Model Selection Statistics

When PROC GLMSELECT and PROC REG create candidate models, they can use various model selection statistics to decide whether to include or exclude variables at each step in the process. For a polynomial term, these model selection statistics are used to determine its appropriate degree.

These model selection statistics are as follows:

- the significance level for individual variables
- the coefficient of determination (R square)
- the adjusted coefficient of determination (adjusted R square)
- Mallows' C_p statistic
- Akaike's information criterion (AIC and AICC), and
- Schwarz's Bayesian criterion (SBC)

Most of these model selection statistics are based on the assumption that you want to create a model that minimizes the unexplained variability (the mean square error) with the smallest possible number of variables (a principle known as parsimony). PROC REG uses only the significance level (or *p*-value) for each individual variable to determine whether the variable enters the model (in forward selection), leaves the model (in backward elimination), or either (in stepwise selection).

By default, PROC GLMSELECT uses the SBC for variable selection. In PROC GLMSELECT, to use the significance level instead of the SBC for variable selection, you specify the SELECT=SL option in the MODEL statement. Then you use the SLENTRY= and SLSTAY= options to specify the significance levels for selection.

The default significance levels are as follows:

- For forward selection, SLENTRY=.50.
- For backward elimination, SLSTAY=.10.
- For stepwise selection, SLENTRY=.15 and SLSTAY=.15.

PROC GLMSELECT also enables you to specify model selection statistics other than the significance level. These additional model selection statistics include the following: the coefficient of determination (R square), the adjusted coefficient of determination (adjusted R square), Mallows' C_p statistic, Akaike's information criterion (AIC) and its corrected version (AICC), and Schwarz's Bayesian criterion (SBC).

Let's look at these model selection statistics one by one. R square (the coefficient of determination) measures the proportion of variability in the response variable that is explained by the predictor variables. However, because R square never decreases when you include more terms in the model, R square is not a good criterion to use to compare models that have a different number of predictor variables.

The adjusted R square is similar to R square, but it penalizes for additional terms in the model. Therefore, when you compare models that have a different number of predictor variables, it is more appropriate to use the adjusted R square. In the formulas for R square and adjusted R square, the following are defined: SSE is the sum of the squared error. SST is the total sum of squares corrected for the mean. df_E is the degrees of freedom associated with the sum of the squared error. df_T is the degrees of freedom associated with the total sum of squares. *n* is the total number of observations. *p* is the total number of parameters in the model, including the intercept.

Mallows' C_p is a simple indicator of model misspecification. This statistic helps to detect and avoid model bias, which refers to either underfitting the model (leaving out important terms) or overfitting the model (including too many terms). In the formula, the following are defined: *p* is the number of parameters in the model being evaluated, including the intercept. *n* is the total number of observations. MSE is the mean squared error. Mallows' C_p compares the mean squared error of the full model to models with a subset of the predictors. According to Mallows, when $C_p > p$, it usually means that the model is underspecified. Therefore, Mallows recommends looking for the first model (that is, the model with the fewest parameters) where Mallows' C_p is $\leq p$. For additional details, see Mallows, C.L. (1973).

The information criteria consist of Akaike's information criterion (AIC), finite-sampled corrected AIC (known as AICC), and Schwarz's Bayesian criterion (SBC). The information criteria are designed to assess the precision of fit of the model against the number of parameters in the model. In the context of multiple linear regression, information criteria measure the difference between a given model and the “true” underlying model. Akaike introduced the concept of information criteria as a tool for optimal model selection. Akaike’s Information Criterion (AIC) is a function of *n* (the

number of observations), the SSE (sum of the squared error), and *p* (the number of parameters in the model, including the intercept). The AIC tends to select models with a larger number of parameters. The AICC is the AIC with a correction for finite sample sizes. The AICC is recommended instead of the AIC when *n* is small. Schwarz developed a model selection criterion that is derived from the Bayesian modification of the AIC criterion. However, the SBC has a larger penalty for additional parameters in the model. For additional details about the model selection statistics, click the Information button.

Close