# Directions in Statistical Methodology for Multivariable Predictive Modeling

Frank E Harrell Jr
University of Virginia
Seattle WA 19May98

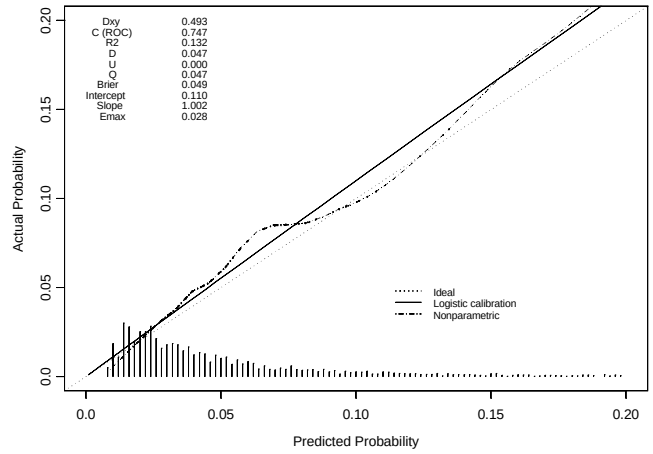
# Overview of Modeling Process

- Model selection
- Regression shape
- Diagnostics - lack of fit
- Quantify precision / Inference
- Validation - quantify accuracy
- Presentation
- Harrell *et al.* Stat in Med 1996, 1998

# Accuracy Measures

- No binning (ECDF, not histogram)
- Summary indexes well dev. for binary Y
  - ROC area (concordance prob.)
  - Brier quadratic prob. score, $R^2$
- Nonparametric calibration plot
  - Uses nonparametric regression (loess)
  - Summarize using average |pred. - obs.|
- For survival models, rank correlation index can allow censoring; other indexes needed

# Nonparametric Calibration



## Empirical Modeling Tools

- Nonparametric regression (loess)
  - single X or generalized additive model
  - can transform both X and Y simultaneously
- Restricted cubic splines (piecewise cubic poly. with linear tails)
- Tensor spline surfaces in multiple predictors

## Adequacy of Biomath. Models

- Can embed pre-specified model into an empirical model, e.g.

  $E(Y|X) = 1 - \exp(x^2/h + \text{spline}(x))$
- Test coefficients of spline(x) for significant additional predictive information

## Model Uncertainty / Selection

- When many models tested, "best" model may be rated optimistically
- Standard errors, P-values assuming "final" model pre-specified are inappropriate
- Need to add model uncertainty to confidence intervals
- Averaging competing models may improve prediction

## Bayesian Modeling

- Incorporate prior beliefs about shapes of effects of Xs, parameter values
- Smoothing toward prior knowledge of mechanisms
- Exact inference without reliance on large-sample $\chi^2$ or Gaussian distributions

## Penalized MLE

- An empirical Bayesian approach
- Discounts parameter estimates toward 0
- Estimate how many d.f. can be reliably estimated C
  - Cross-validation or effective AIC
- Estimate that many effective d.f.

## Bootstrap for S.E., C.L.

- Large sample normal theory may be inadequate for nonlinear models
- Nonparametric bootstrap can estimate S.E. of individual parameter estimates or of predicted values
- Involves sampling with replacement from original data, re-fitting model many times

## Multiple Observations / Subject

- May not affect model parameter estimates very much
- Will deflate variances (C.I. too narrow)
- Can correct variances using “cluster sandwich” estimator or cluster bootstrap
- May want to include subject’s track record to date as a predictor

## Bootstrap Model Validation

- Validate summary indexes ($R^2$)
- Calibration plot corrected for overfitting

## Time-Dependent Covariables

- Excellent for experimentally controlled conditions
- Useful for understanding patterns of risk if covariables are “internal”
- E.g. understand how to use VGE grade profile
- Not very useful for prospective use

## Other Models

- Family of logistic regression models for ordinal response
  - Uses ordering of Y
  - No grouping of Y required
  - Decide whether time until symptom more important than its severity
- Accelerated failure time models
- Other log-logistic formulations
- Nonparametric regression

## Other Models, Continued

- Berridge & Whitehead Stat in Med 1991
- Combines ordinal logistic model and Cox survival model to jointly model time until and severity of event

## Summary

- Explosion of statistical methodology for model fitting, diagnostics, validation
- Avoid categorization - use smoothers
- Software (e.g., S-PLUS) is beginning to keep up with stat. developments
- Opportunities for merging empirical and biomathematical modeling

## Missing Predictor Values

- Omitting incomplete records causes bias and decrease in precision
- Imputing missing values with a mean results in bias if predictors are correlated
- Customized imputation models
- Full likelihood or Bayesian approach

## Outline

- Overview of modeling process
- Measures of predictive accuracy
- New empirical modeling tools
- Testing adequacy of biomath. Models
- Model selection / uncertainty
- Bayesian modeling
- Penalized maximum likelihood estimation

## Outline

- Handling missing predictor values
- Robust s.e., C.L. using bootstrap
- Multiple observations per subject
- Bootstrap model validation
- Time dependent covariables
- Other models