Mallows' C_p Statistic

For a subset model with p explanatory variables, this statistic is defined as

$$C_p = (\mathrm{SSE}_p/s^2) - (n - 2p)$$

where $s^2 = \mathrm{MSE}$ for the full model (i.e., is the model containing all k explanatory variables of interest). SSE_p is the residual sum of squares for the subset model containing p explanatory variables *counting the intercept* (i.e., the number of parameters in the subset model). Usually C_p is plotted against p for the collection of subset models of various sizes under consideration. Acceptable models in the sense of minimizing the total bias of the predicted values are those models for which C_p approaches the value p (i.e., those subset models that fall near the line $C_p = p$ in the above plot).

To understand what is meant by *unbiased predicted values*, consider the full model to be

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

and the subset model to be of the form

$$\mathbf{y} = X_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}$$

where $X = (X_1,\ X_2)$ and $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1',\ \boldsymbol{\beta}_2')$ are conformable partitions of X and $\boldsymbol{\beta}$ from the full model. Let the ith predicted value from the full model be $\hat{y}_i$ and that from the subset model be denoted by $\hat{y}_i^*$. The mean squared error of a fitted value for the full model is given by the expression:

$$\mathrm{mse}(\hat{y}_i) = \mathrm{var}(y_i) + [E(\hat{y}_i) - E(y_i)]^2$$

where $[E(\hat{y}_i) - E(y_i)]$ is called the *bias* in predicting the observation y_i using $\hat{y}_i$. If it is assumed that the full model allows unbiased prediction, the bias term must be zero; that is,

$$E(\hat{y}_i) - E(y_i) \ = \ 0$$

The mean squared error of a fitted value for the subset model is given by the expression

$$\mathrm{mse}(\hat{y}_i^*) = \mathrm{var}(y_i) + [E(\hat{y}_i^*) - E(y_i)]^2$$

which gives the bias in predicting y_i using the subset model fitted value to be

$$E(\hat{y}_i^*) - E(y_i)$$

Under the assumption that the full model is "unbiased", this bias term thus reduces to

$$E(\hat{y}_i^*) - E(\hat{y}_i)$$

The statistic C_p, as defined above, is a measure of the total mean squared error of prediction (MSEP) of a subset model scaled by σ^2, given by

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} \text{mse}(\hat{y}_i^*).$$

C_p has been constructed so that if the subset model is unbiased (i.e., if $E(\hat{y}_i^*) - E(\hat{y}_i) = 0$), then it follows that

$$\begin{aligned} C_p &= \frac{\text{SSE}_p}{s^2} - (n-2p) \\ &\approx \frac{(n-p)\sigma^2}{\sigma^2} - (n-2p) \\ &= p \end{aligned} \tag{1}$$

Recall that p here denotes the total number of parameters in the subset model (i.e., including the intercept). Thus only those subset models that have C_p values close to p must be considered if *unbiasedness* in the sense presented earlier is a desired criterion for selection of a subset model.

However, the construction of the C_p criterion is based on the assumption that s^2, the MSE from fitting the full model, is an unbiased estimate of σ^2. If the full model happens to contain a large number of parameters (β's) that are possibly *not* significantly different from zero, this estimate of σ^2 will be inflated (i.e., larger than the estimate of σ^2 obtained from a model in which more variables are significant). This is because the variables that are not contributing to significantly decreasing the SSE are still counted toward the degrees of freedom when computing the MSE in the full model. If this is the case, C_p will not be a suitable criterion to use for determining a good model. Thus, models chosen by other methods that have lower MSE values than the models selected based on the C_p must be considered competitive.