Introduction to Linear Regression, Estimation, and Inference

POL 682 Syllabus

Christopher Weber, PhD

Table of contents

Instructor

Christopher Weber, PhD School of Government and Public Policy University of Arizona 333 Social Sciences Building chrisweber@arizona.edu

Office Hours: T 12:00PM-1:30PM (and by appointment).

Meeting Location

Tuesdays 2:00PM - 4:30PM

SS 332

Course Objectives

The purpose of this course is to introduce students to linear regression, inference, and estimation tools and techniques. Although many research questions are complex, regression allows researchers to test difficult hypotheses using a series of relatively straightforward techniques. While this class will review and critique *Ordinary Least Squares*, linear regression is integral to the vast majority of advanced statistical methods explored in POL 683.

The topics covered are not exhaustive. This course establishes the steps involved in deriving the ordinary least squares (OLS) estimator, applying linear regression to inference, and reviewing the assumptions necessary for the classical linear regression model.

The first portion focuses on derivation, assumptions, data manipulation, and inference. The classical linear model is the Best Linear Unbiased Estimator only under specific assumptions. The second half examines how these assumptions are violated and possible solutions, including heteroskedasticity, autocorrelation, collinearity, model specification, measurement error, outliers, and endogeneity.

In the final weeks, we introduce non-linear models, including regression with categorical dependent variables, endogeneity, clustering, and measurement error. Throughout, we use an *open science* framework with R for statistical computing and git/github for collaboration and reproducibility.

Learning Outcomes:

1. Demonstrate an understanding of core statistical concepts pertaining to linear regression, estimation, and inference.
2. Critically evaluate statistical research in political science, and communicate statistical concepts effectively.
3. Develop skills in R programming, git, and GitHub for data analysis, collaboration, and scientific reproducibility.
4. Apply linear regression methods to analyze political science data and interpret results.

Three Points:

1. It is best to learn statistics by practicing statistics. In this class, you will complete homework assignments, exams, and a final paper applying these methods. Please budget adequate time for studying and completing assignments. Typically, cramming is ineffective for learning complicated mathematical concepts.
2. This class develops your ability to evaluate and communicate statistical research, particularly linear regression, in a critical manner. With strong statistical knowledge, it becomes easier to differentiate rigorous from poor arguments.
3. This class introduces concepts and tools commonly used in political science research to advance scientific transparency and reproducibility. You will learn R programming, `git`, and `GitHub` for data analysis and collaboration, and `shiny` for interactive applications. While the benefits to these tools are not always immediately apparent, they are invaluable for accurate, transparent, and reproducible research.

Required Materials

Textbook:

Fox, John. 2008. *Applied Regression Analysis and Generalized Linear Models.* Thousand Oaks, California: Sage.

Course Notes:

Course Notes

Additional Readings:

Journal articles available through the University of Arizona Library via JSTOR and ejournals.

Recommended References:

- Greene, William H. 2018. *Econometric Analysis*, 8th Edition. New York: Pearson.
- Berry, William and Stanley Feldman. 1985. *Multiple Regression In Practice.* Newbury, CA: Sage.
- Eliason, Scott. 1993. *Maximum Likelihood Estimation: Logic and Practice.* Thousand Oaks, CA: Sage.
- Fox, John and Sanford Weisberg. 2011. *An R Companion to Applied Regression* Thousand Oaks, CA: Sage.
- Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models.* Cambridge: Cambridge University Press.
- Muenchen, Robert A. and Joseph Hilbe. 2010. *R for Stata Users.* London: Springer.

Statistical Software

There are many statistical software packages available, all of which will perform the functions explored in this class. We will be relying on *R*, a free statistical program available for Mac, Windows, and Linux. R may be downloaded from the CRAN website (http://www.r-project.org). The CRAN site also includes a variety of user's manuals, which I suggest you consult early in the semester (if you are unfamiliar with the R language).

We will explore a variety of R functions throughout the semester. I will assume that you know the basics, such as opening a dataset, accessing variables, and creating new variables. I will also assume that you can run and estimate simple linear regression models, manipulate/recode data, and access saved objects.

Open Science

This class strives to advance an open science framework, a shift in the basic and social sciences to make research, data, code, and publications open and accessible. In this class, we will discuss scientific techniques that promote collaboration, transparency, and reproducibility. This means we will generally rely on free tools – like R – public repositories and develop research protocols that are transparent and well documented. In practice, this means:

- We will regularly use open source software, like R.
- We will develop open source code, by posting code on GitHub.
- We will generate modular computer code.
- We will generate code that produces reproducible examples.
- We will generate clear documentation.

Collaboration

I tend to find that the most useful way to learn statistics is to practice statistics. I suspect this class will be far easier if you take the time (i.e., a few hours every week) to implement the procedures we discuss in class. I will make most of my code for this class available through github. We'll spend some time on basic github operations in this course, so there is no need to set up anything in advance (simply be sure you have a github acct.)

GitHub

GitHub is a tool for version control and collaboration. You should create a github user account during the first week, in addition to acquainting yourself with the service. Please also read some of the online tutorials regarding github and how to set it up with Rstudio or R. If you use an email other than your University of Arizona email, please let me know.

DataCamp

In this class you will be asked to complete assigned DataCamp modules to supplement class material. DataCamp is free for this course via our institutional license. You will receive an invitation email—register using your UA account. More information: https://www.datacamp.com/groups/education

Course Expectations

- Come to class prepared and ready to learn
- Arrive on time; late arrivals are disruptive
- Follow the University's Class Attendance Policy: https://catalog.arizona.edu/policies/classatten
- See Religious Observance Policy: https://deanofstudents.arizona.edu/religiousobservanceandpractice
- Refrain from disruptive behavior (talking, phone use, etc.); students engaging in such behavior will be asked to leave
- All work must comply with the Student Code of Academic Integrity: https://deanofstudents.arizona.edu/codeofacademicintegrity
- Library plagiarism resources: https://library.arizona.edu/help-with-research/avoiding-plagiarism
- For accommodations based on disability: Contact Disability Resources at 520-621-3268
- Faculty own intellectual property for course materials; student notes are for personal use only
- FERPA privacy information: https://registrar.arizona.edu/ferpa
- Incomplete and withdrawal policies: https://catalog.arizona.edu/policies/grade

Grading Policy

Grade	Range
A	90-100%
B	80-89%
C	70-79%
D	60-69%
F	Below 60%

Component	Percent
Assignments (5 × 4%)	20%
Midterm Paper	15%

Component	Percent
Midterm Exam	20%
Final Exam	25%
Final Paper	20%

Graded Work Descriptions

Written Work (35% total)

Written projects comprise 35% percent of your grade. For both the midterm and the final project, you will apply the techniques from this class to an actual data. The midterm paper consists of the literature review and analytic plan for your final paper. You should complete an 8-10 page paper that reviews the research in the relevant area, and then describes your empirical strategy. Be sure to outline the data you will use, the variables you will include, and the models that you propose to estimate. It is advised to meet with me periodically throughout the semester to discuss your project.

The final paper should be a complete research paper, including introduction, literature review, theory/hypotheses, data/methods, results, and conclusion. The paper should build on your midterm paper. For both projects, you are free to choose the topic, but you must be able to provide me with the data used in your report. Your code should be shared on GitHub. If you or your advisor has proprietary data which I cannot access, you should not use this. I must be able to verify that you did all the necessary calculations honestly and accurately, which requires me being able to access any data you use. Also, if you rely on your own data, you must adhere to the University of Arizona's Institutional Review Board requirements.

In your analysis, the dependent variable must be continuous (typically 7+ categories)

The paper should include all the elements of a research paper. That said, my primary interest rests in your analyses. I ask that you be quite thorough when presenting the results. It is not sufficient to just run a regression and present the point estimates. Your paper should include detailed interpretation, robustness checks, and post-estimation clarification (which often come in the form of graphics). It should be a high quality paper; if you would not be comfortable presenting it at a conference, you should not feel comfortable handing it in for a grade. *Do not just cut and paste R output. Tables and figures should be presented in a professional manner. I will deduct points for cut-and-pasted output from R.*

Your statistical models should be theoretically informed. I do not require a full-fledged introduction/literature review, though you should spend a page or two briefly describing the relevant literature. It is essential to include the formal hypotheses you will test, which should relate to this one/two page review. I will need to verify that your empirical tests match your expectations. Please follow American Psychological Association (APA) style or American

Journal of Political Science (AJPS) style. The references should appear at the end of your manuscript. Please do not place the references in footnotes.

Strive for professional, journal-quality presentation.

Exams (45% total)

There will also be a midterm exam (20%) and a final exam (25%). The final will be cumulative, and will cover all material from the semester. The midterm and final should be completed in-class. You may use your notes and the course materials to complete the both exams. The exam must be completed independently.

Assignments (10%)

Complete all assigned modules by the deadline.

Course Schedule

Week 1 (1/20): Introduction to Linear Regression

- **Topics**: The OLS estimator
- **Reading**: Fox, Chapter 2, 3 (1/2)
- **Slides**: Introduction_to_Linear_Regression.qmd | RPubs

Week 2 (1/27): Estimation and Properties

- **Topics**: Properties of the OLS estimator
- **Reading**: Fox, Chapter 3 (1/2), Chapter 5
- **Slides**: ols_estimator.qmd, OLS_Derivation.qmd, gauss_markov_slides.qmd | RPubs
- **Assignment 1**: Complete DataCamp modules, `Introduction to Git` and `Version Control with Git`

Week 3 (2/3): Comments on Model Fit

- **Topics**: R-squared and its interpretation
- **Reading**:
  - Achen, C. H. (1990). What does explained variance explain? Reply. *Political Analysis*, 2, 173–184.
  - King, G. (1990). Stochastic variation: A comment on Lewis-Beck and Skalaban's the R-squared. *Political Analysis*, 2, 185–200.
  - Lewis-Beck, M. S., & Skalaban, A. (1990). The R-squared: Some straight talk. *Political Analysis*, 2, 153–171.

- **Assignment 2**: Create a Github repository for your class project. Include a README file describing the project and the structure of the repository. Share the repository with the instructor (git: crweber9874).

Week 4 (2/10): Statistical Inference

- **Topics**: Theoretical approaches to estimation and inference
- **Reading**: Fox, Chapter 6
- **Assignment 3**: Complete DataCamp modules, `Simple Linear Regression` and `Technical Conditions in Linear Regression`

Week 5 (2/17): Inference, Continued

- **Topics**: Inference in linear regression
- **Reading**: Fox, Chapter 7, 8
- **Assignment 4**: Complete DataCamp module, `Predictions and Model Objects`

Week 6 (2/24): Multiple Variables

- **Topics**: Inference in linear regression
- **Reading**: Fox, Chapter 9 (Optional: Chapter 10)
- **Recommended DataCamp Module**: `Introduction to Linear Algebra`
- **Midterm Project is Due**

Week 7 (3/3): Midterm Exam

Week 8 (3/17): Assumptions of the Linear Model

- **Topics**: GLM and Heteroskedasticity
- **Reading**: Fox, Chapter 12

Week 9 (3/24): Collinearity

- **Topics**: Multicollinearity detection and solutions
- **Reading**: Fox, Chapter 13

Week 10 (3/31): Influence

- **Topics**: Outliers and influential observations
- **Reading**: Fox, Chapter 11

Week 11 (4/7): Panel Data

- **Topics**: Fixed effects, random effects, panel and time series data: Autocorrelation
- **Reading**: Fox, Chapter 16 (First Half)

Week 12 (4/14): Panel Data, Continued

- **Topics**: Fixed effects, random effects, panel and time series data: Autocorrelation
- **Reading**: Fox, Chapter 16 (Second Half)

Week 13 (4/21): Generalized Linear Models

- **Topics**: Introduction to maximum likelihood and the generalized linear model
- **Reading**: Fox, Chapter 15
- **Assignment**: Complete DataCamp modules, `GLMs, an extension of your regression toolbox`

Week 14 (4/28): Simulation and Inference

- **Topics**: Extension to maximum likelihood and the generalized linear model
- **Reading**: Fox, Chapter 15

Week 15 (5/5): Simulation and Inference

- **Reading**: Fox, Chapter 21
- **Assignment 5**: Complete DataCamp modules, `Data wrangling`

May 6, 2026: Final Paper is Due

May 11, 2026: Final Exam, 3:30-5:30 in SS 332

This syllabus is subject to change at the instructor's discretion with advance notice.