PSCI 8357:
Statistics for Political Research II

Vanderbilt University
Spring 2016
R 9:35 a.m.–12:05 p.m.
Commons Center 349
http://bkenkel.com/psci8357

Professor Brenton Kenkel
brenton.kenkel@vanderbilt.edu
Office hours: Tuesdays, 10:00 a.m.–12:00 p.m.
Commons Center 324

TA: Matt DiLorenzo
matthew.d.dilorenzo@vanderbilt.edu
Office hours: TBD
Commons Center 300

Overview

This course will prepare you to conduct empirical research in political science, with a focus on linear regression models. You should come away from this course an informed consumer and user of the most prevalent statistical techniques in political science. You will also learn to appreciate the connections between statistical practices, research ethics, and the ongoing crisis of confidence in the sciences.

Grading

Your grade will be based on:

- Weekly problem sets (40%).
- Midterm exam (20%).
- Final paper (30%).
- Peer review of final paper drafts (10%).

I will not accept late assignments except in case of a documented family or medical emergency.

Software

All analysis will be conducted in R. You must write and submit your problem sets in R Markdown format, which allows you to embed R code and its output, including graphs, directly in a document. You will submit assignments by pushing to a GitHub repository. I will not accept assignments by email. Don't even think about printing them out. We will discuss homework submission policies—and, along the way, the basics of Git and GitHub—in the first class or recitation. There will be a separate handout laying out the details.

If you are not yet comfortable with the basics of R, here are some hands-on tutorials I recommend completing before the start of the semester:

- Swirl: `http://swirlstats.com`
- Data Camp: `https://www.datacamp.com`

Some other useful resources on R include:

- Jared Lander, *R for Everyone*: `http://amzn.com/0321888030`
- Patrick Burns, *The R Inferno*: `http://www.burns-stat.com/pages/Tutor/R_inferno.pdf` (best consulted when you run into weird error messages or other unexpected behavior)
- R wizard Hadley Wickham's guide to advanced programming in R: `http://adv-r.had.co.nz` (especially the coding style guide `http://adv-r.had.co.nz/Style.html`)

Collaboration Policy

Your work in this course must be the product of your own intellectual labor. Although it is important to learn how to collaborate and co-author, at this stage of your academic careers it is even more crucial that you *personally* comprehend the basic principles of statistics and data analysis. I expect you to follow these rules:

- You may work with, at most, one other student on each assignment. If you choose to do so, you must include a note at the beginning of your assignment specifying who you worked with.

- You must write everything you turn in. You may not directly copy any wording or code written by another student. As a corollary, this means I expect you to understand—and thus be able to answer questions about—all code you turn in.
- No collaboration of any kind is allowed on the midterm.

I consider any violation of these guidelines a violation of the university's Honor Code, and I will deal with such a violation accordingly.

Books

Two textbooks are required:

- Jeff Leek, *The Elements of Data Analytic Style*. E-book available from https://leanpub.com/datastyle.
- Jeffrey M. Wooldridge, *Introductory Econometrics: A Modern Approach*.[1]

I recommend, but do not require, supplementing the selections from Wooldridge with their corresponding treatments in any or all of these more advanced books:

- William H. Greene, *Econometric Analysis*.
- Jack Johnston and John DiNardo, *Econometric Methods*.
- Russell Davidson and James G. MacKinnon, *Estimation and Inference in Econometrics*.
- Michael Kutner, Christopher Nachtsheim, and John Neter, *Applied Linear Regression Models*.
- Frank E. Harrell, Jr., *Regression Modeling Strategies*.
- Cosma Shalizi, *Advanced Data Analysis from an Elementary Point of View*. E-book available from http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/.

The last three of these are particularly useful supplements to Wooldridge, since they're by statisticians rather than econometricians.

[1] Chapter numbers in the syllabus correspond to the 6th edition, but any edition is fine.

Schedule

This schedule is tentative and is subject to change.

Basics

January 14: Working with Data

- Topics
  - Best practices for data management
  - Common data issues
  - Culling data from multiple sources
  - Exploratory data analysis and visualization
- Readings
  - Leek, *Elements of Data Analytic Style*, whole book.
  - Svend Juul, "Take Good Care of Your Data" (typescript, University of Aarhaus, 2004).
  - Hadley Wickham, "Tidy Data," *Journal of Statistical Software* 59, no. 10 (2014).
  - Greg Wilson et al., "Best Practices for Scientific Computing," *PLoS Biology* 12, no. 1 (2014). See also Greg Wilson, "Good Enough Practices for Scientific Computing" (typescript, Software Carpentry Foundation, 2014).

January 21: (Re-)Introduction to Regression

- Topics
  - The linear model in matrix form
  - Deriving the estimator
  - Unbiasedness and the Gauss-Markov theorem
- Readings
  - Wooldridge, chapter 3: "Multiple Regression Analysis: Estimation."
  - Carl P. Simon and Lawrence Blume, *Mathematics for Economists*, chapter 8: "Matrix Algebra."

  - David A. Freedman, "Statistical Models and Shoe Leather," *Sociological Methodology* 21 (1991): 291–313.

January 28: Making Inferences

- Topics
  - Refresher on p-values, standard errors, confidence intervals
  - Consistency and asymptotic normality (and when these properties break down)
  - Perils of the "statistical significance filter"
  - *p*-hacking
- Readings
  - Wooldridge, chapter 4: "Multiple Regression Analysis: Inference."
  - Wooldridge, chapter 6, section 4: "Prediction and Residual Analysis."
  - John P. A. Ioannidis, "Why Most Discovered True Associations Are Inflated," *Epidemiology* 19, no. 5 (2008): 640–648.
  - Neal S. Young, John P. A. Ioannidis, and Omar Al-Ubaydli, "Why Current Publication Practices May Distort Science," *PLoS Medicine* 5, no. 10 (2008).
  - Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn, "False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant" *Psychological Science* 22, no. 11 (2011): 1359–1366.
  - Andrew Gelman and Eric Loken, "The Statistical Crisis in Science," *The American Scientist*
  - Christie Aschwanden, "Science Isn't Broken" (online, FiveThirtyEight, 2015).
  - XKCD, "Extrapolating": `https://xkcd.com/605/`
  - XKCD, "Significant": `https://xkcd.com/882/`
  - XKCD, "P-Values": `http://xkcd.com/1478/`

Beyond the Standard Assumptions

February 4: Specification and Misspecification

- Topics
  - Refresher on quadratic and interaction terms
  - Omitted variable bias
  - Collinearity
  - Generalized additive models
- Readings
  - Wooldridge, chapter 6, sections 2: "More on Functional Form"
  - Wooldridge, chapter 7: "Multiple Regression Analysis with Qualitative Information."
  - Bear F. Braumoeller, "Hypothesis Testing and Multiplicative Interaction Terms," *International Organization* 58, no. 4 (2004): 807–820.
  - Kevin Clarke, "The Phantom Menace: Omitted Variable Bias in Econometric Research," *Conflict Management and Peace Science* 22, no. 4 (2005): 341–352.
  - Christopher H. Achen, "Toward a New Political Methodology: Microfoundations and ART," *Annual Review of Political Science* 5 (2002): 423–450.
  - Trevor Hastie and Robert Tibshirani, "Generalized Additive Models," *Statistical Science* 1, no. 3 (1986): 297–310.

February 11: Non-Constant Variance

- Topics
  - When (and why) heteroskedasticity is a problem
  - Heteroskedasticity of known form: generalized least squares
  - Heteroskedasticity of unknown form: Huber-White "robust" standard errors
- Readings
  - Wooldridge, chapter 8: "Heteroskedasticity."
  - Gary King and Margaret E. Roberts, "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It," *Political Analysis* 23, no. 2 (2015): 159–179.

February 18: Panel Data

- Topics
  - Notation for panel data
  - Unobserved heterogeneity
  - Difference in differences
- Readings
  - Wooldridge, chapter 13: "Pooling Cross Sections across Time."

February 25: Panel Data, continued

- Topics
  - Fixed and random effects
  - Hausman test
  - Clustered standard errors
- Readings
  - Wooldridge, chapter 14: "Advanced Panel Data Methods."
  - Brent R. Moulton, "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units," *Review of Economics and Statistics* 72, no. 2 (1990): 334–338.
  - A. Colin Cameron and Douglas L. Miller, "A Practitioner's Guide to Cluster-Robust Inference," *Journal of Human Resources* 50, no. 2 (2015): 317–372.

March 3: Nonlinear Models: A Brief Overview

Take-home midterm sometime this week—exact timing TBD.

- Topics
  - Linear probability model
  - Logistic regression
- Readings
  - Wooldridge, chapter 7, section 5: "A Binary Dependent Variable: The Linear Probability Model."
  - Wooldridge, chapter 8, section 5: "The Linear Probability Model Revisited."

  - Wooldridge, chapter 17, section 1: "Logit and Probit Models for Binary Response."
  - Nathaniel Beck, "Is OLS with a Binary Dependent Variable Really OK?: Estimating (Mostly) TSCS Models with Binary Dependent Variables and Fixed Effects" (typescript, New York University, 2011).

Causal Inference

March 17: Introduction to Causal Inference

Turn in final paper proposals.

- Topics
  - The Neyman-Rubin model
  - OLS for treatment effect estimation
  - Post-treatment bias
  - Endogeneity as a threat to causal inference
- Readings
  - Paul W. Holland, "Statistics and Causal Inference," *Journal of the American Statistical Association* 81, no. 396 (1986): 945–960. Also read the comments and rejoinders in the same issue.
  - Paul R. Rosenbaum, "The Consquences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment," *Journal of the Royal Statistical Society (Series A)* 147, no. 5 (1984): 656–666.
  - Paul R. Rosenbaum, "Choice as an Alternative to Control in Observational Studies," *Statistical Science* 14, no. 3 (1999): 259–278.

March 24: Instrumental Variables

- Topics
  - Requirements for an instrument
  - Two-stage least squares
- Readings
  - Wooldridge, chapter 15: "Instrumental Variables Estimation and Two Stage Least Squares"

  - Daron Acemoglu, Simon Johnson, and James A. Robinson, "The Colonial Origins of Comparative Development: An Empirical Investigation," *American Economic Review* 91, no. 5 (2001): 1369–1401.
  - Edward Miguel, Shanker Satyanath, and Ernest Sergenti, "Economic Shocks and Civil Conflict: An Instrumental Variables Approach," *Journal of Political Economy* 112, no. 4 (2004): 725–753.
  - Joshua D. Angrist and Alan B. Krueger, "Does Compulsory School Attendance Affect Schooling and Earnings?" *Quarterly Journal of Economics* 106, no. 4 (1991): 979–1014.

March 31: Instrumental Variables, continued

- Topics
  - Weak instruments and other pitfalls
  - Sample selection bias and the Heckman estimator
- Readings
  - Wooldridge, chapter 17, section 5: "Sample Selection Corrections."
  - James J. Heckman, "Sample Selection Bias as a Specification Error," *Econometrica* 47, no. 1 (1979): 153–161.
  - Larry M. Bartels, "Instrumental and 'Quasi-Instrumental' Variables," *American Journal of Political Science* 35, no. 3 (1991): 777–800.
  - John Bound, David A. Jaeger and Regina M. Baker, "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak," *Journal of the American Statistical Association* 90, no. 430 (1995): 443–450.
  - Allison J. Sovey and Donald P. Green, "Instrumental Variables Estimation in Political Science: A Readers' Guide," *American Journal of Political Science* 55, no. 1 (2010): 188–200.

Advanced Topics

April 7: Computationally Intensive Methods

Turn in initial drafts of final papers.

- Topics
  - In-sample vs out-of-sample prediction error
  - Bootstrap, jackknife, and cross-validation
- Readings
  - Wooldridge, chapter 6, section 3: "More on Goodness-of-Fit and Selection of Regressors."
  - Gary King, "How Not to Lie with Statistics: Avoiding Common Mistakes in Quantitative Political Science," *American Journal of Political Science* 30, no. 3 (1986): 666–687.
  - Bradley Efron and Gail Gong, "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *The American Statistician* 37, no. 1 (1983): 36–48.
  - Bradley Efron and Robert Tibshirani, "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science* 1, no. 1 (1986): 54–75.
  - A. Colin Cameron, Jonah B. Gelbach, and Douglas L. Miller, "Bootstrap-Based Improvements for Inference with Clustered Errors," *Review of Economics and Statistics* 90, no. 3 (2008): 414–427.
  - Michael D. Ward, Brian D. Greenhill, and Kristin M. Bakke, "The Perils of Policy by *p*-Value: Predicting Civil Conflicts," *Journal of Peace Research* 47, no. 4 (2010): 363–375.

April 14: Model Selection

Turn in peer reviews.

- Topics
  - Non-nested model tests
  - Bootstrap and cross-validation for model selection
  - Freedman's paradox
  - LASSO
- Readings
  - Kevin Clarke, "Testing Nonnested Models of International Relations: Reevaluating Realism," *American Journal of Political Science* 45, no. 3 (2001): 724–744.
  - Leo Breiman, "Statistical Modeling: The Two Cultures," *Statistical Science* 16, no. 3 (2001): 199–215.

  - Philip A. Schrodt, "Seven Deadly Sins of Contemporary Quantitative Political Analysis," *Journal of Peace Research* 51, no. 2 (2013): 287–300.
  - Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman, *The Elements of Statistical Learning*, chapter 7: "Model Assessment and Selection."
  - David A. Freedman, "A Note on Screening Regression Equations," *The American Statistician* 37, no. 2 (1983): 152–155.
  - Robert Tibshirani, "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society (Series B)* 58, no. 1 (1996): 267–288.

April 21: Missing Data

- Topics
  - Missing completely at random vs missing at random vs nonignorable missingness
  - Multiple imputation
- Readings
  - Donald B. Rubin, "Inference and Missing Data," *Biometrika* 63, no. 3 (1976): 581–592.
  - Joseph L. Schafer, "Multiple Imputation: A Primer," *Statistical Methods in Medical Research* 8, no. 1 (1999): 3–15.
  - James Honaker and Gary King, "What to Do about Missing Values in Time-Series Cross-Section Data," *American Journal of Political Science* 54, no. 2 (2010): 561–581.