Goodness-of-Fit Statistics

Sum of Squares Due to Error

This statistic measures the total deviation of the response values from the fit to the response values. It is also called the summed square of residuals and is usually labelled as SSE.

$$\text{SSE} = \text{Sum}_{(i=1 \text{ to } n)}\{w_i\,(y_i - f_i)^2\}$$

Here y_i is the observed data value and f_i is the predicted value from the fit. w_i is the weighting applied to each data point, usually $w_i = 1$.

A value closer to 0 indicates that the model has a smaller random error component, and that the fit will be more useful for prediction.

R-Square

This statistic measures how successful the fit is in explaining the variation of the data. Put another way, R-square is the square of the correlation between the response values and the predicted response values. It is also called the square of the multiple correlation coefficient and the coefficient of multiple determination.

R-square is defined as

$$\text{R-square} = 1 - [\text{Sum}_{(i=1 \text{ to } n)}\{w_i\,(y_i - f_i)^2\}] / [\text{Sum}_{(i=1 \text{ to } n)}\{w_i\,(y_i - y_{av})^2\}] = 1 - \text{SSE/SST}$$

Here f_i is the predicted value from the fit, y_{av} is the mean of the observed data y_i is the observed data value. w_i is the weighting applied to each data point, usually w_i=1. SSE is the sum of squares due to error and SST is the total sum of squares.

R-square can take on any value between 0 and 1, with a value closer to 1 indicating that a greater proportion of variance is accounted for by the model. For example, an R-square value of 0.8234 means that the fit explains 82.34% of the total variation in the data about the average.

If you increase the number of fitted coefficients in your model, R-square will increase although the fit may not improve in a practical sense. To avoid this situation, you should use the degrees of freedom adjusted R-square statistic described below.

Note that it is possible to get a negative R-square for equations that do not contain a constant term. Because R-square is defined as the proportion of variance explained by the fit, if the fit is actually worse than just fitting a horizontal line then R-square is negative. In this case, R-square cannot be interpreted as the square of a correlation. Such situations indicate that a constant term should be added to the model.

Degrees of Freedom Adjusted R-Square

This statistic uses the R-square statistic defined above, and adjusts it based on the residual degrees of freedom. The residual degrees of freedom is defined as the number of response values *n* minus the number of

fitted coefficients *m* estimated from the response values.

$$v = n-m$$

v indicates the number of independent pieces of information involving the *n* data points that are required to calculate the sum of squares. Note that if parameters are bounded and one or more of the estimates are at their bounds, then those estimates are regarded as fixed. The degrees of freedom is increased by the number of such parameters.

The adjusted R-square statistic is generally the best indicator of the fit quality when you compare two models that are nested – that is, a series of models each of which adds additional coefficients to the previous model.

$$\text{adjusted R-square} = 1 - \text{SSE}(n-1)/\text{SST}(v)$$

The adjusted R-square statistic can take on any value less than or equal to 1, with a value closer to 1 indicating a better fit. Negative values can occur when the model contains terms that do not help to predict the response.

Root Mean Squared Error

This statistic is also known as the fit standard error and the standard error of the regression. It is an estimate of the standard deviation of the random component in the data, and is defined as

$$\text{RMSE} = s = (\text{MSE})^{½}$$

where MSE is the mean square error or the residual mean square

$$\text{MSE}=\text{SSE}/v$$

Just as with SSE, an MSE value closer to 0 indicates a fit that is more useful for prediction.