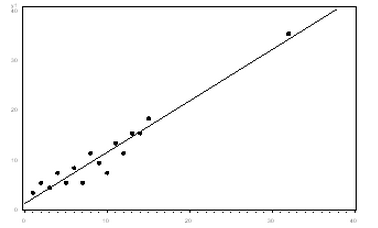
Outliers: To Drop or Not to Drop

Outliers are one of those statistical issues that everyone knows about, but most people aren’t sure how to deal with.  Most parametric statistics, like means, standard deviations, and correlations, and every statistic based on these, are highly sensitive to outliers.  And since the assumptions of common statistical procedures, like linear regression and ANOVA, are also based on these statistics, outliers can really mess up your analysis.

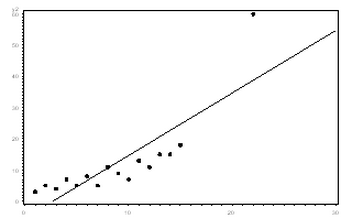
I think the best way to start is to ask whether the outliers even make sense, especially given the other variables you've collected.

Despite all this, as much as you’d like to, it is NOT acceptable to drop an observation just because it is an outlier.  They can be legitimate observations and are sometimes the most interesting ones.  It’s important to investigate the nature of the outlier before deciding.

1. If it is obvious that the outlier is due to incorrectly entered or measured data, you should drop the outlier:
2. If the outlier does not change the results but does affect assumptions, you may drop the outlier.  But note that in a footnote of your paper.

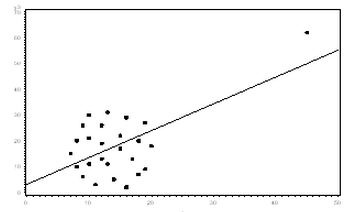


1. More commonly, the outlier affects both results and assumptions.  In this situation, it isnotlegitimate to simply drop the outlier.  You may run the analysis both with and without it, but you should state in at least a footnote the dropping of any such data points and how the results changed.



1. If the outlier creates a significant association, you should drop the outlier and should not report any significance from your analysis.

In the following graph, the relationship between X and Y is clearly created by the outlier.  Without it, there is no relationship between X and Y, so the regression coefficient does not truly describe the effect of X on Y.



**So in those cases where you shouldn’t drop the outlier, what do you do?**

One option is to try a transformation.  Square root and log transformations both pull in high numbers.  This can make assumptions work better if the outlier is a dependent variable and can reduce the impact of a single point if the outlier is an independent variable.

Another option is to try a different model.  This should be done with caution, but it may be that a non-linear model fits better.

Sometimes outliers are bad data, and should be excluded, such as typos. Sometimes they are Wayne Gretzky or Michael Jordan, and should be kept.

Outlier detection methods include:

* Univariate -> boxplot. outside of 1.5 times inter-quartile range (IQR) is an outlier.
* Bivariate -> scatterplot with confidence ellipse. outside of, say, 95% confidence ellipse is an outlier.
* Multivariate -> Mahalanobis D2 distance

Mark those observations as outliers.

Run a logistic regression (on Y=IsOutlier) to see if there are any systematic patterns.

Remove ones that you can demonstrate they are not representative of any sub-population.

**Statistical Distance Measures**

There are two statistical distance measures that are specifically catered to detecting outliers and then considering whether such outliers should be removed from your linear regression.

* The first one is Cook's distance. You can find a pretty good explanation of it at Wikipedia: <http://en.wikipedia.org/wiki/Cook%27s_distance>

The higher the Cook's distance is the more influential (impact on regression coefficient) the observation is. The typical cut-off point to consider removing the observation is a Cook's distance = 4/n (n is sample size).

* The second one is DFFITS which is also well covered by Wikipedia: <http://en.wikipedia.org/wiki/DFFITS>.

The typical cut-off point to consider removing an observation is a DFFITS value of 2 times sqrt(k/n) where k is number of variables and n is the sample size.

Both measures usually give you similar results leading to similar observation selection.