TITLE: Regression Modeling Strategies

NAME AND ADDRESS OF PRESENTER:
Frank E Harrell Jr
Professor of Biostatistics
Vanderbilt University School of Medicine
2525 West End Avenue, Suite 1100
Nashville TN 37203

`fh@fharrell.com`

ABSTRACT: All standard regression models have assumptions that must be verified for the model to have power to test hypotheses and for it to be able to predict accurately. Of the principal assumptions (linearity, additivity, distributional), this course will emphasize methods for assessing and satisfying the first two. Practical but powerful tools are presented for validating model assumptions and presenting model results. This course provides methods for estimating the shape of the relationship between predictors and response using the widely applicable method of augmenting the design matrix using restricted cubic splines. Even when assumptions are satisfied, overfitting can ruin a model's predictive ability for future observations. Methods for data reduction will be introduced to deal with the common case where the number of potential predictors is large in comparison with the number of observations. Methods of model validation (bootstrap and cross–validation) will be covered, as will auxiliary topics such as modeling interaction surfaces, efficiently utilizing partial covariable data by using multiple imputation, variable selection, overly influential observations, collinearity, and shrinkage, and a brief introduction to the R `rms` package for handling these problems. The methods covered will apply to almost any regression model, including ordinary least squares, longitudinal models, logistic regression models, ordinal regression, quantile regression, longitudinal data analysis, and survival models. Statistical models will be contrasted with machine learning so that the student can make an informed choice of predictive tools.

OUTLINE:

1. Introduction; Advantages of prediction over classification
2. Hypothesis Testing vs. Estimation vs. Prediction vs. Classification
3. How Many Degrees of Freedom does a Data Mining Procedure Actually Have?
4. Advantages of regression models and contrasts with machine learning
5. Regression Model Notation

6. Model Formulations
7. Interpreting Model Parameters
   (a) Nominal Predictors
   (b) Interactions
8. Relaxing Linearity Assumption for Continuous Predictors
   (a) Categorization is not an alternative
   (b) Simple Nonlinear Terms
   (c) Splines for Estimating Shape of Regression Function and Determining Predictor Transformations
   (d) Cubic Spline Functions
   (e) Restricted Cubic Splines
   (f) Choosing Number and Position of Knots
   (g) Nonparametric smoothers and regression trees
   (h) Advantages of Splines over Other Methods
9. New Directions in Predictive Modeling
10. How to Make the Choice of Statistical Models vs. Machine Learning
11. Multiple Degree of Freedom Tests of Association
12. Assessment of Model Fit
    (a) Regression Assumptions
    (b) Modeling and Testing Interactions
13. Missing Data
    (a) Types of Missing Data
    (b) Problems Caused by Simple Solutions
    (c) Multiple Imputation
14. Multivariable Modeling Strategy
    (a) Why and How To Pre-specify Model Complexity
    (b) Problems Caused by Ordinary Stepwise Variable Selection
    (c) Collinearity
    (d) Shrinkage
    (e) Data Reduction
    (f) Overly Influential Observations
    (g) Some Useful Modeling Strategies for
        i. Prediction
        ii. Estimation
        iii. Hypothesis Testing
15. Overview of the Bootstrap
16. Model Validation
    (a) Cross-validation

   (b) Bootstrap
17. Graphical Methods for Interpreting Complex Regression Fits
18. Detailed Case Studies
   (a) Generalized Least Squares and Bayesian Semiparametric Proportional Odds Models for Longitudinal Data
   (b) Ordinal Regression for Continuous Y: Predicting glycohemoglobin (and pre-diabetes) from body size characteristics using NHANES data
   (c) Binary Logistic Regression: Survival Patterns of Passengers on the Titanic
   (d) Survival Modeling

More details are found at `hbiostat.org/doc/rms`.

Target Audience: Statisticians and persons from other quantitative disciplines who are interested in multivariable regression analysis of univariate responses, in developing, validating, and graphically describing multivariable predictive models and in covariable adjustment in clinical trials. The course will be of particular interest to applied statisticians and developers of applied statistics methodology, graduate students, clinical and preclinical biostatisticians, health services and outcomes researchers, econometricians, psychometricians, and quantitative epidemiologists. A good command of ordinary multiple regression is a prerequisite.

LEARNING OUTCOMES: Students will

1. be able to fit multivariable regression models:
   (a) accurately
   (b) in a way the sample size will allow, without overfitting
   (c) uncovering complex non–linear or non–additive relationships
   (d) testing for and quantifying the association between one or more predictors and the response, with possible adjustment for other factors
   (e) making maximum use of partial data rather than deleting observations containing missing variables
2. be able to validate models for predictive accuracy and to detect overfitting and understand problems caused by overfitting.
3. learn techniques of "safe data mining" in which significance levels, confidence limits, and measures such as R^2 have the claimed properties.
4. learn how to interpret fitted models using both parameter estimates and graphics
5. learn about the advantages of semiparametric ordinal models for continuous Y

6. learn about some of the differences between frequentist and Bayesian approaches to statistical modeling
7. learn differences between machine learning and statistical models, and how to determine the better approach depending on the nature of the problem
8. gain an appreciation for how study design and causal inference needs to drive model formulation

Content and Instructional Methods: Extensive and tested handouts will be given to students. The course will be informal enough for students to be able to ask questions throughout the day. The style will be a mixture of lecture and presentation of moderately comprehensive case studies. Handouts make heavy use of graphics to facilitate learning. The presentation and handouts show output from R functions but software use is not covered in detail in the course. Students who are interested in later using free R software to run examples presented in the case studies may do so by installing the R `rms` package written by the presenter, available at `www.r-project.org`.

PRESENTER: Dr. Harrell is Professor of Biostatistics, Founding Chair of the Department of Biostatistics of Vanderbilt University School of Medicine, and was Expert Statistical Advisor, Office of Biostatistics, Center for Drug Evaluation and Research, US FDA from 2016-2020. Prior to starting the new department in 2003 he was Chief of the Division Biostatistics and Epidemiology in the Department of Health Evaluation Sciences, University of Virginia School of Medicine. Prior to coming to the University of Virginia in 1996 he was in the Division of Biometry at Duke University Medical Center for 17 years. He received his Ph.D. in biostatistics from the University of North Carolina, Chapel Hill in 1979, where he studied under P.K. Sen. Dr. Harrell's interests include statistical modeling and model validation, statistical computing and graphics, reproducible research, survival analysis, clinical trials, health services and outcomes research, medical diagnostic and prognostic models, bootstrapping, missing data, and Bayesian modeling. He is an associate editor of *Statistics in Medicine*, a member of the editorial board for *American Heart Journal*, and a member of the Scientific Advisory Board, for *Science Translational Medicine*. He is author of the book *Regression Modeling Strategies, Second Edition* (Springer, 2015) and teaches courses in biostatistical modeling. He is a Fellow of the American Statistical Association and was the recipient of the ASA's WJ Dixon award for excellence in statistical consulting in 2014. He is active on Twitter under `@f2harrell` and is active on `stats.stackexchange.com`. He leads `datamethods.org` for in-depth discussion of data-related methodologies with clinical investigators, health services researchers, medical decision makers, epidemiologists, and other researchers.

TEXTBOOK: Harrell, F.E.(2015): Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis, second edition. New York: Springer.

SOFTWARE: R (not used "live")