**Methods in Biostatistics IV**
**140.654**

4th Quarter, 2007-2008

Instructor: Karen Bandeen-Roche, Tel. Extension 5-1166
e-mail kbandeen@jhsph.edu; Room E-3624

Teaching Assistants: Yong Chen - yonchen@jhsph.edu
Marie Thoma - mthoma@jhsph.edu
Hao Wu - hwu@jhsph.edu

## I. COURSE DESCRIPTION

Biostatistics 140.654 is a course in generalized linear regression analysis. Foundational topics of the course include: generalized linear models and their uses; maximum likelihood estimation and inference; and model assumptions, diagnosis, and interpretation. Specific topics include: logistic and Poisson regression, grouped and individual-level data, analysis for unmatched and matched case-control studies, analysis for cohort studies, and introductory survival analysis.

## II. COURSE OBJECTIVES - Biostatistics 140.654 acquaints students with:

- the definition, statistical assumptions, and interpretation of generalized linear regression models, specifically including logistic and Poisson regression; as well as loglinear modeling

- maximum likelihood (ML), conditional likelihood, and partial likelihood estimation, including the iteratively reweighted least squares implementation of ML

- standard methods for making inferences on model parameters, including Wald testing and confidence interval construction, and likelihood ratio / deviance testing

Students will develop skills to:

- build and fit generalized linear regression models and survival analyses using standard statistical software;

- diagnose model appropriateness for description, inference, and prediction;

- analyze case-control, rate, & cohort data, recognizing special features of each;

- sensibly interpret fits and inference for statistical and scientific importance.

## III. COURSE REFERENCES

Textbook: **FEH**: Harrell, F.E. (2001), Regression Modeling Strategies, With Applications to Linear Models, Logistic Regression, and Survival Analysis, New York: Springer.

Suggested Supplemental References

Breslow, N.E. & Day, N.E. (1980), The Analysis of Case-Control Studies, Oxford University Press.

Breslow, N.E. and Day, N.E. (1987), Design and Analysis of Cohort Studies, Oxford Univ. Press.

Cox, D.R. and Snell, E.J. (1981), Applied Statistics, Principles and Examples, New York: Chapman and Hall.

Dobson A.J. (1983), An Introduction to Generalized Linear Models, New York: Wiley.

**Hosmer, D.W. & Lemeshow, S. (2000), Applied Logistic Regression, 2nd edition, New York: Wiley.**
http://www3.interscience.wiley.com/cgi-bin/bookhome/109855848

**McCullagh P. and Nelder J.A. (1989), Generalized Linear Models, 2nd. Ed., Chapman and Hall.**

Santner T.J. and Duffy D.E. (1989), The Statistical Analysis of Discrete Data, New York: Springer-Verlag.

## IV. ADMINISTRATION

### A. Instruction schedule

| **Type** | **Instructor** | Time/Place |
|---|---|---|
| Lecture | Bandeen-Roche | Tu/Th 10:30-12:00<br>Room W4030 |
| Lab | All | Tu 12:15-1:15 ; W4030 |
| Office Hours | Bandeen-Roche | Thursday **1:15-2:30 - ?**, E3624 |
| | Chen, Thoma, Wu (rotating) | Monday **3:00-4:00 PM - ?** Location TBA |

### B. Course requirements and evaluation

Homeworks 40%

> Same policy as for Biostatistics 653–Best 3 out of 4 EXCEPT

that Homework 4 MUST BE SUBMITTED

Project (1) and Final (1) Exam 60%
> Weighted to higher of: 100% Final **OR** 50% Final, 50% Project
> e.g., Project is optional, EXCEPT...
> ... project is **mandatory** for Biostatistics degree students

Guaranteed grades are as for Biostatistics 140.653. Curve may also be implemented.

There will be no extra or make-up credit, except as may occasionally be offered on homework assignments or exams

C. Project

A data analysis project may be submitted for 30% course credit. The primary analytic outcome(s) should be binary or counted, so that the project will draw primarily on Biostatistics 654. The project consists of selecting a data set (preferably related to your own research or field), posing a substantive question of interest, analyzing the data to address the question, and writing up findings in a report. The report should include:

1. Introduction/Background (1-2 pages): describe (i) the scientific problem of interest; (ii) how the data set you will analyze arose and why it well addresses the scientific problem; (iii) motivation for specific potential confounders, mediators or effect modifiers; (iv) references to other work.

2. Aims (1/2 page): motivation and statement of the specific question(s) that you will address in your analysis. This section should make clear whether the primary goal is descriptive, inferential, or predictive; state any hypotheses.

3. Methods (2 pages): (i) operationalization of the problem within a statistical model or sequence of models; (ii) description of analyses to be applied, including how each addresses the scientific question(s) or ensures meaningful interpretation.

4. Analysis (2-3 pages text, plus supporting tables/graphs): a report of analyses conducted, including description/graphs and formal inference.

5. Conclusion (1-2 pages): summary/interpretation of findings, discussion of study limitations and implications.

**GRADING CRITERIA**: Each section of the project will be graded as A, B, or C level on the criteria: clear/engaging narrative, correctness, completeness. The analysis section will count for 50% of score and the other sections equally for the other 50%. I will deduct credit for a trivial project topic; if you are concerned whether your project has sufficient content, please discuss it beforehand with Dr. Bandeen-Roche.

DUE DATE**: 12:00 noon, May 15**.

D. Ethics policy: homework assignments

Please study together, and feel free to talk to one another about homework assignments. The mutual instruction that student colleagues give each other by doing this is among the most valuable that can be achieved. However, it is expected that homework assignments will be implemented and written up independently. Specifically, please do not share analytic code or output. Please do not collaborate on write-up and interpretation. Please do not access or use solutions from any source before your homework assignment is submitted for grading. Thanks.

E. Ethics policy: project

The project must be your own work. Papers that involve research in collaboration with others is permissible provided that all colleagues are acknowledged, you conduct all analyses you report independently, and you write up the work **entirely** on your own. The paper must follow ethical standards of scientific publication. Please cite references appropriately. Any narrative that is not your own must be placed in quotes and attributed to the source. Thanks in advance.

F. Late policy

Course requirement due dates for the term are provided below; occasionally they are modified for all based on course progress. Homeworks must be submitted on time to receive credit. Except for extraordinary crises, there will be no exceptions.

In general exams must also be taken at the scheduled time. At the instructor's discretion, exceptions will be made for unforeseen personal illness, family health emergency or other crisis, or for unavoidable conflicting trips **that are agreed at least three weeks in advance** of the exam at issue.

V. SCHEDULE

| | |
|---|---|
| March 25: | Background to Generalized linear models<br>Weighted least squares<br>Robust variance estimation<br>Transformation<br>Motivation: Why more than linear regression<br>Reading: Weisberg §5.1; Ch. 7 |
| March 27: | Introduction to Generalized linear models<br>Overview<br>Formulation/link functions<br>Maximum likelihood estimation, inference<br>Deviance<br>Reading: FEH Ch. 9; Article (McCullagh) |
| April 1: | Logistic regression: description<br>The logistic function<br>Parameter interpretation:<br>Simple<br>Multiple: Main, interactions<br>Nonlinear / smooth curves<br>Grouped, individual models<br>Reading: FEH Chapter 10.1 |
| April 3: | Multiple logistic regression–fitting & inference<br>ML fitting<br>Iteratively reweighted least squares<br>Wald inference<br>Inference using nested models, deviances<br>Deviance test distribution<br>Reading: FEH Chapter 10.2-3 |
| April 8: | Multiple logistic regression–model diagnosis<br>Goodness of fit<br>Leverage and influence<br>Residual checking<br>Case Study, part I<br>Reading: FEH Chapter 10.4-7 |
| April 9: | HOMEWORK 1 DUE, 5:00 PM, BIOSTAT OFFICE |

| | |
|---|---|
| April 10: | Multiple logistic regression: prediction; extensions<br>  Sensitivity/Specificity<br>  Receiver Operating Characteristic (ROC) curve<br>  Polytomous, ordinal logistic regression<br>Reading: FEH Chapter 10.8-9, 13; Articles (ROC) |
| April 15: | Model building<br>  Method overview<br>  Bias/variance tradeoff: AIC, BIC<br>  Case Study<br>Reading: FEH Chapter 11 |
| April 17: | Analysis of Event Counts: Poisson regression<br>  Poisson regression<br>  Negative binomial regression<br>Reading: Article |
| April 18: | HOMEWORK 2 DUE, 5:00 PM, BIOSTAT OFFICE |
| April 22: | Case-control studies<br>  Odds ratio equivalence<br>  Unmatched fitting, interpretation<br>  Example<br>Reading: H&L Chapter 6; article |
| April 24: | Matched case-control studies<br>  Setup: nuisance parameters<br>  Conditional logistic regression<br>  Fitting/Inference<br>Reading: H&L Chapter 7; article |
| April 29: | Cohort study analysis<br>  Incidence: beyond the logit link/collapsibility<br>  Censoring<br>  Rate/Cohort studies with Poisson regression<br>Reading: Article |
| May 1: | REVIEW |
| May 2: | HOMEWORK 3 DUE, 5:00 PM, BIOSTAT OFFICE |
| May 6: | EXAM |

May 8: Loglinear models
- Model
- Interpretation

Reading: Article

May 13: Loglinear models
- Estimation
- Hierarchical framework

Reading: Article

May 15: Causality versus association
- Paradigms defining causality
- Potential outcomes
- Propensity scoring

Reading: Articles (Holland; Rubin)

**May 15:** PROJECTS DUE, 12:00 NOON

**May 15**: HOMEWORK 4 DUE, 12:00 NOON