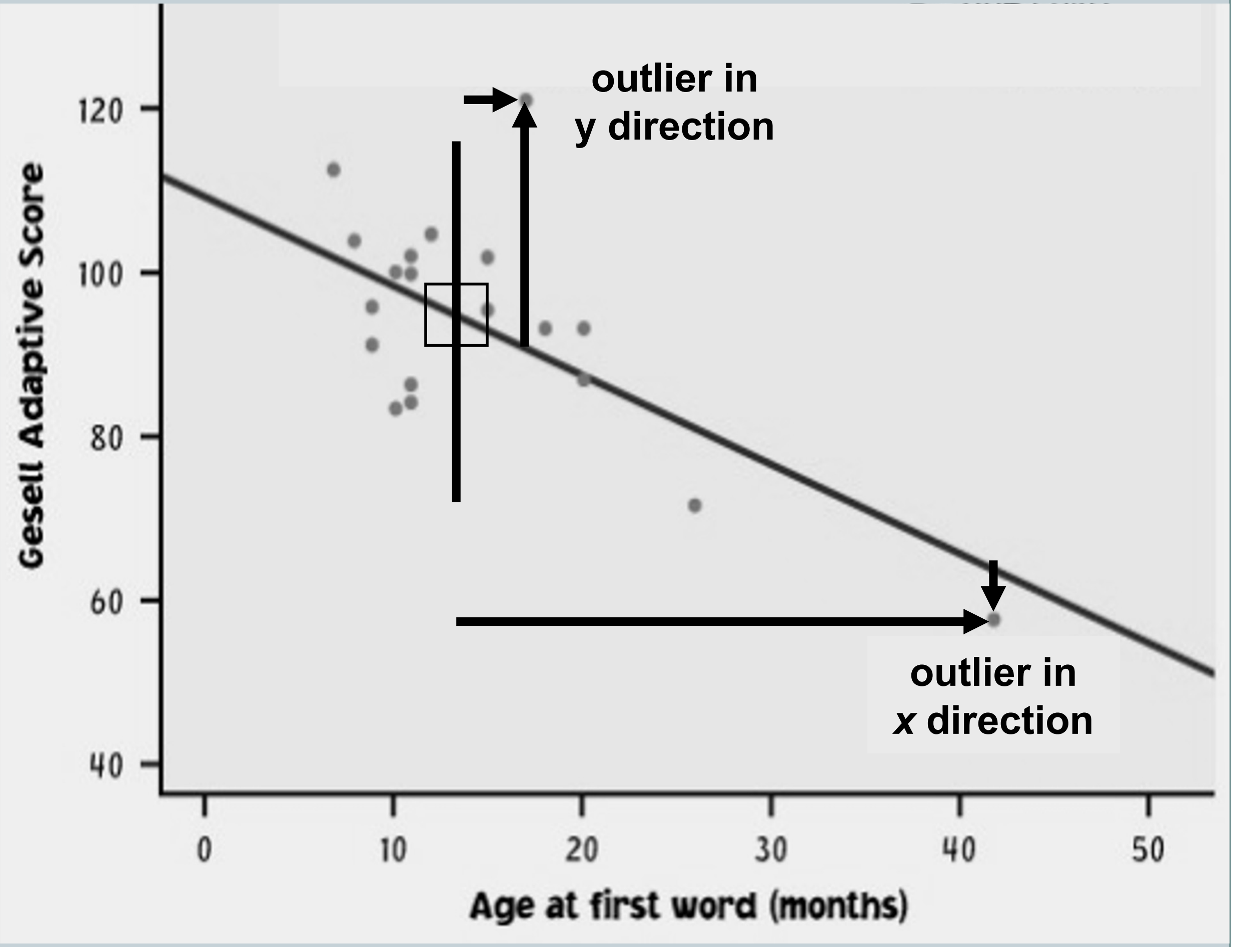
Regression Diagnostics

Outliers & Influential Points

Outlier:
An observation that lies outside the overall pattern of observations.

"Influential observation":
An observation that markedly changes the regression if removed. This is often an *outlier* on the x-axis.



Outliers

Broadly speaking, an ***outlier*** is a data point that is distinct or deviant from the bulk of the data.

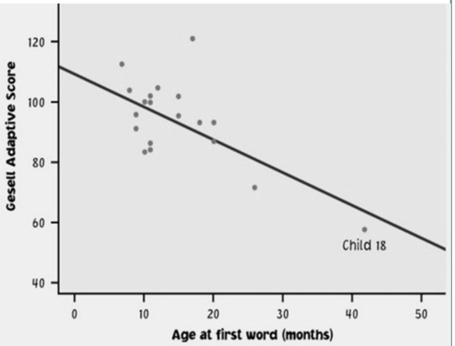
Their relatively low probability of natural occurrence indicates that they are most likely due to error.

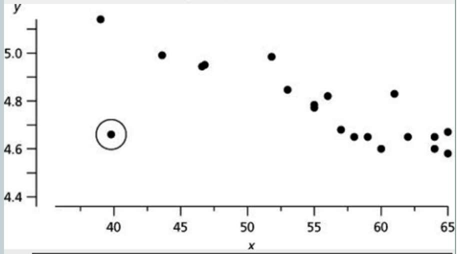
- Measurement, recording, administration of treatment, etc.

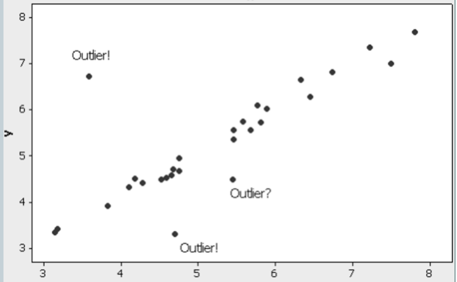
Of course, outliers can also occur in the absence of error. These are known as ***true outliers*** in contrast to the ***false outliers*** described above.

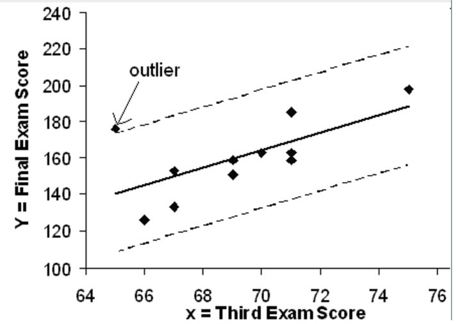
The distinction is important because *true outliers* have something to teach us.

- Correct model?
- Violation of assumptions?
- Are there observations with an undue influence on the results?

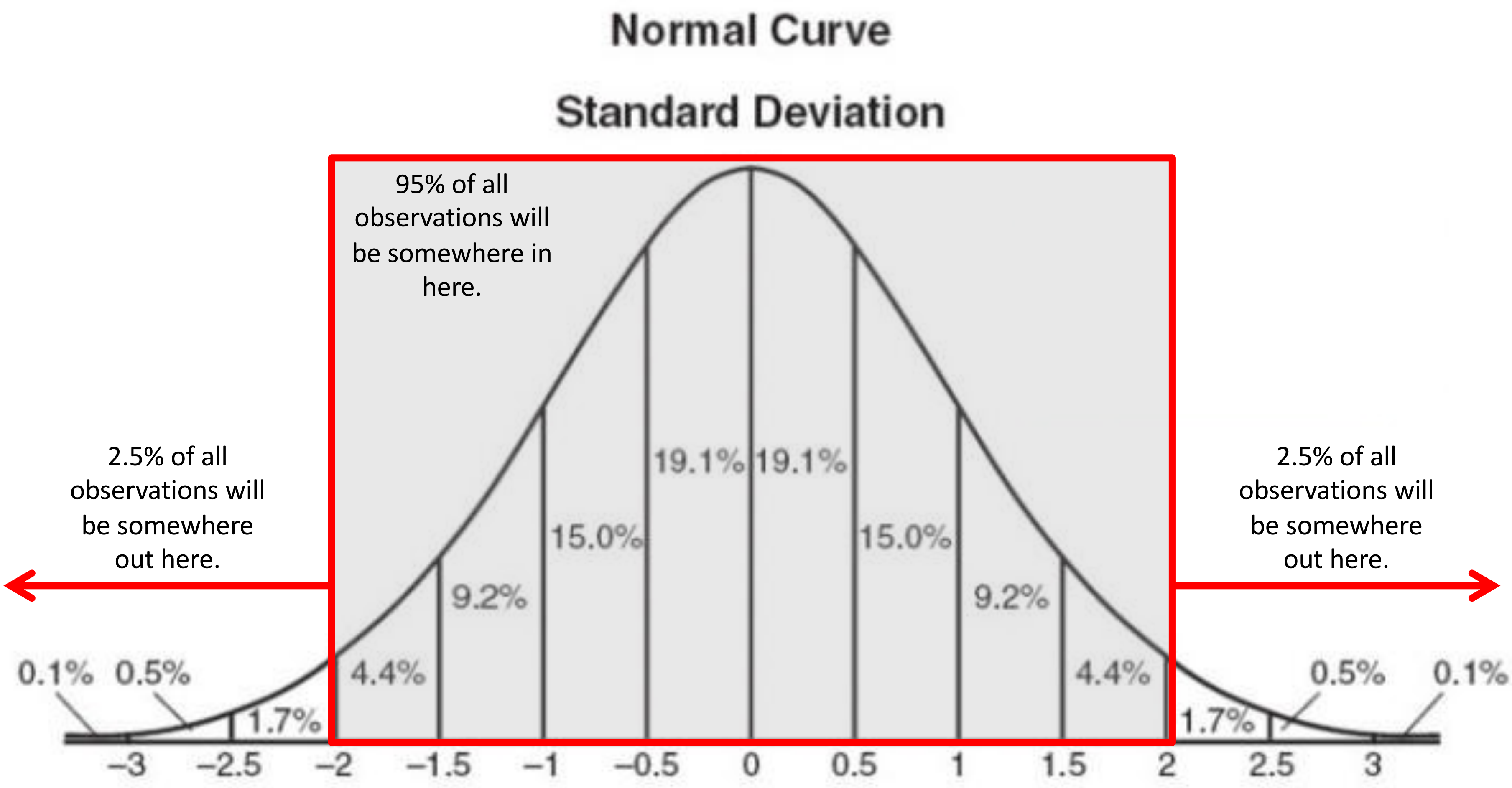








Univariate Outliers

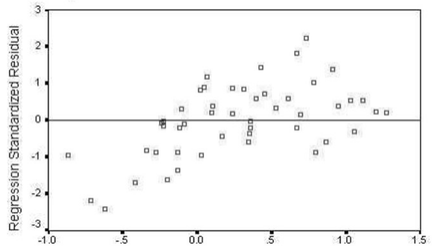


95% of all observations will be somewhere in here.

2.5% of all observations will be somewhere out here.

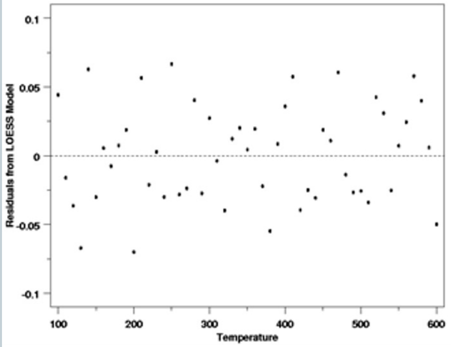
2.5% of all observations will be somewhere out here.

Detection of Outliers



Standardized Residuals (ZRESID):

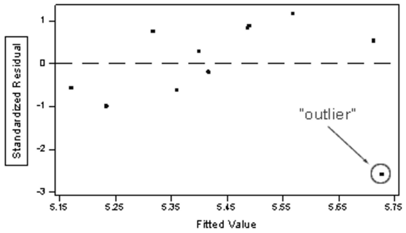
We discussed the *residual plot* in the previous lecture. Although it is not necessary to standardize the residuals to visualize the data and evaluate whether or not a linear fit is suitable, it is useful.



Standardizing residuals can be accomplished by using a *z* transformation.

$$z = \frac{X - \bar{X}}{s_{y|x}}$$

Residuals Versus the Fitted Values
(response is Alcohol)



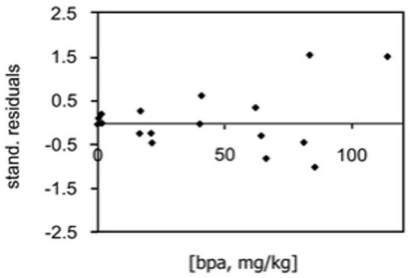
Where:

z is the standard score.

X is the raw score of a given individual (i.e., their residual).

$\bar{X}$ is the mean, and specifically 0 for residuals.

$s_{y|x}$ is the standard error of the estimate: $s_{y|x} = \sqrt{\frac{\sum(Y-\hat{Y})^2}{N-k-1}} = \sqrt{\frac{SS_{res}}{N-k-1}}$



Using this method, one can easily establish a rule by which to identify outlying points (e.g., *z* > |2.0|).

Detection of Outliers

Studentized Residuals (SRESID):

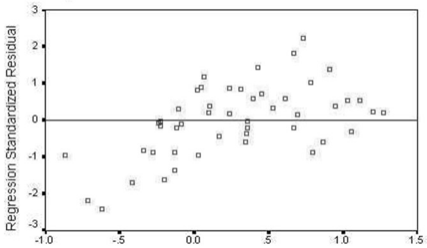
Z-score transformations make the assumption that every residual has the same amount of variance as every other residual (*homoskedasticity*).

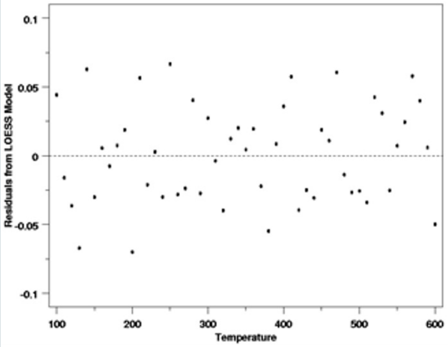
This is not always true (e.g., *heteroskedasticity*), and so an alternative method for evaluating residuals if this is a concern is presented below.

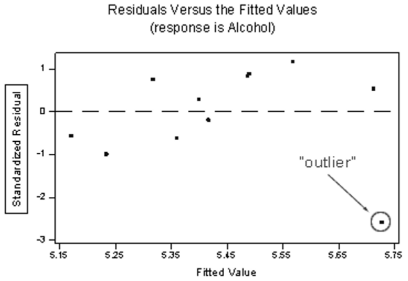
Rather than dividing each residual by the standard error of the estimate, one may divide by the ***estimated standard deviation***:

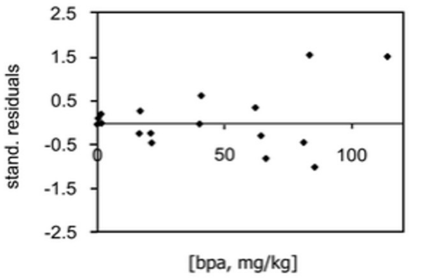
$$s_{e_i} = s_{y|x}\sqrt{1-\left[\frac{1}{N}+\frac{(X_i-\bar{X})^2}{\sum x^2}\right]}$$

When the assumptions of the linear model are reasonably met, then the ***studentized residuals*** will follow a *t* distribution.









Detection of Outliers

<u>When an Outlier is detected:</u>

Can the outlier be traced to a reparable cause (e.g., incorrect entry into dataset)?

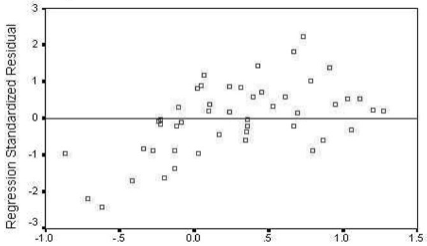
- If yes, then simply fix the data point and repeat your analysis to verify.

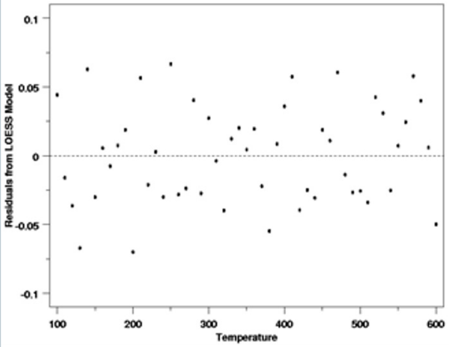
What if an outlier cannot be "repaired?" Can nothing can be done to rescue the data?

- In this case, you may opt to remove this subject from further analysis.

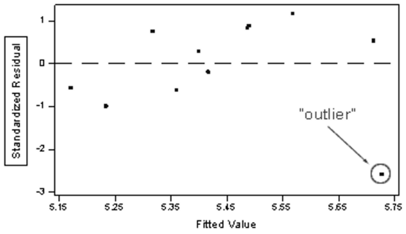
The univariate statistics have changed, so do you report the original mean, standard deviation, etc. or the new ones?

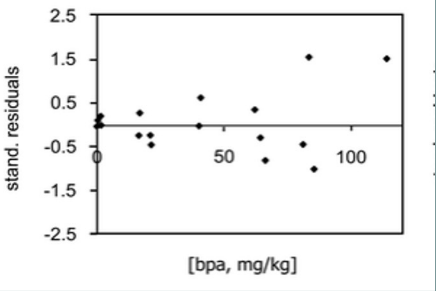
- You report the new statistics as the statistics that reflect your data.
- The original statistics may be reported in the section wherein you describe the rule for detecting the outliers.





Residuals Versus the Fitted Values
(response is Alcohol)





Influence Analysis

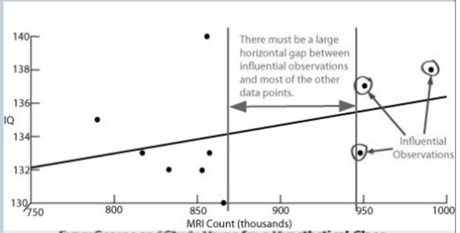
Broadly speaking, an ***Influential Observation*** may be defined as:

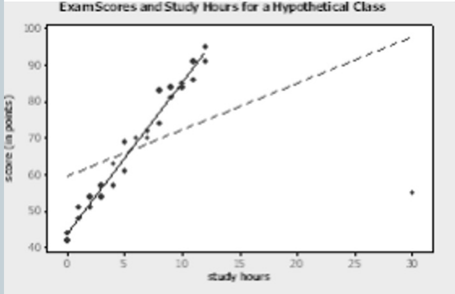
> *“one which, either individually or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates than is the case for most of the other observations.”*

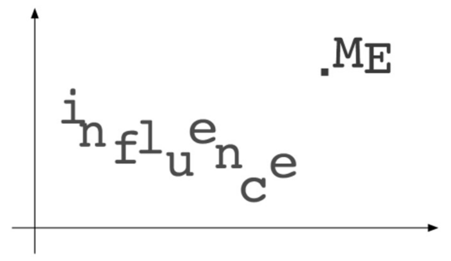
As mentioned earlier, an outlier is not necessarily an influential point.

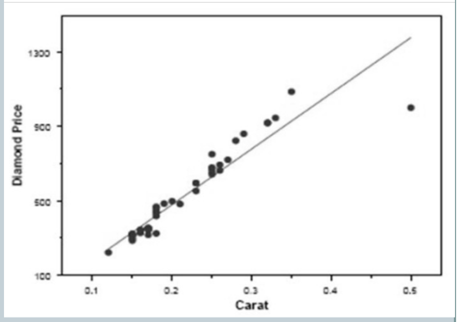
- An outlier, although relatively deviant, may have little impact on the resulting regression line.

- Influential points may not be appreciably deviant, as based on an analysis of the residuals, but which may greatly affect the terms in the regression equation (which is why it is seldom detected by an analysis of the residuals).









Influence Analysis

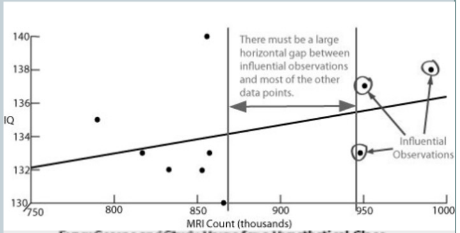
Leverage:

One may liken the effect of an influential observation to that of a lever, wherein greater pulling power in a certain direction is achieved.

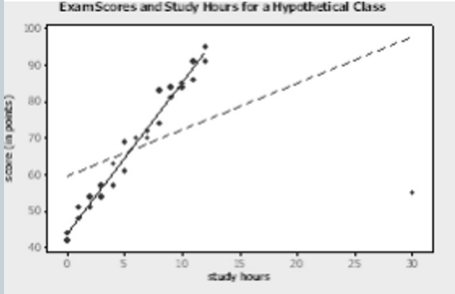
Therefore, by measuring the leveraging power of each point and establishing a rule, one can detect influential observations.

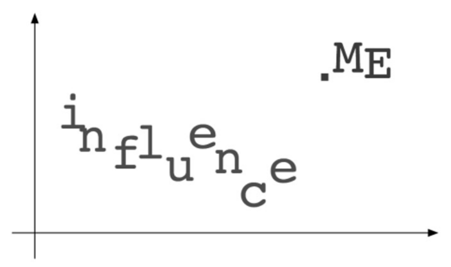
In simple linear regression, *leverage* may be calculated in the following way:

$$h_i = \frac{1}{N} + \frac{(X - \bar{X})^2}{\sum x^2}$$



There must be a large horizontal gap between influential observations and most of the other data points.


ExamScores and Study Hours for a Hypothetical Class



Leverage:

One may liken the effect of an influential observation to that of a lever, wherein greater pulling power in a certain direction is achieved.

Therefore, by measuring the leveraging power of each point and establishing a rule, one can detect influential observations.

In simple linear regression, *leverage* may be calculated in the following way:

$$h_i = \frac{1}{N} + \frac{(X - \bar{X})^2}{\sum x^2}$$

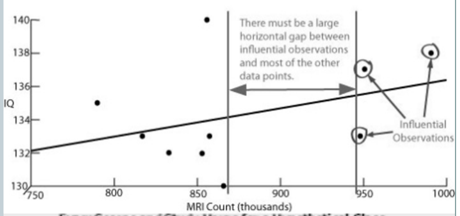
Influence Analysis

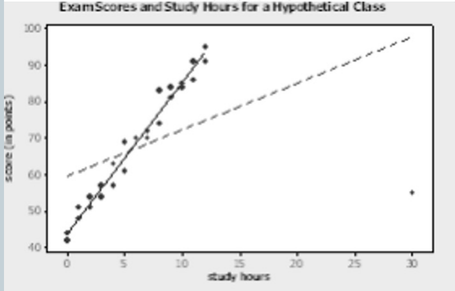
Leverage has several properties:

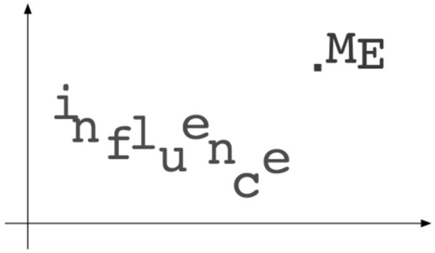
1. Leverage is a function solely of scores on the independent variable(s).
2. Other things equal, the larger the deviation of X_i from $\bar{X}$, the larger the leverage.
3. The maximum value for leverage is 1.0.
4. The average leverage for a set of scores is equal to $(k+1)/N$.

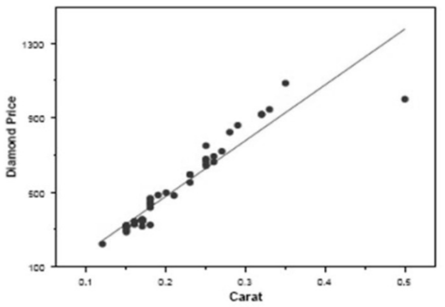
As a general rule of thumb, an influential observation may be considered to be one whose leverage, h_i, is greater than twice the average leverage of the data set, i.e.:

$$h_i > 2\,(k+1)/N$$









Influence Analysis

<u>When an Influential Observation is detected:</u>

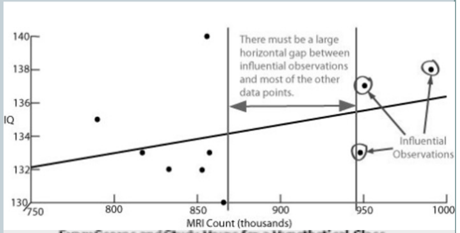
When an influential observation is detected, the procedures regarding how to deal with it are more complicated than those for dealing with outliers.

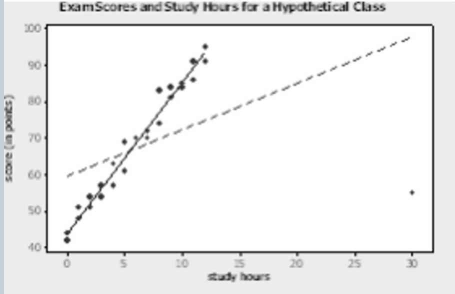
First, scrutinize the attributes of the influential individual.

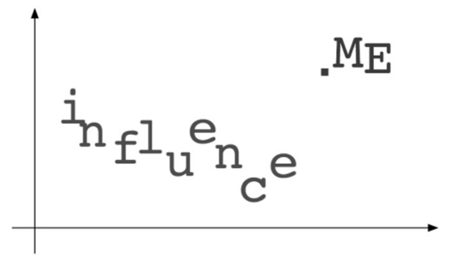
- How does this subject differ from the other subjects?

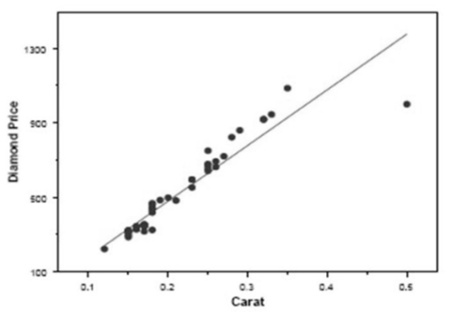
Second, understand that influential observations may reflect random error, or they may serve as clues.

- Increase sample size
- Replication!









Cook’s Distance (Cooks’ D)

Data points with high leverage can perturb the accuracy of a regression line, and therefore deserve closer examination.

- ***Cook's distance*** measures the *effect* of deleting a given observation.

Specifically, Cook’s D is designed to identify an influential observation whose influence is due to its status on the independent variable, the dependent variable, or both.

It is calculated in the following way:

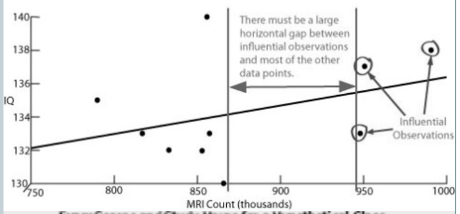
$$D_i = \left[\frac{(SRESID)_i{}^2}{k+1}\right]\left[\frac{h_i}{1-h_i}\right]$$

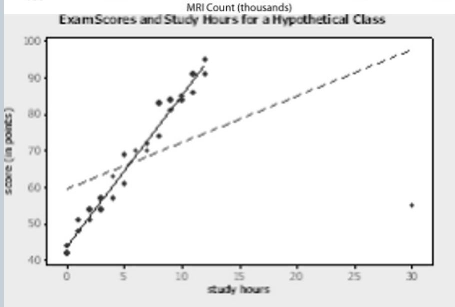
Where:

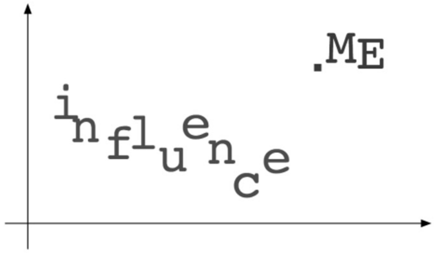
$SRESID$ is studentized residual for a given individual

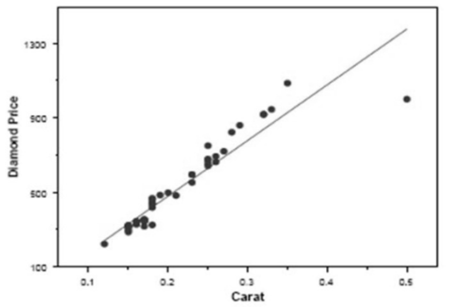
h_i is the leverage of a given individual

k is the number of independent variables









Cook’s Distance (Cooks’ D)

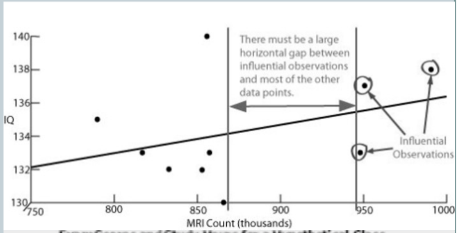
How to use Cook’s D:

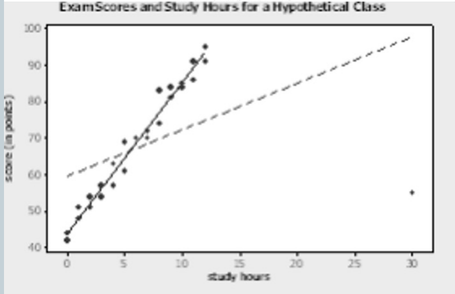
Statistical tests can be performed on Cook’s D, but more simply, one may search for relatively large values.

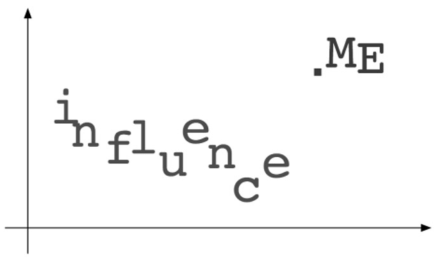
- Cook’s D is large when *SRESID* is large, when h_i is large, or when both are large.
- Obvious influential observations will leave large gaps between themselves and the rest of the data.

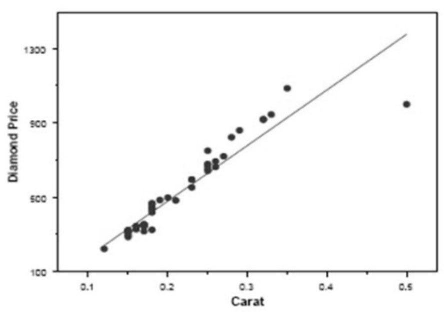
Cook’s D is more sensitive to detecting influential observations than leverage in isolation.

- A given point may have low leverage, but a large Cook’s D.









Recap: Outliers & Influential Points

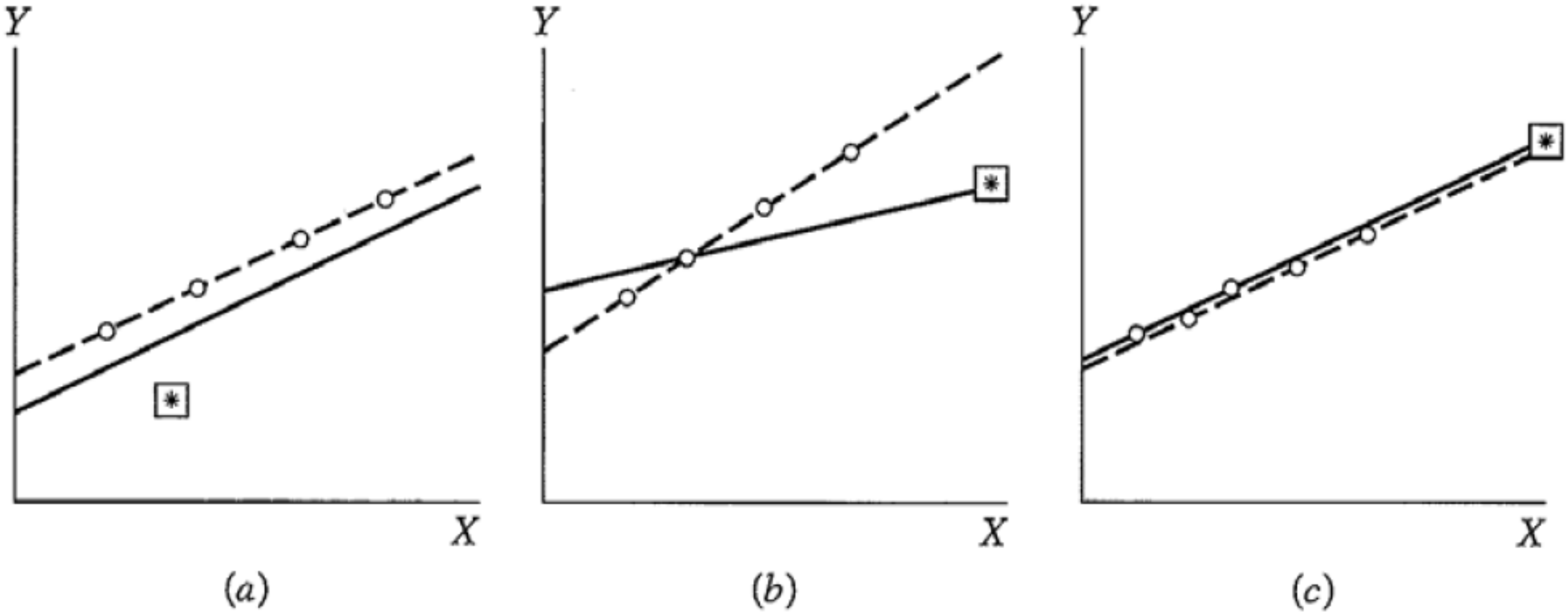


FIGURE 13.4 In each subfigure, the solid line gives the OLS line for all the data and the broken line gives the OLS line with the outlier, denoted by an ⊡, omitted. In (*a*), the outlier is near the mean value of *X* and has low leverage and little influence on the regression coefficients. In (*b*), the outlier is far away from the mean value of *X* and has high leverage as well as substantial influence on the regression coefficients. In (*c*), the outlier has high leverage but low influence on the regression coefficients because it is in line with the rest of the observations.

Source: Adapted from John Fox, op. cit., p. 268.

[42] Adapted from John Fox, *Applied Regression Analysis, Linear Models, and Related Methods* Sage Publications, California, 1997, p. 268.

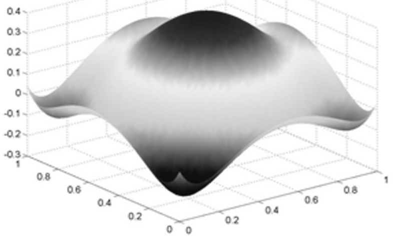
DFBETA

Leverage and Cook's D are *global indices*, in that, they can detect influential observations, but do not provide any information with regard to what effect they have on our estimates of parameters.

Presumably, if an influential observation is deleted from the data set, the regression line will change significantly.

This has consequences for our estimation of the parameters which would describe the same regression line for the population.

To this end, we use a term known as ***DFBETA*** to *measure the change in the estimate of the parameter* (i.e., α or β) as a consequence of deleting subject i.



DFBETA

To calculate DFBETA for a given observation:

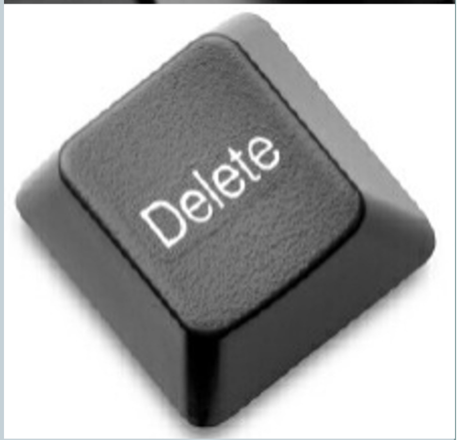
1. Plot a regression line for the original data, all subjects included.
2. Delete the influential observation.
3. Re-plot the regression line using the remaining data.
4. Note the changes that have occurred in the estimation of each of the parameters (α and β).

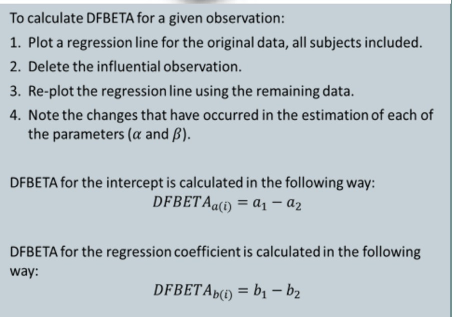
DFBETA for the intercept is calculated in the following way:

$$DFBETA_{a(i)} = a_1 - a_2$$

DFBETA for the regression coefficient is calculated in the following way:

$$DFBETA_{b(i)} = b_1 - b_2$$




To calculate DFBETA for a given observation:

1. Plot a regression line for the original data, all subjects included.
2. Delete the influential observation.
3. Re-plot the regression line using the remaining data.
4. Note the changes that have occurred in the estimation of each of the parameters (α and β).

DFBETA for the intercept is calculated in the following way:

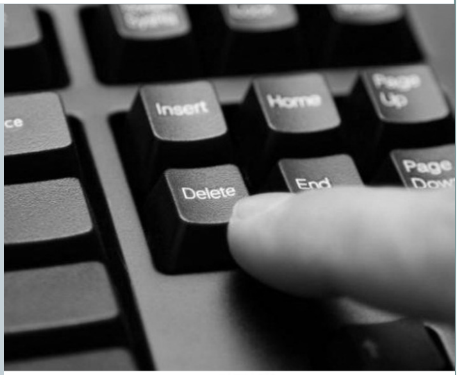
$$DFBETA_{a(i)} = a_1 - a_2$$

DFBETA for the regression coefficient is calculated in the following way:

$$DFBETA_{b(i)} = b_1 - b_2$$

DFBETA

As you may have gathered, this is a laborious process since one must recalculate the regression line for each of the subjects…



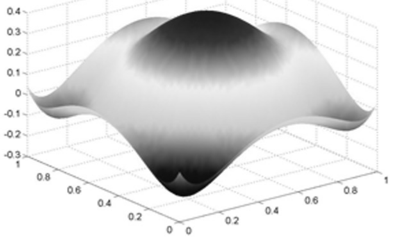
Fortunately, an alternative approach based on the results obtained from a *single* regression analysis, in which all the data are used, is available.

For a:

$$DFBETA_{a(i)} = \left[\left(\frac{\sum X^2}{N\sum X^2 - (\sum X)^2}\right) + \left(\frac{-\sum X}{N\sum X^2 - (\sum X)^2}\right) X_i\right] \frac{e_i}{1 - h_i}$$

For b:

$$DFBETA_{b(i)} = \left[\left(\frac{-\sum X}{N\sum X^2 - (\sum X)^2}\right) + \left(\frac{N}{N\sum X^2 - (\sum X)^2}\right) X_i\right] \frac{e_i}{1 - h_i}$$



DFBETA

After DFBETA values have been obtained, finding the new regression equation is easy.

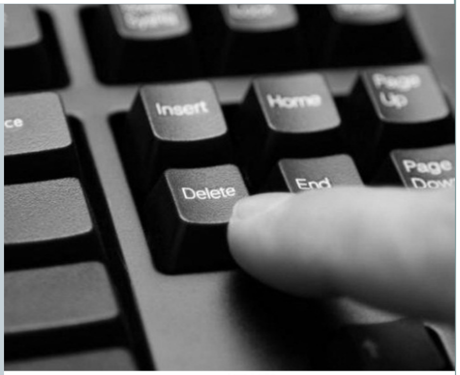
Recall the following equalities:

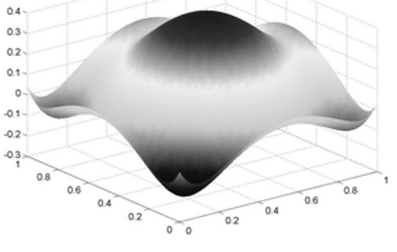
$$DFBETA_{a(i)} = a_1 - a_2$$

$$DFBETA_{b(i)} = b_1 - b_2$$

Using the DFBETA approach gives you information about *how influential* a point is.

Ultimately, you may use this information to determine the value in deleting or keeping any given individual in your data set.



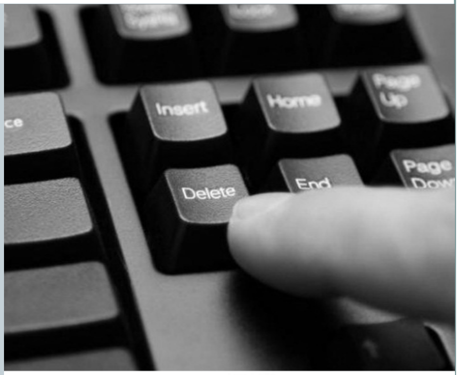


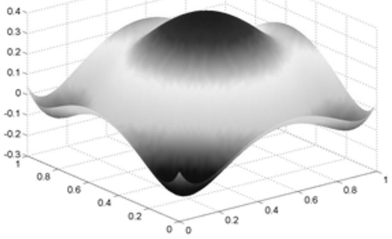
Standardized DFBETA

On that note, what constitutes a large DFBETA?

Unfortunately, there is no simple answer to this question since the magnitude of DFBETA is dependent on the units of measurement used in the study.

To get around this problem, then, one may prefer to use standardized scores.





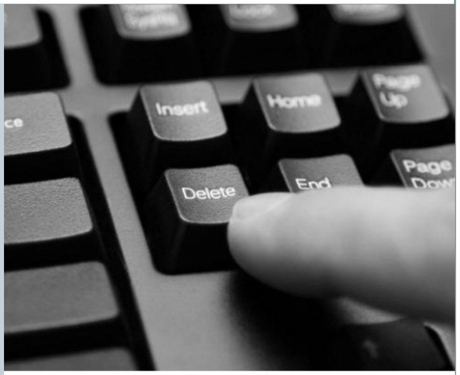
Standardized DFBETA

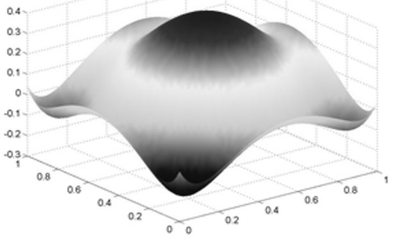
For the intercept, a, the standardized DFBETA is found by:

$$DFBETAS_{a(i)} = \frac{DFBETA_{a(i)}}{\sqrt{MS_{res(i)}\left[\frac{\sum X^2}{N\sum X^2 - (\sum X)^2}\right]}}$$

For the regression coefficient, b, the standardized DFBETA is found by:

$$DFBETAS_{b(i)} = \frac{DFBETA_{b(i)}}{\sqrt{MS_{res(i)}\left[\frac{N}{N\sum X^2 - (\sum X)^2}\right]}}$$





Standardized DFBETA

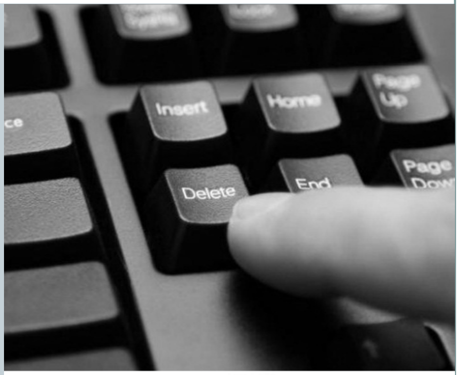
Note that in each formula the term, $MS_{res(i)}$, appeared.

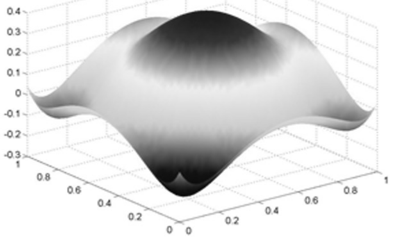
This term is based on an analysis in which a given subject, *i*, is deleted.

- As many regressions as there are subjects would, again, be required to calculate DFBETAS for everyone...

To avoid this, $MS_{res(i)}$ can be calculated in the following way:

$$MS_{res(i)} = \frac{SS_{res} - \left(\frac{{e_i}^2}{1 - h_i}\right)}{N - k - 1 - 1}$$





Standardized DFBETA

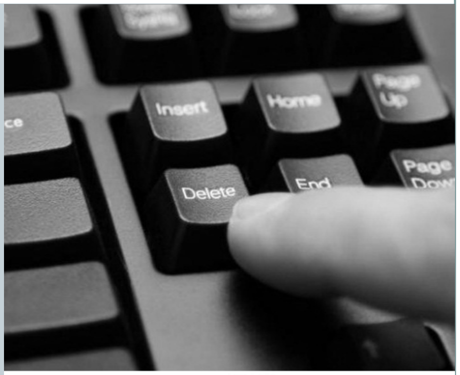
After the DFBETA has been standardized (DFBETAS) a criterion by which to find influential points can be constructed.

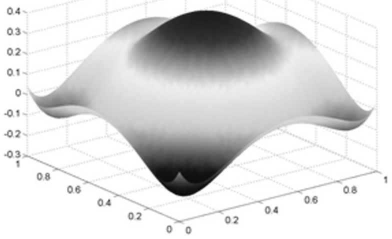
Although there is no *universal* agreement on what a “good” cutoff is, there are some general suggestions, conventions, etc.

One of the most common procedures, for example, is to cut values whose DFBETAS exceeds $2/\sqrt{n}$.

Others will choose one cutoff for small sample sizes and another for larger sample sizes (e.g., $2/\sqrt{n}$ and 1.00 for small and large samples, respectively).

Still, others will apply a universal cutoff, independent of sample size.





Remedies

"To delete or not to delete, that is the question."

-Wilbur Shakesmann

When an influential point is detected, it is important to keep in mind that it was only "detected" because of your criterion.

- In other words, a different criterion, a more conservative one, may not have labeled said point as "influential."

The urge to "correct" your data set by removing the errant point may be strong, but before you do so, you must carefully weigh the consequences.

- Is it *more misleading* to retain the errant point and report the data as a whole?
- Or is it *more misleading* to remove the errant point and only explain the data you understand?