36-720: Model Selection & Model Evaluation

Brian Junker

September 26, 2007

- Model Selection—Overview
- Stepwise Procedures—Overview
- The Initial Model(s)
- Several Stepwise Criteria
- Staying Within Model Families
- All-Subsets Methods
- Residuals, Influence, etc.
- Examples

Model Selection—Overview

- Model selection "should be" for a purpose, e.g. *generically,*
  - Description (of dependence structure)
  - Prediction (of responses in regression)

  When we have *specific information* about how the model will be used, this should influence our model selection.
- Basically we are going for a version of the *bias vs. variance/uncertainty tradeoff*:
  - More complex models tend to have less bias in prediction, parameter estimation, etc.
  - Less complex models tend to have less variance/uncertainty in prediction, parameter estimation, etc.
- We will focus on choosing a good *descriptive* model. In this setting, the bias/variance tradeoff leads to considering the *smallest model in some class of models that is consistent with the data*.

- Most model-selection procedures are *exploratory*, especially because there is typically not much control for multiple comparisons.
- There always is *model uncertainty*. The best procedures tend to produce a *set of candidate models* which can be
  - Inspected for common features ("Hey, all these models say $A \perp\!\!\!\perp B \mid C$!").
  - Inspected for "plausibility" given background knowledge about the data generation process, how the model will be used, etc.

  It is a good idea to compare the results of several model-selection heuristics!
- Candidate models can/should also be checked for reasonable residuals, influence/leverage, etc.

Stepwise Procedures—Overview

- Three flavors of this heuristic
  - *Forward Selection* starts with a small *initial model* that doesn't fit the data, adds terms to the model until you get a fitting model.
  - *Backward elimination* starts with a large *initial model* that fits the data, deletes terms until any further deletion produces a model that no longer fits.
  - *Bi-directional* or *stepwise* procedures iterate between backwards and forwards until no more changes are possible.
- Stepwise procedures lead to a *single final model* that depends on
  - The initial model chosen
  - The order in which terms are added or deleted (the path through model space)

- As with all model selection procedures, we select models within a particular family of models
  - Hierarchy Principle
  - Graphical Models
  - Decomposable Models

  Often easiest to do by focusing on *graph edges*, that is, two-way interactions, and then "promoting" the models to hierarchical/graphical/decomposable form.

- For log-linear models *we should always keep terms that are needed to be consistent with the sampling model:*
  - Poisson Sampling: no special terms needed
  - Multinomial Sampling: keep the intercept, corresp. to $n_{++\ldots}$
  - Product-Multinomial: keep terms corresponding to the individual multinomial totals.

The Initial Model(s)

- It is a good idea to bound the search by selecting a smallest and a largest model you will consider. These can be [0] and [all factors], but often we can do better.
- *All s-factor interactions:* Successively fit the models
  - Independence: [1][2][3]$\cdots$
  - All 2-way interactions: [12][13][14][23][24]$\cdots$
  - All 3-way interactions: [123][124][134][234]$\cdots$
  - All 4-way interactions: [1234][1235]$\cdots$

  until you have a largest non-fitting model and/or a smallest fitting model. Model search then proceeds from (one of) these two.
- *Test each term last.* This is essentially equivalent to inspecting *t*-statistics in R's `summary(...)` output. Bi-directional stepwise might proceed from the model by deleting "non-significant" terms.

- *Marginal and Partial Association tests.*
  - The Partial Association Test for an s-factor term compare the "all s-way interactions" model, vs. the same model with that s-factor term deleted.

    E.g. in a 4-way table the partial association test for [123] compares [123][124][134][234] with [124][134][234].
  - The Marginal Association Test for an s-factor term compare the saturated model with the "no s-way interaction" model, on a table that has been collapsed to just those s factors.

    E.g. in a 4-way table the marginal association test for [123] compares [123] with [12][13][23] (omitting/collapsing over [4]).
  - The initial model might be chosen to be

    (a) All marginally significant terms

    (b) All partially significant terms

    (c) (a) $\cup$ (b)

    (d) (a) $\cap$ (b)

Several Stepwise Criteria

- Forward Selection vs. Backward Elimination
- Coherence
- Staying Within Model Families

Forward Selection vs. Backward Elimination

- *Backward Elimination*

  If we start from the saturated model (or some other model that fits) we are always comparing $\mathcal{M}_0 \subset \mathcal{M}_1$ where $\mathcal{M}_1$ is known to fit (but may be overparameterized) and $\mathcal{M}_0$ may or may not fit. *LR test statistics inherit their null distribution from $\mathcal{M}_0$, and are designed to detect misfit when $\mathcal{M}_1$ fits but $\mathcal{M}_0$ does not.*

- *Forward Selection*

  We always start from a model that is too small to fit well (often [0], e.g.). Therefore we are comparing $\mathcal{M}_0 \subset \mathcal{M}_1$ where $\mathcal{M}_0$ definitely does not fit, and $\mathcal{M}_1$ may or may not fit. *LR test statistics inherit their null distributions from $\mathcal{M}_0$, but are not really designed to detect greater misfit in $\mathcal{M}_0$ when neither $\mathcal{M}_0$ nor $\mathcal{M}_1$ fits.*

For these reasons, some people prefer *backward elimination*. However, Edwards (2000, Ch 6) points out that the two heuristics often yield similar results.

Coherence

The *coherence principle* says that

- If $\mathcal{M}_0 \subset \mathcal{M}_1$ and the fit of $\mathcal{M}_0$ is accepted, then so is $\mathcal{M}_1$.
- If $\mathcal{M}_0 \subset \mathcal{M}_1$ and the fit of $\mathcal{M}_1$ is rejected, then so is $\mathcal{M}_0$.

In *Backward Elimination*, suppose we test $\mathcal{M}$ vs $\mathcal{M} \setminus [JK]$ and reject $\mathcal{M} \setminus [JK]$.

- No submodel of $\mathcal{M} \setminus [JK]$ need be considered
- Subsequently, every model should have $[JK]$ in it.

Once [JK] is in the model, it stays in for good.

In *Forward Selection*, suppose we test $\mathcal{M}$ vs $\mathcal{M} \cup [JK]$ and we accept $\mathcal{M}$.

- Coherence would require [JK] to never be considered again.
- However, $\mathcal{M}$ is known not to fit. Maybe [JK] *would* be involved in some well-fitting model $\mathcal{M}'$, it just doesn't help with $\mathcal{M}$.

Once [JK] is out of the model, it stays out for good.

For these reasons, some people prefer the combination of coherence with backwards selection. Some "All subsets procedures" (e.g. EH, to be described below) also use coherence to prune the model space.

Staying Within Model Families

- R has a stepwise procedure but it is termwise in $X\beta$, and hence will not stay within hierarchical interaction models.
- MIM can check graphicality and decomposability at each stage in stepwise variable selection.
- Wermuth (1976): *Backward Elimination among Decomposable Models.*
  1. Start with a large decomposable model $\mathcal{M}$ (e.g. saturated).
  2. Test all the edges $[JK] \in \mathcal{M}$ that are *not* involved in more than one clique. Delete the least significant nonsignificant edge.
  3. Replace $\mathcal{M}$ with the new model, and go back to (2) until there are no more nonsignificant edges.

  Theorem: If $\mathcal{M}_0 \subset \mathcal{M}_1$ are both decomposable, there is a sequence of decomposable models formed by single edge-removals going from $\mathcal{M}_1$ to $\mathcal{M}_0$.
- Edwards & Havránek (1985): Considering all possible edge removals produces *Backward Elimination among Graphical Models.*

All-Subsets Methods

- If we have Q possible terms in the model, compare all 2^Q models using a measure that doesn't depend on nesting, like BIC = $G^2 - p\log(n_+)$, AIC = $G^2 - 2p$, (C_p), etc. *Only feasible if Q is small.*

- The *Edwards-Havránek (EH) procedure* keeps three lists, to search for the best model in the family $\mathcal{F}$:

$$\begin{aligned} W_A &= \{ \text{ models with an accepted submodel } \} \\ W_R &= \{ \text{ models with a rejected supermodel } \} \\ I &= \mathcal{F} \setminus (W_A \cup W_R) \end{aligned}$$

  At each stage, EH reduces I by testing either the minimal or maximal models in I (whichever list is shorter), until I is empty. Then the minimal accepted models can be read from W_A. See Edwards (2000, Sect. 6.2).

  EH is basically an all-subsets procedure with pruning by coherence.

Residuals, Influence, etc.

Residuals

- *Pearson Residuals*

$$X^2 = \sum_c \frac{(n_c - \hat{m}_c)^2}{\hat{m}_c} = \sum_c \tilde{r}_c^2 \,, \quad \tilde{r}_c = \frac{n_c - \hat{m}_c}{\sqrt{\hat{m}_c}}$$

- *Deviance Residuals*

$$G^2 = 2\sum_c n_c \log(n_c/\hat{m}_c) = \sum_c r_c^{*2} \,, \quad r_c^* = \text{sgn}(n_c - \hat{m}_c)\sqrt{2n_c \log(n_c/\hat{m}_c)}$$

- *Standardized Residuals*

$$s_c = \frac{n_c - \hat{m}_c}{\sqrt{\widehat{\text{Var}}\,(n_c)}} \,, \text{ e.g. for Multinom., } s_c = \frac{n_c - \hat{m}_c}{\sqrt{n_+ \hat{p}_c(1-\hat{p}_c)}} = \frac{n_c - \hat{m}_c}{\sqrt{\hat{m}_c(1-\hat{p}_c)}}$$

More generally, we can think of the algorithm for fitting a (poisson) GLM (Christensen, Ch 10, esp. sect 10.7) as an (iterative) weighted least squares fit of the model

$$\log m = X\beta$$

where X is the design matrix (determined by the shape of the table and the interactions included in the model; see Christensen Ch. 10), weighted by the expected cell counts.

In this case the appropriate "hat matrix" is

$$H = X(X^T D(\hat{m})X)^{-1} X^T D(\hat{m}) \, ,$$

where X is the design matrix of the GLM and $D = diag(\hat{m})$. Then as in ordinary linear regression, the *standardized residuals*

$$s_c = \frac{n_c - \hat{m}_c}{\sqrt{\hat{m}_c(1 - h_{cc})}}$$

will be approximately asymptotically $N(0, 1)$.

Cook's Distance

Again exploiting the connection between GLM and weighted regression, we can construct a *Cook's distance* measure for deleting the q^{th} cell in the table. If the log linear model is

$$\log m = X\beta$$

then there are good approximations for getting efficiently from $\hat{\beta}$ (the full MLE) to $\hat{\beta}_{(q)}$ (the MLE with cell q missing). We can then calculate an approximate *Cook's Distance*

$$C_q = \frac{(\hat{\beta} - \hat{\beta}_{(q)})^T X^T D(\hat{m}) X(\hat{\beta} - \hat{\beta}_{(q)})}{p} = \frac{1}{p} \sum_c \hat{m}_c \left[\log(\hat{m}_c / \hat{m}_{c(q)})\right]^2$$

where $\hat{m}_{c(q)}$ is the fitted cell count for each cell (including $c = q$!) computed from $\hat{\beta}_{(q)}$, and p is the df of the model.

Cook's distances can be compared to $\frac{1}{p}\chi^2_p$ to see whether cell q has substantial influence on the fit. Since it doesn't hold marginal totals fixed, Cook's distance is most appropriate for Poisson sampling.

Examples

(see in class examples!)