Print

Model Selection Statistics

Mallows' C_p Statistic

Another representation of the computational formula for C_p is as follows:

$$C_p = \frac{SS(Residual)}{MS(Residual)} + 2p - n$$

where *SS(Residual)* is the residual sum of squares for the model with *p* - 1 variables and *MS(Residual)* is the residual mean square when using all the independent variables.

When the model is correctly specified the residual sum of squares is an unbiased estimate of $(n-p)\sigma^2$, and C_p is an unbiased estimate of $\frac{(n-p)\sigma^2}{\sigma^2} + 2p - n = p$. So C_p is approximately equal to *p* when the model is correctly specified. When important variables are omitted from the model, the residual sum of squares is increased by the amount of variability that can be explained by those terms if they were included in the model. Therefore, C_p increases and $C_p > p$. (Rawlings, Pantula, and Dickey 1998)

Information Criteria

PROC GLMSELECT uses the definitions of AIC and AICC described in Hurvich and Tsai (1989). PROC REG uses an earlier definition of AIC (Akaike 1969 and Judge 1980).

Close