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:


As an alternative, you might consider:


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 S² 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


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


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 to 13, 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


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


**VIII. Additional Topics as Time Permits**

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

Gujarati, 16, 17

19(2), 119–122.


