STA 6505 ANALYSIS OF CATEGORICAL DATA Spring 2010

Topics: This course surveys methods for the analysis of categorical response variables, from the maximum likelihood (frequentist) perspective. The main subject areas covered are descriptive and inferential statistics for two-way and three-way contingency tables, generalized linear models for discrete responses, binary regression models (emphasizing logistic regression), multi-category logit models for nominal and ordinal responses, loglinear models for contingency tables, and matched pairs.

Instructor: Alan Agresti (I am an emeritus faculty member at UF but I am teaching here part-time during spring semester 2010.)

Office: Griffin-Floyd 204

Phone: (352) 273-2981

E-mail: aa@stat.ufl.edu

Office Hours: Monday and Wednesday 3-5 pm, and by appointment.

Teaching assistant: Steve Chung, stevec@stat.ufl.edu, office hours Tuesday and Thursday 11 am to 1 pm or by appointment, 115A Griffin-Floyd.

Course Homepage: www.stat.ufl.edu/∼aa/sta6505/index.html

Course Text: *Categorical Data Analysis*, second edition, by A. Agresti (Wiley, 2002). This is on 2-hour reserve for this course at the Science library. New and used copies are available for purchase at the UF bookstore or over the Internet. (My royalties from sales of new copies of the text for this course are donated to UF.) The website for the text is www.stat.ufl.edu/∼aa/cda/cda.html.

1. Exam Schedule

Exam	Date
1	Monday evening, February 22
2	Thursday evening, April 15

The exams are not cumulative, and each will count for 1/3 of the final grade, the other 1/3 being based on homework. The exams, although intended as one-hour exams, will be given in the evening so that students do not feel time pressure. Make-up exams will not be given except for medical or family emergencies and *must be approved before the time of the exam*. Because of the extra evening periods for the two exams, there will be no class on April 16 and on one or two other days to be announced.

Students have the option of substituting a project for the second exam. There are three options for this. The first option is to analyze a set of data containing at least one categorical response variable using the modeling methods of this course, preparing a written report about the analysis (at most 5 double-spaced pages, plus appendix with printouts), describing (a) nature of data and purposes of modeling, (b) models fitted, (c) model checking, (d) interpretations and conclusions. The second option is to prepare a report on the Bayesian approach to modeling categorical response variables, focusing on (a) binary regression models, (b) ordinal models, or (c) nominal models. The third option is to prepare a report on non-model-based methods for classification and prediction of categorical responses, such as CART and other methods summarized in books on data mining and statistical learning (such as *The Elements of Statistical Learning* by Hastie, Tibshirani and Friedman, available at http://www-stat.stanford.edu/∼tibs/ElemStatLearn). For options (2) and (3), the student should conduct a review of the statistics literature in research journals that deals with this topic, and summarize what he/she feels to be useful results in a report that is a maximum of 6 double-spaced pages. The project must be turned in by Thursday, April 15. The two best projects will receive a bonus of 10 points (of the 100 maximum) and be asked to make a 20-minute presentation to the class about their project on the final class day, April 21.

2. Homework

Homework problems are listed on the outline of topics in Section 4 of this syllabus. For each chapter please keep a neat, organized file of solutions and hand them in to the TA on the dates specified in class lectures. A sample of the exercises will be graded, and the total grade on the homework assignments will count toward 1/3 of the final course grade. *No homeworks will be accepted after the due date*, but we will drop your lowest homework grade before calculating your homework average. To provide you with feedback about your solutions, brief outlines of the solutions to many of the homework problems are available in a pdf file at

http://www.stat.ufl.edu/∼aa/restricted/solutions-cda2-hw.pdf

Your solutions must be complete, not merely repeating the posted outline solutions, and must show the use of software for exercises that require it by attaching relevant computer output. You are permitted to work together with other students on problems with which you have difficulties, but the solutions handed in should be yours alone.

3. Software

I'll give class examples primarily using SAS and occasionally R. The website for the text has a link to

http://www.stat.ufl.edu/∼aa/cda/software.html

where there is information about various software for categorical data analysis. That site has a link

https://home.comcast.net/∼lthompson221/Splusdiscrete2.pdf

to a detailed manual prepared by Dr. Laura Thompson showing how to use R and S-Plus to conduct all the analyses in the text. I highly recommend this resource if you would like to use R for statistical analyses of categorical data. There is also a link there to a website of Dr. Chris Bilder, whose link to R has examples of the use of R for many methods for categorical data (organized in terms of my lower-level text, *An Introduction to Categorical Data Analysis*). Another useful site is web.stat.ufl.edu/∼presnell/Courses/sta4504-2000sp/R/ set up by Dr. Brett Presnell from when he taught STA 4504/5503 at UF.

4. Outline of Topics and Homework Problems

Topics	Text Pages	Homework
1. Introduction: Distributions and Inference		
Discrete distributions	1-9	1, 3, 12
Inference for categorical data	10-26	7(a-d), 11, 17, 18, 30, 33, 34
2. Describing Contingency Tables		
Probability structure	36-43	21
Comparing proportions	43-47	3, 4, 8, 10, 11, 23a, 24
Stratified tables	47-54	12, 16, 29
3. Inference for Contingency Tables		
Deriving large-sample normal distributions	70-78, 577-580	22, 24, 26, Ch 14: 5, 7, 8, 9
Chi-squared tests of independence	78-86	3, 4, 29, 31, 34, 35
Exact tests for small samples	91-100	13, 40, 42, 43
4. Introduction to Generalized Linear Models		
Generalized linear models	115-119	17, 30
GLMs for binary data	120-125	1, 2, 5, 19, 20, 29
Inference and fitting GLMs	143-145	22, 28, 34
5. Logistic Regression		
Interpreting parameters	165-171	15, 28, 29, 30, 32, 33(a-b)
Inference for logistic regression	172-177	1, 4
Categorical and multiple predictors	177-192	8, 9, 12
Fitting logistic regression models	192-196	36, 37, 38
6. Building and Applying Logistic Regression Models		
Model selection	211-219	22, 23
Diagnostics	219-230	5, 6
Inference in stratified tables	230-236	7(a-d)
Power	236-245	26
Probit and complementary log-log link	245-250	14, 28, 29, 30, 32
7. Models for Multinomial Responses		
Baseline-category logit models	267-272	1, 3, 26
Cumulative logit models	274-282	5, 7, 9, 29, 31
8. Loglinear Models		
Loglinear models for two-way tables	314-318	14, 15
Loglinear models for three-way tables	318-324	18, 19, 21
Inference for loglinear models	324-326, 333-343	1, 6, 29, 31, 32, 38b
Loglinear – logit connection	330-333	9
10. Models for Matched Pairs		
Comparing dependent proportions	409-413	1, 21, 22

5. References

Textbooks:

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis*, MIT Press.

Lloyd, C. J. (1999). *Statistical Analysis of Categorical Data*, Wiley.

McCullagh, P., and Nelder, J. (1989). *Generalized Linear Models*, Chapman and Hall.

Santner, T., and Duffy, D. (1990) *The Statistical Analysis of Discrete Data*, Springer-Verlag.

Some research articles that supplement course material:

Agresti, A. (1992), "A Survey of Exact Inference for Contingency Tables," (with discussion) *Statistical Science*, 7, 131-177.

Birch, M.W. (1963), "Maximum Likelihood in Three-Way Contingency Tables," *J. Roy. Statist. Soc.*, B, 25, 220-233.

Breslow, N., and Clayton, D.G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *J. Amer. Stat. Assoc.*. 88, 9-25.

Cochran, W.G. (1954), "Some Methods of Strengthening the Common χ^2 Tests," *Biometrics*, 10, 417-451.

Cox, D.R. (1958a). "The Regression Analysis of Binary Sequences," *J. Roy. Statist. Soc.*, B, 20, 215-242.

Goodman, L.A. (1970), "The Multivariate Analysis of Qualitative Data: Interactions Among Multiple Classifications," *Journal of the American Statistical Association*, 65, 226-256.

Goodman, L.A. (1979), "Simple Models for the Analysis of Association in Cross-Classifications Having Ordered Categories," *Journal of the American Statistical Association*, 74, 537-552.

Grizzle, J.E., Starmer, C.F., and Koch, G.G. (1969), "Analysis of Categorical Data by Linear Models," *Biometrics*, 25, 489-504.

Liang, K.-Y., and Zeger, S.L., and Qaqish, B. (1986). "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13-22.

McCullagh, P. (1980), "Regression Models for Ordinal Data" (with discussion), *Journal of the Royal Statistical Society*, B, 42, 109-142.

Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society*, A, 135, 370-384.

Neyman, J. (1949), "Contributions to the Theory of the χ^2 Test," *Proceedings of the First Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, pp. 230-273, Berkeley: U. of California Press.