Statistics 262: Intermediate Biostatistics

Model selection

Jonathan Taylor & Kristin Cobb
Today’s class

- Model selection.
- Strategies for model selection.
- Model selection in survival analysis.
Model selection

- Up to now, we had a fixed model which we presumed was “good”, in both regression and survival models.

- In applied settings, we may be faced with MANY covariates, some of which may not be a priori related to the outcome of interest.

- “Model selection” is the process of “building” a model from possible many covariates.
Model selection: goals

- Which main effects do we include?
- Which interactions do we include?
- The previous two steps define a “collection” of models: we need an “algorithm” to “choose” a model from this collection.
Model selection: general

- This is an “unsolved” problem in statistics: there are no magic procedures to get you the “best model.”

- In some sense, model selection is “data mining.”

- Data miners / machine learners often work with very many predictors.
Model selection: strategies

- Model selection can be done according to a fixed set of rules (as in Hosmer & Lemeshow): “purposeful” selection of variables.
- Alternatively, can be done algorithmically.
- To automate the procedure, we first need a criterion or benchmark to compare two models. We also need a search strategy.
1. First, fit all models with only one predictor variable.

2. Fit a model with all variables that had $p$-value less than 0.2 in previous step.

3. Add variables not added in previous step, one at a time to the multivariable model to see if we have missed any important confounders.
Continuous covariates

1. Plot #1: \((X_i, M_i)\) where \(M_i\) are the martingale residuals, add a lowess smooth to get an idea of the functional relationship.

2. Plot #2: \((X_i, \log(\delta_{i,sm}/\hat{H}_{i,sm}))\).

3. Above, \(\delta_{i,sm}\) is a smoothed version of the plot \((X_i, \delta_i)\) and \(\hat{H}_{i,sm}\) is a smoothed version of the plot \((X_i, \hat{H}_i)\).
WHAS: length of stay

Martingale residual

Length of stay
WHAS: length of stay

Smoothed log(c/H) vs. Length of stay

Statistics 262: Intermediate Biostatistics – p.10/?
WHAS: age

Length of stay

Martingale residual
Interactions

1. Interactions should be added after determining scales of continuous covariates.

2. In the “purposeful search”, H&L suggest only including plausible interactions, and include all interactions significant at a level of 0.05 or lower.

3. Suggest following “rule” of including both terms of an interaction, even if one of the main effects is not significant.
Stepwise search (likelihood tests)

- Includes / deletes variables based on partial likelihood tests.
- Choose $p_E < p_R$ ($E$ for entry, $R$ for reject).
- Works for many different regression model selection – not just survival analysis.
Stepwise search (likelihood tests)

- Start with a model with no coefficients, just an intercept.
- For each covariate \( \{X_1, \ldots, X_p\} \) fit model with just one covariate.
- If any covariate has \( p \)-value less than \( p_E \) choose covariate with smallest \( p \)-value.
Stepwise search (likelihood tests)

- Assume we have chosen a variable, $X_i$
- For each covariate $X_j \in \{X_1, \ldots, X_p\} \setminus \{X_i\}$ fit a model with two covariates: $X_i, X_j$.
- If any partial likelihood tests have $p$-values less than $p_E$ choose covariate with the smallest $p$-value.
Assume we now have two variables $X_{i1}, X_{i2}$.

Test whether any variable can be dropped, by looking at partial likelihood tests: drop variable with largest $p$-value $> p_R$.

After dropping (or not), check to see if any variable can be added with $p$-value less than $p_E$.

Repeat until no variables can be added or dropped.
Variations

- Begin at an initially “well-chosen” model, possibly including “treatment” + some confounders.
- Search only forward or only backward.
If the number of predictors (including interactions) is not too large, then it is possible to fit all models of a certain size. Models are compared (within a fixed size) on the basis of the score test $\chi^2$ statistic. This identifies “best model” of a given size, but does not allow you to compare models of different sizes.
PROC PHREG DATA=WHASMOD;
MODEL LENFOL*CENSOR(0) = AGE SEXNUM CPK SHONUM CHFNUM LENSTAY /
  SELECTION=STEPWISE DETAILS;
RUN;
PROC PHREG DATA=WHASMOD;
MODEL LENFOL*CENSOR(0) = AGE SEXNUM CPK SHONUM CHFNUM LENSTAY /
  SELECTION=BACKWARD DETAILS;
RUN;
PROC PHREG DATA=WHASMOD;
MODEL LENFOL*CENSOR(0) = AGE SEXNUM CPK SHONUM CHFNUM LENSTAY /
  SELECTION=FORWARD DETAILS;
RUN;
PROC PHREG DATA=WHASMOD;
MODEL LENFOL*CENSOR(0) = AGE SEXNUM CPK SHONUM CHFNUM LENSTAY / SELECTION=SCORE BEST=3;
RUN;
PROC PRINT DATA=SELOUT;
RUN;
An alternative stepwise search uses the AIC (Akaike Information Criterion).

\[ AIC(M) = -2 \log L(M) + 2 \times \#M \]

where \(-2 \log L(M)\) is partial likelihood test statistic comparing to null model and \(#M\) is the number of parameters in \(M\).

\(AIC\) is a criterion that penalized larger models, with equal fit. It can be used to compare models of different size.
AIC also defined for many other models: method is not specific to survival analysis data.

If we replace $2 \cdot df_M$ with $\log n \cdot df_M$ then we get Schwarz’s Bayesian Information Criterion (BIC).

Other penalties have been proposed in the literature, but AIC and BIC are the most popular.
Stepwise AIC – search

- Start with a model with no coefficients, just an intercept.
- For each covariate \( \{X_1, \ldots, X_p\} \) fit model with just one covariate.
- If any models have lower AIC, choose the one with the lowest AIC.
Stepwise AIC – search

- Next, try to add one of the remaining variables, also try to drop each of the variables in the model.

- If any models have lower AIC, choose the one with the lowest AIC.

- Repeat until no further additions / deletions will decrease the AIC – converges to a local minimum of AIC.
Begin at an initially “well-chosen” model, possibly including “treatment” + some confounders.

Search only forward or only backward.
Caveats

- In choosing a model automatically, even if the “full” model is correct (unbiased) our resulting model may be biased – a fact we have ignored so far.

- Inference \((F, \chi^2\) tests, etc) is not quite correct for biased models.

- Diagnostics are still necessary! Just because an algorithm tells us it is good doesn’t mean it is!
Alternatives: penalized models

- Alternatively: we can choose “biased” models by imposing constraints.
- Here, “large $\beta$” is interpreted as “complex model”. Goal is really to penalize “complex” models, i.e. Occam’s razor.
- Idea is to “accept” a little bias to get a “better model”: bias-variance tradeoff.
Ridge regression

- Assume that columns \((X_j)_{1 \leq j \leq p}\) have zero mean, and length 1.

- Minimize

\[
L_{p,\lambda}(\beta) = -2 \log L_p(\beta) + \lambda \sum_{j=1}^{p} \beta_j^2
\]

where \(L_p\) is the Cox partial likelihood.

- Corresponds (through Lagrange multiplier) to a quadratic constraint on \(\beta\)’s. LASSO, another penalized regression uses \(\sum_{j=1}^{p} |\beta_j|\).