Criteria for Model Selection

1. R-squared Value

The **R-squared value** is defined by

$$R^2 = \frac{\mathrm{SSM}}{\mathrm{SST}} = 1 - \frac{\mathrm{SSE}}{\mathrm{SST}}.$$

Adding an additional regression parameter will always increase the R-squared value. The real question is if the increase in accuracy is worth the decrease in parsimony. In practice, R-squared often increases dramatically for the first few added regression parameters and then levels off as more parameters are added. Often the best choice for the number of regression parameters is where R-squared levels off and fails to increase significantly.

2. Adjusted R-squared value

The **adjusted R-squared value** takes the number of independent variables into account. It is defined by

$$\bar{R}^2 = 1 - \frac{\mathrm{MSE}}{\mathrm{MST}} = 1 - \frac{\frac{\mathrm{SSE}}{\mathrm{DFE}}}{\frac{\mathrm{SST}}{\mathrm{DFT}}} = 1 - \frac{(n-1)\mathrm{SSE}}{(n-p)\mathrm{SST}}$$

Usually $\bar{R}^2$ will initially increase as independent variables are added, but at a certain point will reach its maximum, then decrease slightly as more variables are added.

3. Information Criteria for Model Selection

Information criteria are measures of the tradeoff between the uncertainty in the model and the number of parameters in the model. These criteria measure the difference between the model being evaluated and the "true" model that is being sought. The general form of these criteria is

$$C = n \ln\left(\frac{\mathrm{SSE}}{n}\right) + q,$$

where $n \ln(\text{SSE}/n)$ represents the uncertainty in the model and q is a penalty term for the number of parameters. Here are some popular information criteria for regression model selection:

- **Akaike Information Criterion**

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

  R uses AIC as its criterion for stepwise selection.

- **Corrected Akaike Information Criterion**

$$\text{AICC} = 1 + \ln\left(\frac{\text{SSE}}{n}\right) + \frac{2(p+1)}{n-p-2}$$

- **Schwartz Baysian Criterion**

$$\text{SBC} = n \ln\left(\frac{\text{SSE}}{n}\right) + p \ln n$$

  Comparing AIC and SBC, we notice that $2p < p \ln n$ if $e^2 = 7.34 < n$. Because the penalty term is smaller for SBC than for AIC, SBC tends to favor models with fewer parameters than AIC does.

- **Hannan-Quinn Criterion**

$$\text{HQC} = n \ln\left(\frac{\text{SSE}}{n}\right) + p \ln \ln n$$

- **Sawa Baysian Information Criterion**

$$\text{SBIC} = n \ln\left(\frac{\text{SSE}}{n}\right) + \frac{2(p+2)n\hat{\sigma}^2}{\text{SSE}} - \frac{2n^2\hat{\sigma}^4}{\text{SSE}^2}$$

  where $\hat{\sigma}^2$ is the full error variance from fitting the full model.

4. Mallows CP Statistic

The Mallows CP statistic statistic (also denoted $C(p)$ or C_p) is defined as

$$\mathrm{CP} = \frac{\mathrm{SSE}}{\hat{\sigma}^2} + 2p - n, \tag{1}$$

where again $\hat{\sigma}^2$ is defined as the full error variance from fitting the full model. Using the assumption that the full model is unbiased, we can show that (1) is approximately equal to

$$\frac{(n-p)\sigma^2}{\sigma^2} - (n - 2p) = p.$$

Thus the "best" model should be one with CP close to p. One criticism of CP is that $\hat{\sigma}^2$ is not always a good approximation of the true error variance.