

Fall Term - 2005

Instructor: Germán Rodríguez
241 Wallace Hall, 258-4872
grodri@princeton.edu

Contents

This course deals with statistical models for the analysis of quantitative and qualitative data, of the types usually encountered in social science research. The statistical methods studied are the general linear model for quantitative responses (including multiple regression, analysis of variance and analysis of covariance), binomial regression models for binary data (including logistic regression and probit models), and Poisson regression models for count data (including log-linear models for contingency tables and hazard models for survival data). All of these techniques are covered as special cases of the Generalized Linear Statistical Model, which provides a central unifying statistical framework for the entire course.

Approach

The course is taught at an intermediate statistical level. The emphasis is on understanding and applying statistical concepts and techniques, rather than proving theorems. However, the course assumes familiarity with basic concepts in probability theory, statistical estimation and testing theory, and statistical methodology up to multiple regression analysis, at least at the level of a serious introductory course such as WWS507c. Some familiarity with matrix algebra and calculus is necessary. Computer literacy is essential, as we make extensive use of the computer. We recommend using Stata, a general-purpose statistical package available on Windows and other platforms, but students are free to use other software packages such as R/S-Plus or SAS.

Requirements

Course requirements consist of required readings, six problem sets, and two partial exams, one near the middle and another at the end of the term. Most of the material of the course is covered in formal lectures. A set of lecture notes is available on the web, and these can be supplemented with optional readings. The problem sets deal mostly with analysis of small datasets using Stata. The two partial exams emphasize the application of techniques and the interpretation of results. Final grades are calculated as a weighted average of the grades received during the term, using weights of 40% for the problem sets and 30% for each of the two partial exams.

List of Lectures

The following is a tentative list of the topics to be covered in each of the lectures scheduled for this term. The overall pace and/or the distribution of lectures within each topic may be altered if an adjustment seems advisable during the course of the term.

Thursday, September 15	Introduction and overview of the course. Responses and predictors. Factors and covariates. The generalized linear model. Review of likelihood theory.
Tuesday, September 20	Linear models. Ordinary least squares estimation. Testing the general linear hypothesis: t-tests and F-tests. Simple linear regression.
Thursday, September 22	Multiple linear regression. Interpretation of the coefficients. Gross and net effects. Hierarchical anova for multiple regression. Partial and multiple correlation.
Tuesday, September 27	Analysis of variance models. One-way anova and regression with dummy variables. Two-way anova. The additive model. Main effects and interactions.
Thursday, September 29	Analysis of covariance models. The additive model. The assumption of parallelism. Models with different intercepts and different slopes. Interpretation.
Tuesday, October 4	Regression diagnostics. Analysis of residuals. Influential observations, leverage and influence. Q-Q plots. Box-Cox transformations.
Thursday, October 6	Binary data. The binomial distribution. Grouped and ungrouped data. Odds and log-odds. The logit transformation. Logistic regression.
Tuesday, October 11	Maximum likelihood estimation and testing in logistic regression models. The comparison of two groups. The odds ratio. Comparison of several groups. The one-factor model. The one-variate model.
Thursday, October 13	Regression models for binary data. Models with two predictors. The two-factor model. Main effects and interactions. Analysis of covariance models. Linearity and additivity.
Tuesday, October 18	Multifactor models. Model selection. Additive models and models with interactions.
Thursday, October 20	Alternative links for binary data. Probit analysis. The c-log-log link. Regression diagnostics with binary data.
Tuesday, October 25	First Partial Exam
Thursday, October 27	The analysis of panel data. Random effects and fixed effects. Intraclass correlation. Extensions to binary data.
Tuesday, November 8	Count data. The Poisson distribution. The log link. Maximum likelihood estimation and testing in Poisson regression. The Poisson deviance. Modelling heteroscedastic counts.
Thursday, November 10	Models for rates of events. Exposure and the use of an offset in the linear predictor. Extra-Poisson variation and the negative binomial model.
Tuesday,	Contingency tables. Equivalence of binomial, multinomial and Poisson

November 15	sampling models. Models for two-way tables. Independence and homogeneity.
Thursday, November 17	Models for three-way tables. Complete, block and conditional independence. Uniform association and the model with no three-factor interaction. Higher dimensional tables.
Tuesday, November 22	Multinomial response models. Multinomial logits. Independence of irrelevant alternatives. Random utilities and the conditional logit model.
Tuesday, November 29	Hierarchical logits. Sequential binary choice and continuation ratio models. Equivalence with logit models.
Thursday, December 1	Models for ordered categorical data. Ordered logits and probits. Latent variable formulation and interpretation of the coefficients.
Tuesday, December 6	Survival and event history models. The survival and hazard functions. Censoring mechanisms. The likelihood function for non-informative censoring.
Thursday, December 8	The proportional hazards model. The baseline hazard. Relative risks. Time-varying covariates. Time-varying effects and models with interactions.
Tuesday, December 13	Semi-parametric models. The piece-wise exponential model. Equivalence with Poisson regression and with models for contingency tables.
Thursday, December 15	Discrete time models and equivalence with logistic regression. Unobserved heterogeneity. Topics in survival analysis.
Wednesday, January 18	Second Partial Exam

Supplementary Readings

The material of this course is covered in detail in the lecture notes. The following references are pointers to more detailed supplementary discussions, classified by subject.

Linear Models

Weisberg, S. (1985). *Applied Linear Regression*. 2nd Edition. New York: John Wiley and Sons. [QA278.2W44]. My favorite regression text, with good coverage of the basics and a lucid presentation of regression diagnostics.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. Thousand Oaks: Sage Publications. [HA31.3.F69]. A nice discussion aimed at sociologists and other social scientists, with plenty of examples. Chapter 15 has a bit on logit and probit models, and introduces generalized linear models.

Generalized Linear Models

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. London: Chapman and Hall. [QA276.M38]. The "bible" on generalized linear models, absolutely brilliant but rather on the terse side. Aimed at the more advanced statistics student.

Hardin, J. and Hilbe, J. (2001). *Generalized Linear Models and Extensions*. College Station, Texas: Stata Press. A more applied book covering the fundamentals and including worked out analyses using Stata.

Other General Books

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press. A comprehensive treatise that will be particularly useful to economists, covering the models for cross-sectional data discussed in the course as well as extensions for longitudinal data.

Long, J. S. and Freese, J. (2001). *Regression Models for Categorical Dependent Variables Using Stata*. College Station, Texas: Stata Press. A nice discussion of models for binary, ordinal, nominal, and count data with emphasis on post-estimation aids to interpretation and effective use of Stata.

Powers, D. A. and Xie, Y. (2000). *Statistical Methods for Categorical Data Analysis*. New York: Academic Press. [QA278.P694]. Covers a wider range of models that you might think from the title, and includes many examples, in a discussion aimed at social scientists.

More Specialized Texts

Hosmer, D.W. and Lemeshow, S. (1989, 2000). *Applied Logistic Regression*. New York: John Wiley and Sons. [QA278.2H67]. A more detailed discussion of logistic regression models.

Agresti, A. (1990, 2002). *Categorical Data Analysis*. New York: John Wiley and Sons. [QA278.A35]. An excellent book on models for contingency tables.

Cameron, A. C. and Trivedi, P. K. (1998). *Regression Analysis of Count Data*. Cambridge: Cambridge University Press. [QA278.2C36]. One of very few books on Poisson regression, with extensions to negative binomial and related models.

Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data*. London: Chapman and Hall. [QA276.C665]. An excellent book on survival analysis, brief and to the point, by the statistician who gave us proportional hazard models.