

Fall Term - 2011

Instructor: Germán Rodríguez
grodri@princeton.edu
241 Wallace Hall

Teaching Assistant: Elizabeth Sully
esully@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), models for count data (including Poisson regression and negative binomial models) and models for survival data (focusing on piecewise exponential models fitted via Poisson regression). 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. The date of the final exam will be set by the WWS later in 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.
Thursday, October 6	Regression remedies. Transforming the response. The Box-Cox family of transformations. Transforming the predictors.
Tuesday, October 11	Binary data. The binomial distribution. Grouped and ungrouped data. Odds and log-odds. The logit transformation. Logistic regression.
Thursday, October 13	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.
Tuesday, October 18	Regression models for binary data. Models with two predictors. Main effects and interactions. Multifactor models. Model selection.
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	Count data. The Poisson distribution. The log link. Maximum likelihood estimation and testing in Poisson regression. The Poisson deviance. Modelling heteroscedastic counts.
Tuesday, November 8	Models for rates of events. Exposure and the use of an offset in the linear predictor.
Thursday, November 10	Extra Poisson variation. The negative binomial model. Zero-inflated models for counts
Tuesday, November 15	Multinomial response models. Multinomial logits. Independence of irrelevant alternatives. Random utilities and the conditional logit model.
Thursday, November 17	Hierarchical logits. Sequential binary choice and continuation ratio models. Equivalence with logit models.
Tuesday, November 22	Models for ordered categorical data. Ordered logits and probits. Latent variable formulation and interpretation of the coefficients.
Tuesday,	Survival and event history models. The survival and hazard functions. Censoring

November 29	mechanisms. The likelihood function for non-informative censoring.
Thursday, December 1	The proportional hazards model. The baseline hazard. Relative risks. Time-varying covariates. Time-varying effects and models with interactions.
Tuesday, December 6	Semi-parametric models. The piece-wise exponential model. Equivalence with Poisson regression and with models for contingency tables.
Thursday, December 8	Discrete time models and equivalence with logistic regression. Unobserved heterogeneity. Topics in survival analysis.
Tuesday, December 13	The analysis of panel data. Random effects and fixed effects. Intraclass correlation.
Thursday, December 15	Fixed and random effect models for binary and count data. Hierarchical models.
TBA	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. My favorite regression text, with good coverage of the basics and a lucid presentation of regression diagnostics.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*, 2nd Edition. Thousand Oaks: Sage Publications. A nice discussion aimed at sociologists and other social scientists, with plenty of examples. This second edition has expanded the treatment of generalized linear models in Chapters 14 and 15, a change reflected in a new title.

Generalized Linear Models

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, 2nd Edition. London: Chapman and Hall. 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. (2007). *Generalized Linear Models and Extensions*, 2nd Edition. College Station, Texas: Stata Press. A more applied book covering the fundamentals and including worked out analyses using Stata.

Other General Books

Long, J. S. and Freese, J. (2006). *Regression Models for Categorical Dependent Variables Using Stata*, 2nd Edition. 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.

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. A very nice and accessible discussion of regression modeling with extensions into causal inference and multilevel models, with a Bayesian flavor and examples using R and WinBugs.

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.

Powers, D. A. and Xie, Y. (2000). *Statistical Methods for Categorical Data Analysis*. New York: Academic Press. 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

Agresti, A. (2002). *Categorical Data Analysis*, 2nd Edition. New York: John Wiley and Sons. 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. A comprehensive discussion of 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. An excellent book on survival analysis, brief and to the point. The first author is the statistician who gave us proportional hazard models.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression*. New York: John Wiley and Sons. A more detailed discussion of logistic regression models with applications.