* Ex: See example in R code.

* Diagnostics for Influential observations

* Here, we will be concerned with detecting influential observations - observations whose presence in the data have a distorting effect on parameter estimates and possibly the entire analysis.

* Importantly, we will be concerned w/ both outliers / influential points in the x and y direction as both can affect our analysis.

* Our analysis will be based on the hat matrix

$H := X(X^T X)^{-1} X^T$, which is an idempotent projection matrix which directly maps $Y \to \hat{Y}$:

$$\hat{Y} = HY.$$

Properties of H:

① $H = H^2 = HH^T$
② $(I - H)(I - H)^T = (I - H)$
③ $var(\hat{Y}) = H\sigma^2 \Rightarrow var(\hat{y}_i) = h_{ii}\sigma^2$
④ $0 \le h_{ii} \le 1$
⑤ $\sum_i h_{ii} = p$

- Property ⑤ suggests that a "typical" value of h_{ii} will be about p/n.

- Data points for which h_{ii} is close to 1 correspond to obs. for which ~~e_i is very small~~ the varience of e_i is very small $\Rightarrow$ $\hat{y}_i$ is close to y_i $\Rightarrow$ if h_{ii} is large (close to 1) then the i^{th} obs. distorted the regression line to pass close to the i^{th} observation.

- Thus, points with large h_{ii} are data points of high leverage.

- Show picture of high leverage points

- Rule of thumb (based on work of Belsley, Kuh & Welsch (1980)): points with

$$h_{ii} > 2p/n$$

are considered high-leverage data points.

(Go back to an example!)

- Deletion Diagnostics (testing for high-leverage)

- To analyze the effect of the i^{th} observation on the regression ~~of~~ line, we consider what the regression line would have looked like if that ~~value~~ observation was deleted.

In particular, define

$\hat{y}_{i(i)}$:= predicted value of y_i based on the model fit with obs. i omitted.

Then, we analyze the deletion residual:

$$d_i = y_i - \hat{y}_{i(i)}$$

Large values of d_i suggest that obs. i is influential! Through some work, one can show that

$$d_i = \frac{e_i}{1 - h_{ii}}$$ ⇒ we don't actually have to refit the model!

- Define $S^2_{(i)}$:= MSE of the regression if the i^{th} observation is omitted. Then, one can show that

$$s^2(d_i) = \frac{S^2_{(i)}}{1 - h_{ii}}$$

Furthermore, d_i and $s_{(i)}^2$ are independent and

$$d_i^* = \frac{d_i}{s_{(i)}} \sim t_{n-p-1}, \qquad (*)$$

we call d_i^* the <u>externally studentized residual</u>.

~~Using $(*)$, we can~~

Furthermore, we can calculate $s_{(i)}^2$ without refitting b/c of the following property:

$$s_{(i)}^2 = \frac{n-p}{n-p-1} s^2 - \frac{e_i^2}{(n-p-1)(1-h_{ii})}$$

which gives the "nice" form for d_i^*:

$$d_i^* = e_i \left[\frac{n-p-1}{(1-h_{ii})(n-p)s^2 - e_i^2} \right]^{1/2}$$

Using $(*)$, we can formally test whether observation i is an outlier. Namely, we conduct the following test:

$$H_0: E[d_i^*] = 0 \quad \text{vs.} \quad H_1: E[d_i^*] \neq 0$$

and reject H_0 if $|d_i^*| > t_{n-p-1, 1-\frac{\alpha}{2}}$.

- $\boxed{\text{Give an example in R}}$

- DFFITS:

Another possible value that provides a diagnostic for the influence of obs. i is the DFFITS measure, given by the following:

$$(DFFITS)_i = d_i^* \sqrt{\frac{h_{ii}}{1-h_{ii}}}$$

This diagnostic acts as a combined measure of influence that takes into account

i) the leverage of the data point
ii) the size of the residual

There is no distributional theory here but a rule of thumb is to consider obs. i influential if

$$(DFFITS)_i > 2\sqrt{p/n}$$

- Graphical Methods for Assessing Influence

- Main idea: half-normal envelope plots.

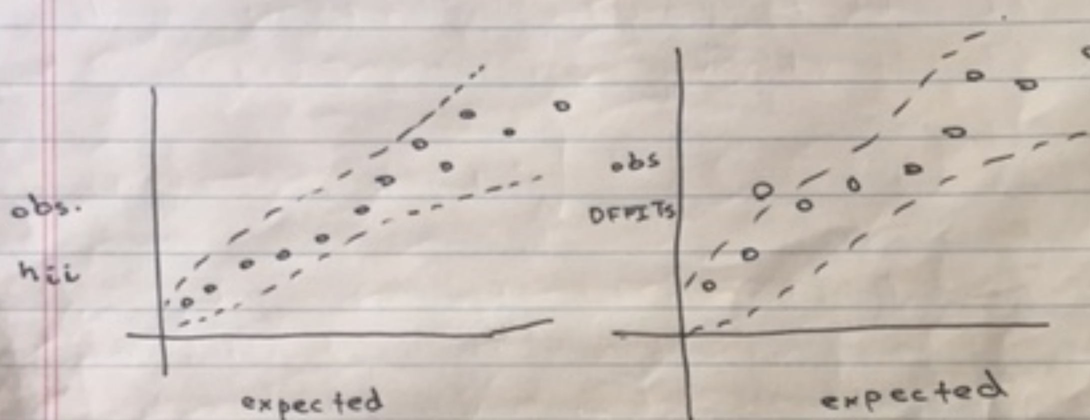
To generate, do the following:

Repeat N times to obtain CI's

1) Simulate $2n+1$ $N(0,1)$ values
2) Calculate the n largest values

Now, order the deletion residuals and DFFITS values and plot them against the 95 % CIs of the largest N(0,1) values.

Plot should look like:



Points outside the envelope plots suggest points of high influence.

Then, one can formally test using d_i^*.

* <u>Remedial Measures</u>

* In general, you should go back and investigate the data points are influential and check for mis-entry or imputation.

* If there are only a few (out of many data points), sometimes it's easier to simply omit the data points for analysis.