STAT6038 Week 11 Lecture Notes

Rui Qiu

2017-05-17

1 Wednesday's Lecture

1.1 Indicator Variables (continued yet again)

`prostate.lm3`

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon_i$$

where $Y =$`lcavol`, continuous response
$X_1 =$`lpsa`,
....

1.2 Model Selection

A "good" model is one which we can use to address the research question, which may:

- involve certain variables, which we must include in the model, so we can observe and/or "control for" the effects of these variables.
- other variables (included in the data) may also be included in the model, if they help to explain some of variation (i.e. they turn out to be "significant")
- Ultimately, the research question may require some predictions; preferably, predictions that hold general validity.

Note, if have already chosen some scale for the variables in the model and a particular form for the model, we can then experiment with models that include other X variables in the data as predictors, as well as derived variables ($X^2, \log X$, interaction terms involving $X, \dots$)

If we have k possible predictors, then the number of candidate models is $0(2^k)$ as a minimum (as we can also allow for different orders of the predictors), i.e. $k = 1$, 2 possible models; $k = 10$, 1024 models; $k = 20$, 2^{20} models.

For observational covariates (optional X's) we use:

Principle of Parsimony (Occarm's Razor: Of two similar models, we will tend to prefer the simpler one (especially if there is no significant different between them).

2 Thursday's Lecture

2.1 Model Selection Criteria

In general, we will favor models with:

- less unexplained variation, i.e. smaller MSE ($\hat{\sigma}^2$) or smaller RSE ($\hat{\sigma} = s$).
  Note: $\hat{\sigma}^2$ is Mean Square Residual/Error from ANOVA table. $\hat{\sigma}$ is 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 model for Y with models for $\log Y$ (as they are not nested).
- larger R^2 (R^2 is a standardised measure)
  $$R^2 = 1 - \frac{\text{SS}_{\text{Error}}}{\text{SS}_{\text{Total}}}$$
  BUT:
  - no obvious point of comparison i.e. how big should R^2 be?
  - does not protect against our-fitting as each additional X will increase (at least not decrease) the R^2.
- larger **adjusted** R^2, which does adjust for the degree of freedom involved
  $$\bar{R}^2 = 1 - \frac{\text{MS}_{\text{Error}}}{\text{MS}_{\text{Total}}} = R^2 - (1 - R^2) \cdot \frac{\text{df}_{\text{regression}}}{\text{df}_{\text{error}}}$$
  where degree of freedom of regression is k, degree of freedom of error is $n - p$.
  Note this can be shown to be directly equivalent to preferring models with more significant overall F-tests, i.e.
  $$F_{\text{statistic}} = \frac{\text{MS}_{\text{Error}}}{\text{MS}_{\text{Total}}}$$
  and associated p-value of the overall F statistic does have an obvious point of comparison $F_{k,n-p}(1 - \alpha)$.

2.2 Model Selection Criteria (continued)

Other options:

$$\text{PRESS}_p = \sum_{i=1}^{n} e_{i,-i}^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2 = \sum_{i=1}^{n} r_i^2.$$

where $e_{i,-i}$ is the deletion or PRESS residual (standardised) i.e. internally studentised residual sum of squares.

$\longrightarrow$ Based on the idea of **cross-validation** $\Longrightarrow$ it is an example of "leave-one-out" or n-fold cross-validation (see pages 33, 34 of chapter 2).

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

$\longrightarrow$ can also compare PRESS_p with s^2 $\longrightarrow$ problems with outliers if $\text{PRESS}_p \gg s^2$.

2.3 Yet more Model Selection Criteria

Mallow's C_p $\longrightarrow$ 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 variance predictions.

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

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

$\longrightarrow$ requires some "independent" estimates 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.

$\longrightarrow$ prefer models where $C_p = p$ (i.e. 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 i.e. smaller values of p for which $C_p = p$.