PSCI 8357: Statistics II

Professor Brenton Kenkel
Vanderbilt University
Spring 2023

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 important statistical modeling techniques in political science.

General Information

Place and time. Stat II meets in Commons 349 from 9:30–10:45 a.m. on Tuesdays and Thursdays. A weekly recitation will be held at a time TBD.

Contact info. You can reach me by email at `brenton.kenkel@vanderbilt.edu`.

Office hours. My office hours are Mondays from 1:00–2:30 p.m. in Commons 326.

TA. The TA for Stat II is Martín Gou. His office hours are TBD. You can also email him at `fernando.martin.gou@vanderbilt.edu`.

Grading

Your grade will be based on:

- Weekly problem sets (30%).
- Midterm exam (20%) in class on Thursday, March 9.
- Data analysis paper (10% proposal, 15% initial draft, 25% final paper).
    1. Proposal due Monday, February 27, at 5:00 p.m.
    2. Initial draft due Monday, April 3, at 5:00 p.m.
    3. Final paper due Friday, April 28, at 5:00 p.m.

Late assignments will not be accepted except in case of a documented family or medical emergency. It is much better to turn in work that is imperfect or incomplete than to turn in nothing at all.

Books

Garner access to the following books:

- Jeff Leek, *The Elements of Data Analytic Style*. E-book available from `https://leanpub.com/datastyle`. [EDAS]
- Joshua D. Angrist and Jörn-Steffen Pischke, *Mostly Harmless Econometrics*. [MHE]

  You may also want to grab the companion book *Mastering 'Metrics*, a less technical introduction to the same material.
- William H. Greene, *Econometric Analysis*, 7th edition.

  Earlier or later editions are fine, but the chapter numbers in the syllabus might not match.

Schedule and Readings

The most important thing to read each unit is the corresponding chapter of the lecture notes, posted online at `https://bkenkel.com/pdaps`. Next most important are the readings listed in **bold**. All others listed are for additional reference or edification.

1 Principles of Programming and Data Management

Data analysis as programming. Reproducibility via scripts. Control structures. Functions. Best practices for handling and sharing data. Tidy data principles.

Bowers (2011); **Wilson et al. (2014)**; **EDAS (entire book)**; Wilson et al. (2016); Healy (2016); Wickham (2014).

2 Matrix Algebra (A Crash Course)

Matrix notation. Basic operations. Inversion and invertibility. Systems of linear equations, and their solutions.

Simon and Blume ch. 8 (handout); Greene Appendix A ("Matrix Algebra").

3 Reintroduction to the Linear Model and OLS

Conditional expectation. The linear model in matrix form. The ordinary least squares estimator.

MHE 3.1.1–3.1.2; Green ch. 2–3 ("The Linear Regression Model" and "Least Squares").

4 OLS Inference

Sampling distribution of OLS. Joint hypothesis tests.

MHE 3.1.3; Green ch. 4 ("The Least Squares Estimator").

5 Non-Spherical Errors

Heteroskedasticity. Weighted least squares. "Robust" standard errors. Autocorrelation, briefly.

MHE 8.1; White (1980); Greene ch. 9, 20 ("The Generalized Linear Model and Heteroskedasticity" and "Serial Correlation"); King and Roberts (2015); Aronow (2016).[1]

6 The Statistical Crisis in Science

Publication bias. p-hacking. The garden of forking paths.

Ioannidis (2008); **Simmons, Nelson and Simonsohn (2011)**; **Gelman and Loken (2014)**.

[1] If you read the King and Roberts paper, you *must* also read the Aronow paper.

7 Clustered and Panel Data

Grouped data notation. Random and fixed effects. Hausman test. "Cluster-robust" standard errors.

MHE 8.2; **Moulton (1990)**; Cameron and Miller (2015); Greene ch. 11 ("Models for Panel Data").

8 Binary Outcomes

Maximum likelihood estimation. Logit and probit. Interpreting nonlinear models. Average marginal effects. Comparison to linear probability model.

MHE 3.4.2; **Hanmer and Kalkan (2013)**; Greene ch. 14, 17 ("Maximum Likelihood Estimation" and "Discrete Choice").

9 Reintroduction to Causal Inference

Conditional independence. The fundamental problem of causal inference. Potential outcomes model. Covariate selection. Why you can't "sign the bias".

MHE 3.2; **Holland (1986) and responses in same issue**; Manski (1990); Freedman (1991); Rosenbaum (1984); Rosenbaum (1999); Heckman (2005).

10 Instrumental Variables

Conditions for an instrument. Wald estimator. Two-stage least squares.

MHE 4.1; **Angrist and Krueger (1991)**; **Acemoglu, Johnson and Robinson (2001)**; **Miguel, Satyanath and Sergenti (2004)**; Bartels (1991); Bound, Jaeger and Baker (1995); Sovey and Green (2011); Greene ch. 8 ("Endogeneity and Instrumental Variable Estimation").

11 Predictive Modeling

In- versus out-of-sample error. Cross-validation. Ridge regresion. LASSO.

Hastie, Tibshirani and Friedman (2009, ch. 7); **Breiman (2001)**; Tibshirani (1996).

12 Computational Techniques

Parametric and nonparametric bootstrap. Jackknife. Clustered variants. Applications to nonlinear model inference.

Efron and Gong (1983); King, Tomz and Wittenberg (2000); Greene ch. 15 ("Simulation-Based Estimation and Inference and Random Parameter Models").

13 Missing Data

Varieties of missingness. Problems with listwise deletion. Multiple imputation.

Rubin (1976); **Little (1992)**; Schafer (1999).

References

Acemoglu, Daron, Simon Johnson and James A Robinson. 2001. "The Colonial Origins of Comparative Development: An Empirical Investigation." *The American Economic Review* 91(5):1369–1401.

Angrist, Joshua D and Alan B Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106(4):979–1014.

Aronow, Peter M. 2016. "A Note on "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It".".

Bartels, Larry M. 1991. "Instrumental and "Quasi-Instrumental" Variables." *American Journal of Political Science* 35(3):777.

Bound, John, David A Jaeger and Regina M Baker. 1995. "Problems with Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogeneous Explanatory Variable is Weak." 90(430):443.

Bowers, Jake. 2011. "Six Steps to a Better Relationship with Your Future Self." *The Political Methodologist* 18(2):2–8.
URL: `http://tpm.blogs.rice.edu/files/2013/09/tpm_v18_n2.pdf`

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures." *Statistical Science* 16(3).

Cameron, A Colin and Douglas L Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50(2):317–372.

Efron, Bradley and Gail Gong. 1983. "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation." *The American Statistician* 37(1):36.

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

Gelman, Andrew and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102(6):460.
URL: `http://www.americanscientist.org/issues/feature/2014/6/the-statistical-crisis-in-science`

Hanmer, Michael J. and Kerem Ozan Kalkan. 2013. "Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models." *American Journal of Political Science* 57(1):263–277.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning*. 2 ed. New York: Springer.
URL: `http://statweb.stanford.edu/~tibs/ElemStatLearn/`

Healy, Kieran. 2016. "The Plain Person's Guide to Plain Text Social Science.".
URL: `http://plain-text.co`

Heckman, James J. 2005. "The Scientific Model of Causality." *Sociological Methodology* 35(1):1–97.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–960.

Ioannidis, John P A. 2008. "Why Most Discovered True Associations Are Inflated." *Epidemiology* 19(5):640–648.

King, Gary and Margaret E Roberts. 2015. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It." *Political Analysis* 23(2):159–179.

King, Gary, Michael Tomz and Jason Wittenberg. 2000. "Making the Most of

Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* pp. 347–361.

Little, Roderick J A. 1992. "Regression With Missing X's: A Review." *Journal of the American Statistical Association* 87(420):1227.

Manski, Charles F. 1990. "Nonparametric Bounds on Treatment Effects." *The American Economic Review* 80(2):319–323.

Miguel, Edward, Shanker Satyanath and Ernest Sergenti. 2004. "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." *Journal of Political Economy* 112(4):725–753.

Moulton, Brent R. 1990. "An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units." *The Review of Economics and Statistics* 72(2):334.

Rosenbaum, Paul R. 1984. "The Consquences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment." 147(5):656–666.

Rosenbaum, Paul R. 1999. "Choice as an Alternative to Control in Observational Studies." *Statistical Science* 14(3):259–278.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63(3):581–592.

Schafer, Joseph L. 1999. "Multiple imputation: a primer." *Statistical Methods in Medical Research* 8(1):3–15.

Simmons, Joseph P, Leif D Nelson and Uri Simonsohn. 2011. "False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant." *Psychological Science* 22(11):1359–1366.

Sovey, Allison J and Donald P Green. 2011. "Instrumental Variables Estimation in Political Science: A Readers' Guide." *American Journal of Political Science* 55(1):188–200.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." 58(1):267–288.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48(4):817.

Wickham, Hadley. 2014. "Tidy Data." *Journal of Statistical Software* 59(10).

Wilson, Greg, D A Aruliah, C Titus Brown, Neil P Chue Hong, Matt Davis, Richard T Guy, Steven H D Haddock, Kathryn D Huff, Ian M Mitchell, Mark D Plumbley, Ben Waugh, Ethan P White and Paul Wilson. 2014. "Best Practices for Scientific Computing." *PLOS Biology* 12(1):e1001745.

Wilson, Greg, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt and Tracy K Teal. 2016. "Good Enough Practices in Scientific Computing.".