PS 813: Multivariate Statistical Inference
Spring 2020

Dave Weimer
263-2325
weimer@lafollette.wisc.edu

Mondays/Wednesdays, 8:00 a.m. to 9:15 a.m.
Ingraham 223

Office Hours: Mondays and Wednesdays 9:30 a.m. to 11:30 a.m., 215 North Hall
Other times welcome by appointment.

Effective participation in the social sciences requires familiarity with the basic elements of multivariate statistics. As social scientists rarely have the opportunity to study phenomena or behavior through controlled experiments, empirical tests of hypotheses derived from theory must often be coaxed either from data collected without the benefit of random assignment or from data that "happen" to be available as a byproduct of some non-research process. It is usually necessary, therefore, to use multivariate techniques to attempt to control statistically for at least those observable factors that cannot be controlled through random assignment. Absent familiarity with these basic techniques, social scientists cannot critically evaluate empirical results in their substantive areas of interest. Without some facility for actually using the techniques, they are less likely to be able to contribute in an important way to the testing of hypotheses implied by theories or even to the description of complicated phenomena.

Our objective is to prepare for the roles of consumer and producer of multivariate statistical analysis. Because it is commonly used, intuitively appealing, and fairly flexible, we focus primarily on the basic linear regression model. It also provides a frame of reference for considering other techniques that we will learn. We will try to develop appropriate practical use and intuitive understanding rather than an ability to prove theorems. At the same time, however, we must be careful to develop an adequate theoretical base to allow continued learning beyond the course. Consequently, although we will cover relatively few formal proofs in class, we will go through a number of derivations to convey key points and increase capability for continued learning after the course.

We will adopt a pace that is consistent with developing firm conceptual foundations. Consequently, we may or may not cover all the advanced topics listed on the syllabus by the end of the course.

Mathematics

Applying some basic concepts and techniques drawn from calculus and linear algebra enables us to develop a deeper understanding of multivariate estimation and inference. I assume that you have a familiarity with basic differential calculus but not necessarily with matrix algebra. We will spend several classes covering the latter after we have completed our introductory tour of bivariate regression.

Statistical Computing

A number of course assignments will require you to use the Stata statistical package or an alternative such as R. Enough guidance will be provided for the assignments. However, I highly recommend that you concurrently take PS 881 (1 credit), which will develop your statistical computing skills in more depth. It will also help you develop effective data handling skills that will be useful as you begin your own research projects.

Course Requirements

Examinations: Midterm (20 percent) on **March 11**; Final (50 percent) to be scheduled during the examination period.

Assignments: Approximately weekly assignments will be in a variety of formats: problem sets, computing exercises, Monte Carlo experiments, and memoranda tied to data analyses (20 percent).

Project: Attempt to answer a disciplinary or policy question by applying techniques learned in course to data that you have assembled (10 percent). Due at noon on **May 1**.

This is a 3-credit course that meets 150 minutes per week.

Possible Texts

Although there is no required text, I recommend that you have one available for reference. If you already have an econometrics text, then no need to purchase an additional one. In case you don't have a text, the following text is available on reserve at the College Library:

> Damodar N. Gujarati, *Basic Econometrics* 4th (New York: McGraw-Hill, 2003).

As an alternative, you might consider:

> William H. Greene, *Econometric Analysis* 7th (New York: Macmillan Publishing Company, 2011).

Greene provides a much more comprehensive survey of the theory underlying the commonly used basic techniques. If you are planning on doing methods as a field and you already have some mathematical confidence, then Greene might be a reasonable investment. Otherwise, I recommend Gujarati, which comes closest to the level at which we will approach material in class.

In any event, I attempt to make lectures self-contained so the primary use of either text is to get a second view. Therefore, if you already have a comparable text, then you do not necessarily have to purchase either of these texts.

Readings and exercises are available on Canvas.

Outline of Topics

I. ***Introduction*** (class 1)

Overview

Problem Set 1 (Review of concepts from mathematical statistics and summation notation)

II. ***Bivariate Regression***

History
Fitting curves to data
Ordinary least squares (OLS)
Hypothesis testing, power, confidence intervals
Properties of least squares estimators
Maximum likelihood estimators (MLEs)

Gujarati, 1 to 6

Phil Cook's lessons on presenting statistical analysis.

Problem Sets 2 (line fitting—due at beginning of section), 3 (practice with S^2 notation—due midway), and 4 (derivation practice—due at end of section)

Data Exercise 1

III. ***Multivariate Regression***

Review of matrix notation and operations
Gauss-Markov theorem and BLUE estimators
Properties of estimators
Statistical inference

Gujarati, Appendix B, C, 7, 8

David Weimer and Aidan Vining (2011) *Policy Analysis: Concepts and Practice* 5th Ed. (Englewood Cliffs, N.J.: Prentice-Hall), Chapter 17: "Revising the Lead Standard for Gasoline," 424–447.

Problem Sets 5 (prepare for beginning of section) and 6 (matrix review)

Monte Carlo Exercise 1 (team activity)

Play with binormalm (download from Mathematical Exercises)

IV. *Model Specification*

Non-linear models, Cobb-Douglas models, interaction terms, indicator variables
Analysis of residuals
Specification error

Gujarati, 9

Thomas Brambor, William Roberts Clark, and Matt Golder (2006) Understanding Interaction Models: Improving Empirical Analysis. *Political Analysis* 14(1), 63–82.

Data Exercises 2 and 3

V. *Pathologies and Treatments*

Multicollinearity
Heteroscedasticity and generalized least squares (GLS)
Feasible GLS
Autocorrelation
Aggregation bias
Measurement error

Gujarati, 10 to13, 17

Problem Set 7 (after heteroscedasticity section)

VI. *Models with Discrete Dependent Variables*

Contingency table analysis
Linear probability models, logit, and probit
Ordered probit, multinomial and conditional logit

Gujarati, 15

R. Michael Alvarez and Jonathan Nagler (1998) When Politics and Models Collide: Estimating Models of Multiparty Elections. *American Journal of Political Science* 42(1), 55–96.

Chunrong Ai and Edward C. Norton (2003) Interaction Terms in Logit and Probit Models. *Economic Letters* 80(1), 123–129.

Tue Tjur (2009) Coefficients of Determination in Logistic Regression Models—A New Proposal: The Coefficient of Discrimination. *American Statistician* 63(4), 366–372.

Data Exercise 4

Monte Carlo Exercise 2 (team activity)

VII. *Simultaneous Equation Models*

Identification
Estimation: instrumental variables; two-stage least squares; three-state least squares

Gujarati, 18 to 20

Joshua D. Angrist and Alan B. Kruger (2001) Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives* 15(4), 69–85.

Michael P. Murray (2006) Avoiding Invalid Instruments and Coping with Weak Instruments. *Journal of Economic Perspectives* 20(3), 111-132.

John G. Richards, Aidan R. Vining and David L. Weimer (2010) Aboriginal Performance on Standardized Tests: Evidence and Analysis from Provincial Schools in British Columbia," *Policy Studies Journal* 38(1), 47–67.

Joseph V.Terza (2018) Two-Stage Residual Inclusion Estimation in Health Services Research and Health Economics. *Health Services Research* 53(3), 1890-1899.

VIII. *Additional Topics as Time Permits*

Panel data
Censored data
Seemingly unrelated regressions
Selection models
Hierarchical models
Regression discontinuity

Gujarati, 16, 17

Nathaniel Beck (2011) Of Fixed-Effects and Time Invariant Variables. *Political Analysis*

19(2), 119–122.

Nathaniel Beck and Jonathan N. Katz (1995) What to Do (and Not to Do) with Time-Series Cross-Section Data. *American Political Science Review* 89(3), 634–647.

Marianne Bertrand, Esther Duflo, and Sendhil Mullainathan (2004) How Much Should We Trust Difference-in-Difference Estimates? *Quarterly Journal of Economics* 119(1), 249–275.

Charles H. Franklin (1989) Estimation across Data Sets: Two-Stage Auxiliary Instrumental Variables Estimation (2SAIV). *Political Analysis* 1 (1989), 1–24.

Taylor Grant and Matthew J. Lebo (2016) Error Correction Methods with Political Time Series. *Political Analysis* 24(1), 3–30

Guido W. Imbens and Thomas Lemieux (2008) Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics* 142(2), 615–635.

Ranjit Lall (2016) How Multiple Imputation Makes a Difference. *Political Analysis* 24(4), 414–433

Charlisle Rainey (2016) Dealing with Separation in Logistic Regression Models. *Political Analysis* 24(3), 339–355.

W. Robert Reed and Rachel Webb (2010) The PCSE Estimator is Good——Just Not as Good as You Think. *Journal of Time Series Econometrics* 2(1), Article 8, 1–24.

Curtis S. Signorino (1999) Strategic Interaction and the Statistical Analysis of International Conflict. *American Political Science Review* 93(2), 279–297.

Joseph V. Terza, Anirban Basu, and Paul J. Rathouz (2008) Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling**.** *Journal of Health Econ*. 27(3), 531–543.

Michael Tomz, Joshua A. Tucker, and Jason Wittenberg (2002) An Easy and Accurate Regression Model for Multiparty Elections. *Political Analysis* 10(1), 66–83.