STAT6038 Week 12 Lecture Notes

Rui Qiu

2017-05-24

1 Wednesday's Lecture

1.1 Nested model F test (revisited)

Candidate models of log(survival)

size $p = 4$, model 13 in big table

$$Y = \beta_0 + \beta_1 \text{clot} + \beta_2 \text{prog} + \beta_3 \text{enzyme} + \epsilon$$

size $p = 5$, model 25 with base ... and addition ...

$$Y = \beta_0 + \beta_1 \text{clot} + \beta_2 \text{prog} + \beta_3 \text{enzyme} + \beta_4 \text{prog}^2 + \epsilon$$

size $p = 6$, model 39 with base ... and addition ...

$$Y = \beta_0 + \beta_1 \text{clot} + \beta_2 \text{prog} + \beta_3 \text{enzyme} + \beta_4 \text{prog}^2 + \beta_5 \text{prog:enzyme} + \epsilon$$

...

size $p = 8$, model 51

Y = full model

$= \beta_0 + \beta_1 \text{clot} + \beta_2 \text{prog} + \beta_3 \text{enzyme} + \beta_4 \text{liver} + \beta_5 \text{prog}^2 + \beta_6 \text{prog:enzyme} + \beta_7 \text{enzyme}^2 + \epsilon$

Nested model F-test for (51) vs (13)

$H_0: \ \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$
vs
$H_A: \ \text{not all } \beta_4, \beta_5, \beta_6, \beta_7 = 0$ (at least one $beta_j \neq 0, j = 4, 5, 6, 7$)
In variance term, we are testing:

$$\frac{\sigma^2_{\text{addition}}}{\sigma^2_{\text{error}}} = 1 \text{ vs. } \frac{\sigma^2_{\text{addition}}}{\sigma^2_{\text{error}}} > 1$$

Using the `cbind()` approach in R:

$$\begin{aligned} F &= \frac{\text{MS}_{\text{addition}}}{\text{MS}_{\text{error}}} \sim F_{?,n-p} = 46 \\ &= 1.092 \end{aligned}$$

p-value = 0.3717. So do not reject H_0, additions not needed.

Even the nested F test is not totally consistent with what happens if we refine models one step (one variable) at a time!

1.2 Model Refinement

Sensible systematic approaches to model selection involve changing the model one term at a time.

1.2.1 Forward Selection

1. Start with a base model.
   This could be a null (intercept only) model or **a model that already includes important research and control variables.**

2. Add the most promising of the optional candidate predictors.
   This would be the covariate which is most highly correlated with Y or **it could be one of a group which is correlated with Y but which is also promising based on the underlying science.**

3. Look at the last sequential F test for this newly added term **and/or at the added variable plot to see if some transformation is required.**

4. If the test in step 3 is significant, retain this new term in the model; if not, do not retain the term.

5. Repeat step 2,3,4, and 5 with the next most promising candidates until all candidates have been examined and included or rejected.

Note: most of these steps can be automated (by programming in some computer package), however, the underlined bits above require judgement and will probably be omitted in most automated processes.

1.2.2 Stepwise Refinement

We can combine Forward Selection and Backward Elimination in two ways:

1. Forwards stepwise refinement

   (a) Start with a base model and use forward selection to select the next candidate to add to the model.

   (b) Each time the model changes in step 1, use backward elimination to check that all the terms in the expanded model are necessary.

   (c) Repeat steps 1 and 2 until there are no more changes to the model

2. Backwards stepwise refinement

   (a) Start with full model, use backward elimination to remove one variable at a time.

   (b) Use forward selection to check if any of the omitted variables should be added back in.

   (c) Repeat steps 1 and 2 until no more changes.

2 Thursday and Friday's Lecture

2.1 A "Testing-Based" Procedure for Model Selection in R

Reference: Chapter 10 in Faraway (2e)

All the methods in the last lecture can be applied "manually" (and require only a little effort with relatively small datasets), **but** will very large datasets they can be automated.

Various approaches have been suggested, most of which initially tried to use model selection criteria discussed earlier (such as adjusted R^2), but this is an area of current research and is changing rapidly.

The `stepwise()` command from S-Plus was used in the example in the lecture notes, but this has not been implemented (in R), as it has been superseded by approaches which use a set of "information criteria" (**AIC, BIC (Bayesian information criteria), DIC (Deviance information criteria), FIC (Focused information criteria)**); and the calculation of these measures is beyond the scope of this course.

One approach, currently available in R is the `step()` function, which uses Akaike's Information Criteria (AIC).

Suppose that we have a statistical model M of soma data x. Let k be the number of estimated parameters in the model. Let $\hat{L}$ be the maximized value of the likelihood function for the model; i.e. $\hat{L} = P(x|\hat{\theta}, M)$, where $\hat{\theta}$ are the parameter values that maximize the likelihood function. Then the AIC value of the model is the following:

$$\text{AIC} = 2k - 2\ln(\hat{L})$$

Given a set of candidate models for the data, the preferred model is the one with the minimum AIC value. AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of estimated parameters. The penalty discourages **overfitting**, because increasing the number of parameters in the model almost always improves the goodness of the fit.

For ordinary linear models, AIC can be shown to be equivalent to Mallow's C_p (and both of them are only really valid for large samples). AIC is definitely measured on a different scale to Mallow's C_p.....

2.2 The Modelling Process

1. Try to understand ...
2. ...
3. ...
4. ...
5. Examine the ANOVA table to consider what are the important terms in the model AND decide if the model is an adequate model for answering the research question.
   `anova()`
6. Are there