POS 7113: Quantitative Research Methodology II

Tulane University
Center for Inter-American Policy & Research
Richardson Building,
Second Floor, Room M
New Orleans, LA 70112

Last updated: February 28, 2018.
Download last version here.
Not intended as a definitive version

Professor: Hector Bahamonde
e:hbahamonde@tulane.edu
w:www.HectorBahamonde.com
Class meetings: DAY TIME.
Location: TBA.
Office Hours: Make an appointment here.

Overview and Objectives

What is the influence of income on the probability of voting? Is there any systematic relationship between democracy and trade? Is it true that correlation does not imply causation? To consider these kind of statements in a rigorous way, we will learn a technique called *ordinary least regression*, the workhorse of applied quantitative methodology in social sciences. Additionally, the course provides in-depth, hands-on training in statistical techniques and research methodologies used across the social and behavioral sciences. At the end of the course you will become an educated consumer and producer of quantitative analysis.

This **graduate-level course** offers an introduction to linear models in political science. This course will hopefully prepare you for the things you will encounter when you (attempt to) publish quantitative work with linear models. This course will provide you with a systematic approach to assessing, fixing and presenting your linear model results. Though we focus almost exclusively on linear models, some nonlinear models will also be introduced.

During the semester, we will focus on a number of assumptions, applications, violations to those assumptions, and their corresponding solutions. It is the intention of this course to be both theoretically relevant and practically oriented at the same time. As you will quickly realize, you will "learn by doing." Besides learning quantitative methodology, you will also learn a number of computer skills, mainly **R** and LaTeX. Therefore, if you are not already with those tools, try to be patient with your self.

I hope this course catches your attention, in the expectation that you continue taking more methods and/or computing courses. Most of all, I hope you see how applied statistics are both an art and a science. **Welcome!**

Methodology

This is a 42-hours **independent study** which will take place between the weeks of **July 2** and **August 20**. In these **8 weeks**, the student will meet the professor **twice a week**. Each meeting will cover one item from the schedule section . Since there are 16 topics to be covered, they will be distributed in 8 weeks, in $\frac{16}{8} = 2$ weekly meetings. Each meeting will last 2.5 hours. 2.5 hours × 2 weekly meetings × 8 weeks = 40 hours. There are two hours to spare during the 8-weeks period, which will be used to solve additional questions, and computational issues. The emphasis of each meeting will cover both the assigned readings, and the applied-computational components. During the meetings, applied examples will be given, particularly, regarding applications to political science.

Prerequisites

There are no formal prerequisites for this course. Only good predisposition to learn and spend quality-time doing the readings and problem sets. That said, basic math, calculus, probability and linear algebra is assumed at the level covered in any "Math Camp," or equivalently at the level of *Essential Mathematics for Political and Social Research* (Gill, 2006).

Software

It is not enough to just *read* about statistics. The best way to learn applied statistics is by *doing*. For this reason, in this course you will sit in the driving seat and actually estimate (a lot of) statistical models using **R**. **R** is free and compatible with essentially all kinds of machines. One of **R**'s main virtues from the grad-student point of view is that the base package and all of the add-ons (called "packages" in **R**) are free. You can download the base package of R from the *Comprehensive R Archive Network* (*CRAN*) website. You can find a virtually endless set of resources for **R** on the Internet, including this Getting Started With R page. If you are completely new to **R**, you should complete this online short course, Try R. Finally, I suggest a **R** interphase called `RStudio`. RStudio makes **R** easier to use. It includes a code editor, debugging and visualization tools.

Quantitative methodology is not only about estimating the "best" model (if such a thing even exists) but also about presenting your results in an effective way. For this reason, all problem sets and final presentations should be carried out in LaTeX. LaTeX is a high-quality typesetting system; it includes features designed for the production of technical and scientific documentation. LaTeX is the *de facto* standard for the communication and publication of scientific documents. Download the last version here.

You are free to use other softwares such as `STATA`, `MathLab`, `SAS`, etc. However, I might provide assistance only for the former. That said, all in-class examples are going to be taught in **R**, which is by far the most common statistical tool nowadays.

Course Texts

- Fox, John. (2015) Applied Regression Analysis and Generalized Linear Models. 3rd ed. Thousand Oaks, CA: Sage Publications, Inc.
- Fox, John and Sanford Weisberg. (2011) An R Companion to Applied Regression. 2nd ed. Thousand Oaks, CA: Sage Publications, Inc.

Course Learning Objectives

Upon successful completion of this course, you will be able to:

- Acquire an understanding of the main tools related to linear regression.
- Use the linear method up to the point where you feel comfortable doing analysis in your own research.
- Consume *critically* methodological-oriented literature.

Requirements and Evaluations

1. **Student Project, Date**: 50 %.

   An empirical paper that applies methods learned in this class to a research question of their choice. The paper should be 10-15 pages in length and focus on the research question, data, empirical strategy, results, and conclusions. Literature reviews, background, lengthy motivations, etc. should be omitted or may be included as an appendix.

   You should be able to convince me that your empirical strategy is solid. For that, you should provide graphs, tables, robustness checks, among other tools.

   The paper should be very clear about the kind of model used, and it should use proper mathematical notation. The paper should be written in LaTeX. Documents in *Word* will not be accepted.

   You also need to submit a copy of your **R** code. I should be able to replicate your exact results. That is, I should be able to hit "run" in *RStudio* and obtain the exact same output. Failure to replicate your results will discount points out of grade.

   Students are free to choose any topic they want. As long as it has a strong data-analytical component, it's fine by me. I strongly advice you to get the data in advance. As you will quickly realize, the real world does not work as when you do homeworks, e.g. the data you will originally find, will be messy, with lots of inconsistencies, variables will not be labeled, among other programs. My advice: *start early*. One alternative I propose is to find a dataset that interests *you* in the ICPSR repository.

2. **Problem Sets, Date**: 30%.

   This is a methodological course, developing skills in understanding and applying statistical methods. You can only learn statistics by doing statistics and therefore the homework for this course is extensive, including weekly homework assignments. The assignments consist of analytical problems, computer simulations, and data analysis.

   In this section you will learn how to recode a variable, merge datasets, summarize variables, plot a distribution, among other more theoretical issues. And while we will not focus on mathematical proofs, you will spend time thinking about and working on the mechanics of linear regression.

3. **Final Exam**: 20 %.

   There will be a single, take-home final exam in the course that will require that you apply the knowledge gained in the course to particular methodological problem. The format of the exam will be discussed by the end of the semester. However, you can expect the test to be an exercise in guided replication and primarily involves data analysis and interpretation. Exam questions will be drawn both from the readings and lectures. The final exam is set by the registrar. Hence, both place and time are TBA. There will not be exceptions. We will also schedule an in-class review session for **date**.

Grading

This course will be grade according to the following scale: A: $\geq$ 93, A-: 90-92, B+: 87-89, B: 83-86, B-: 80-82, C+: 77-79, C:73-76, C-: 70-72, D+: 67-69, D:63-66, D-: 60-62, and F: $\leq$ 59.

Disputing Grades

I am happy to go over any exam or paper with you. Request for re-grading, though, must be done in writing. Please refer to my re-grading policy.

Academic Integrity

In accordance with Tulane University policy on Academic Integrity, you are expected to fully comply with the school's `policies`.

Students with Disabilities

Students with disabilities who require accommodation should check with the `Goldman Center for Student Accessibility`.

Absence from Exams

There will be no make-up exams unless you have a *documented* **medical** emergency. If at all possible, I need to be notified before the exam of your inability to take it. Absence from an exam because of travel plans will not be excused. Make travel plans accordingly.

Office Hours

I have an open-doors policy, feel free to stop by my office at any time. However, you might want to minimize the risks that I am not there, or can't meet you that day. I advice you then to `schedule time with me` using my automatic scheduler. I think fixed office hours do not work because ... well, they are *fixed*. I prefer flexibility. Hence, you can see me any day/time that's available during the week. Do not send me a reminder as I will receive an alert: If the time spot is available, I am happy to see you there.

Schedule

Each entry represents a single topic. Readings are designated either as required (⋆) or supplemental (-). This should serve as a nice reference to which you can return if the intricacies of a particular topic have faded from your memory.

1. Preliminary Material

   ⋆ Fox (2015), Chapters 1 & 2.

   ⋆ Fox and Weisberg (2011), Chapters 1 & 2.

2. OLS II: Effective Presentation

   □ Factors and contrasts; quasi-variances and graphical displays

   □ Interactions and effect displays

   □ Standardization and relative importance

   ⋆ Armstrong II (2013)

   ⋆ Berry, Golder and Milton (2012)

   ⋆ Silber, Rosenbaum and Ross (1995)

   - Brambor, Clark and Golder (2006)

- Braumoeller (2004)
- Firth and Menzes (2004)
- Kam and Franzese (2007)

3. Self-guided Replication Computational Exercise

□ Factors and contrasts.
□ Interactions.
□ Relative Importance.

4. Linearity: Diagnostics, Transformations and Polynomials

□ Diagnosing linearity through residual plots.
□ Fixing non-linearity with data transformations and polynomials.
□ Linearity and ordinal variables.

⋆ Fox (2015) Chapters 4 & 12 (Sections 12.3-12.5)
⋆ Fox and Weisberg (2011) Chapter 3
⋆ Jacoby (1999)
- Box and Tidwell (1962)
- Breiman and Friedman (1985a,b), Pregibon and Vardi (1985), Buja and Kass (1985), Fowlkes and Kettering (1985)

5. Re-sampling Techniques and Regression

□ Bootstrapping and Jackknifing.
□ Cross-validation.

⋆ Fox (2015) Chapter 21
- Stone (1974)
- Efron and Tibshirani (1993)
- Davison and Hinkley (1997)
- Ronchetti, Field and Blanchard (1997)

6. Model Selection

□ Theoretical issues in model searching and post-data model construction.
□ Model selection criteria and multi-model inference.
□ Subset selection models.

⋆ Fox (2015) Chapter 22
⋆ Leamer (1983)
⋆ Burnham and Anderson (2004)
⋆ Leamer and Leonard (1983)
⋆ Box (1976), Box and Hunter (1962)
- Freedman (1991b,a), Berk (1991), Blalock (1991), Mason (1991)

  - Miller (2002), Breiman (1992), Breiman and Spector (1992)

7. Non-Linearity, Smoothing and Splines

   □ Nonparametric Smoothing - Lowess.
   □ Inference for regression smoothers.
   □ Regression Splines.
   □ Generalized Additive Models.

   ⋆ Fox (2015) Chapters 17 & 18
   ⋆ James et al. (2013) Chapter 7
   - Keele (2008) Chapters 2-6

8. Flexible Models: Tree-based Regression, Multivariate Adaptive Regression Splines

   □ Fundamentals of flexible models.
   □ Automatic variable selection.
   □ Inference and effects in statistical learning models.
   □ When (and when not) to use these kinds of models.

   ⋆ Montgomery and Olivella (forthcoming)
   ⋆ James et al. (2013) Chapter 7
   ⋆ Berk (2016) Section 3.14

9. Self-guided Replication Computational Exercise

   □ Non-linearity transformations.
   □ Polynomials.
   □ Smoothers and splines.
   □ Trees and other flexible methods.

10. Regression Discontinuity Designs

    ⋆ Cattaneo, Idrobo and Titiunik (2017).
    ⋆ Calonico, Cattaneo and Titiunik (2015) .
    ⋆ Keele (2015).
    ⋆ Sekhon and Titiunik (2016).

11. Finite Mixture Models

    ⋆ Imai and Tingley (Forthcoming).
    ⋆ Grun and Leisch (2008).
    ⋆ Grun and Leisch (2007).

12. Missing Data and Multiple Imputation

    □ Whats the problem with missing data?
    □ When can we fix it?

- □ How do we impute the data and use those imputations?

- ⋆ Fox (2015) Chapter 20
- ⋆ van Buuren and Groothuis-Oudshoorn (2011)
- ⋆ Honaker and King (2010)
- ⋆ Cranmer and Gill (2013)
- ⋆ Akande, Li and Reiter (forthcoming)
- ⋆ Xia and Yang (2016)
- ⋆ Resseguier, Giorgi and Paoletti (2011)
- - Schafer (1997)
- - Rubin (1987)

13. Self-guided Replication Computational Exercise

    - □ RDD.
    - □ Mixture Models.
    - □ Missing Data and Multiple Imputation.

14. Outliers and Influential Data: Diagnostics

    - □ Outliers, leverage and influential data.
    - □ Hat values, standardized residuals, Cook's D.
    - □ M-estimation (and extension) and iterative re-weighted least squares.
    - □ Diagnostics for outliers revisited.

    - ⋆ Fox (2015) Chapter 11.
    - ⋆ Fox and Weisberg (2011) Chapter 6 (pp 101-201).
    - ⋆ Andersen (2008).
    - ⋆ Fox (2015) Chapter 19.
    - - Cantoni and Ronchetti (2001).
    - - Rousseeuw and Leroy (1987).
    - - Jasso (1985, 1996), Kahn and Udry (1986).

15. Non-constant error variance and collinearity: Diagnostics and Fixes

    - □ Residual plots.
    - □ ML transformations of Y.
    - □ Weighted least squares.
    - □ Heteroskedastic linear regression.
    - □ Robust standard errors.

    - ⋆ Fox (2015) Chapters 12 & 13.
    - ⋆ Fox and Weisberg (2011) Chapters 3 & 6.
    - ⋆ Long and Ervin (2000).
    - ⋆ King and Roberts (2015).

- Harvey (1976).
- Cribari-Neto (2004), Cribari-Neto, Souza and Vasconcellos (2007), Cribari-Neto and da Silva (2011).

16. Critiques of the Linear Regression Model

□ How important are the assumptions behind OLS Regression?

□ How should we appropriately use regression models?

□ The importance of sampling to inference.

⋆ Berk (2004).

References

Akande, Olanrewaju, Fan Li and Jerome Reiter. forthcoming. An Empirical Comparison of Multiple Imputation Methods for Categorical Data. The American Statistician.

Andersen, Robert. 2008. Modern Methods for Robust Regression. Thousand Oaks, CA: Sage.

Armstrong II, David A. 2013. factorplot: Improving Presentation of Simple Contrasts in GLMs. The R Journal 5(2):4-15.

Berk, Richard A. 1991. Toward a Methodology for Mere Mortals. Sociological Methodology 21:315-324.

Berk, Richard A. 2004. Regression Analysis: A Constructive Critique. Thousand Oaks, CA: Sage.

Berk, Richard A. 2016. Statistical Learning from a Regression Perspective, 2nd ed. Switzerland: Springer.

Berry, William, Matt Golder and Daniel Milton. 2012. Improving Tests of Theories Positing Interaction. Journal of Politics 74(3):653-671.

Blalock, Hubert M. 1991. Are There Really Any Constructive Alternatives to Causal Modeling? Sociological Methodology 21:325-335.

Box, George E. P. 1976. Science and Statistics. Journal of the American Statistical Association 71(356):791-799.

Box, George E. P. and William G. Hunter. 1962. A Useful Method for Model-Building. Technometrics 4(3):301-318.

Box, George and P.W. Tidwell. 1962. Transformation of the Independent Variables. Technometrics 4:531-550.

Brambor, Thomas, William Clark and Matt Golder. 2006. Understanding Interaction Models: Improving Empirical Analyses. Political Analysis 14(1):63-82.

Braumoeller, Bear F. 2004. Hypothesis Testing and Multiplicative Interaction Terms. International Organization 58(4):807-820.

Breiman, Leo. 1992. The Little Bootstrap and Other Methods for DImensionality Se- lection in Regression: X-Fixed Prediction Error. Journal of the American Statistical Association 87(419):738-754.

Breiman, Leo and Jerome H. Friedman. 1985a. Estimating Optimal Transformations for Multiple Regression and Correlation. Journal of the American Statistical Association 80(391):580-598.

Breiman, Leo and Jerome H. Friedman. 1985b. Estimating Optimal Transformations for Multiple Regression and Correlation: Rejoinder. Journal of the American Statistical Association 80(391):614-619.

Breiman, Leo and Philip Spector. 1992. Submodel Selection and Evaluation in Regres- sion. The X-Random Case. Internation Statistical Review 60(3):291-319.

Buja, Andreas and Robert E. Kass. 1985. Estimating Optimal Transformations for Multiple Regression and Correlation: Comment. Journal of the American Statistical Association 80(391):602-607.

Burnham, Kenneth P. and David R. Anderson. 2004. Multimodel Inference: Understand- ing AIC and BIC in Model Selection. Sociological Methods and Research 33(2):261- 304.

Calonico, Sebastian, Matias D. Cattaneo and Rocio Titiunik. 2015. rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs. The R Journal 7(1):38-51.

Cantoni, Gustavo E. and Elvezio Ronchetti. 2001. Robust Inference for Generalized Linear Models. Journal of the American Statistical Association 96:1022-1030.

Cattaneo, Matias D., Nicolas Idrobo and Rocio Titiunik. 2017. A Practical Introduction to Regression Discontinuity Designs.. Working Manuscript.

Cranmer, Skyler J. and Jeff Gill. 2013. We Have to Be Discrete About This: A Non- Parametric Imputation Technique for Missing Categorical Data. British Journal of Political Science 43:425-449.

Cribari-Neto, Francisco. 2004. Asymptotic Inference Under Heteroskedasticity of Un- known Form. Computational Statistics and Data Analysis 45:215-233.

Cribari-Neto, Francisco, Tatiene C. Souza and Klaus L.P. Vasconcellos. 2007. Inference Under Heteroskedasticity and Leveraged Data. Communications in Statistics - Theory and Methods 36(10):1877-1888.

Cribari-Neto, Francisco and Wilton Bernardino da Silva. 2011. A New Heteroskedasticity-consistent Covariance Matrix Estimator for the Linear Regression Model. Advances in Statistical Analysis 95(1):129-146.

Davison, Anthony C. and D.V. Hinkley. 1997. Bootstrap Methods and Their Applications. Cambridge: Cambridge University Press.

Efron, Bradley and Robert Tibshirani. 1993. An Introduction to the Bootstrap. New York: Chapman & Hall.

Firth, David and Renee X. De Menzes. 2004. Quasi-Variances. Biometrika 91(1):65-80.

Fowlkes, E.B. and J.R. Kettering. 1985. Estimating Optimal Transformations for Mul- tiple Regression and Correlation: Comment. Journal of the American Statistical As- sociation 80(391):607-613.

Fox, John. 2015. Applied Regression Analysis and Generalized Linear Models, 3rd edition. Thousand Oaks, CA: Sage, Inc.

Fox, John and Sanford Weisberg. 2011. An R Companion to Applied Regression, 2nd ed. Thousand Oaks, CA: Sage.

Freedman, David A. 1991a. A Rejoinder to Berk, Blalock and Mason. Sociological Methodology 21:353-358.

Freedman, David A. 1991b. Statistical Models and Shoe Leather. Sociological Methodology 21:291-313.

Grun, Bettina and Friedrich Leisch. 2007. Fitting Finite Mixtures of Generalized Linear Regressions in R. Computational Statistics & Data Analysis 51(11):5247-5252.

Grun, Bettina and Friedrich Leisch. 2008. FlexMix Version 2: Finite Mixtures with Concomitant Variables and Varying and Constant Parameters. Journal of Statistical Software 28(4):1-35.

Harvey, Andrew C. 1976. Estimating Regression Models with Multiplicative Heteroskedasticity. Econometrica 44(3):461-465.

Honaker, James and Gary King. 2010. What to Do about Missing Values in Time-Series Cross-Section Data. American Journal of Political Science 54(2):561-581.

Imai, Kosuke and Dusting Tingley. Forthcoming. A Statistical Method for Empirical Testing of Competing Theories. American Journal of Political Science x(x):xxx-xxx.

Jacoby, William G. 1999. Levels of Measurement and Political Research: An Optimistic View. American Journal of Political Science 43(1):271-301.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. An Introduction to Statistical Learning with Applications in R. New York, NY: Springer.

Jasso, Guillermina. 1985. Marital Coital Frequency and the Passage of Time: Estimating the Separate Effects of Spouses' Ages and Marital Duration, Birth and Marriage Cohorts, and Period Influences. American Sociological Review 50(2):224-241.

Jasso, Guillermina. 1996. Is It Outlier Deletion or is it Sample Truncation? Notes on Science and Sexuality. American Sociological Review 51(5):738-742.

Kahn, Joan R. and J. Richard Udry. 1986. Marital Coital Frequency: Unnoticed Outliers and Unspecified Interactions Lead to Erroneous Conclusions. American Sociological Review 51(5):734-737.

Kam, Cindy and Robert J. Franzese. 2007. Modeling and Interpreting Interactive Hy- potheses in Regression Analyses. Ann Arbor: University of Michigan Press.

Keele, Luke J. 2008. Semi-parametric Regression for the Social Sciences. New York: Wiley & Sons, Inc.

Keele, Luke J. 2015. Geographic Boundaries as Regression Discontinuities. Political Analysis 23:127-155.

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.

Leamer, Edward E. 1983. Let's Take the Con Out of Econometrics. The American Economic Review 73(1):31-43.

Leamer, Edward E. and Herman Leonard. 1983. Reporting the Fragility of Regression Estimates. The Review of Economics and Statistics 65(2):306-317.

Long, J. Scott and Laurie H. Ervin. 2000. Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. The American Statistician 54(3):217-224.

Mason, William M. 1991. Freedman is Right as Far as He Goes, but THere is More, and It's Worse. Statisticians Could Help. Sociological Methodology 21:337-357.

Miller, Alan. 2002. Subset Selection in Regression, 2nd edition. Boca Raton, FL: Chapman & Hall/CRC.

Montgomery, Jacob and Santiago Olivella. forthcoming. Tree-based Models for Political Science Data.

American Journal of Political Science .

Pregibon, Daryl and Yehuda Vardi. 1985. Estimating Optimal Transformations for Multiple Regression and Correlation: Comment. Journal of the American Statistical Association 80(391):598-601.

Resseguier, Noemie, Roch Giorgi and Xavier Paoletti. 2011. Sensitivity Analysis When Data Are Missing Not-at-random. Epidemiology 22(2):282.

Ronchetti, Elvezio, Christopher Field and Wade Blanchard. 1997. Robust Linear Model Selection by Cross-validation. Journal of the American Statistical Association 92(439):1017-1023.