Course Description

The course examines major techniques of multivariate statistical analysis in the social sciences, with emphasis on the applications in policy studies and political science. The course begins with a brief review of research design and fundamentals of statistical analysis, and then moves on to major multivariate techniques commonly used in the social sciences. The course covers logic and application of causal analysis using multivariate cross-tabulation and hypothesis testing, bivariate and multivariate regression, regression diagnostics, Generalized Linear Models (GLMs) for categorical and limited dependent variables, duration models, and panel data analysis. The objective is for you to become familiar enough with these models to understand how, when and why to use them. Emphasis will therefore be on empirical applications and a large portion of class time (probably about 40% of classes) will be devoted to hands-on use and interpretation of these methods in the computer lab. The course stresses applications and interpretations over mathematical foundations, and in particular it will emphasize the applications of multivariate statistics to real data using the statistical programs, Stata and R.

The main topics this semester are regression diagnostics, GLMs for categorical and limited dependent variables, duration analysis, and techniques for causal inference from panel data. Additional and related topics will be covered as necessary. The main tool through which you will familiarize yourself with these methods is Monte Carlo analysis, which will be previewed and presented in the first month of class.

Grades will be based on five parts: class participation during discussion sessions (10%), homework assignments (35%), a final project and presentation (25%), the circulation of replication materials for the final project (5%), and a take-home final exam (25%).

PREREQUISITE: PUBP511/GOVT 511 or equivalent. The background required for the course is a good introduction to probability and statistical inference, and at least one good regression course. Additionally, some familiarity with linear algebra and calculus is assumed, but we can review these topics as necessary.
Course Objectives

1. Knowledge and understanding
   • Students should be able to read and understand journal articles that involve basic multivariate statistical analyses.
   • Students should be able to critically assess policy arguments, comments, reports and other materials that use statistical analysis.
   • Where relevant, students should be able to make policy remarks and draw policy implications/conclusions based on the findings of various policy studies that apply statistical techniques.

2. Develop quantitative skills
   • Students should be able to conduct basic multivariate statistical analyses to analyze various policy issues.
   • Students should be able to properly interpret the outputs of various statistical analyses covered in this course.
   • Students should acquire basic programming skills necessary to conduct basic multivariate statistical analysis in Stata and R.
   • Students should learn how to find, collect, organize and clean datasets necessary to analyze topics of interest.
   • Students should understand the importance of producing reproducible research and learn how to prepare replication programs to this end.

3. Professional development and leadership
   • Students should learn to work effectively in a team setting through homework assignments and the course project.
   • Students should learn how to make professional presentations through a course project.
   • Students should learn how to seek assistance when needed and communicate effectively with classmates and instructors.
   • Students should learn how to deliver assignments and outputs in a timely fashion.

Course Materials

Required Books

   • Buy used if you can.
   • This is very reasonably priced, especially for the electronic versions.
Required Software

1. Stata, at least version 11 and at least the “Intercooled” version
   - Stata is probably the most widely used program for statistical analyses in the social sciences. It is a powerful tool for data management, analysis, and display and there are myriad resources in the extensive manual and in various online communities (e.g., http://stackoverflow.com/questions/tagged/stata, http://www.statalist.org/).
   - Stata is commercial software, currently in version 14. You may use any version greater than 11 for the material covered in this class. Stata is available in the computers in the classroom, but it is recommended that you purchase a 6 month or 1 year license (http://www.stata.com/order/new/edu/gradplans/student-pricing/) for use on your personal computers.

2. R and RStudio
   - R is a statistical environment and high-level programming language for data analysis and visualization. R is free and open source. Unlike Stata and most other statistical packages, it operates by assigning values to objects in the workspace. This makes R more intuitive for those with general programming experience, but it increases the learning curve for those who have exclusive experience in Stata, SPSS, or SAS, for example. Nevertheless, it is much easier to program Monte Carlo analyses in R than in Stata.
   - You can download R from the Comprehensive R Archive Network Network (https://cran.r-project.org/)
   - RStudio is a nice user-friendly interface for R and is highly recommended and also free (https://www.rstudio.com/)

Recommended Books


**Grading Policy**

<table>
<thead>
<tr>
<th>Category</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homework assignments</td>
<td>35%</td>
</tr>
<tr>
<td>Final project and presentations</td>
<td>25%</td>
</tr>
<tr>
<td>Replication files for final project</td>
<td>5%</td>
</tr>
<tr>
<td>Discussion participation</td>
<td>10%</td>
</tr>
<tr>
<td>Take-home final exam</td>
<td>25%</td>
</tr>
</tbody>
</table>

**Letter Grade Distribution**

<table>
<thead>
<tr>
<th>Grade</th>
<th>Points</th>
<th>Grade</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+</td>
<td>&gt;= 97.00</td>
<td>73.00 - 76.99</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>93.00 - 96.99</td>
<td>70.00 - 72.99</td>
<td>C-</td>
</tr>
<tr>
<td>A-</td>
<td>90.00 - 92.99</td>
<td>67.00 - 69.99</td>
<td>D</td>
</tr>
<tr>
<td>B+</td>
<td>87.00 - 89.99</td>
<td>63.00 - 66.99</td>
<td>D</td>
</tr>
<tr>
<td>B</td>
<td>83.00 - 86.99</td>
<td>60.00 - 62.99</td>
<td>D</td>
</tr>
<tr>
<td>B-</td>
<td>80.00 - 82.99</td>
<td>&lt;= 59.99</td>
<td>F</td>
</tr>
<tr>
<td>C+</td>
<td>77.00 - 79.99</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Class Schedule**

*This is a DRAFT - subject to change and completion*

**Week 01, 01/23: Course Overview/Statistical Inference**

- Course Overview / Introduction
- Review of statistical inference/hypothesis testing
- Lab Session
  - Introduction to Stata (overview, structure of commands)
  - Getting help
  - Using and saving datasets/Data management
  - Programming .do files

**Week 02, 01/30: Linear Models for Continuous Outcomes I**

- Review of linear models
  - Estimation with OLS
- Hypothesis tests
- Simple linear regression
- Multiple linear regression

- Lab Session
  - Estimating linear models in Stata, using OLS
- HW 1 DISTRIBUTED (on writing and distributing a Stata .do file that works on any computer)

Week 03, 02/06: Linear Models for Continuous Outcomes II, Maximum Likelihood

- Regression Assumptions and Diagnostics
- Estimation with MLE
- Introduction to Generalized Linear Models
- Likelihood Theory and Estimation
- Lab Session
  - Estimating linear models in Stata, using MLE
  - Introduction to R (overview, structure of commands)
  - Getting help
  - Using and saving datasets/Data management
  - Programming R scripts

Week 04, 02/13: Monte Carlo Analysis/Advanced OLS

- Multiplicative Interactions
- What to do when you have violations
- Article discussion
- HW2 DISTRIBUTED (on writing and distributing an R program that works on any computer)

Week 05, 02/20: Models for Binary Outcomes I

- Linear probability model
- Logit
- Probit

Week 06, 02/27: Models for Binary Outcomes II

- Interpretation/estimating quantities of interest
- Hypothesis testing and goodness of fit
- HW3 DISTRIBUTED (estimation and interpretation on real data/Monte Carlo)
Week 07, 03/06: Models for Ordered/Nominal Outcomes

Week 08, 03/13: Spring Break – NO CLASS

Week 09, 03/20: Models for Event Counts I

Week 10, 03/27: Models for Event Counts II
  - HW4 DISTRIBUTED (estimation and interpretation on real data/Monte Carlo)

Week 11, 04/03: Models for Duration Outcomes

Week 12, 04/10: Endogeneity and Omitted Variable Bias
  - IV

Week 13, 04/17: Models for Inference in Panel Data
  - HW5 DISTRIBUTED (estimation and interpretation on real data/Monte Carlo)

Week 14, 04/24: Class Presentations

Week 15, 05/01: Class Presentations

Week 16, 05/08: Take-home Final Exam DUE