Regression Modeling Strategies

Frank E. Harrell Jr.
School of Medicine, Department of Biostatistics
Vanderbilt University

Regression models are frequently used to develop diagnostic, prognostic, and health resource utilization models in clinical, health services, outcomes, pharmacoeconomic, and epidemiologic research, and in a multitude of non-health-related areas. Regression models are also used to adjust for patient heterogeneity in randomized clinical trials, to obtain tests that are more powerful and valid than unadjusted treatment comparisons.

Models must be flexible enough to fit nonlinear and non-additive relationships, but unless the sample size is enormous, the approach to modeling must avoid common problems with data mining or data dredging that result in overfitting and a failure of the predictive model to validate on new subjects.

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 short 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.

The morning sessions will cover modeling and model validation methods; the afternoon sessions will cover the overall modeling strategy, examples of displaying effects of predictors and of presenting models graphically to non-statisticians, and case studies.

The first part of the course presents the following elements of multivariable predictive modeling for a single response variable: using regression splines to relax linearity assumptions, perils of variable selection and overfitting, where to spend degrees of freedom, shrinkage, imputation of missing data, data reduction, and interaction surfaces. Then a default overall modeling strategy will be described. This is followed by methods for graphically understanding models (e.g., using nomograms) and using re-sampling to estimate a model's likely performance on new data. Then the freely available R/S-Plus Design library will be overviewed. This library facilitates most of the steps of the modeling process. Two case studies will be presented: an exploration of the survival status of Titanic passengers and using least squares to model the 1992 presidential election results.

The methods covered in this course will apply to almost any regression model, including ordinary least squares, logistic regression models, and survival models.

Instructor: Dr. Harrell is Professor and Chair of the Department of Biostatistics in the School of Medicine at Vanderbilt University in Nashville, Tennessee. 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 has been involved in statistical computing since 1969 and is the author of many R/S-Plus functions and SAS procedures. Since 1973 he has been involved in medical applications of statistics, especially in the area of survival analysis and clinical prediction modeling. He is the author of the book on which the course is based, Regression Modeling

Strategies. He is an editorial consultant for the Journal of Clinical Epidemiology, is on the editorial board of Statistics in Medicine, is co-managing editor of the new journal Health Services and Outcomes Research Methodology, and is a consultant to FDA.