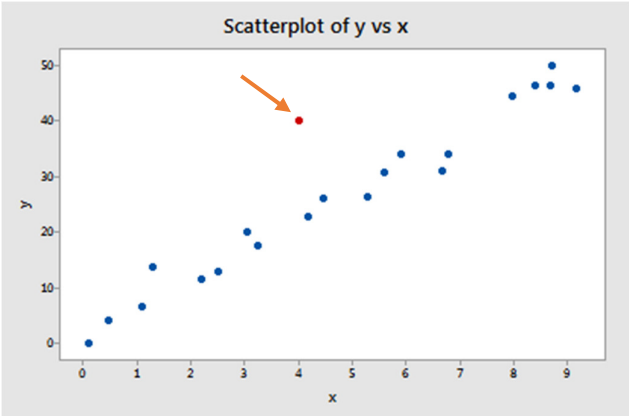
How to handle outliers in regression

We know that an outlier data, by its nature, is very different from all the others collected under a study. When we have a set of replicated measurements, we can apply easily the Grubb's, Dixon's or any other outlier statistics to confirm if any one of the extreme values is a suspect value or possibly an outlier which is to be omitted before conducting further statistical analyses.

But, it is harder to deal with them in regression statistics, referring to say, the construction of a calibration curve. The outlying points may not be necessarily at the extremely low or high positions.

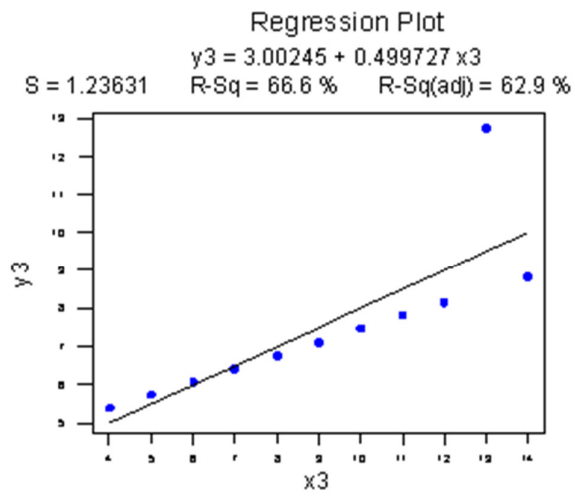
How are we then going to deal with the red spot as shown on Figure 1 below?



Let's first recall the meaning of **residuals**.

A regression residual is represented by equation $(y_i - \hat{y}_i)$ where y_i is observed or experimental value and $\hat{y}_i$, predicted or calculated value from the regression equation $y = a + bx$. It represents errors on the model.

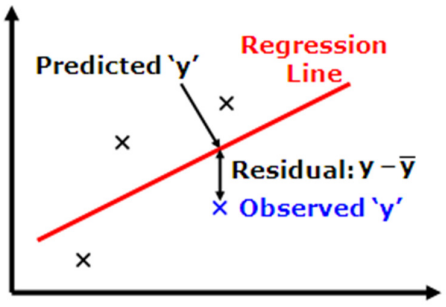
The least- squares method minimizes the sum of the squares of the y-residuals, so a suspect point with a large y-residual can have a significant effect on the calculated slope b and intercept a of the linear regression line, and thus on the analytical information derived from the latter. The illustration in Figure 2 highlights this point.



In cases where an obvious error such as a transcription mistake or an instrument malfunction has occurred, it is of course permissible to reject the resulting measurement (and, if possible, to repeat it). However, if there are suspect measurements for which there are no obvious sources of error or explanation, three distinct approaches are available:

1. The use of a significance test or similar method to decide whether a measurement should be accepted or rejected;
2. The use of median-based methods, in which suspect or outlying values are discounted; and
3. The use of robust methods, in which such values may be included in our calculations, but given less weight, i.e. importance, in plotting the regression line.

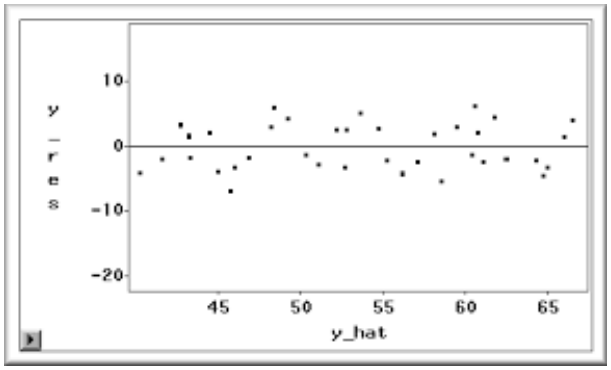
Simple outlier statistics cannot be directly applied to the points forming regression lines. This is because, although the individual y_i-values in a calibration experiment are assumed to be independent of one another, the residuals are not independent of one another, as the sum of residuals is always zero. See Figure 3 below.



A large residual indicates possible error and the presence of an outlier. However, it is not permissible to treat the residuals as if they were a conventional set of replicate measurements, and apply a familiar test such as the Grubbs' test to identify any outliers.

Of course, if we were to have a large number of y_i-values which is not generally met in our routine analytical work, the above mentioned prohibition can be relaxed.

Most computer programs handling regression data provide residual diagnostics routines. Some of these are quite simple, including plots of the individual residuals against y_i-values. See Figure 4 below which shows a satisfactory distribution of residuals.



When one of the residuals on the scatter plot shows obviously large as compared with the others, that particular y-value might be an outlier.

A very simple statistical approach is to compare the regression models with and without the suspected value. If the suspected data does not exert a large influence over the model, then we would expect the adjusted predicted y_i-values to be very similar to the original predicted y_i-values when the suspected data is included. We can say that the model in question is 'stable' regardless whether the suspected value is included or not.

A more advanced method is the estimation for each point of Cook's squared distance, SD^2 (sometimes abbreviated to 'Cook's distance'), first proposed in 1977. This is an example of an influence function, i.e. it measures the effect that rejecting the calibration point in question would have on the regression coefficients.

For a straight line graph, it can be calculated from:

$$CD^2 = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_j^{(i)})^2}{2s^2_{y/x}}$$

where

$\hat{y}_j$ is a predicted y-value obtained when ALL the data points are used

$\hat{y}_j^{(i)}$ is the corresponding predicted y-value obtained when the i^{th} point is omitted

$s^2_{y/x}$ is the standard error calculated using ALL the data points.

When values of CD^2 is greater than 1, we can justify to omit the suspected point from the calibration regression.

In practice, the Cook's square distance method turns out to be better at identifying some types of outlier than others: outliers in the middle of a data set are less readily detected than those at the extremes. Other popular and effective methods in handling outliers in regression include the alternative non-parametric and robust methods.