PA 541: Advanced Data Analysis I

Spring 2021

Michael D. Siciliano

email: sicilian@uic.edu

Class: Monday's 3:00 – 5:50, online via Zoom

Office Hours: Happy to meet at any time. Email me to set up a convenient time.

Teaching Assistants: Saman Shafiq (sshafi7@uic.edu) & Evgenia Kapousouz (ekapou2@uic.edu)

I. LEARNING AND COURSE OBJECTIVES:

This course provides an introduction to modern econometric and statistical techniques useful for applied researchers and analysts. The main focus will be on your ability to appropriately develop and interpret statistical models based on a strong conceptual understanding of modeling assumptions and limitations. Less emphasis will be placed on mathematical proofs and derivations. Topics to be covered in this course include standard OLS regression, regression with qualitative predictors, proper handling of data issues, limited dependent variable models, panel data methods, and data visualization. The course will also cover data wrangling techniques (obtaining, cleaning, transforming, and merging data) that often consume much of a researcher's time but are not often a central focus of statistics and econometric classes.

Academic Integrity: As an academic community, UIC is committed to providing an environment in which research, learning, and scholarship can flourish and in which all endeavors are guided by academic and professional integrity. All members of the campus community—students, staff, faculty, administrators—share the responsibility of insuring that these standards are upheld so that such an environment exists. Instances of academic misconduct by students shall be handled pursuant to the Student Disciplinary Policy. The Student Disciplinary Policy is available online at https://dos.uic.edu/wp-content/uploads/sites/262/2018/10/DOS-Student-Disciplinary-Policy-2018-2019-FINAL.pdf.

Special Needs: UIC and the Department of Public Administration are committed to maintaining a barrier-free environment so individuals with disabilities can fully access programs, services and all activities on campus. The Office of Disability Services works to ensure the accessibility of UIC programs, classes, and services to students

with disabilities. Services are available for students who have documented disabilities, including vision or hearing impairments and emotional or physical disabilities. Students with disability/access needs or questions may contact the Office of Disability Services at (312) 413-2183 (voice) or (312) 413-0123 (TTY only). Please feel free to contact me if you need any special accommodations.

Diversity and Inclusion: It is my goal that people from diverse backgrounds and perspectives be included in and served by this course, the Department of Public Administration, and the University of Illinois at Chicago. I believe that the diversity that the students bring to this class is a resource that can be used to improve student learning and perspectives. Course materials and activities are designed to be respectful of all types of diversity: gender identity, sexuality, disability, age, socioeconomic status, ethnicity, race, nationality, religion, values, and culture. It is my intent that this class be an environment where all students feel safe to express their perspectives and opinions. Please let me know if something said or done by me, the TA, guest lecturers, or other students, is troubling or causes discomfort or offense. If this occurs, please feel free to discuss the situation with me privately, to raise your concern in class, or to let me know about the issue through a trusted sources (e.g., another student or faculty member or your academic advisor).

Campus Advocacy Network: Under the Title IX law you have the right to an education that is free from any form of gender-based violence and discrimination. Crimes of sexual assault, domestic violence, sexual harassment, and stalking are against the law and can be prevented. For more information or for confidential victim-services and advocacy contact UIC's Campus Advocacy Network at 312-413-1025 or visit http://can.uic.edu/. To make a report to UIC's Title IX office, contact Rebecca Gordon, EdD at TitleIX@uic.edu or (312) 996-5657.

II. SOFTWARE REQUIREMENTS

We will be using the R programming language. R is quickly becoming the statistical package of choice for social scientists and offers several advantages over other software programs. R provides broad coverage and availability of new, cutting-edge statistical applications (no need to wait for a new release of the software) as well as in other methodological areas that may be of interest to researchers (i.e. text analysis, spatial analysis, network analysis, QCA, etc…). Learning R will also facilitate your understanding of the literature in your field as more and more people are reporting their results and providing replication files in R. An active and engaged group of users provide quality support available through listservs, Stack Exchange, etc. R allows you to easily write your own functions to perform tasks unique to your data that would otherwise be quite time intensive. Finally, and importantly, R is free.

The R code used to produce the data analysis examples and graphics in my lectures will be made available to you. This allows you to read through the lecture notes while simultaneously working in R to reproduce all of the outcomes. Over the years teaching this course, this has proven to be a useful way to learn data analysis techniques.

III. TEXTS

There are two textbooks assigned for this class. The second book by Teetor is optional.

1. Wooldridge, J. M. (2016). *Introductory econometrics: a modern approach*: Cengage Learning. [***Required: Earlier editions are fine***]
2. Long, J.D. & Teetor, Paul (2019). *R cookbook*: O'Reilly [***Optional***]

I will not be assigning any specific readings out of the Long and Teetor text. This book will serve as a reference guide for you as you learn to use R. Most weeks, I will assign two sets of readings: one covering the method and another detailing the application of the method in scholarly publications. I will also share a folder through Blackboard with a number of free, online R resources.

IV. REQUIREMENTS AND GRADES

1. Midterm and Final

A midterm and final exam will be given. Both exams will be open book and open notes.

2. Problem sets

Most weeks you will have a problem set to work on. I usually give problem sets that span two weeks of classes and I expect to assign around 4 problem sets throughout the semester (depending on our progress and student needs). The problem sets will focus on applying the methods to real data. I encourage you to work with your classmates on these assignments but to ultimately complete and write-up the results individually. Please see UIC's policy on academic integrity.

3. Data Analysis Paper

The final component of this course is a relatively short data analysis paper (3,000 – 5,000 words). For this paper you will use existing datasets from online resources or repositories such as ICPSR, Harvard Dataverse, World Values Survey, IES, Census Data, Pew Internet and American Life Project, Current Population Survey, Open Data Portals (e.g., https://data.cityofchicago.org/) , etc….or from a researcher willing to share their data. You will be

required to develop testable hypotheses using methods covered in this course and to situate your hypotheses within the current literature regarding your topic of interest. The aim of this paper is to give you a more complete experience of the data analysis process. You will go from initial data screening and cleaning to final model output and interpretation. For this assignment you will write your paper as if preparing it for submission to a journal or as a policy report for a think tank. In an appendix, you will provide more of the technical details regarding how you cleaned the data, tested model assumptions, etc. Additional details on the paper will be discussed in class. You should begin thinking about this paper and searching for applicable datasets early on in the semester. You will work in groups of three or four on this paper. I will assign these groups during the third week of class.

Your grade in this course will be calculated as follows:

Component	Percentage of overall grade	Due date
Midterm	25%	March 8
Final	25%	April 26
Problem Sets (4)	20%	Ongoing
Data Analysis Paper	20%	May 5
Class Participation	10%	Ongoing

V. WEEKLY SCHEDULE (readings subject to change)

**SCHEDULE OF CLASSES AND ASSIGNMENTS IS SUBJECT TO CHANGE. IN THE EVENT OF ANY CHANGE IN ASSIGNMENT, TOPIC, OR DUE DATE, I WILL POST NOTICE IN THE ANNOUNCEMENTS ON BLACKBOARD AND UPLOAD A CORRECTED SYLLABUS.*

Week 1 - January 11

Topics: Course intro and review. Overview of the R programming language and the R-Studio IDE. Introduction to the Tidyverse.

*Please download R and RStudio onto your laptops before class. I will make sure that everyone is up and running with R at the end of the class. If you already have R and RStudio installed, please update to the most recent version.

R Readings:

Wickham, H. and Grolemund, G. (2017). *R for Data Science*: Chapter 5: Data Transformation
https://r4ds.had.co.nz/transform.html

R Tutorials:

R Studio Primers: https://rstudio.cloud/learn/primers/1 Work through both the visualization basics and the programming basics. Start with the programming basics.

Week 2 – January 18

*NO CLASS *****Martin Luther King Day******

Review R Resources.

Week 3 – January 25

Topics: General discussion of statistical models. The nature of econometric data. Simple regression and simple regression assumptions. Using R for data management; useful functions in R.

Methods Readings:

Wooldridge Chapter 1 and 2

Week 4 – February 1

Topics: Multiple regression - mechanics and interpretation.

Methods Readings:

Wooldridge Chapter 3

Application Readings (just skim):

Moynihan, D. P., Pandey, S. K., & Wright, B. E. (2012). Prosocial Values and Performance Management Theory: Linking Perceived Social Impact and Performance Information Use. *Governance, 25*(3), 463-483.

Week 5 – February 8

Topics: Testing hypotheses about model parameters; testing exclusion restrictions.

Methods Readings:

Wooldridge Chapter 4

Application Readings (just skim):

Stensöta, H. O. (2012). Political Influence on Street-Level Bureaucratic Outcome: Testing the Interaction between Bureaucratic Ideology and Local Community Political Orientation. *Journal of Public Administration Research and Theory, 22*(3), 553-571

Week 6 - February 15

Topics: Intro to data visualization.

R Readings:

Wickham, H. and Grolemund, G. (2017). *R for Data Science*: Chapter 3: Data Visualization: https://r4ds.had.co.nz/data-visualisation.html

Healy, Kieran (2019). *Data Visualization*. Chapter 3: Make a Plot: https://socviz.co/makeplot.html#makeplot

Wilkes, Claus (2019). *Fundamentals of Data Visualization*. Sections 1-5: https://clauswilke.com/dataviz/introduction.html [Note: these are very short sections]

Week 7 - February 22

Topics: Using qualitative/categorical predictors. Interpreting main effects and interactions with categorical predictors.

Methods Readings:

Wooldridge Chapter 7

Application Readings (just skim):

Brewer, G. A. (2003). Building Social Capital: Civic Attitudes and Behavior of Public Servants. *Journal of Public Administration Research and Theory, 13*(1), 5-26.

Week 8 – March 1

Topics: Non-linear relationships. Checking the validity of regressions assumptions. Dealing with heteroskedasiticty. Issues with highly correlated predictors.

Methods Readings:

Skim Wooldridge Chapter 8

Zuur, A. F., Ieno, E. N., & Elphick, C. S. (2010). A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution, 1*(1), 3-14.

Application Readings (just skim):

Yang, K., Hsieh, J. Y., & Li, T. S. (2009). Contracting Capacity and Perceived Contracting Performance: Nonlinear Effects and the Role of Time. *Public Administration Review, 69*(4), 681-696

Week 9 – March 8

****Midterm**

Week 10 - March 15

Topics: Model specification and data issues. Log Models. Data screening and cleaning. Methods for handling outliers and missing data.

Methods Readings:

Wooldridge Chapter 9

Pardoe, I. (2012) Applied Regression Modeling, Chapter 5

Application Readings:

None

Week 11 - March 22

*NO CLASS *****SPRING BREAK******

Week 12 - March 29

Topics: Limited dependent variables and the generalized linear model. Logistic regression.

Methods Readings:

Moore, David S. and McCabe, George P., "Introduction to the Practice of Statistics", Chapter 16 – Logistic Regression. P. 1-16

Application Readings (just skim):

Shingler, J., Van Loon, M. E., Alter, T. R., & Bridger, J. C. (2008). The Importance of Subjective Data for Public Agency Performance Evaluation. *Public Administration Review, 68*(6), 1101-1111.

Week 13 – April 5

Topics: Panel data analysis - Part I. Pooled cross-sectional analysis and difference in difference models.

Methods Readings:

Wooldridge Chapter 13

Application Readings (just skim):

Huang, J., Chaloupka, F. J., & Fong, G. T. (2013). Cigarette graphic warning labels and smoking prevalence in Canada: a critical examination and reformulation of the FDA regulatory impact analysis. *Tobacco Control*.

Week 14 - April 12

Topics: Panel data analysis - part II. Fixed and random effects models.

Methods Readings:

Wooldridge Chapter 14

Dougherty, Christopher. (2011) "Introduction to Econometrics", 4th ed. Oxford University Press: New York. Chapter 14 - Introduction to Panel Methods.

Application Readings (just skim):

Pihl, A. M., & Basso, G. (2019). Did California Paid Family Leave Impact Infant Health? *Journal of Policy Analysis and Management, 38*(1), 155-180. doi:doi:10.1002/pam.22101

Week 15 - April 19

Topics: Introduction to Directed Acyclic Graphs

Methods Readings:

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science, 1*(1), 27-42. doi:10.1177/2515245917745629

Shrier, I., & Platt, R. W. (2008). Reducing bias through directed acyclic graphs. *BMC Medical Research Methodology, 8*(1), 70. doi:10.1186/1471-2288-8-70

OPTIONAL: Here are some useful links and tutorials: http://dagitty.net/learn/

Week 16 - April 26

****Final Exam**

FINALS WEEK

*****Data Analysis Paper Due on May 5th**