<u>26 November 2020</u>

Best Subset Selection Methods

Cp statistic

<u>Remarks</u>

1. If a model is underfitted, $\hat{\sigma}^2$, $\hat{\underset{\sim}{\beta}}$ and $\hat{y}$ at a new observation of $\underset{\sim}{x}_0$ are not unbiased estimators.
2. If a model is overfitted, $\hat{\sigma}^2$, $\hat{\underset{\sim}{\beta}}$ and $\hat{y}$ at a new observation of $\underset{\sim}{x}_0$ are still unbiased estimators but with larger variance.
3. Normally, we choose a larger value for α_{IN} than α_{OUT} to avoid underfitting a model.
4. The Lack of Fit test in Chapter 1 is a test for testing whether the model is underfitted if there are repeated measurements of y for the same $\underset{\sim}{x}$.

Consider the mean square error of $\hat{y}(\underset{\sim}{x}_i)$

$$\sum_{i=1}^{n} \frac{\mathrm{MSE}(\hat{y}(\underset{\sim}{x}_i))}{\sigma^2} \quad = \quad p' + \frac{\mathrm{E}(\hat{\sigma}^2_{p'}) - \sigma^2}{\sigma^2}(n - p')$$

Then, Cp is defined as its estimate, i.e.,

$$Cp \quad = \quad 2p' - n + \frac{ResS.S._{p'}}{\hat{\sigma}^2_{\text{full model}}}$$

<u>Remarks</u>

1. For each p', calculate C_p for the model with smallest $ResS.S._{p'}$.
2. Find the model with the smallest C_p.
3. Find a model with the smallest mean square error on predicted values of y.