Biost 533 / Stat 533
Theory of Linear Models

Syllabus
Spring, 2009

Instructor : Scott S. Emerson, M.D., Ph.D.
Professor of Biostatistics
Office : HSB F673
Phone : 616-6678
Email : semerson@u.washington.edu
Office hours : MW 9:00 - 10:00
(or by appointment)

Assistant : Michael Sachs (sachsmc@u.washington.edu)
Office hours : TBA

Time and Place : Lectures : MW 1:00 - 2:20 HSB K069

Class Web Pages: http://www.emersonstatistics.com/533/

The web page will be used to post homeworks, notices, handouts, etc. I urge you to check this site regularly. Questions that are submitted to me (via email or otherwise) that I think might be of general interest will have their answers posted on the web page, as well.

Prerequisites : matrix algebra, Stat 513, and Biost 515 or Stat 423 (or permission of instructor)

Texts:
Required : Seber and Lee, *Linear Regression Analysis*
Wiley, 2003

Attendance : Lectures : Highly recommended

Assignments : Written problem sets approximately weekly

Homework problems requiring a written solution will be due approximately weekly. Students are encouraged to seek help from the instructor, the TA or other students with the written homework problems. However, the work that is handed in should reflect only that student's work. That is, obtaining help from other students in order to learn the METHODS of solution is allowed, but copying another student's answer is NOT. Assignments handed in late will not be accepted. We reserve the right to grade only selected portions of the written homework.

Grading :

Written homeworks	40%
Midterm	30%
Final examination	30%

Course Objectives

It is assumed that students are familiar with matrix algebra and the basic properties of the multivariate normal distribution. The primary goal of this course is to develop the theory that is the basis for the analysis of data using regression models that involve linear predictors. Emphasis will be placed on that theory which is crucial to the application of linear regression analysis to a dataset and the theory that generalizes to other forms of regression. Less attention will be paid to the theoretical results that are generally only of interest when programming software to perform analyses.

At all times we are interested in the theoretical results that allow the greatest generalization of regression analysis. Hence, although at times we will develop theory based on strong parametric assumptions (i.e., when pretending that the regression models accurately model the data), we will ultimately want to be able to describe the behavior of those parametrically derived estimates in other settings (i.e., to be able to use theory to investigate the extent to which a regression method might be able to model a scientific question in a more distribution-free manner).

At the end of Biost/Stat 533, a student should have made significant progress toward being able to:

1. Identify the most common probability and statistical analysis models that involve regression on linear predictors (with emphasis on linear regression and analysis of variance). In particular:
   a. Understand distinctions between the use of common regression models when a) modeling data across covariate groups, and b) modeling an approach to answering scientific questions.
   b. Describe the difference between the scientific and statistical interpretation of models using linear predictors that contain only fixed effects and those that might also contain random effects.
   c. Describe the identifiability of regression parameters in specific models.
2. For least squares estimates (including ordinary, weighted, and general least squares):
   a. State the optimality properties of least squares estimates for data having arbitrary probability distributions, as well as when response variables are normally distributed. Describe the ways that optimality is affected by heteroscedasticity and/or dependence among the observed responses.
   b. State an asymptotic distribution of least squares estimates and the assumptions necessary for those asymptotic results to hold.
   c. State the small sample distribution of least squares estimates when response variables are normally distributed.
3. Describe the similarities and differences between least squares estimates of regression parameters and estimates that are derived from other methods, such as likelihood, quasi-likelihood, partial likelihood, and generalized estimating equations.
4. Provide valid statistical inference using regression parameter estimates. In particular:
   a. Identify testable hypotheses and estimable functions.
   b. Construct test statistics and estimators suitable for inference about those hypotheses and estimable functions. Discuss the use of ANOVA table and expected mean squares in the general linear model.
   c. Describe the assumptions necessary for the proposed inference to be statistically valid.
   d. Describe the impact that other predictors, both modeled and unmodeled, can have on the sampling distribution for a given slope parameter estimate. Discuss theoretical results that suggest optimal approaches to experimental design including randomization, blocking, and stratification.
   e. Describe the ways in which inference with fixed effects differs from inference in the presence of random effects.
5. Be able to provide theoretical justification for the behavior of regression parameter estimates under“model misspecification”, which might include
   a. Departures from assumptions about the independence / dependence of observations.

b. Departures from assumptions about linearity of functionals across covariate groups.

c. Departures from assumptions about distribution of observations within covariate groups.

d. Predictors measured with error.

e. Outlying or influential observations.

I welcome student suggestions regarding ways in which these goals can be best achieved. If you have questions regarding the content or structure of the class, please feel free to talk (or write) to me at any time during the quarter.