Leverage

If the data set contains outliers, these can affect the least-squares fit. The $i$'th fitted value $\hat{y}_i$ can be written as a linear combination of the observations,

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j ,$$

where

$$h_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum_{j=1}^{n} (x_j - \bar{x})^2} .$$

The $n \times n$ matrix $H$ with $h_{ij}$ as its $(i,j)$'th entry is called the **hat matrix**. The $i$'th diagonal element of $H$, 

\[
    h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{n \sum_{j=1}^{n} (x_j - \bar{x})^2}
\]

is called the **leverage** of the data point \((x_i, y_i)\), and measures the impact that \(y_i\) has on \(\hat{y}_i\). Note that the further \(x_i\) is from \(\bar{x}\), the larger \(h_{ii}\), and therefore the more sensitive \(\hat{y}_i\) is to changes in \(y_i\). So points with very large and very small \(x\) values have more leverage than points with intermediate \(x\) values.

If for some reason a point with high leverage also happens to be far from the least squares line which would be fitted to the remaining data points (i.e., if the point is an outlier), then we may need to take some action, e.g., delete the point, reconsider
whether the model is reasonable, see if there was a recording error, etc. It can be shown that the $h_{ii}$ are all between 0 and 1, and that $\sum h_{ii} = 2$. In practice $h_{ii}$ is considered large if it exceeds $4/n$.

**Influence Diagnostics**

An observation is **influential** if the estimates change substantially when the point is omitted.

- Leverage depends only on the $x$’s, not on the $y$’s.
- A point with high leverage may or may not be influential.
- A point with low leverage may or may not be influential.
- Looking at residuals may not reveal influential
points, since an outlier, particularly if it occurs at a point of high leverage, will tend to drag the fitted line along with it and therefore it may have a small residual. This phenomenon is called **masking**.

A more direct measure of the influence of \((x_i, y_i)\) is given by **Cook’s D statistic**, 

\[
D_i = \sum_{j=1}^{n} \frac{(\hat{y}_{(i)j} - \hat{y}_j)^2}{(2s^2)} , \quad i = 1, \ldots, n ,
\]

where \(\hat{y}_{(i)j}\) is the prediction for \(y\) at \(x = x_j\) without the point \((x_i, y_i)\). Jobson recommends that observations with \(D_i > 1\) should be examined carefully.

**Eg:** For the baseball example, we first examine the Minitab "Fitted Line Plot".
This gives a scatterplot, together with the fitted line, and (an option for) 95% confidence and prediction intervals. Note that the confidence intervals are wider at the ends. The point for Minnesota (Case 9) has a leverage of .1945, which does not exceed $4/n = .29$, and therefore would not be considered extremely high. It has a Cook’s D of .65, which does not exceed 1, and so would not be considered an outlier by this criterion. But the unusualness of Minnesota is partially masked by Cleveland, Milwaukee and Toronto. If we leave out all four teams, the results change drastically. In general, Cook’s D can be "fooled" by multiple outliers.