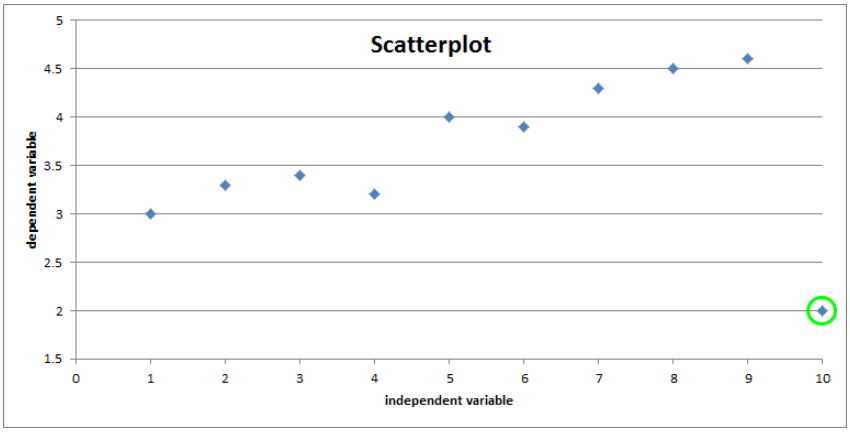
A caution in applying the Method of Least Squares in regression analysis — Outliers

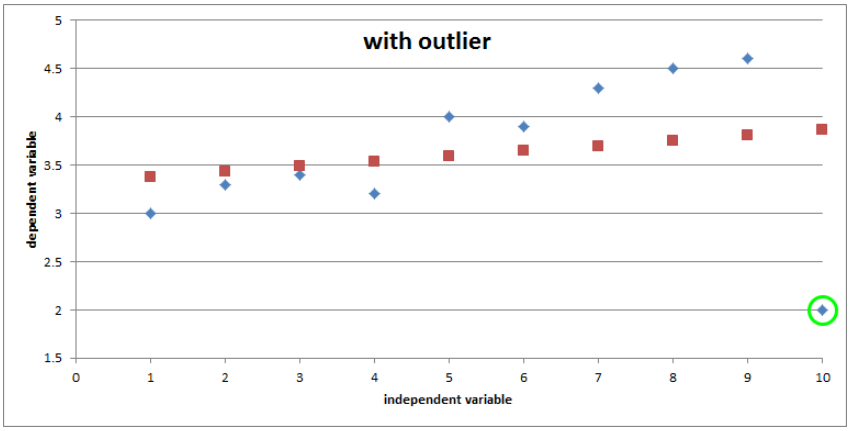
Applying the Method of Least Squares typically involves a mix of mathematics and good judgment. Specifically data may contain ***outliers*** – data points that could be anomalous due to mismeasurement or other factors. Though you would never want to discard data simply due to inconvenience (especially if the data may indicate something significant), it is often necessary to discard one or several outliers in order to produce more appropriate predictions.

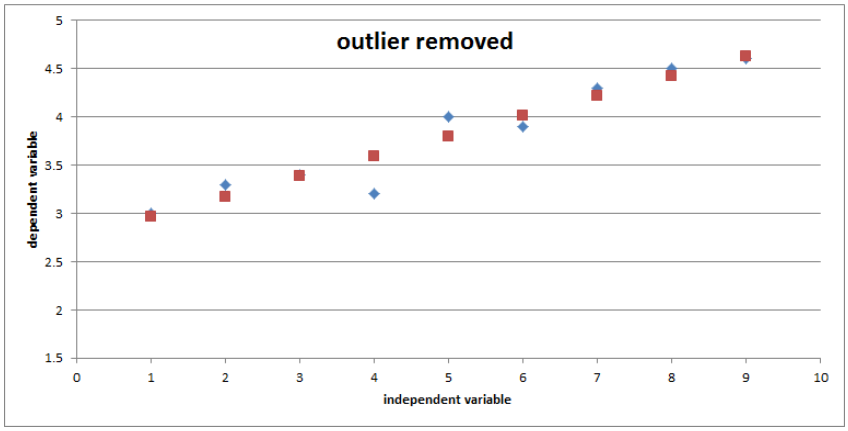
For example, suppose we have data in the form of ordered pairs $\{(x_i, y_i);\ 1 \le i \le n\}$ and a scatterplot suggests a straight line relationship between the independent (explanatory) variable x and the dependent variable y.

x	*y*
1	3
2	3.3
3	3.4
4	3.2
5	4
6	3.9
7	4.3
8	4.5
9	4.6
10	2



The scatterplot suggests that the circled point (10, 2) is an outlier. You may have an intuitive sense that if you tried to approximate the given data with a straight line, minimizing the sum of the squares of vertical distances between the given data and a theoretical line might be greatly affected by this one outlier. This becomes quite evident when you go through the calculations. Shown at right are two graphs – the first one containing the outlier and the second one omitting the outlier. The raw data is shown in blue and the corresponding points and the regression line are shown in red.





The square of the ***correlation coefficient*** (r^2) is generally accepted as a measure of how well the regression line matches the given data – with a value of 1 indicating a perfect fit. In this example, we have $r^2 = .04$ for the data including the outlier, and $r^2 = .92$ with the outlier removed.