

REGRESSION ANALYSIS

Course Objectives: This course provides an introduction to the theory, methods, and practice of regression analysis. The goal is to provide students with the skills that are necessary to: (1) Read, understand, and evaluate the professional literature that uses regression analysis; (2) design and carry out studies that employ regression techniques for testing substantive theories; and (3) prepare to learn about more advanced statistical procedures.

This course will not dwell on statistical theory. But, neither will it take a superficial, “cookbook” approach to methodology. Instead, we will concentrate on: The importance of evaluating empirical relationships between variables as a component of the theory-testing process; the utility of regression analysis for doing so; the nature of the basic regression model; and the development of the regression estimators. We will see that the regression model depends very heavily on several assumptions. Therefore, we will examine these assumptions in detail, considering why they are necessary, whether they are valid in practical research situations, and the consequences of violating them in particular applications of the regression techniques. These formal, analytic treatments will be counterbalanced by the frequent use of substantive examples and class exercises. Again, the overall course objective is NOT to turn you into a statistician— rather, we are trying to maximize your research skills as a political scientist.

Course Prerequisites: Any course of this type must assume a working knowledge of elementary statistical concepts and techniques. Students should be familiar with such ideas as descriptive statistics, graphical displays for univariate data, probability, sampling distributions, statistical inference, confidence intervals, and hypothesis testing. An understanding of these basic concepts is absolutely essential before moving on to the more complicated matters that will comprise the majority of the course material. Therefore, I assume that everyone has taken at least one prior course in introductory statistics and/or data analysis (e.g., PLS 801).

Course requirements: Regular attendance and active class participation is expected. This is a mandatory component of the course: Statistical knowledge is cumulative, and gaps in the early material will always have detrimental consequences later on. Homework assignments will be given frequently (about once a week). Some of these will be problems requiring pencil-and-paper calculations. But, most of the assignments will be computer-based data analysis exercises. All of them are intended to familiarize you with the various concepts and techniques introduced in class and in the readings. Assignments will not be graded for correct answers. But, they will be checked for completion, and comments will be provided. There will be two examinations in this course. The midterm will be given in class, and the final (which is cumulative) will be a take-home.

The course grades will be determined as follows:

Homework Assignments	20%
Class Participation	10%
Midterm Examination	30%
Final Examination	40%

TEXTBOOKS

The required course texts are:

McClendon, McKee. (1994) *Multiple Regression and Causal Analysis*. Prospect Heights, IL: Waveland Press.

Berry, William D. and Stanley Feldman. (1985) *Multiple Regression in Practice*. Beverly Hills, CA: Sage Publications.

The recommended course texts are:

Gujarati, Damodar N. (2003) *Basic Econometrics (Fourth Edition)*. Boston, MA: McGraw-Hill.

Wooldridge, Jeffrey M. (2006) *Introductory Econometrics: A Modern Approach (Third Edition)*. Mason, OH: Thomson South-Western.

The following books are supplemental:

Fox, John (2002) *An R and S-Plus Companion for Applied Regression*. Thousand Oaks, CA: Sage Publications.

Kennedy, Peter. (2003) *A Guide to Econometrics (Fifth Edition)*. Cambridge, MA: MIT Press.

The McClendon text and the Berry and Feldman monograph are both very reasonably priced. Taken together, they cover most of the course topics in an accessible manner. For these reasons, they are listed as the basic “required” reading material for the class.

However, the treatments in McClendon and in Berry and Feldman are fairly elementary and non-technical. Therefore, students planning to carry out further work in methods (e.g., taking more advanced courses, declaring a minor field in Methodology within the MSU Political Science Ph.D. program, etc.) are advised to do the additional readings in either of the two recommended course texts. The latter provide more detailed coverage of the material along with explicit derivations of important statistical concepts.

The supplemental texts are optional; they are intended to provide additional information that may be helpful for carrying out the work in this course. The Fox book covers computing issues involving **R** software in the context of regression analysis and linear models. The Kennedy book provides an alternative, and relatively basic, introduction to many of the topics that will be covered during the semester. Its level of sophistication generally lies somewhere in between those of the required and recommended course texts. We will discuss the required and supplemental texts in greater detail in class, and I will be happy to talk with anyone about the pros and cons of each one.

COURSE WEB SITE

The Home Page for this course is located at the following URL:

<http://www.polisci.msu.edu/jacoby/msu/pls802>

The contents of this website will evolve and expand as the course proceeds through the subject matter. You should regard the site as an information resource. It will contain the syllabus, copies of handouts, datasets, assignments, computing and software resources, lists of important concepts, study materials for examinations, and links to other interesting and useful sites on the Worldwide Web.

COMPUTING AND SOFTWARE

Statistical methods almost invariably require repetitive calculations, applied to large amounts of data. At the same time, graphical displays of quantitative information require a precision in their rendering that is almost impossible to achieve using pencil and paper. For these reasons, computers and statistical software are absolutely necessary for employing modern statistical techniques in an effective manner. They will be closely integrated into the course material.

Most of our work (including class examples, demonstrations, homework, and examinations) will rely on the **R** computing environment (or its commercial alternative, **S-Plus**). Along with its superb functionality and flexibility, **R** is also attractive because it is available free of charge. You can download the software from the **R** website, <http://www.r-project.org/>. This site also provides a great deal of useful information about **R** and many useful links to additional material (e.g., manuals, FAQ's, newsletter, etc.). Installation on your own computer is very easy; if you are offered any choices during the process you should use the defaults.

Two important advantages of **R** are its extensible nature and active user community. Many researchers have written functions and procedures (called "packages") that extend **R**'s capabilities far beyond the features that are available in its basic implementation. We will use a number of these packages throughout the course (particularly the "lattice" and "car" packages), and they will be introduced as the need arises. All of the packages are either included in the basic **R** installation (e.g., lattice), or they can be downloaded from the web (e.g., car). Most of them are available at the *Comprehensive R Archive Network* (CRAN) site, <http://cran.r-project.org/> or one of its mirror sites around the world.

You may also use other statistical software in this course (e.g., STATA, SAS, SPSS, SYSTAT, etc.), as long as it has the analytical routines and capacities that are required to complete the assignments and examinations. Note, however, that most other software packages do *not* have all of the capabilities required for this course! Therefore, students interested in alternative software for assignments and examinations *must* check with me early in the semester!

TOPICS AND READING ASSIGNMENTS

I. Introduction and Preliminary Material

A. Causality and the Nature of Statistical Models

Read: McClendon (1994), pages 1-19
Gujarati (2003), pages 15-32
Wooldridge (2006), pages 1-19
Kennedy (2003), pages 1-10

B. Basic Concepts: Functional Dependence; Linear Transformations; Linear Combinations; the Properties of Statistical Estimators

Read: McClendon (1994), pages 20-28
Wooldridge (2006), pages 707-802

C. Basics of the **R** computing environment

Read: Fox (2002), pages 1-84

Fox, John (2005) “The **R** Commander: A Basic-Statistics Graphical User Interface to **R**.” *Journal of Statistical Software* 14 (9). Available on the course website and also on the journal’s website, at <http://www.jstatsoft.org>.

D. Graphical Displays for Bivariate and Multivariate Data

Read: Fox (2002), pages 85-106

Becker, R. A. and W. S. Cleveland (1996) “Trellis Graphics User’s Manual.” Unpublished manuscript. Available on the course website. Note that all of this information has also been reproduced in the **S-Plus** documentation.

Becker, R. A.; W. S. Cleveland; M. Shyu; S. P. Kaluzny (1995) “A Tour of Trellis Graphics.” Unpublished manuscript. Available on the course website.

E. Scatterplot Smoothing and Nonparametric Regression

Read: To be announced.

II. The Descriptive Linear Regression Model

A. Bivariate Regression and Correlation

Read: McClendon (1994), pages 28-45, 53-59

Gujarati (2003), pages 58-65, 81-93

Wooldridge (2006), pages 24-46

Fox (2002), pages 119-123

B. The Multiple Regression Model

Read: McClendon (1994), pages 60-83, 94-109, 116-118

Gujarati (2003), pages 202-207, 212-215

Wooldridge (2006), pages 73-88

Fox (2002), pages 123-124

III. Statistical Inference for the Linear Regression Model

A. Regression Assumptions and Properties of the Least Squares Estimator

Read: McClendon (1994), pages 133-146

Berry and Feldman (1985), pages 9-12

Kennedy (2003), pages 11-59

Gujarati (2003), pages 37-52, 65-81, 100-114, 207-212

Wooldridge (2006), pages 50-66, 89-95, 106-109, 123-126, 176-181, 187-190

B. Confidence Intervals and Hypothesis Tests for Bivariate Regression Models

Read: McClendon (1994), pages 147-154

Gujarati (2003), pages 119-140, 142-145

Wooldridge (2006), pages 126-147

C. Statistical Inference for Multiple Regression

Read: McClendon (1994), pages 157-174
Berry and Feldman (1985), pages 12-18
Gujarati (2003), pages 248-273
Wooldridge (2006), pages 147-167, 214-218
Kennedy (2003), pages 60-80

D. Interpretation and Model Specification Issues in Regression Analysis

Read: McClendon (1994), pages 45-49, 154-157
Berry and Feldman (1985), pages 18-26
Gujarati (2003), pages 506-523
Wooldridge (2006), pages 95-99, 105-106, 192-197, 206-214
Fox (2002), pages 124-125
Kennedy (2003), pages 81-109

MIDTERM EXAMINATION: TUESDAY, MARCH 1

IV. The Multiple Regression Model in Matrix Form (If Time Permits)

Read: McClendon (1994), pages 119-132
Gujarati (2003), pages 926-947
Wooldridge (2006), pages 819-832
Fox (2002), pages 145-148

V. Categorical Independent Variables

A. Dummy Variables

Read: McClendon (1994), pages 198-214, 223-226
Gujarati (2003), pages 297-298, 301-306
Wooldridge (2006), pages 230-244
Fox (2002), pages 126-133
Kennedy (2003), pages 248-252

B. Multiplicative Terms and Interaction

Read: McClendon (1994), pages 271-287
Gujarati (2003), pages 310-312
Wooldridge (2006), pages 244-252
Fox (2002), pages 133-136, 149-154
Kennedy (2003), pages 252-258

C. A Brief Introduction to Analysis of Variance (ANOVA)

Read: McClendon (1994), pages 226-229
Gujarati (2003), pages 298-301
Fox (2002), pages 136-144

VI. Multicollinearity and Its Effects

Read: Berry and Feldman (1985), pages 37-50
Gujarati (2003), pages 341-370
Wooldridge (2006), pages 101-105
Fox (2002), pages 216-224
Kennedy (2003), pages 205-217

VII. Outliers, Unusual, and Influential Observations

Read: Bollen, Kenneth A. and Robert W. Jackman (1990) "Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases." In John Fox and J. Scott Long (Editors), *Modern Methods of Data Analysis*. Newbury Park, CA: Sage.
Fox (2002), pages 191-201
Kennedy (2003), pages 372-388

VIII. Functional Forms, Nonlinearity, and Transformations

Read: McClendon (1994) 230-270
Berry and Feldman (1985), pages 51-72
Wooldridge (2006), pages 304-310
Fox (2002), pages 106-117, 210-216
Kennedy (2003), pages 109-128
Jacoby, William G. (2000) "Loess: A Nonparametric, Graphical Tool for Depicting Relationships Between Variables." *Electoral Studies* 19: 577-613.

IX. Nonnormal and Nonconstant (Heteroskedastic) Errors

Read: McClendon (1994), pages 174-195
Berry and Feldman (1985), pages 73-88
Gujarati (2003), pages 387-428
Wooldridge (2006), pages 181-185
Fox (2002), pages 201-210
Kennedy (2003), pages 133-139

X. Measurement Error and Regression Analysis

Read: Berry and Feldman (1985), pages 26-37
Gujarati (2003), pages 524-528
Wooldridge (2006), pages 318-325
Kennedy (2003), pages 157-163
Jacoby, William G. and Sandra K. Schneider (2007) "Dependent Variable Measurement Error in Regression Models: Complacency, Caution, and Correction." Unpublished manuscript.

XI. Dichotomous Dependent Variables: A Brief Look

A. The Linear Probability Model

Read: Gujarati (2003), pages 580-593
Wooldridge (2006), pages 252-258
Fox, John (1997) *Applied Regression Analysis, Linear Models, and Related Methods*.
Thousand Oaks, CA: Sage Publications, pages 438-443

B. The Logistic Regression and Probit Models

Read: Gujarati (2003), pages 593-616
Wooldridge (2006), pages 582-595
Fox (1997), pages 443-466
Fox (2002), pages 155-177
Kennedy (2003), pages 259-262

XII. Nonindependent Disturbances and Time Series Data: A Brief Look

Read: Gujarati (2003), pages 441-490
Wooldridge (2006), pages 341-375
Fox (1997), pages 369-385
Kennedy (2003), pages 139-156, 163-179

XIII. Good Statistical Practice in Political Science Research

Read: Wooldridge (2006), pages 678-698
Kennedy (2003), pages 389-417
Berk, Richard A. (2004) "What to Do." Chapter 11 in *Regression Analysis: A Constructive Critique*. Thousand Oaks, CA: Sage.

FINAL EXAMINATIONS DUE BY THURSDAY, MAY 1, 5:00 p.m.