POLI 784 - Intermediate Statistics, Spring 2014

Tuesday and Thursday, 9:30-10:45, New East Rm 0102

Weekly Lab (POLI 891-001)
Thursday 1:00-1:50, Cobb Hall Rm 0024

Instructor

Dr. Thomas Carsey
E-mail: carsey@unc.edu
Office: 317 Hamilton Hall
Phone: 962-1207 (but e-mail is a MUCH better way to contact me)
Office Hours: Tuesday and Thursday 8:00-9:30, and by appointment

Teaching Assistant

Brice Acree
E-mail: bdlacree@live.unc.edu
Office: Hamilton Hall 300
Office Hours: Wednesday 3-4 p.m., Thursday 11 a.m. -1 p.m., and by appointment

Course Description

This course focuses on the linear regression model as estimated via Ordinary Least Squares (OLS). We will examine the properties of OLS, the assumptions underlying the model, the consequences of violating these assumptions, how you can detect such violations, and how you might begin to respond to them. This course builds on what students learned in POLI 780 (Scope and Methods) and POLI 783 (Statistics), and assumes that all students have completed these or equivalent courses.

This is a great course to teach and to take! Your previous coursework has provided you with many of the basic building blocks to begin doing real systematic quantitative social science research. We will actually do such research in this class. We will devote ourselves in this class to quantitative analysis, but the skills you refine in this course apply to systematic empirical research of all stripes. Thus, while this course appears to be a continuation of your statistical training alone, it has as much to do with theory and design as it does with crunching numbers. Statistical analysis is an important tool used by social scientists, but for most of us, the real goal is to learn something about some social or political process, not just something about mathematics.

This course will help you become increasingly skilled users and critical consumers of research employing quantitative methods. You will not be experts by the end of this course – not even close. However, you will have built a very strong foundation for future coursework, you will be better-positioned to read more of the literature, and you will be well-versed in using OLS and the linear model in your own research. OLS and the linear model provide the basis for most published empirical political science research, and the jumping off point for most of the more sophisticated methods you will see and use. As a result, this course is extremely important in your Ph.D. training. I strongly encourage you to push yourself to get as much out of this course as you possibly can.

In order to engage the scholarly community in virtually every subfield of political science, one needs an understanding and working knowledge of statistical methods. Statistical analysis of data is

certainly not the only, or even necessarily the best, approach to conducting research. However, every area in social science makes use of statistical methods. Furthermore, the general logic of the methods we will explore extends beyond large-N quantitative studies.

I have trouble thinking about a course in statistics that is not mathematical at some level, so of course we will be doing math in this class. However, I do not think the math will be a barrier to anyone (particularly anyone who took POLI 783). We will work through the math with the goal of providing a deeper understanding of the concepts under consideration, but it is that understanding, and not the math itself, that is the primary goal. I often use the word 'intuition' to describe the level of understanding that I want students to have regarding statistical methods. That intuition is not a substitute for the math, nor do I mean that you should be satisfied with some sort of general sense of what is going out without understanding the math. What I do mean is that understanding the logic of quantitative analysis runs deeper than just a set of mathematical rules and formulas. We will continually ask why a practicing political scientist would want to know about the statistical topic at hand. If I fail to make it clear at any point in the semester why we are learning what we are learning, you should press me on it.

I believe in learning by doing. Thus, we will have regular assignments. Ultimately, social scientists need to be able to formulate theories of social or political processes, translate theories into testable hypothesis, develop models that capture the theory and permit the testing of hypotheses, apply appropriate methods, interpret the results, and return to the theory in order to evaluate it. I think about this process as trying to move seamlessly back and forth between words, pictures/figures, and equations. This is one of the hardest parts of becoming a successful social scientist just getting the statistics right is only one aspect of the process. The intuition I noted in the previous paragraph is a critical part of this larger process, as it provides a mechanism to facilitate translating our theories of social and political processes into statistical models that can be evaluated without loosing something in the translation. Of course, getting the stats right is a critical part for many scholars, and it is the central task of this course, but I want to make sure that students do not view learning about methods as something different and separate from learning and thinking theoretically about politics.

When doing research, a good rule of thumb is to think about trying to satisfy three types of reviewers: 1) the substantive expert, 2) the methods expert, and 3) a friend, college undergraduate, or elderly relative that is neither of #1 or #2. While this course is mostly about methods, learning methods tools in isolation of trying to satisfy the other two types of reviewers would be a mistake.

Course Requirements

There is one required text for the course:

Applied Regression Analysis and Generalized Linear Models, Second Edition, by John Fox and published by Sage.

The text is no-doubt expensive, but it is worth having. It will be on your shelf and return to for years to come. This particular text is highly regarded and widely used for similar courses.

I have also listed two recommended books for this class:

An R Companion to Applied Regression, 2nd Edition, by John Fox and Sanford Weisberg, Sage.
Monte Carlo Simulation and Resampling Methods for Social Science by myself and Jeffrey J. Harden, Sage.

The Fox and Weisberg book is a companion to Fox's textbook. The value of this companion text is that it will help you work through doing OLS analysis in `R` . I will have more to say about using `R` below. The book by Jeff and I will help you with doing simulations that explore regression assumptions, and we will also explore some of the specific resampling topics covered in later chapters.

This course is often taught using an econometrics text. I have used *Basic Econometrics* by Damodar N. Gujarati and Dawn C. Porter (McGraw-Hill) many times, and others use popular texts by Greene and Wooldridge. Such texts give more attention to formal proofs, mathematics, and demonstrating the properties of various estimation methods asymptotically than does the Fox book. In contrast, the Fox book gives a bit more attention to overall model fit and evaluation, a number of issues that arise in applied settings, and is more informed by applied statistics rather than econometrics. In particular, the Fox book is better organized to provide a foundation for the next course in our department's sequence, POLI 787, which focuses on generalized linear models and maximum likelihood estimation.

We will not read every chapter in Fox, but we will read many. It is pretty readable as statistics texts go, but it is still a statistics text. You are strongly encouraged to read the assigned material before coming to each class, and you would certainly benefit from reading it again afterward. There is no substitute for just hammering away at this material, and I can tell you that the better you understand the material in this course, the better off you will be down the road in other courses (both substantive and methodological), in writing papers, in writing your dissertation, in publishing, in getting a job, and in getting tenure. Class time will be devoted to nailing down the basics and making sure you know when and when not to use the methods we discuss. There will always be more material to cover than we have time for in class, so in that sense, you will always be left wanting for more.

Regular class attendance is expected and required – it will be obvious who is and is not here every day. For a class like this, it is imperative that you keep up with the readings, assignments, and lectures. Thus, missing class is really not an option. I also expect you all here on time and ready to go on time for every class meeting. We simply have too much work to do to proceed any other way.

Finally, there is a 1-credit hour lab/workshop associated with this course that all students are required to take (POLI 891-001, listed under my name). The lab sessions will be led primarily by the TA. Most lab sessions will be devoted to practical issues associated with managing and analyzing data. This will include some support to help with the assignments for the course, and will also include a good deal of training in `R` and some exposure to STATA (More on `R` below).

Assignments and Grading

We will have 2 exams in the class, a paper/project, and several assignments. You will also receive a grade for the lab, though that grade will be folded into your performance in the class overall and you will be given the same single grade for both the 3-credit class and the 1-credit lab.

20% Midterm exam
30% Final Exam
10% Course assignments
10% Lab (participation/assignments, etc.)
25% Paper/project
5% Paper comments

I reserve the right to make minor adjustments to final course grades based on overall performance in the class. However, as a rule of thumb, if you are scoring in the 90s, you are doing well in the course. If you are scoring in the 80s, you are making satisfactory progress but missing more than you should. If you are scoring in the 70s or below, you are not really doing satisfactory work. I will not accept late papers or late assignments unless a compelling reason is provided to me in advance or a serious unanticipated problem arises (NOTE: computer or printer problems do not qualify). Anticipate having problems with the assignments and the paper and plan accordingly.

There will be an assignment handed out for the main course nearly every week. You can and should collaborate on the assignments, but you need to learn the material for yourself. I do NOT want to see identical assignments turned in by students. Use each other as a resource, but NOT as a crutch. Some of the assignments will require some computational work by hand, but most will require use of a computer. All assignments that require use of the computer will be conducted in using the statistical software `R` and written using LaTeX, both of which are freely available.

Using `R`

You can download `R` online at: `http://cran.r-project.org/` and you can learn more about `R` in general at the R-project homepage at: `http://www.r-project.org/`. `R` is best thought of as a statistical computing environment rather than as software. `R` is not a point-and-click program. There are some Graphical User Interfaces (GUI's) available for `R` , but we won't be using them. Instead, you will be writing text files, called script files in `R` , that send `R` a series of commands to execute. Learning `R` can be a bit more challenging than learning a point-and-click program, but it is much more powerful, flexible, and is increasingly the computing environment of choice for those doing statistical work across a wide range of disciplines including Political Science. More importantly, our goal in this class is to learn about statistics, NOT about software. Programming in `R` is a far-superior way to learn about statistics than is using a point-and-click program.

There is no substitute for reading the documentation for `R` . I STRONGLY recommend that you begin with the manual called "An Introduction to R." This document provides the core basics to understanding `R` as a statistical computing environment. You can find the manual by clicking the "Manuals" link on the CRAN homepage. The direct link to the .pdf file is here: `http://cran.r-project.org/doc/manuals/R-intro.pdf` This manual is also downloaded and stored on your computer when you install `R` .

You can interact with `R` directly, but is it much better to use a text editor to create a file of operations that you execute. `R` comes with a simple text editor, but there are many other free editors that make working with `R` much easier. I prefer a tool called RStudio (`http://www.rstudio.com/`) because it is easy to use, it has some nice features, and it works on Mac, Windows, and Linux machines.

Springer books (`http://www.springer.com`) has an entire series of books in their Use `R` series that are designed to be practical applications of `R` for users. Many of these can be accessed through the UNC library online for free. One in particular that is quite useful is *Data Manipulation with `R`* by Phil Spector. Another is *A Beginner's Guide to `R`* by Zuur et al.

There are also some very helpful short reference documents for `R` commands that you might want to print and keep handy, located at: `http://www.rpad.org/Rpad/R-refcard.pdf` and at: `http://www.psych.upenn.edu/~baron/refcard.pdf`. John Fox has a couple of very useful websites for

materials on R . First, the website for his book is: http://socserv.mcmaster.ca/jfox/Books/Companion/index.html. Second, he taught a two-day workshop on R here at UNC a few semesters ago and created a website for that, which is located at: http://socserv.mcmaster.ca/jfox/Courses/R-course/index.html.

The Odum Institute (http://www.odum.unc.edu/) has an online version of an R short course available (http://www.odum.unc.edu/odum/contentSubpage.jsp?nodeid=665) that is very good. I strongly encourage you to go through that course no later than the first week of class.

Using LaTeX

LaTeXis a document processing environment that is also free and open source. Working in LaTeXis a lot like working with html code for webpages. You write LaTeXdocuments using a text editor. You add formating features to the document using various codes that are available in the base LaTeXsystem or in various packages that you load as part of the document. You then ask the text editor to use the functions available in LaTeXto compile your document. The result is a PDF (or other) file that looks clean and professionally produced. The value of LaTeXover word processors like MS Word is that you have complete control over how your final document looks. This is often trival for memos or letters, but can be quite helpful when doing scientific writing. LaTeXmakes it much easier to write mathematical formulas, include figures, construct tables, and build your references. There are a number of features that make it particularly compatible with using R to conduct your statistical analysis.

LaTeXis available from several sources online. TeX Live works for all three common platforms. A version called MacTEX is a version of Tex Live specially designed for installation on a Mac. MiKTeX and its newer version, proTeXt, are distributions for Windows machines. You can read more about this online here: http://latex-project.org/ftp.html and here: http://www.ctan.org/ and here: http://www.tug.org/

There are many free LaTeXeditors available. You can see a list of them here: http://en.wikipedia.org/wiki/Comparison_of_TeX_editors. The one I currently use is called TeXstudio (http://texstudio.sourceforge.net/). I like it because it has a number of built-in features, it syncs your TEX and PDF documents easily, and it runs on Macs, Windows, and Linux systems. It is very similar to Texmaker (which runs on all platforms) and TeXnicCenter (which only runs on Windows).

The Odum Institute has an online version of a LaTeXcourse available as well (http://www.odum.unc.edu/odum/contentSubpage.jsp?nodeid=665). Again, you should work through this course by the end of the first week in class.

We will provide support and direction with R and LaTeX, but you need to take the responsibility yourself to learn the tools you need to do your work. It is O.K. to ask each other questions when working on assignments and such, but ultimately you have to know how to do this stuff on your own. Your learning will be greatly enhanced by banging through the assignments, and that will no doubt be reflected on the exams. I would rather see you make your own mistakes on the assignments and learn from them as opposed to copying correct answers from others but not really understanding what you are doing.

The Paper

The paper should be a piece of original quantitative analysis. For this class, you should pursue a paper that is a replication and extension of an existing published paper (or book chapter). This will make it easier for you to present the literature review and theory sections for your paper since they will be closely tied to the paper you replicate. Your emphasis for the paper will be on properly analyzing data to test your hypotheses. You should plan to use an OLS model for your paper, but I can work with you if you need to employ some other method. If you have concerns about the appropriateness of OLS for your paper, or if you need assistance in developing a paper topic, you should consult with me early and often in the semester. There is no formal page length, but for most of you I expect the paper will constitute 14-20 pages of text. You should model your paper after the quantitative papers you have seen in journals like *APSR*, *AJPS* or *JOP*, with the caveat that the front-end of your paper (everything up to the Data/Methods section) will be shorter than the typical journal article (because you are doing a replication), and that you will be asked to provide a bit more detail in the back half of your paper regarding the analyses, tests, etc. that you performed.

By "replication," I mean that your first task will be to reproduce the findings exactly as shown in the published paper or chapter you are looking at. This DOES NOT mean simply contacting the author or using a so-called replication data set in which all of the coding, modeling, and estimation decisions/commands are already done for you. Rather, it means going back to the primary (electronic) data source if possible and proceeding from there. For example, suppose you are replicating a study by Bill Smith that uses survey data from the American National Election Studies (ANES) series. Rather than asking Bill Smith to send you any computer files that record all of his coding decisions that you simply have to run, I want you to download the original NES data, locate the proper variables, make any coding changes, and perform the analysis. In other words, your first task is an independent replication/verification of Bill Smith's reported analysis. If that is not possible, we can talk about contacting the original author and other strategies for proceeding.

By "extension," I mean that once you have replicated the results of an existing study, you will then build upon that analysis in some way. This might involve using a different coding of a variable, adding additional variables, considering different (maybe non-linear) model specifications, or adding additional data. Whatever extension you attempt, however, must be derived from a clear theoretical proposition and/or a clear methodological critique. In other words, don't toss in an interaction term, "just to see what happens." Remember, this is not just an exercise in number crunching you are writing a paper with the goal of answering a theoretically motivated research question. However, by going through the process of trying to replicate another scholar's study, I hope you will learn the value of documenting every step of the research process.

All students will read drafts of two other student papers in the course and provide written comments to them. Seeing the work of others and how others react to your work helps you to improve your skills as a researcher. You might be a bit nervous about sharing your work with others, let alone receiving their written comments and providing such comments yourself. It is O.K. to be nervous, but it is also time to start getting used to this. It is better to begin this among friends and colleagues before you have to deal with anonymous reviewers.

Finally, I anticipate holding a public poster presentation event for the class near the end of the term. More will be said about that in class as the semester unfolds.

I do not expect perfect papers ready for submission to *APSR* by the end of the semester. In fact, whether the paper is ever publishable or not is not the goal of this assignment. However, I do expect

your best professional effort. The only way that I and your classmates can help you to improve is if you do the best you can on your own with your first draft so our advice can focus on how to push beyond that. Don't worry - this will be fun!

The course schedule lists several due dates associated with the paper. I expect you to provide me with at least what is asked for on that date. If you give me more, I will read more. The only part of the paper assignment that is graded is the final version of the paper you submit to me at the end of the course. Thus, being lax with these intermediate deadlines does not directly hurt your grade, but it does limit my ability to be helpful to you and it would leave you behind schedule and scrambling to catch up. Hitting these deadlines signals your effort on this project. Of course, you are free to talk with me at any point along the way about your paper.

Communication

I make every effort to communicate my expectations, your responsibilities, and the information covered in this course. I will send e-mails to the entire class. I maintain a Sakai website for the class, and I will make announcements and issue some reminders in class. Note that I will only send e-mail out to your UNC e-mail accounts as listed on the course roster. I will not keep track of any other e-mail addresses that you might use. The best ways communicate with me outside of class are to come to my office hours or sending me an e-mail. I do my best to be responsive to my students. It is important for you to stay in touch, particularly if any problems arise. I don't like to change exam schedules. If a situation arises where I need to give a make-up exam, I reserve the right to give it during the final week of the semester. I reserve the right to give a make-up exam that differs substantially from the normal exam in order to protect the integrity of the exam process. I or any faculty member will be much more understanding if you communicate with us early and up front.

A Note on Academic Honesty

In order for me to evaluate your work fairly, you have to do your own work. It is much easier to study, work hard, and complete your own assignments than it is to try and figure out some way to "beat the system" without getting caught. Cheating, plagiarism, and all other forms of academic dishonesty are pretty easy to spot and come with severe consequences. All students should familiarize themselves with the Academic Honor Code at UNC (`http://honor.unc.edu/honor/code.html`). Students caught cheating in any form in this course may receive an F for the course and may be turned over for further disciplinary action by the University. By taking this class, you have committed to comply with all aspects of the Honor Code regarding all aspects of this course.

Students with Disabilities

Students with disabilities needing academic accommodation should; (1) contact the office of Learning Disabilities at UNC (`http://www.unc.edu/depts/lds/index.html`), (2) bring a letter to the instructor indicating the need for accommodation and what type. This should be done during the first week of class.

Responsibilities

The success of this course depends upon all of us meeting our responsibilities. Myself and the TA are responsible for being prepared each week to present and discuss course material, for challenging you academically and stimulating your curiosity, and for being available for and responsive to your questions and inquiries. You are responsible for being prepared each week as well, for asking questions when you are confused and actively engaging the material, for doing your own work, for meeting the course requirements, and for pushing yourselves to get the most out of this course that you can. Ultimately, this is your education and you should take responsibility for it.

Course Schedule

The schedule below serves as a guideline for the semester. As we proceed, we may discover that some topics take a bit longer than expected to cover while others take less time. We may also add or change a few of the topics along the way. Readings associated with each topic are listed on the schedule and should be read by you prior to coming to class. It may be the case that additional readings will be assigned during the semester. Those readings will be provided for you either in class or online. Announcements regarding such changes will be made in class and distributed to students via e-mail. However, the dates for the exams will NOT change, nor with the due date for the paper.

DAY	TOPIC
Jan. 9	Introduction and Overview of R and LaTeX **Read:** "An Introduction to R" and look at the links in the syllabus; Read Carsey and Harden Chap. 3; Install LaTeX, R, and appropriate tools like TeXstudio and RStudio on your laptops.
Jan. 14	Introduction to Regression **Read:** Fox, Chap. 1,2; Carsey and Harden, Chap. 1, 4
Jan. 16	Introduction to Data Analysis **Read:** Fox, Chap. 3
Jan. 21	Regression and OLS **Read:** Fox, Chap. 5, Chap. 15 section 15.1
Jan. 23	Regression and OLS (continued); Causality and Control
Jan. 28	Inference and OLS Regression **Read:** Fox, Chap. 6
Jan. 30	Inference and OLS Regression (continued); Model Fit **Read:** Carsey and Harden, Chap. 9 section 9.3 **Paper Assignment:** 1-page statement of project, including access to data, plan for extension, and copy of paper to be replicated.
Feb. 4	Estimation and Inference, Matrix Form **Read:** Fox, Chap. 9 section 9.1 through 9.4
Feb. 6	Estimation and Inference, Matrix Form (continued)
Feb. 11	Dummy Variable Regression, and Interaction Terms **Read:** Fox Chap. 7
Feb. 13	Interaction Terms **Read:** Brambor et al. "Understanding Interaction Models ..." Political Analysis (2006) 14:63-82.

DAY	TOPIC
Feb. 18	Interaction Terms (continued)
Feb. 20	ANOVA **Read:** Fox, Chap. 8
Feb. 25	Outliers, Omitted Variables, Measurement Error, and Over-fitting **Read:** Fox,Chap. 11; Carsey and Harden Chap. 5 sections 5.5.3, 5.3.4
Feb. 27	Outliers, Omitted Variables, Measurement Error, and Over-fitting (continued)
March 4	Catch-up and Review
March 6	**Midterm Exam**
March 11-13	Spring Break – No Class
March 18	Heteroscedasticity **Read:** Fox Chap. 12 section 12.2; Carsey and Harden Chap. 5 section 5.3.1
March 20	Heteroscedasticity and bootstrapping **Read:** Fox Chap. 21; Carsey and Harden Chap. 8 section 8.4 **Paper Assignment:** Summary of theory for replication and extension, preliminary analysis, and results.
March 25	Serial Correlation and Time series **Read:** Fox, Chap. 16; Carsey and Harden Chap. 5 section 5.3.5
March 27	Serial Correlation and Time Series
April 1	Multicollinearity **Read:** Fox, Chap. 13; Carsey and Harden Chapter 5 section 5.3.2
April 3	Non-Normality and Median Regression **Read:** Fox, Chap. 12 section 12.1, Chap. 19 section 19.3; Carsey and Harden Chap. 5 section 5.3.7
April 8	Pooled Time Series **Read:** TBA
April 10	Pooled Data **Read:** Carsey and Harden Chap. 5 section 5.3.6
April 15	Endogeneity, SEMS, and 2SLS **Read:** TBA
April 17	Missing Data **Read:** Fox Chap. 20 **Paper Assignment:** Complete Draft Due
April 22	Introduction to GLMs and MLE **Read:** Fox Chap. 14 and 15
April 24	Introduction to GLMs and MLE (continued) **Paper Assignment:** Poster Session during Lab
April 28	(Friday) Final Paper Due, 5:00 p.m.
May 2	**Final Exam:** 8:00 a.m. in New East 0102 (allow up to 3 hours)