

Sociology 5601: ADVANCED DATA ANALYSIS

Jane Menken 303-492-2144,1697
218A Ketchum, 202 IBS#3 (1424 Broadway)
menken@colorado.edu

M-W 1:15-2:30
Ketchum 8
Office Hours: Wednesday 10-12 and by appointment

This course emphasizes *using* statistics on real data. It is intended to help you review general linear regression models and extend consideration to residual analysis, curvilinearity, and interactions. It then moves on to consideration of limited dependent variables -- through logit and probit analysis of 2-category dependent variables and then multiple-category dependent variables. We will use both ordered logit and multinomial logit models. Thus this class is intended to be a practicum in advanced multivariate analysis.

Students may bring their own quantitative data set to the class, and analyze it using the techniques covered during the semester. For those without their own data, I can sometimes recommend a data set that matches student interests, but more often suggest using either the General Social Survey (GSS) or our data from Bangladesh. Using data of their choice, students may then employ the UNIX workstations in Ketchum Hall to run statistical programs (using STATA) and generate computer output. Students may also obtain STATA for use on their own home computers.

The course is intended to help develop understanding of the logic of various statistical techniques based on linear models (regression and its extensions) and ability to use them appropriately. We will also consider the difficulties involved in the analysis of non-experimental data (observational studies) and the pitfalls in making inferences from such data. Examples of empirical studies from various areas of application will be discussed.

We will use the computer lab in 117 Ketchum for about every third session, so that everyone can gain experience using STATA. I understand that many students are far more familiar with SPSS than STATA. Both Fred Pampel and I have found that STATA is keeping up with advances in statistics far more than SPSS and is far easier to use than SAS. It is also quite easy to graph material in STATA -- and we both emphasize the importance of *looking* at one's data and believe that graphical presentation of results frequently serves to enhance their communication.

The course emphasizes the match between theoretical reasoning, substantive research problems, and statistical results. The most important and perhaps the most difficult skill to teach involves application of the statistical techniques to real research problems. With that in mind, we will devote considerable class time to presentation and discussion of research studies from social science journals, the relevant substantive issues they raise, and the ability of the statistical techniques they use to address the substantive issues. As the semester goes on, I'll ask students to find and discuss in class articles that use the techniques being studied.

The topics cover those aspects of the general linear regression model studied in SOCY 5021 that present common problems in actual research. Among other things, the regression model assumes 1) no measurement errors in the independent variables and normally distributed, independent error terms, 2) linear and additive relationships, and 3) continuous dependent variables. To deal with these limiting assumptions, the course examines 1) analysis of regression residuals to evaluate assumptions about the error term, 2) transformations to model various types of nonlinear and non-additive relationships, and 3) logistic and probit regression models and their extensions for use with categorical dependent variables. The quantitative research literatures in Sociology and other disciplines employ each of these techniques extensively.

The course assumes knowledge of basic multivariate regression and analysis of covariance.

ASSIGNMENTS

In this advanced class, students need to spend less time on usual terminology and none on hand calculations. Instead, they can focus on using the computer to generate appropriate statistics and on understanding and interpreting the output. To facilitate this, I will introduce and illustrate use of the STATA program as needed for analysis, and will assign five short homework assignments involving the interpretation of computer output or published tables. Students can then apply the computer programs to analysis of their own data. I assign three papers based on appropriate analysis of the data and interpretation of the statistical results.

GRADING

I divide the course into three sections, each of which requires completion of a short paper and one or two homework assignments. The first paper should emphasize theory, data, and measurement and comprise about 5-7 pages. Each subsequent paper should revise this previous material and add new material based on the most recent statistical analyses. Each paper contributes 25% (75% together) to the total grade. The five homework assignments contribute 5% each (25% together) to the total grade. I will pass out instructions for both the homework and paper assignments during the semester, but can make a few comments about the papers now.

The papers should be clearly written, as if for a professional audience, with a tight connection between theory and results. One needs considerable practice to write clear, organized, and theoretically meaningful prose when describing statistical results. We will discuss writing issues often throughout the semester, but the real learning will come from your efforts to rewrite, revise, edit, and (perhaps most importantly) organize your papers until they read smoothly, proceed logically, and highlight the substantive meaning of the statistical results.

By the end of the semester you should have completed an empirical research paper of around 20 pages that explores in some detail and with a variety of techniques the quantitative relationships among concepts and variables of interest to you. The assignments in combination may serve as a preliminary, exploratory analysis for a more detailed comp paper or research study. Indeed, if all goes well, the work in this course may result in paper potentially publishable in a social science journal.

READINGS

You will need to use whatever regression text you used for Socy5021. The one used most recently is

McClendon, McKee J. 1994. *Multiple Regression and Causal Analysis*. Itasca, IL: F.E. Peacock Publishers, Inc.

We will also use:

Pampel, Fred C. 2000. *Logistic Regression: A Primer*. Quantitative Applications in the Social Sciences #132. Thousand Oaks, CA: Sage Publications.

I would recommend buying this book. We can get a discount for a large order, and, since it's not used for several weeks, we can make this order.

I am also looking for a good publication on ordered and multinomial logit -- but have been disappointed in the search thus far. I will make recommendations as we move into the semester.

I regularly prepare notes and will either xerox them for you or make them available on the course website well in advance of their use.

We will also consult selected articles and chapters among those listed at end of syllabus - or added in the course of the semester.

NOTE ON EMAIL: As many of you already know, I use email extensively for communication with students about courses and all other matters. *I prefer having all assignments emailed to me at the address given above - either in the text of an email message or as a PC-compatible attachment in Word or WordPerfect. PLEASE give the attachment an understandable name, e.g. ted.wk1 or ted.95, that identifies person(s) and date.*

SCHEDULE

Wk1	Aug28	Computer lab	
	Aug30	Regression review 1	
Wk 2	Sep4	Labor Day - no classes	
	Sep6	Regression review 2	
Wk3	Sep11	Regression review 3	
	Sep13	Computer lab	
Wk4	Sep18	Non-linear models I: Dummy variables and interactions	
	Sep2	Search strategies	
Wk5	Sep25	Regression diagnostics I	HW #1 due
	Sep27	Computer lab	
Wk6	Oct2	Regression diagnostics II	Paper #1 due
	Oct4	Non-linear models II - curvilinearity	
Wk7	Oct9	Computer lab	
	Oct11	Odds	HW #2 due
Wk8	Oct16	Logistic regression logic	
	Oct18	Logistic regression interpretation	
Wk9	Oct23	Estimation and model fit	HW #3 due
	Oct 25	Computer lab	
Wk10	Oct30	Probit analysis	
	Nov1	Event history analysis I - the life table	Paper #2 due
Wk11	Nov6	Event history analysis II - the hazard model	
	Nov8	Computer lab	
Wk12	Nov13	Event history analysis III - estimation	HW #4 due
	Nov15	Event history analysis IV - interpretation	
Wk13	Nov20	Extensions of logit regression I - multinomial logistic	no class Nov22
Wk14	Nov27	Extensions II - estimation of multinomial logistic	HW #5 due
	Nov29	Extensions II - estimation of multinomial logistic	
Wk15	Dec4	Extensions III - interpretation of results	
	Dec6	Extensions IV - ordered logit	
Wk16	Dec11	Computer lab	
	Dec12	Extensions V- interpretation of ordered logit	Paper #3 due

READINGS

A. Preface. Theory and Statistics

McCloskey, Donald N. 1990. "Formalism in the Social Sciences, Rhetorically Speaking." *The American Sociologist* 21:3-19.

Freedman, David A. "Statistical Models and Shoe Leather." *Sociological Methodology* 21:291-314.

Berk, Richard A. 1991. "Toward a Methodology for Mere Mortals." *Sociological Methodology* 21:315-324.

B. Regression Review

McClendon, McKee J. 1994. *Multiple Regression and Causal Analysis*. Itasca IL: Peacock. Chapters 3-5.

STATA manuals

Lewis-Beck, Michael S. 1980. *Applied Regression: An Introduction*. Newbury Park CA: Sage.

Berry, William D. 1993. *Understand Regression Assumptions*. Newbury Park CA: Sage.

C. Residual Analysis

Bollen, Kenneth A. and Robert W. Jackman. 1985. "Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases." *Sociological Methods and Research* 13:510-545.

Hamilton, Lawrence C. 1992. *Regression With Graphics: A Second Course in Applied Statistics*. Pacific Grove CA: Brooks/Cole. Chapter 4.

Glantz, Stanton A. and Bryan K. Slinker. 1990. *Primer of Applied Regression and Analysis of Variance*. New York: McGraw-Hill. Chap. 4.

Dietz, Thomas, R. Scott Frey, and Linda Kalof. 1987. "Estimation with Cross-National Data: Robust and Nonparametric Methods." *American Sociological Review* 52:380-390.

STATA manuals

D. Non-Linear and Non-Additive Models

McClendon, Chapters 6-7

Jaccard, James, Robert Turrisi, and Choi K. Wan. 1990. *Interaction Effects in Multiple Regression*. Newbury Park CA: Sage. Chapters 1-4

Hardy, Melissa A. 1993. *Regression With Dummy Variables*. Newbury Park CA: Sage. Chapters 1-4.

E. Analysis of Categorical Dependent Variables: Logistic Regression

Morgan, Philip S. and Jay D. Teachman. 1988. "Logistic Regression: Descriptions, Examples and Comparisons." *Journal of Marriage and the Family* 50:929-936.

STATA manuals

Menard, Scott. 1995. Logistic Regression. Newbury Park CA: Sage. Chapters 4-5.

Allison, Paul D. 1982. A Discrete-Time Methods for the Analysis of Event Histories. @ Pp. 61-98 in Samuel Leinhardt (ed.), Sociological Methodology 1982. San Francisco: Jossey-Bass.

F. Analysis of Categorical Dependent Variables -- Multiple Categories

To be added