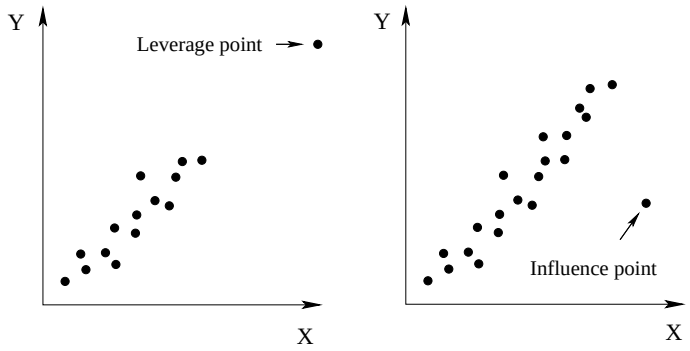
Importance of Detecting Influential Observations

DEFINITION: A LEVERAGE point is an observation, that has an unusual predictor value (very different from the bulk of the observations). An INFLUENCE POINT is an observation whose removal from the data set would cause a large change in the estimated reggression model coefficients.



A leverage point may have no influence if the observation lies close to the regression line. A point has to have at least some leverage in order to be influential. If a data set contains one or more influential points, then the parameters of the regression equation may be determined in large part by those influential values and less by the majority of the data. This is why identifying influential and leverage points is important when fitting a model. Sometimes a fitted model is not conform with theory that we know about the problem. A slope may have the wrong sign, predictors that are known to be important are not significant etc. These problems can be caused by one or more influential observations.

Leverage

The location of a point in $\mathbf{x}$-space can be important in determining whether the point is a leverage point. Remote points have more impact on the model parameters. The hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

is helpful in identifying leverage points. We have seen previously that $\mathbf{H}$ determines the covariance matrices of $\hat{\mathbf{y}}$ and $\mathbf{e}$

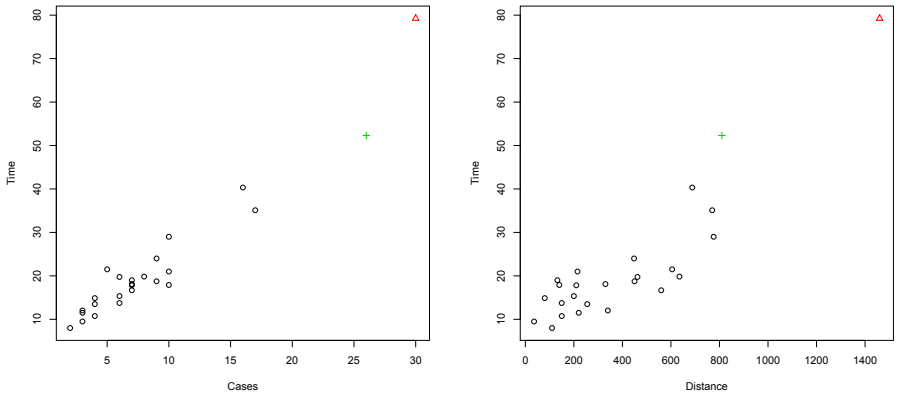
$$Var(\hat{\mathbf{y}}) = \sigma^2\mathbf{H}, \qquad Var(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$$

The elements h_{ij} of the hat matrix may be interpreted as the leverage that the i^{th} observation y_i exerts on the j^{th} fitted value $\hat{y}_j$. The diagonal entries h_{ii} of $\mathbf{H}$ can be seen as a measure for how far the i^{th} observation lies from the center of the $\mathbf{x}$-space. Thus, large diagonal values of the hat-matrix mean that the points they correspond to are potential influential points. It can be shown that the average of the diagonal elements of the hat matrix is $(k+1)/n$. Thus, a rule of thumb is to consider any point for which h_{ii} exceeds $2(k+1)/n$ a leverage point.

Not all leverage points are influential points. To check, consider the residuals (for instance the PRESS residuals) of the points. Observations with large **H**-matrix diagonal entries *and* large residuals are good candidates for influential points. To confirm whether a point is influential, fit the model with and without the point and decide whether there is a large change in relevant model parameters.

Example: Delivery Time Data

Recall, that for the Delivery Time data, we had $k = 2$ and $n = 25$. Thus $2(k+1)/n = 0.24$ can be used as our cutoff for leverage points. Computing the diagonal entries of the hat-matrix **H** in R shows that the observations 9 and 22 have h_{ii} entries that are larger than this cutoff. You can create scatter plots of the response against each individual predictor that show the position of these points. How much influence do these points have on the model?



To answer that question, refit the model with and without the points (9 in red, 22 in green) and compare the model parameters:

Model	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	R^2
9 and 22 in	2.34	1.62	0.014	0.9596
9 out	4.45	1.50	0.010	0.9487
22 out	1.92	1.79	0.012	0.9564
9 and 22 out	4.64	1.46	0.011	0.9072

Removing observation 9 has a larger influence percentage-wise on the slope for DISTANCE than on the slope for CASES. Deleting only observation 22 has only minor effects on the model parameters. Deleting both observations has a similar effect to just deleting observation 9.

Measures of Influence: Cook's D

To decide whether a point is influential, both its location in $\mathbf{x}$-space and its response value y have to be considered. In the late 1970's Dennis Cook suggested a measure for the influence each point exerts in linear regression. This measure, called the COOK'S DISTANCE of observation i is computed by comparing the parameter estimates obtained when using all points $\hat{\boldsymbol{\beta}}$ and the parameter estimates obtained when deleting the i^{th} observation y_i: $\hat{\boldsymbol{\beta}}_{(i)}$.

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{(k+1)\,MS_{Res}} = \frac{(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})'(\hat{\mathbf{y}}_{(i)} - \hat{\mathbf{y}})}{(k+1)\,MS_{Res}}$$

Recall, that $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$. Points with large Cook's distance D values have a large influence on the least squares parameter estimate vector $\hat{\boldsymbol{\beta}}$. How large is large? D_i does not have an easy to understand named distribution. Usually, the threshold used for Cook's distance is 1. That is points with $D_i > 1$ are considered to be influential. Points with $D_i > 0.5$ could be considered "mildly influential".

Another way to look at Cook's distance is to rewrite D_i as

$$D_i = \frac{r_i^2}{k+1}\frac{Var(\hat{y}_i)}{Var(e_i)} = \frac{r_i^2}{(k+1)}\frac{h_{ii}}{(1-h_{ii})}$$

"measure of unusualness"

"measure of leverage"

where r_i are the studentized residuals. One can show that $h_{ii}/(1-h_{ii})$ is a measure of the distance of the vector $\mathbf{x}_i$ from the centroid of the remaining data. Thus, D is made up of two components: The residual component describes how unusual y_i is in the current model and the other component describes how unusual $\mathbf{x}_i$ is. Cook's distance combines measures of large residuals and location of the data point in $\mathbf{x}$-space.

Recall, that if you ask R for `plot(fit)` of some fitted linear model, you get four different plots. The fourth of these plots shows residuals (on the y-axis) against leverage (h_{ii} on the x-axis). Cook's distance values of 1 and 0.5 are shown as dashed grey lines in that plot.

Example: Delivery Time Data

Values of Cook's distances are most easily computed by hand using the function that involves the studentized residuals r_i and hat matrix diagonal entries h_{ii}. In R, the command `cooks.distance()` obtains the same information for a linear model. The two observations with the largest Cook's distance measures are again observations 9 and 22. The Cook's distance measure for observation 9 is 3.419, for example. The value of Cook's distance for observation 22 is 0.451 (< 0.5). Thus, we can declare observation 9 for influential ($D_9 > 1$) while we would not declare observation 22 for influential based on Cook's distance (22 is borderline moderately influential).

Groups of Influential Observations

The discussion of which points constitute leverage or influential points can be extended to groups of two or more points that influence the model in a similar way. Cook's distance can be extended to assess the simultaneous influence of a group of m points - simply leave out all m points simultaneously and re-calculate the regression parameters. There can be situations in which several data points are jointly influential, while individual points are not. Other author's suggest the use of cluster analysis to find groups of similar observations in a multivariate problem.

Treatment of Influential Observations

You have now seen several methods with which to identify influential and leverage points in multiple linear regression. But what do you do with those points once they've been identified? Should they be excluded from the analysis? This decision is similar to our previous discussion of outliers. Whether or not single observations should be excluded depends on

- How much data we have total (and whether we can afford to loose any)
- Whether we have an explanation (based on the experiment the data has been generated with) why single observations are very different
- How costly it would be to generate more data (for the same $\mathbf{x}$-levels)
- Does the fitted model become much easier to explain if the influential point(s) are removed?

A compromise between retaining and deleting influential observations is to use a weighted estimation technique that gives relatively less weight to the influential observations. These so-called ROBUST regression techniques down-weight observations proportional to their residuals or influence measures.

In practice, if severely influential points are identified in a regression problem and if it is not practical to repeat the experiment for these predictor values, then more than one model is usually fitted (with and without each severely influential point) and all the resulting models are presented as alternatives to the experimenters. It is the statisticians job to bring the existence of influential points to the experimenter's notice. But it is generally not the statistician's job to make the final decision of whether or not these points should be removed.

If the answers to important questions that should be addressed using the model would change based on whether or not outliers are included or excluded, then the statistican should describe how these answers would change in their analysis and name the influential values (by row identifying names wherever possible).