STA 6934: Categorical Data Methods, Featuring Logistic Regression


Instructor: Alan Agresti
Office: Griffin-Floyd 204
Phone: 392-1941, ext. 234
E-mail: aa@stat.ufl.edu
Office Hours: Tuesday and Thursday 9-10:30 am, and by appointment.

Course description: This 1-credit short course will present a non-technical introduction to the most useful methods for analyzing categorical response variables, with emphasis on logistic regression. Students will learn how to fit and interpret logistic regression and other generalized linear models pertaining to discrete data. The course also covers specialized models for ordinal responses and repeated measurement data as well as some older but still important methods such as chi-squared, Mantel-Haenszel, McNemar, and exact tests. Several examples will be given of the analysis of data sets and the use of software (with primary emphasis on SAS, using PROC GENMOD) for conducting the analyses.

No previous knowledge of categorical data analysis is necessary. Models will be taught by analogy with regression and analysis of variance (ANOVA) models for continuous responses, so good background in these methods is needed, such as STA 6208 or STA 6166 or STA 6127. Although some background in statistical theory (such as maximum likelihood methods) is also helpful, it is not necessary since theoretical derivations are only briefly discussed.

Course text: A photocopy of overhead transparencies used in the lectures is available at University Copy Center (1620 W. University Avenue). Lectures will be based on material in the texts "Categorical Data Analysis" (2nd edition, Wiley 2002) and its lower-level version "An Introduction to Categorical Data Analysis," but these are not required.

Grading: The grade will be based on a written report, maximum length 3 pages typed double-spaced, to be handed in at the final lecture for the course. This can consist of any of three forms:

1. A statistical analysis based on modeling a data set containing a categorical response variable: You should do this project as a team with 1-2 other students in the class. If you do not have ready access to a dataset, you can create a dataset from the General Social Survey using the WWW at http://www.icpsr.uminich.edu/gss/. (Guidelines for doing this are available at my homepage for the course STA 4504,
www.stat.ufl.edu/~aa/sta4504/index.html) The report should include sections directed toward (1) Description of data and statement of questions to be addressed, (2) Specification of relevant model, (3) Interpretations of results of model fitting. Be careful not to overemphasize significance testing at the expense of inference about the size of effects (e.g., confidence intervals) and not to include too many predictors for the amount of data you have. Include, in a separate appendix (not included in the page limit), edited copies of relevant parts of your computer printouts and (if possible) the data.

2. A report on the use of logistic regression in a published article: Explain how to interpret the results of the analyses, and critically evaluate the choice of methodology. This can be done solely or as a team with 1-2 other students. Attach a copy of the article as an appendix to your report.

3. A report on a published statistics research paper: This can be done solely or as a team with 1-2 other students. Present a summary of the main points in one of the following articles:


Course Outline

1. Two-Way Contingency Tables
   - Odds Ratios and Other Ways of Comparing Proportions
   - Testing Independence: Chi-Squared
   - SAS (FREQ, GENMOD)

2. Logistic Regression for Binary Response Variables
   - Generalized Linear Models for Nonnormal Responses
   - Logistic Regression - Interpreting Parameters
   - Logit Model Analoges of ANOVA for Qualitative Predictors
   - Checking Goodness of Fit
   - Maximum Likelihood Inference (Wald, Score, Likelihood-Ratio)
   - SAS (GENMOD and LOGISTIC)

3. Logistic Model-Building
   - Strategies in Model Selection
   - Analysis of Residuals
   - Power and Sample Size Considerations
   - Analyzing Rates and Survival (Poisson Regression)
   - SAS (GENMOD)

4. Generalized Logit Models for Ordinal Responses
   - Cumulative Logit Models
   - Other Ways to Use Ordered Categories
   - SAS (LOGISTIC, GENMOD, FREQ)

5. Analysis of Repeated Categorical Measurement Data
   - Matched Pairs and Dependent Proportions
   - Logit Models for Binary Repeated Categorical Measurement
   - Generalized Estimating Equations (GEE) Methodology
   - Generalized Linear Mixed Models (Random Effects)
   - SAS (GENMOD, FREQ, NLMIXED)

6. Small-Sample Methods (time permitting)
   - Fisher’s Exact Test
   - Conditional Methods for Contingency Tables and Logistic Regression
   - StatXact and LogXact, SAS (FREQ)