Summary of residual diagnostics and influence measures

Residuals: residuals(model) in R

$$e_i = Y_i - \hat{Y}_i \qquad i = 1, 2, \cdots n$$

These are the observed estimates of the errors for each sample observation.
They are measured on the same scale (and in the same units) as the Y variable.

Deletion Residuals:

$$PRESS_i = e_{i,-i} = Y_i - \hat{Y}_{i,-i} = \frac{e_i}{1 - h_{ii}}$$

Where $\hat{Y}_{i,-i}$ is the predicted value from a regression fit to the dataset, with the i^{th} observation deleted, with all the X variables set at the x_i values for the i^{th} observation. h_{ii} is the leverage value for the i^{th} observation, the i^{th} diagonal element of the hat matrix $X(X^TX)^{-1}X^T$.

Standardised Residuals – First type

Internally Studentised Residuals: rstandard(model) in R

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}} = \frac{e_{i,-i}}{s/\sqrt{1-h_{ii}}}$$

Where $s = \sqrt{MS_{Error}}$ is the residual standard error (with $n-p$ degrees of freedom)

These have an asymptotic standard normal distribution, so are a sensible choice to use in residual plots, so that we can interpret the scale using the empirical rule, if the sample size is sufficiently large. Under the empirical rule we expect to find:

- two thirds (67%) of the residuals within ± 1 standard errors of the model (the model is the x-axis or horizontal line where $r = 0$ on a residual plot);
- 95% of the residuals within ± 2 standard errors of the model; and
- almost all of the residuals within ± 3 standard errors of the model.

Standardised Residuals – Second type

Externally Studentised Residuals: rstudent(model) in R

$$t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}} = \frac{e_{i,-i}}{s_{-i}/\sqrt{1-h_{ii}}}$$

Where s_{-i} is the residual standard error for a regression fit to the dataset, with the i^{th} observation deleted (which has $n-p-1$ degrees of freedom).

Each t_i could be used as a test against a Student's t distribution with $(n-p-1)$ degrees of freedom to check for potentially influential outliers. The null hypothesis for each test is $H_0: \Delta_i = 0$, where Δ_i is the mean (location) shift in the model caused by the i^{th} observation.

However, be wary, just because an observation might "fail" this test does not necessarily mean it is definitely an outlier – these are heuristics or "rules of thumb" rather than formal statistical tests. It is also important to consider what we are trying to do with the model and the research question we are trying to address.

Influence Measures

Leverage Values: hatvalues(model) in R

$$h_{ii} = \text{the } i^{th} \text{ diagonal element of the hat matrix } X(X^TX)^{-1}X^T \quad Note: \sum_{i=1}^{n} h_{ii} = p$$

Here $p = k + 1$ is the number of parameters in the model (1 more than k, the number of variables, for models with an intercept term). Note that leverage is purely a function of the values of the X variables and does not depend on the Y_i value for the i^{th} observation.

For an observation to be considered high leverage, the lecture notes suggest a cut-off of more than twice the average leverage ($h_{ii} > 2\,p/n$).

Again, a word of caution, just because the i^{th} observation has high leverage, does not mean it is automatically highly influential in the fit of the model, as the Y_i value for that observation may still be following the general trend of the rest of the data.

Influence Measures continued

Deletion measure of the change in fitted values: dffits(model) in R

$$DFFITS_i = \frac{Y_i - \hat{Y}_{i,-i}}{s_{-i}\sqrt{h_{ii}}} = t_i \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

The DFFITS can be interpreted as the approximate number of standard deviations by which the predicted or fitted value will change if the i^{th} observation is deleted.

As a standardised measure, the DFFITS have "t-like" behaviour in that if the sample size is large enough (i.e. in large samples), any observation with:

$$|DFFITS_i| > 2\sqrt{p/n}$$

should be regarded as potentially influential.

Influence Measures continued

Deletion measure of the change in parameter estimates: dfbetas(model) in R

$$DFBETAS_i = \frac{b_j - b_{j,-i}}{s_{-i}\sqrt{c_{jj}}} \quad j = 1, \ldots, k$$

Where $c_{jj}$is the j^{th} diagonal element of the matrix $(X^T X)^{-1}$ and $\hat{\beta}_{j,-i} = b_{j,-i}$ is the least-squares estimates of β_j from a regression with the i^{th} observation deleted. Note that the "raw" deletion coefficients $\hat{\beta}_{j,-i} = b_{j,-i}$ are given using the function dfbeta() in R, but like normal partial regression coefficients, these are on the same scale as the Y variable. The function dfbetas() in R gives the above standardised deletion coefficients, which like the DFFITS have "t-like" behaviour in that if the sample size is large enough (i.e. in large samples), any observation with any coefficient:

$$\left|DFBETAS_{j,i}\right| > 2/\sqrt{n}$$

should be regarded as potentially influential.

Influence Measures continued

Cook's Distance: cooks.distance(model) in R

$$D_i = \frac{(b_i - b_{-i})^T (X^T X)(b_i - b_{-i})}{ps^2} = \left(\frac{r_i^2}{p}\right)\left(\frac{h_{ii}}{1 - h_{ii}}\right)$$

Where b_i is the vector of estimated partial regression coefficients from the "full" and b_{-i} is the vector of partial regression coefficients from the model with the the i^{th} observation deleted.

Cook's distance can be seen as measuring the overall effect of the i^{th} observation on parameters, whereas the DFBETAS measure the effects on each of the individual parameters.

Influence Measures continued

Cook's Distance continued:

$$D_i = \frac{(b_i - b_{-i})^T (X^T X)(b_i - b_{-i})}{ps^2} = \left(\frac{r_i^2}{p}\right)\left(\frac{h_{ii}}{1 - h_{ii}}\right)$$

The lecture notes suggest that as this measure has the form of a standardised ratio of variances (the distance/variance of the change in coefficients to the variance of the partial regression coefficents), the Cook's distances have "F-like" behaviour and an F distribution with p and $(n - p)$ degrees of freedom could be used as a "cut-off", but this is not really these diagnostics measure were designed to be used. My preference is to look for values of Cook's distance that are very large relative to the other observations.

Cook's distance is also a function of both the standard residual value and the leverage values, which is why values of Cook's D can be drawn as a series of curves on the default plot `plot(model, which=5)` in `R`, but the choice of 0.5 and 1 used in this plot are definitely arbitrary choices.

Influence Measures continued

Covariance Ratio: covratio(model) in R

$$COVRATIO_i = \left(\frac{s^2_{-i}}{s^2}\right)^p \left(\frac{1}{1-h_{ii}}\right)$$

Which attempts to measure how the i^{th} observation is affecting the estimate of the residual scale ($s^2 = MS_{Error}$), which is the key estimate in all of the statistical inference using the model.

The lecture notes suggest that as a "rule of thumb" any observation for which $COVRATIO_i$ is OUTSIDE the range $(1 - 3p/n, 1 + 3p/n)$ should be investigated as potentially highly influential.

Finally, examination of a range of the above influence measures will tell you more about the unusual observations in the data, but it is still a matter of judgement about whether or not you need to "treat" any potential outliers (by possibly removing them).