Model Selection Criteria

In general, we will favour models with:

- less unexplained variation

  ie smaller MSE ($\hat{\sigma}^2 = s^2$) or smaller RSE ($\hat{\sigma} = s$)

  Mean Square Residual/Error from ANOVA table — Residual Standard Error from summary(model)

A useful comparison here is the nested model F test which indicates whether the apparent drop in s^2 is significant (for nested models). But s is on the same scale as Y, so we cannot use s to compare models on different scales, for example, we can't compare models for Y with models for $\log Y$ (as they are not nested)

- larger R^2 (R^2 is a standardised measure)

$$R^2 = 1 - \frac{SS_{Error}}{SS_{Total}}$$

BUT:
- no obvious point of comparison ie how big should R^2 be?
- does not protect against over-fitting as each additional X will increase (at at least not decrease) the R^2

- larger adjusted R^2, which does adjust for the df involved

$$\bar{R}^2 = 1 - \frac{MS_{Error}}{MS_{Total}} = R^2 - \underbrace{(1-R^2) \cdot \frac{df_{Regression}}{df_{Error}}}_{\text{adjustment}}$$

$df_{Regression} = k$, $df_{Error} = n-p$

Note this can be shown to be directly equivalent to preferring models with more significant overall F tests

ie $F_{statistic} = \dfrac{MS_{Regression}}{MS_{Error}}$; & associated p-value

& the overall F statistic does have an obvious point of comparison $F_{k,\,n-p}(1-\alpha)$

Model Selection Criteria (cont^{d}.)

Other options:

$$PRESS_p = \underbrace{\sum_{i=1}^{n} e_{i,-i}^2}_{} = \sum_{i=1}^{n}\left(\frac{e_i}{1-h_{ii}}\right)^2 = \sum_{i=1}^{n} r_i^2$$

deletion or PRESS residual (standardised)
ie internally studentised residual
sum of squares

→ based on the idea of <u>cross-validated</u> ⇒ it is an example of "leave-one-out" or n-fold cross-validation (see pages 33 & 34 of chapter 2)

→ as with $\hat{\sigma}^2 = s^2$, models with smaller $PRESS_p$ preferred

→ can also compare $PRESS_p$ with s^2
→ problems with outliers if $PRESS_p >> s^2$

Yet more Model Selection Criteria

Mallow's C_p

→ based on the idea that mis-specifying the model will create a bias in the estimate of σ^2 and that over-fitting will inflate the variances for predictions

(see lengthy argument on pages 35 & 36 of chapter 2 or even better Mallow's original paper)

$$C_p = p + \frac{(n-p)(s^2 - \hat{\sigma}^2)}{\hat{\sigma}^2}$$

→ requires some "independent" estimate of σ^2, called $\hat{\sigma}^2$, but in practice we often use $\hat{\sigma}^2 = s^2$ from "full" model with all predictors included

→ prefer models where $C_p = p$ (ie the bias term is 0), but if we use $\hat{\sigma}^2 = s^2$ from the "full" model then $C_p = p$ is guaranteed for the "full" model, so we also typically prefer simpler models ie smaller values of C_p for which $C_p \doteq p$