Methods in Biostatistics III - 140.653
3rd 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 140.653 introduces linear regression analysis for public health science. Foundational topics include: correlation, regression, and analysis of variance (ANOVA) models and their uses; least squares estimation and inference for parameters; model formulation, checking for adequacy, and interpretation; and making predictions. Topics are introduced using simple linear regression equations, then amplified in the context of multiple linear regression and matrices. Techniques are introduced for: identifying influential points; modeling variable adjustments, effect modification, and nonlinear relationships; and identifying and handling departures from basic model assumptions.

II. COURSE OBJECTIVES - By the end of the course a student should be familiar with:

- the definition and interpretation of the standard linear regression model;
- least squares estimation of parameters;
- appropriate methods for making scientific inferences, statistical assumptions that underlie the methods, and statistical properties of estimators, tests, and prediction strategies
- methods to describe fit of models to observed data.

The student should be able to:

- build regression models that address specific scientific questions using linear, polynomial, spline, and interacting relationships of multiple predictors with outcome variables;
- use models to make inferences about direct associations, confounding, effect modification, and statistical and scientific importance of findings
- correctly interpret and develop predictions from linear regression models;
- evaluate analyses for quality of description, inference, and predictions.

III. COURSE REFERENCES

Textbooks:

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

SW: Weisberg S. (2005), Applied Linear Regression, 3rd. Ed., New York: John Wiley & Sons: http://www3.interscience.wiley.com/cgi-bin/bookhome/109880490/

Suggested Supplemental Books:

Carroll, R. J. and Ruppert, D. (1988), Transformation and Weighting in Regression, New York, Chapman and Hall.

Draper, N. R. and Smith, H. (1998), Applied Regression Analysis, 3rd. Ed., New York: John Wiley & Sons.

Miller, R. G. (1986) Beyond ANOVA, Basics of Applied Statistics, New York: John Wiley & Sons.

Mosteller, F. and Tukey, J. W. (1977), Data Analysis and Regression: A Second Course in Statistics, Reading, MA: Addison-Wesley.

Scheffe', H. (1959), The Analysis of Variance, New York: John Wiley & Sons.

Seber, G. A. F. (1977), Linear Regression Analysis, New York: John Wiley & Sons.

Vittinghoff, E., Glidden, D.V., Shiboski, S.C., and McCulloch, C.E. (2004). Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models, New York: Springer.

IV. ADMINISTRATION

A. Instruction schedule

Type	**Instructor**	Time/Place
Lecture	Bandeen-Roche	Tu/Th 10:30-12:00 Room W4030
Lab	All	Tu 12:15-1:15; W4030
Office Hours	Bandeen-Roche	Th 4:00-5:30; E3624
	Chen, Thoma, Wu (rotating)	Monday 12:15-1:15 Location TBA

B. Course requirements and evaluation

Homework assignments (4) 40%

> **In lieu of late allowance**: Homework score will be calculated using the THREE assignments yielding the highest average score.

Midterm (1) and Final (1) Exam 60% (30% per exam)
(In-class)

Guaranteed grades:

A = 90% on both components
B = 80% on both components
C = 70% on both components

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. Ethics policy: homework assignments

Please feel free to study together and talk to one another about homework assignments. The mutual instruction that student colleagues so give each otheris 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. Thank you.

D. 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. Exceptions will be considered only for extended health, family, or other personal crises.

In general exams must also be taken at the scheduled time. At the instructor's discretion, exceptions will be made for 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

Jan. 22: **Introduction/overview**

Statistical modeling
Regression and correlation
Parameter interpretation: slopes; means
Analytic purposes

Reading: FEH Ch. 1; SW Ch. 1

Jan. 24: **Model and estimation: Simple linear regression**

Statement of model, assumptions
Estimation: Least Squares
"Quality" of estimation: Accuracy, precision

Reading: SW Ch. 2.1-2.4

Jan. 29: **Simple linear regression: Sample characteristics and random component estimation**

Isolated points; influence (“sensitivity”)

Decomposition of variance: ANOVA table

Residual variance estimation

Brief inference introduction

Reading: SW Ch 2.5-2.9

Jan. 31: **Multiple linear regression: Uses**

Multiple predictors

Direct versus total effects

Nonlinear relationships: Polynomials/splines

Categorical predictors: Dummy variables

Reading: FEH Ch. 2; SW Ch. 6.1-2

Feb. 1: **Problem Set 1 due 5:00 PM, Homework Lock Box**

Feb. 5: **Model and estimation: Multiple linear regression**

Statement of model, assumptions

Matrix specification

Least squares in the multiple covariate setting

Gauss-Markov theorem

Introduction to inference: variance components

Reading: SW Ch. 3.1-3.4

Feb. 7: **Inference in multiple linear regression**

t-based inference for individual parameters

Global/F-tests, regions for multiple parameters

Confidence intervals for contrasts, model

Reading: SW Ch. 3.5; scan Ch. 4

Feb. 12: **More on models with multiple covariates**

Adjustment

Effect modification / interaction

Mediation

Multiple comparisons

Reading: Revisit FEH Ch. 2

Feb. 13: **Problem Set 2 due 5:00 PM, Homework Lock Box**

Feb. 14: Case study / review

Feb. 19: **MIDTERM EXAM**

Feb. 21: **Model checking**

- Residual versus predicted plots
- Partial residual plots
- Outliers, influential points
- Standardized, studentized residuals

Reading: SW Ch. 8-9

Feb. 26: **Model checking: Two-stage regression**

- Partial correlation / Adjusted variable plots
- Inference in the face of assumption violations
  - Nonlinearity: transformations
  - Heteroscedasticity: transfrms, weighting
  - Correlation: robust variance

Reading: FEH Ch. 9; SW Ch 3.1; scan Chs 5 and 7

Feb. 28: **Prediction**

- Inference for fitted values; sums of coefficients
- Colinearity
- Multiple R-squared
- Confidence bands / prediction intervals

Reading: SW Ch. 2.8.3; 10.1

Feb. 29: **Problem Set 3 due 5:00 PM, Homework Lock Box**

Mar. 4: **Prediction, continued**

- Overfitting; cross-validation
- Mallows’ CP (bias-variance tradeoff)
- PRESS

Reading: FEH Ch. 5

Mar. 6: **Model building strategies**

- Parsimony
- Role of theory; variable groupings
- Data based methods: AIC, BIC
- Automated methods
- Extrapolation; Propensity scoring

Reading: FEH Ch. 4; SW Ch. 10.2-10.4

Mar. 11: **Case study / review**

Reading: FEH Ch. 7

Problem Set 4 due *in class*

Mar. 13: FINAL EXAM