3.6.2 → Selection Criteria

→ $\boxed{\text{Mallows } C_p = \frac{1}{n}(SSE + 2p \cdot \hat{\sigma}^2_{SE_g})}$ ↳ model that uses all g predictors

or $C_p' = \frac{SSE}{\hat{\sigma}^2_{SE_g}} + 2p - n$

→ $\boxed{AIC = \frac{1}{n}(SSE + 2p \cdot \hat{\sigma}^2_{SE_g})}$

→ The best model by AIC will also be the best model by C_p

one version of these formulas, another is given on the exam tools

→ $\boxed{BIC = \frac{1}{n}[SSE + \ln(n) \cdot p \cdot \hat{\sigma}^2_{SE_g}]}$

→ If $\ln(n)$ is replaced w/ 2, then we get the AIC formula. In other words, the penalty for each additional predictor is relative to the # of observations ⇒ the more obs that are available, the larger the penalty. ⇒ for $n \geq 8$, BIC favors models w/ a smaller p compared to AIC (as well as C_p).

→ other ideas

→ C_p, AIC & BIC have theoretical justifications for being good measures of model quality, whereas R^2_{adj} does not have similar theoretical support

→ For overfitted models due to high dimensions, C_p, AIC, BIC & R^2_{adj} are not reliable b/c they are functions of SSE

→ Cross-validation

Bias → validation set approach > k-fold CV > LOOCV

more bias/variance

Variance → LOOCV > k-fold > validation

→ Subset selection

	Scope of Models	Computationally Intensive	Suitable in High Dimensions
Best subset selection	All	Yes	No
Forward selection	Limited	No	Yes
Backward selection	Limited	No	No