HES 704: Biostatistical Modeling (4 credits)

Frank E. Harrell, Jr. (fharrell@virginia.edu)
Professor of Biostatistics and Statistics
Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences
School of Medicine
924-8712
Office Hour: 10-11a Tuesdays

Thomas Gwise (teg7a@virginia.edu)
Graduate Student in the Department of Statistics
Teaching Assistant
Room 106 Halsey Hall
924-3136
Office Hours: M 1-2p, Tu Th 3:30-5:00p

15 January - 28 April 2003
Monday, Wednesday, Friday 9:00-10:15a
Help Sessions Friday 10:15-11:00a
Health Evaluation Sciences Classroom, 3rd Floor Hospital West

Class Web Page:
http://hesweb1.med.virginia.edu/biostat/teaching/biostat.mod

Biostatistical Modeling is a required course for students in the Community and Public Health and the Clinical Investigation and Patient-oriented Research Tracks of the M.S. in Health Evaluation Sciences program and is recommended for the Informatics in Medicine track. This course covers many aspects of multivariable regression modeling as it is commonly used in prognostic, diagnostic, and epidemiologic modeling and in clinical trials.

Motivation Accurate estimation of patient prognosis or of the probability of a disease or other outcomes is important for many reasons. First, prognostic estimates can be used to inform the patient about likely outcomes of her disease. Second, the physician can use estimates of diagnosis or prognosis as a guide for ordering additional tests and selecting appropriate therapies. Third, outcome assessments are useful in the evaluation of technologies; for example, diagnostic estimates derived both with and without using the results of a given test can be compared to measure the incremental diagnostic information provided by that test over what is provided by prior information. Fourth, a researcher may want to estimate the effect of a single factor (e.g., treatment given) on outcomes in an observational study in which many uncontrolled confounding factors are also measured. Here the simultaneous effects of the uncontrolled variables must be controlled (held constant mathematically if using a regression model) so that the effect of the factor of interest can be more purely estimated. An analysis of how variables (especially continuous ones) affect the patient outcomes of interest is necessary to ascertain how to control their effects. Fifth, predictive modeling

is useful in designing randomized clinical trials. Both the decision concerning which patients to randomize and the design of the randomization process (e.g., stratified randomization using prognostic factors) are aided by the availability of accurate prognostic estimates before randomization. Lastly, accurate prognostic models can be used to test for differential therapeutic benefit or to estimate the clinical benefit for an individual patient in a clinical trial, taking into account the fact that low-risk patients must have less absolute benefit (e.g., lower change in survival probability).

To accomplish these objectives, researchers must create multivariable models that accurately reflect the patterns existing in the underlying data and that are valid when applied to comparable data in other settings or institutions. Models may be inaccurate due to violation of assumptions, omission of important predictors, high frequency of missing data and/or improper imputation methods, and especially with small datasets, overfitting.

Description This four semester hour course will first cover the basics of multivariable regression models including interpretation of regression coefficients, coding of categorical predictors, meaning of linearity assumptions, estimating the relationships between two variables nonparametrically, and coding and interpretation of interaction terms. These concepts will be taught using the ordinary multiple regression model. Other types of regression models are increasingly being used in developing clinical prediction models for diagnosis, prognosis, and other applications in epidemiology, health services research, health economics, and clinical trials. Popular models include logistic models for binary and ordinal responses, which will be the principal models covered. All regression models have assumptions that must be verified for them to have power to test hypotheses and 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 as these methods apply to all regression models. To deal with the linearity assumption, this course provides methods for estimating the shape of the relationship between predictors and response using the widely applicable method of piecewise polynomials. Emphasis will be given to interpreting fitted models using effect plots (e.g., odds ratio charts) and nomograms.

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 introduced, as will auxiliary topics such as modeling interaction surfaces, dealing with missing data, variable selection, collinearity, and shrinkage. All methods covered will apply to almost any regression model.

The course will include detailed case studies in developing, validating, and interpreting clinical prediction and epidemiologic models.

Prerequisites Introduction to Biostatistics (HES 700), Statistical Computing and Graphics (HES 703/STAT 301/501), and a good command of algebra. Before the third class, read Rosner Sections 10.1-10.2 and 10.6.1. Ignore Yates' continuity correction (don't subtract $\frac{n}{2}$ or $\frac{1}{2}$ from any of the formulas given). Yates' correction makes the Pearson χ^2 statistic behave more like Fisher's exact test, i.e., makes it unduly conservative. Also, don't worry about using the χ^2 distribution when an expected cell frequency < 5; this is unduly cautious. We will emphasize χ^2 tests of association, not tests of goodness of fit. In the 2×2 table case, the χ^2 test of association between rows and columns is identical to the normal-theory-based test for differences between two proportions.

Learning Objectives To become familiar with modern methods for fitting multivariable regression models (1) accurately; (2) in a way the sample size will allow, without overfitting; (3) uncovering complex non-linear or non-additive relationships; and (4) testing for and quantifying the association between one or more predictors and the response, with possible adjustment for other factors. Students will be introduced to the bootstrap and will learn how to deal with missing data, how to validate models for predictive accuracy and to detect overfitting, will be able to interpret fitted models using both parameter estimates and graphics, and will be able to critique the literature to determine when models are likely to be unreliable.

Texts

1. Rosner B. *Fundamentals of Biostatistics, 5th Edition.* New York: Duxbury Press, 2000.
2. Harrell FE. *Regression Modeling Strategies with Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York: Springer; 2001. Book available at main UVa bookstore.
3. Katz, Mitchell H. *Multivariable Analysis: A Practical Guide for Clinicians.* Cambridge University Press, 1999.
4. Alzola CF, Harrell FE: *An Introduction to S and to the Hmisc and Design Libraries.* Available from the Biostatistics web page or from the UVa bookstore, 2003.

Two course packs containing lecture notes and articles to read are available from the Copy Shop on Elliewood Avenue.

Recommended Supplemental Reading

1. Feinstein A. *Multivariable Analysis.* New Haven: Yale University Press, 1996.

2. Kleinbaum DG. *Logistic Regression: A Self-Learning Text.* New York: Springer-Verlag, 1994.

3. Collett, D. *Modelling Binary Data.* London: Chapman & Hall, 1994.

Datasets

1. Datasets from Rosner's *Fundamentals of Biostatistics,* 5^{th} *edition.*

2. Other datasets from web page

3. Students are encouraged to find their own datasets for the final project

See `Rosner` and `Datasets` area under `Teaching Materials` on the web page. Several of the Rosner datasets have already been converted to S data frames. These are under `dumpdata.sdd` under the `Rosner` area. You can use these for assignments in which you are not asked to create the data frames.

Class Announcements See `biostat.virginia.edu/discus` for clarifications of assignments and notes, news about updates to the web page, class schedule changes, software bug corrections, and other information.

Class Discussion Group See the **Biostatistical Modeling** section of `biostat.virginia.edu/discus`. This is an excellent way to post questions and answers because all postings and replies are "threaded" (categorized), so you can return to the discussion group weeks later and still benefit from seeing answers regarding a specific topic. The discussion group is an excellent way to keep in touch with the class and even more to ask and answer questions. I hope that all students will use it to

- ask or answer any question whatsoever related to group assignments
- ask or answer any logistical or purely technical questions related to individual work assignments
- ask or answer any questions about modeling or statistical computing concepts that are not directly related to a pending individual work assignment

Be **sure** to check existing topics for posting your message, to avoid creating any unnecessary new topics that will make it more difficult for others to navigate the discussion board.

Software S-PLUS on Microsoft Windows (version 2000, release 3 or version 6.x), UNIX (version 3.4 or 6.x) or Linux (version 6.x), or R on any platform, with add-on S libraries `Hmisc` and `Design`.

Class Format Lecture, interactive computer demonstrations, and occasional computer labs

Labs Computer labs available for doing homework:
ACHS (Academic Computing Health Sciences) 8a-5p Mon-Fri (3rd floor Hospital West by DHES)
Health Sciences Library Computer Lab

Help Sessions Most Fridays 10:15-11:00a

Schedule exceptions No class 3 Mar - 7 Mar (Spring recess).

Grading Assignments are due by 5p on date listed. Projects must be done independently and must include the honor pledge. Small homeworks may be done in groups, with all group participants signing a single copy verifying their participation. Projects count 2 × group homeworks (the largest project may be counted more). Final project counts 4× group homeworks. One letter grade is deducted per day late unless prior arrangements are made with the instructor (these prior arrangements are allowed once or twice per semester for students with excellent attendance records otherwise). Students not having at least a B- average on individual work (projects and quizzes) may be assigned extra individual projects.

Work must be as concise as clear communication will allow. Students must use the S to LaTeX to `PDF` convertor on the class web page to typeset their work. This involves heavily commenting S script or report files with comment lines (lines beginning with #). There is online help for these procedures including example template files.

Exams Weekly quizzes on random days, counting equal to a group assignment. The lowest quiz grade for the semester will not be counted in the final grade. There will be no makeups for quizzes unless prior arrangements have been approved by the instructor. There may be a midterm exam. A quiz lasts about 15 minutes and tests modeling concepts. The concepts from which quiz questions are drawn and which you are responsible to master are listed at
`http://hesweb1.med.virginia.edu/biostat/teaching/biostat.mod/concepts.html`
on the class Web page. It is exceptionally important to understand concepts taught in the course, as students not understanding these concepts are frequently unable to properly analyze data in their other courses or theses, or in their final project for this course.

Be sure that if you do not understand the concepts you take advantage of the Teaching Assistant's and Instructor's office hours, ask questions in class and

the Friday help session, and that you make use of the discussion board. This discussion group is ideal for asking questions about concepts as well as about details.

Assistance Questions for instructor outside class - E-mail or phone for appointment but try to come to the lab help session or during open office hours. First try the Discussion Board, so that other students can see your question and the answers you get from others. Assistance will only be given on projects when the same assistance is available to all students (i.e., during the help session or through discussion group).

Anonymous Feedback to Instructor See the class web page.

Course Outline, Projects, and Approximate Schedule

Numbers to the right of topics indicate sequential lecture numbers.
Hn stands for Harrell Chapter n. Kn stands for Katz Chapter n.

1. Introduction ($H1$) **(1)**
   (a) Course overview and logistics
   (b) Hypothesis testing vs. estimation vs. prediction ($K1$)
   (c) Examples of multivariable prediction problems ($K2$)
   (d) Study planning considerations ($K2, K7.1$)
   (e) Choice of model ($K3.1$)
2. Multiple linear regression and least squares (Rosner Chapter 11, 12.5-12.6)
   (a) Introduction and general concepts **(3)**
   (b) Least squares fitting
   (c) Inferences about parameters **(4)**
   (d) Interval estimation **(5)**
   (e) Assessing goodness of fit
   (f) Multiple regression **(6)**
   (g) `Design` library S functions for multiple regression — Preview of $H6$ **(8)**
   (h) Case study: Lead exposure **(10)**
   (i) Correlation
   (j) ANOVA, Two-way ANOVA and introduction to interaction **(11)**
   (k) Allowing slopes to vary by category of subject (simple interaction)

   Project: Develop multivariable *linear* regression model from a "clean" dataset using least squares **Assigned 12 Due 15**
3. General methods for multivariable models ($H2, K4$)
   (a) Notation for general regression models **(13)**
   (b) Model formulations
   (c) Interpreting model parameters
      i. nominal predictors (Rosner Section 12.5.2, $K8.1 - 8.2$)
      ii. interactions
   (d) Relaxing linearity assumption for continuous predictors **(14)**
      i. nonparametric smoothing
      ii. simple nonlinear terms

9. Case study in data reduction and missing value imputation ($H8$ up until discussion of principal components) (H14.2,14.3)

   Project: Understanding interrelationships of predictor variables, dealing with missing data, developing and validating a multiple regression model using least squares **Assigned 23 Due 28**

10. Maximum Likelihood Estimation ($H9$ up until 9.3, $K8.13 - 8.15$) **(30)**

11. Binary Logistic Model ($H10$, Background reading: Rosner Section 13.1-13.2, 13.3.3, 13.4.1, 13.7) **(30)**

    (a) Model

    (b) Odds ratios

    (c) Student presentations **(32)**

    (d) Special residual plots **(33)**

    (e) Applications of general methods

    (f) Graphically presenting model **(34)**

    (g) Case studies

    Project: Develop and validate a binary logistic regression model **Assigned 32 Due 35**

12. Proportional Odds Ordinal Logistic Models ($H13.1 - 13.3$) **(36)**

    (a) Model

    (b) Odds ratios

    (c) Applications of general methods

    (d) Case study ($H14 - 14.3$)

    Assignment: Interpret an analysis that used a proportional odds ordinal logistic model **Assigned 38 Due 41**

13. Brief Introduction to Survival Analysis (Rosner 14.8-14.11, $H16, K3.3, 7, 8.4-8.6, 10.6$) **(37)**

    (a) Survival data and right-censoring

    (b) log-rank test for comparing two groups

    (c) Cox regression model ($H19.1$) **(38)**

14. Other Case Studies and Labs **(40-41)**