

Diagnostics Report – Formulas & Definitions (p1/2)

Residual ($e_i = y_i - \hat{y}_i$):

Difference between the actual individual value (y_i) and the value predicted from the model (\hat{y}_i).

Leverage ($h_{ii} = x_i^T (X^T X)^{-1} x_i$ where x is factor level and X is design matrix):

Numerical value between 0 and 1 that indicates the potential for a case to influence the model fit. A leverage of 1 means the predicted value at that particular case will exactly equal the observed value of the experiment (residual=0.) The sum of leverage values across all cases (design points) equals the number of coefficients (including the constant) fit by the model. The maximum leverage an experiment can have is $1/k$, where k is the number of times the experiment is replicated. Values larger than 2 times the average leverage are flagged.

Internally Studentized Residual ($r_i = \frac{e_i}{s\sqrt{(1-h_{ii})}}$):

The residual divided by the estimated standard deviation of that residual (dependent on leverage), which measures the number of standard deviations separating the actual from predicted values.

Externally Studentized Residual ($t_i = \frac{e_i}{s_{-i}\sqrt{1-h_{ii}}}$):

This “outlier t” value is calculated by leaving the run in question out of the analysis and estimating the response from the remaining runs. It represents the number of standard deviations between this predicted value and the actual response. Runs with large t values (rule-of-thumb: $|t| > 3.5$) are flagged and should be investigated.

Diagnostics Report – Formulas & Definitions (p2/2)

DFFITS ($DFFITS_i = \frac{\hat{y}_i - \hat{y}_{i,-i}}{s_{-1}\sqrt{h_{ii}}}$, alternatively $DFFITS_i = t_i \left[\frac{h_{ii}}{1 - h_{ii}} \right]^{1/2}$):

DFFITS measures the influence each individual case (i) has on the predicted value (see Myers² page 284.) It is a function of leverage (h). Mathematically it is the studentized difference between the predicted value with and without observation "i". As shown by the alternative formula, DFFITS represents the externally studentized residual (t_i) magnified by high leverage points and shrunk by low leverage points. Note that DFFITS becomes undefined for leverages of one (h=1).

DFBETAS ($DFBETAS_{j,i} = \frac{\beta_j - \hat{y}_{j,-i}}{s_{-1}\sqrt{c_{jj}}}$, c_{jj} is the jth diagonal element of (X'X)⁻¹):

DFBETAS measures the influence each individual case (i) has on each model coefficient (β_j). It represents the number of standard errors that the jth beta-coefficient changes if the ith observation is removed. Like DFFITS, this statistic becomes undefined for leverages of one (h=1). *DFBETAS are calculated for each beta-coefficient, so make sure to use the pull-down menu and click through the terms (the down arrow is a good short-cut key – also, try the wheel if you have one on your mouse).*

Cook's Distance ($D_i = \frac{1}{p} r_i^2 \left(\frac{h_{ii}}{1 - h_{ii}} \right)$):

A measure of how much the regression would change if the case is omitted from the analysis (see Weisberg¹ page 118). Relatively large values are associated with cases with high leverage and large studentized residuals. Cases with large D_i values relative to the other cases should be investigated. Look for mistakes in recording, an incorrect model, or a design point far from the others.

References:

1. Weisberg, Stanford: "Applied Linear Regression", 2nd edition, 1985, John Wiley & Sons, Inc.
2. Myers, Raymond: "Classical and Modern Regression with Applications", 1986, Duxbury Press.