Numerical Model Comparison Criteria

As one searches through myriad possible MLR models potentially describing the relationship between predictors $x$ and a response, $y$, one generally wants a model with a "small" number of predictors (a simple or parsimonious model), a "large" value of $R^2$ and a "small" value of $s_e$. The naive "solution" to the model search problem of just "picking the biggest possible model" is in fact no solution, in light of the potential for "overfitting" a data set (adopting a model with too many "wiggles" that can nicely reproduce the $y$'s in the data set, but that does a very poor job when used to produce even mild extrapolations or interpolations).

For a given prediction problem (and therefore fixed SSTot), $R^2$ and $s_e = \sqrt{MSE}$ are "equivalent" in the sense that one could be obtained from the other (and SSTot). They don't, however, necessarily produce the same ordering of reduced models of some grand MLR model in terms of "best looking" values of the criteria (e.g. the full model has the largest $R^2$ but may not have the smallest $s_e$). There are several other functions of $SSE$ that have been suggested as possible statistics for model diagnostics/selection. Among them are Mallows' $C_p$ and Akaike's Information Criterion (the AIC).

**Mallows' $C_p$** is based on the fact that the average total squared difference between the $n$ values $\hat{y}_i$ from a MLR fit and their (real) means $\mu_i$ can be worked out theoretically. When divided by $\sigma^2$, this is a quantity $\Gamma_p$ that is

a) $p$, when there is a choice of $p$ coefficients $\beta$ in a $(p - 1)$-predictor reduced MLR model that produces correct means for all $n$ data points, and
b) larger than $p$, when there is no such choice.

Mallows' suggestion for comparing reduced versions of a full ($k$-predictor) MLR model, is that for a reduced model with $(p - 1) \leq k$ predictors, one compute

$$C_p = \frac{SSE_{\text{Red}}}{MSE_{\text{Full}}} + 2p - n$$

(an estimate of $\Gamma_p$ if the full $k$-variable MLR model is correct) and look for small $p$ and $C_p$ no more than about $p$. The thinking is that such a reduced model is simpler than the full model and appears to produce predictions comparable to the full model predictions at the $(x_1, x_2, \ldots, x_k)$ vectors in the data set.

**Akaike's Information Criterion** is based on considerations beyond the scope of this exposition. For a MLR model with $k$ predictors, it is

$$AIC = n\ln\left(\frac{SSE}{n}\right) + 2(k + 1)$$

and people look for small values of $AIC$. 