0.1 Equation Derivation

Let's say you have 2 dimensional data x_i, y_i and you would like to do a linear prediction with parameters β to get a $\hat{y}_i$.

$$\hat{y}_i = x\beta \tag{1}$$

Ordinary residuals can be calculated as the difference between the predicted and actual value.

$$e_i = \hat{y}_i - y_i \tag{2}$$

Residuals can be standardized in several different ways with a z-score or t-statistic but here we will look at Studentized residuals (a division of a residual by an estimate of its standard deviation).

$$r_i = e_i/\sigma(e) \tag{3}$$

The standard deviation of a residual can be determined with the leverage (diagonal of the hat matrix) and the SSE (sum of squared errors).

$$\sigma(e) = \sqrt{MSE(1 - h_{ii})} \tag{4}$$

Thus an internally Studentized residual can be calculated with the following equation

$$r_i = \frac{e_i}{\sqrt{MSE(1 - h_{ii})}} \tag{5}$$

In many cases, it is desired rather to compute the deleted Studentized residuals which are found by developing a regression prediction for each point i that has that point discarded from the fitting data:

$$d_i = \hat{y}_{i,d} - y_i \tag{6}$$

where d_i is a deleted residual and $\hat{y}_{i,d}$ is the prediction of point i with that point removed. To Studentize these residuals (known as externally Studenized residuals), we can follow the same approach as before

$$t_i = d_i/\sigma(d) = \frac{d_i}{\sqrt{(MSE_i(1 - h_{ii}))}} \tag{7}$$

where SSE_i is the sum of squared errors of all deleted residuals. As you can imagine, it can be a super big pain to have to refit your curve each time to determine what the deleted residuals are. Luckily there is a way to equate externally Studentized residuals with internal Studentized ones:

$$t_i = r_i\sqrt{\frac{n - p - 1}{n - p - r_i^2}} \tag{8}$$

This formula can then be rearranged to give the following equations

$$t_i = r_i \sqrt{\frac{n-p-1}{n-p-\frac{e_i^2}{MSE(1-h_{ii})}}}$$

$$t_i = r_i \sqrt{MSE(1-h_{ii})} \sqrt{\frac{n-p-1}{(n-p)MSE(1-h_{ii})-e_i^2}}$$

$$t_i = e_i \sqrt{\frac{n-p-1}{(n-p)MSE(1-h_{ii})-e_i^2}}$$

Note that the MSE can be expressed in terms of the SSE.

$$MSE = \frac{SSE}{n-p} \tag{9}$$

This will consequently lead to the following formula for deriving the Studentized deleted residuals in terms of the internally Studentized ones.

$$t_i = e_i \sqrt{\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}} \tag{10}$$

0.2 References

- Slides on regression with above formula.
- YouTube video on leverage and deleted residuals
- Example of a Studentized residuals in action

0.3 Nomenclature

- CDF : Cumulative Distribution Function refers to the probability that a value (X) is less than or equal to a given value. e.g. P(X≤=3) for example. It can be thought of as the integral of the PDF.
- PPF : Percent Point Function (also known as the quantile function) is defined as the inverse of the CDF, i.e. determining what the threshold value that a value will lie at or below a given threshold. e.g. X(P¡0.95).
- PDF : Probability Density Function can be thought of as the derivative of the CDF and is simply the probability distribution itself.
- PMF: Probability Mass Function

0.4 Determining Cut-Off

Since the externally Studentized residuals follow a T-distribution with n-p-1 degrees of freedom, a critical T value can be determined using the PPF (or quantile function). For instance here they indeed they use the function **qt** which refers to the quantile function to compute the cut-off for identifying any outliers.

Similarly this critical value can be computed with a significance value that is adjusted with the **Bonferroni Correction** which is a way of controlling for Type I errors (i.e. false positives) by adjusting the significance value.

0.5 Nomenclature

Symbol	**Meaning**
α	Significance value
a_p	Coefficient p
x_i	X Coordinate of data point i
$\hat{y_i}$	Prediction of regression model for point i
$\hat{y_{i,d}}$	Prediction of regression model for point i for data not including point i
y_i	Sensor value at point i
r_i	Ordinary residual at point i
$r_{i,d}$	Deleted residual at point i
N	Order of polynomial equation
t_i	Deleted residual of point i
n	Total number of data points
p	Number of parameters (2 in this case)
SSE	Sum of squared errors
$h_{i,i}$	Diagonal i of Hat Matrix
σ	Standard deviation
H	Hat matrix
X	Vector of all x values
BC	Bonferroni Critical Value