tell us how influential an outlier may be. Not all outliers are influential. An approach to examining the influence of a particular outlier is to conduct the regression model both with and without the offending observation. A statistic D_{ij} can then be computed as $B_j - B_{j(-i)}$. This is simply the difference between the j -th coefficient in the model when observation i is deleted from the data. This measure can be standardized using the following formula:

$$D_{ij}^* = \frac{D_{ij}}{SE_{-i}(\beta_j)}$$

This standardization makes these statistics somewhat more comparable across variables, but we may be more interested in comparing these statistics across observations for the same coefficient/variable. In doing so, it may be helpful to simply plot all the D statistics for all observations for each variable.

1.2 Compensating for Outliers

Probably the easiest solution for dealing with outliers is to delete them. This costs information, but it may be the best solution. Often outliers are miscoded variables (either in the original data or in your own recoding schemes). Sometimes, however, they are important cases that should be investigated further. Finding clusters of outliers, for example, may lead you to discover that you have omitted something important from your model.

1.3 Cautions

Outliers can be very difficult to detect in high dimensions. You must also be very careful in using the approaches discussed above in looking for outliers. If there is a cluster of outliers, deleting only one of the observations in the cluster will not lead to a significant change in the coefficients, so the D statistic will not detect it.