Regression Modeling and Pre-Modeling Strategies

Frank Harrell

Div. Of Biostatistics & Epidemiology

Dept. of Health Evaluation Sciences

UVa School of Medicine

Dept. of Biostatistics -MCV VCU

26Oct2001

hesweb1.med.virginia.edu/biostat

Outline

- Reasons for multivariable modeling
- Problems to address
- Modeling strategy
- Useful tools
- S-PLUS software
- Examples

Reasons for Modeling

- Hypothesis tests
- Estimates of partial effects
- Understanding shapes of effects
- Predictions for individual subjects
- Getting the "right" estimate of treatment effect in RCTs with non-normal response
- Testing/estimating differential treat. effect
- Estimating absolute treatment effects

Problems to Address

- Missing predictors
- Linearity/additivity assumptions
- Assessment of overfitting
- Fitting without overfitting (shrinkage)
- Redundancies / data reduction
- Model validation without $\Downarrow$ n
- Explaining complex models to non-stat.
- Removing excuses for using best methods

A Modeling Strategy

- Understand the problem and the data
- Assemble candidate Xs ignoring Y
- Characterize missing Ys before discarding
- Study relationships among Xs
  - data reduction
  - possible imputation models
- Characterize missing Xs
  - simultaneous missings
  - which subjects tend to having missings

Strategy, Continued

- Multiply impute missing Xs based on other Xs and Y, rather than ⇓ n, with variance correction (or surrogate splits for CART)
- Major decision: labor-intensive strategy below or fully automated approach
  - MARS, PP, neural network, trees, etc.
  - Advantage: can bootstrap entire process, ⇓ time
  - Possible disadvantage: interpretation, inference

Strategy, Continued

- For each predictor postulate the allowable complexity or nonlinearity (# d.f.)
  - Can use generalized Spearman ρ^2 or Hoeffding *D* to estimate # d.f. a predictor may be worth
- Estimate total # parameters that can be reliably estimated
- Data reduction or shrinkage to $\Downarrow$ d.f.

Strategy, Continued

- Relax linearity assumption using regression splines or nonparametric regression (GAM)
- Add pre-specified interaction effects
- Examine distributional assumptions
- Don't remove insignificant terms or do stepwise variable selection
- Validate model using the bootstrap
  - overfitting-corrected R^2, calibration curve

Strategy, Concluded

- Examine sensitivity of predictor effects to imputation
- Display effects graphically
  - effect shape plots (3d if interactions)
  - inter-quartile-range odds, hazard ratios + C.L.
  - nomogram

In time, all things shall become unclear

Some Useful Tools for Pre-Modeling

- Hierarchical clustering algorithms
  - relationships among predictors
  - simultaneous missings
- Nonlinear principal components/imputation
- Recursive partitioning (CART)
  - Finding tendencies for subjects to have missing Y or Xs
  - Nonparametric imputation model
- Nonparametric regression (loess)

S-Plus Software and References

- Add-on libraries Hmisc and Design
- See hesweb1.med.virginia.edu/biostat/s
- See Harrell *et al*. Stat in Med 1996, 1998 or Regression Modeling Strategies (Springer, 2001) for background and detailed case studies

S-PLUS Functions Demonstrated

- Builtin: bwplot
- Hmisc library: datadensity, hist.data.frame, ecdf, naclus, varclus, transcan, plsmo, summary.formula, fit.mult.impute
- Design library: lrm, anova, summary, bootcov, validate, calibrate, nomogram, Function
- rpart library (Atkinson & Therneau)

What the Critics Say

Interesting, if true

Regression Modeling and Pre-Modeling Strategies

Frank E Harrell Jr
Professor of Biostatistics and Statistics
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
University of Virginia School of Medicine

Multivariable regression models are powerful tools for uncovering strength and shapes of relationships between predictors and response, for making tests of partial association, and for predicting outcomes of future subjects. Successful modeling must address problems such as missing values in the predictors, overfitting (solved by data reduction or shrinkage), relaxing linearity assumptions, fitting nonlinear interactions, and presenting results graphically so that non-statisticians can understand complex models.

There are a number of pre-modeling steps that can help the modeling process including (1) displaying distributions of predictors and response, (2) understanding inter-relationships among predictors, (3) displaying how missing values occur simultaneously among several predictors, and (4) uncovering what kinds of subjects tend to have missing values for some of the predictors. All of these are helpful in deciding whether and how to impute missing values. In addition, nonparametric regression is an excellent exploratory tool.

Once one is ready to model the response, techniques such as regression splines, pooled effect tests (main effects combined with interactions), automatic tests of linearity, shrinkage (penalized maximum likelihood estimation), and various graphical displays of how the predictors affect the response are very useful.

This talk will focus on live demonstrations of various pre-modeling and modeling strategies, using S-PLUS and add-on libraries `Hmisc` and `Design` (available from `Statlib`). Two datasets will be used to demonstrate the methods. One of them contains information on passengers on the Titanic, where we will answer the question "were women and children first?".