ANOVA - Model Selection

Applied Statistics

4 Nov 2010

Model Selection

- Could fit all possible effects into a model
  - BUT: a model that is too big will be difficult to understand
- Instead, remove effects that are not important
- **HOW???**
- A good model should
  - fit the data reasonably well
  - be as simple as possible for its intended purpose (*e.g.* descriptive, explanatory, prediction)
  - be interpretable
- Tradeoff: between *fit* and *complexity* of the model

Criteria for Model Comparison

- F-tests for individual effects
  - **Beware:** the *order* of the terms in the model can make a difference (nonorthogonal designs)
- Information Criteria (AIC, BIC)
  - xIC = Deviance + Complexity
  - *Deviance* = -2 $\times$ log Likelihood = measure of goodness of fit
  - *Complexity*: gives a *penalty* for including more parameters

Information Criteria

- Better model fit $\Rightarrow$ lower deviance
- More parameters $\Rightarrow$ bigger complexity/penalty
- $\Rightarrow$ 'best' model has *lowest value* of the IC
- Akaike Information Criterion (AIC) $= -2 \ln L + 2p$
  - tends to select larger models
- Bayesian Information Criterion (BIC) $= -2 \ln L + p \ln(n)$
  - tends to select smaller models
  - may overpenalize factors with many levels

Choosing a Model

- Compare models using F-tests, AIC, BIC
- If the number of variables is small enough, could *compare all possible models*
- Usually this is not practical, use *automatic procedures*
  - forward selection
  - backward elimination
  - stepwise selection

Marginality Restriction

- Lower order terms are *marginal* to higher order terms
- Need to keep terms in the model that are marginal to other terms
  - if include *polynomial* term *e.g.* x^2, need to also keep x in the model
  - if include *interaction* term, need to keep all primary variables and lower order interactions in the model

Model (Variable) Selection Procedures I

- *Forward Selection*
  - start with *no* variables in the model
  - in successive steps, add in the 'best' unselected variable/term
  - stop when have the best model according to the chosen criterion, *e.g.* F, AIC, BIC
- *Backward Elimination*
  - start with *all* variables/terms in the model
  - in successive steps, take out the 'worst' included variable/term
  - stop when have the best model according to the chosen criterion, *e.g.* F, AIC, BIC

Model (Variable) Selection Procedures II

- *Stepwise Selection*
  - start with the *full* model
  - use Backward Elimination to see if any term can be removed
  - use Forward Selection to see if a term can be added
  - iterate (Backward - Forward - Backward - *etc.*)
  - stop when model doesn't change

Selection Procedures: Problems

- The methods are *automatic*
  - do not take into account scientific knowledge
  - do not take effect size into account – can include a significant variable with an effect size that is not interesting or important
  - can lead to model that are not meaningful or unrealistic
- Not guaranteed to find the optimum
  - Stepwise: try multiple times, starting with a different model each time
- *All models are wrong, but some are useful*

HOWTO: Model Selection

- Use scientific/problem-specific knowledge to suggest important variables/terms for potential inclusion
- Then, can try automatic procedures (stepwise selection, F-tests, *etc.*)
- Observe marginality
- If you use F-tests/ANOVA tables, remember that the order of inclusion of variables matters – try different orders
- Better to use `stepAIC` function in the R package `MASS`
- (see handout, Section 6.8 in the MASS book)

Model Assessment

- Important model *assumptions* :
  - Independent observations
  - Normally distributed errors
  - Constant error variance
  - Additive effects
- If the assumptions do not hold (at least approximately), then the results of the analysis will generally not be meaningful
- $\Rightarrow$ **Check assumptions!!**

Diagnostic Plots

- In addition to the *exploratory plots* you make at the beginning of the analysis, you will also need *diagnostic plots* in the model assessment phase
- There should not be any *structure* in the residuals
- Plot residuals against predicted values, variables in the model, variables *not* in the model (*e.g.* to see if some important variable is left out, assess dependence), normal QQ-plot
- Look for outliers, constant variance, patterns, normality