POLSCI 630: Probability and Regression in Political Science

Seminar: M/W 1:25 – 2:40 PM, Gross Hall 111
Lab: Friday 1:25-2:40, Bunche Lab (Gross 276)

Prof. Edmund Malesky, Ph.D
eddy.malesky@duke.edu
919.660.4318
Office hours: W (10 AM to Noon)
Office location: Gross Hall 221

TA's: Ngoc Phan (ngoc.phan@duke.edu) and Haohan Chen (haohanch@gmail.com)

Course Description & Objectives

This course covers basic techniques in quantitative political analysis. It introduces students to widely-used procedures for regression analysis, and provides intuitive, applied, and formal foundations for regression and more advanced methods covered in later studies. This course will use rudimentary calculus and some matrix algebra. This course relies on R for statistical software.

This course strives to achieve four overarching goals. First, students will become literate in regression analysis. Even though basic ordinary least squares (OLS) is not still commonly used in many contemporary political science analyses, the regression framework—and related interpretations of marginal effects, hypothesis testing, causal identification, forecasting, and bias-efficiency tradeoffs—readily generalizes to more state-of-the-art applications. Second, students will establish a foundation in statistical theory and applied econometrics that will help students move forward with their methods training. Third, students will develop experience working with data on topics related to political science, in the context of in-class examples, lab practicums, take-home problem sets and a final research paper. Fourth, students will practice applying the quantitative methods to analyses of their own research questions. Students will collect and analyze data as part of a final project.

Requirements

Grades in the course will be based on the following items:

- **60%** — Problem Sets. At the end of each lab session (Fridays), students will receive a take-home problem set to complete. The problem sets will ask students to demonstrate mastery of statistical theory, as well as in analyzing data to draw inferences. The problem sets are due by the start of class the following Tuesday. Problem sets should be submitted electronically to the course Box folder in PDF form. In addition, students should submit a copy of their R code, with annotations to allow graders to understand their steps. During each lab session, the instructors will cover any questions about the previous problem set before moving on to the current week's material. The lowest problem set score from the semester will be dropped when calculating the final grade.

- **40%** — Methods paper, 15-25 pages (double spaced). This paper will demonstrate students' technical mastery of the practical aspects of OLS regression in the context of a specific research problem of their own formulation. Students will choose a topic, develop a hypothesis, and test it quantitatively. The format should be of a "research note" with heavy emphasis on the technical details. There does not need to be much emphasis on engaging with the existing literature (although proper attribution should be given when appropriate), and the hypotheses need not be novel. The final paper is due on **December 8, by 5pm**. Final papers should be submitted electronically to the course Box folder in PDF form. Students should also submit a copy of their R code, with annotations that describe the steps being taken. Students should have their dataset available for analysis by **November 17**, as the problem set for that week will call for use of your own data.

Course Policies

Late assignments will be penalized. Each day the assignment is late will result in a drop of a letter grade, e.g., A to B, etc. Problem sets and the final paper will be graded on a 16-point scale as follows:

[15-16] -- A	[93.75-100] – A
[14-15) -- A-	[87.5-93.75) – A-
[13-14) -- B+	[81.25-87.5) – B+
[11-13) -- B	[68.75-81.25) – B
[10-11) -- B-	[62.5-68.75) – B-
[9-10) -- C+	[56.25-62.5) – C+
[7-9) -- C	[43.75-56.25) – C
[6-7) -- C-	[37.5-43.75) – C-
[2-6) -- D	[12.5-37.5) – D
[0-2) -- F	[0-12.5) – F

The Duke community standard is in effect throughout the semester. By taking this course, you affirm that it is a violation of the code to cheat on assignments, to plagiarize, to deviate from the teacher's instructions about collaboration on work that is submitted for grades, to give false information to a faculty member, and to undertake any other form of academic misconduct. You also affirm that if you witness others violating the code you have a duty to report them.

Given the nature of this course, some amount of student collaboration is expected and permitted. Students may work on developing their R syntax together by sharing helpful tips, and students are welcome to compare outputs with one another, with the following stipulations: 1) the sharing of ideas must not be one directional, where one students is doing the work and the other is free riding; and 2) the actual write-up of the work that is handed in must be the work of each individual, with absolutely no copying and pasting from one student's work to another's.

If the analyses in the final paper overlap with a project being developed in another for-credit course, permission from both instructors is needed.

Texts

We will rely on the following texts in this course:

- Wooldridge, Jeffrey. 2016. *Introductory Econometrics: A Modern Approach*, 6th Edition. (If you want to use an earlier edition, make sure the chapters match up to the 6th edition.)
- Joshua D. Angrist & Jörn-Steffen Pischke. 2015. *Mastering 'Metrics: The Path from Cause to Effect.*
-Will H. Moore & David A. Siegel. 2013. *A Mathematics Course for Political & Social Research*.

Schedule

1. Aug 27 - Aug 31: Properties of random variables (Moore & Siegel chs. 10 & 11; Wooldridge ch. 1)
    a. Concepts, measurement and construct validity
    b. Expected value and variance
    c. Central Limit Theorem
    d. Theoretical distributions

2. Sep 3 - 7: Comparisons and inference (Mastering 'Metrics ch. 1)
    a. Counterfactual comparisons, covariance, correlation and cross-tabs
    b. Hypothesis testing
        i. Experimental ideal
        ii. Difference of means
        iii. Correlation coefficient

3. Sep 10-14: Regression model estimation (Wooldridge ch. 2)
    a. Least squares estimators
    b. Predictions (expectations and errors)
    c. Model fit

4. Sep 17 - 21: Regression model interpretation (Wooldridge chs. 3 & 4; Mastering 'Metrics ch. 2)
    a. Marginal effects and intercepts
    b. Hypothesis testing
    c. Multiple regression
    d. Graphical representations
        i. Coefficient plots and distributions
        ii. Substantive effects
        iii. Plotting residuals
        iv. Simulation approaches

5. Sep 24-28: Statistical Theory of OLS (Wooldridge ch. 5)
    a. Matrix representation of multiple regression
    b. Gauss Markov Theorem
    c. Bias and efficiency

6. Oct 1-12 (1.5 weeks): Dummy variables and interactions (Wooldridge chs. 6-7)
    a. Additive and interactive effects
    b. Heterogeneous effects
    c. Dummy variables and fixed effects

   d. Interpretations of interaction models
   e. Graphical representation of conditional marginal effects

7. Oct 15- 19: Limited DVs (Wooldridge ch. 17)
   a. Logit
   b. Probit
   c. Poisson

8. Oct 22 – Nov 2 (2 weeks): Model specification and diagnostics (Wooldridge chs. 8 & 9)
   a. Outliers and leverage
   b. Missing data and imputation
   c. Nonlinear transformations and functional forms
   d. Heteroskedasticitity
   e. Collinearity

9. Nov 5-Nov 9: Omitted Variable Bias and Endogeneity (Mastering 'Metrics ch. 3)
   a. Potential Outcomes Framework
   b. Counterfactuals
   c. Sample selection bias
   d. IV/2SLS regression
   e. Fixed vs. Random Effects

10. Nov 12-Nov 23: Other primers in causal identification (Mastering 'Metrics chs. 4-6)
    a. The experimental ideal
    b. Natural experiments
    c. Difference in differences
    d. Regression discontinuity designs