POLI 784 – Regression Models in Political Science

Monday and Wednesday
3:35–4:50pm
Location: Murphey Hall 202
Weekly Lab: Mondays 9:05–9:55am (Hamilton Hall 351)

Instructor Information

Santiago Olivella
Office: 353 Hamilton Hall
e-Mail: olivella@unc.edu
Office Hours: Mondays and Wednesdays 10:00am–11:30am.

Teaching Assistant Information

Simon Hoellerbauer
Office: Hamilton 300
e-Mail: hoellers@unc.edu
Office Hours: Mondays and Wednesdays 1:45pm–3:15pm

Course Description

This class is designed to introduce you to the linear and generalized linear models, as well as to the the principles of causal inference and the maximum likelihood framework for statistical inference. It is an intermediate statistics course, and it is meant to build on the skills and knowledge developed in an introductory statistics course such as POLI 783. It relies heavily on a solid understanding of basic probability, linear algebra and calculus concepts. In addition to discussing the details of the linear regression model, it also covers models for categorical data – including binary, multinomial, ordered and count outcomes. Finally, the model covers tools for modeling correlated data, relying primarily on random-effects models to do so. The course will also further familiarize you with the **R** programming environment, a powerful and versatile open-source statistics suite, and with LaTeX (through RMarkdown), a similarly flexible document production language. The course is therefore ideal for those who wish to become thorough consumers and apt producers of empirical research.

Learning Objectives

By the end of this course, you should be able to:

- Understand the definition of linear model and its properties (particularly the Gauss-Markov theorem)
- Conduct inference in the context of the LM and use common diagnostic tools to assess it.

- Understand the basics of maximum likelihood estimation (MLE) techniques.
- Understand the optimization problems involved in parameter estimation and statistical inference.
- Estimate and assess parametric models for binary, multinomial, ordered and count outcome variables, and interpret results of these estimations using graphical tools.
- Understand how random-effects models can address issues of non-independence across observation.
- Use **R** to manage data, analyze it, and create summary graphs of these analyses.
- Use LaTeX to produce publication-quality manuscripts.
- Use **R** to write and optimize your own likelihood functions.

Course Prerequisites

Completion of POLI 783 or equivalent course.

Class Structure

Our sessions will be divided into traditional lectures and a weekly lab. While I will make all slides available at end of each lecture, I strongly encourage you to take hand-written notes. During each lab, you will work on a completing a "notebook" – a file with interactive code snippets – with activities designed to help you apply the concepts learned during lectures, as well as develop important professional skills (such as appropriate coding style).

Textbook and recommended books

The course has two required textbooks, but we will draw from multiple sources at different times. Electronic copies of assigned readings coming from a source other than the textbooks will be available on Sakai.

(A) Agresti, Alan. (2015). Foundations of Linear and Generalized Linear Models. New Jersey: Wiley.

(G-H) Gelman, Andrew and Jennifer Hill. (2006). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press.

Software

The **R** programming environment (available for download at http://cran.rstudio.com/) is the *lingua franca* of applied statisticians. Its main appeal is its open-source nature, which allows anyone to expand it and modify it however they want, and you to

use it for free. It is also cross-platform, making it highly portable. For all these reasons, we will learn how to apply all the concepts learned in class in **R**. In addition to the **R** base, we will be using a user interface called RStudio (available for download at http://www.rstudio.com/ide/download/). While **R** code can be written in any plain text editor (like Emacs), RStudio provides a number of useful features in a user-friendly development environment.

You will also be required to submit assignments and your replication/expansion exercise in compiled LaTeX(using RMarkdown). LaTeX is a typesetting language that allows writers to focus on content, while handling common text formatting issues automatically. It is a powerful, portable and extensible document production environment, providing many useful functionalities for researchers – including simple mathematical typesetting and bibliography management (*via* Bibtex). It also happens that RStudio can double as a LaTeX editor/compiler, making it a great environment for all your data analysis and report production needs.

Requirements and Evaluation

Grading in this class will be based on the components described below. Failure to meet the requirements of the course will result in a failing grade.

Pre-Lecture Questions – 10%

To ensure that you get the most of each session, you are required to either a) post at least one question prompted by the study of the reading materials assigned for each week, or b) answer at least one of the current questions posted by your classmates. Questions and/or answers should be posted on Piazza no later than **5:00pm on the Sunday prior to the first lecture of the week**. These will help adjust the lectures according to your needs, and will only be answered explicitly by the instructional team when they haven't been addressed in either the lectures or the lab sessions.

Lab work – 10%

At any point during the first two weeks of classes, you'll be required to complete two courses from DataCamp: "Intermediate R" and "Data Manipulation in R with `dplyr`". Starting the second week of classes, there will be a weekly lab session where you will work through a guided practical exercise. At the end of each session, you will be required to submit the completed notebook for that session to Sakai as proof of attendance. You are required to attend all lab sessions. At the end of the semester, you will receive the same grade for both the main course (POLI 784) and the lab session.

Problem set assignments – 40% (5% each, plus 5% peer assessment)

There will be seven problem sets due throughout the semester. These are time-intensive assignments that will be made available 2 weeks prior to the due date. For each assignment, you will be randomly assigned to a different group of 3 people, with whom you

will be expected to collaborate in the production of a single answer. For each assignment, each individual member will also submit a three-point assessment of the work of all other teammates, which will count towards 5% of each student's grade. While there will be no time to go over solutions during regular class or lab hours, a solution key will be made available immediately after the due date, and you are welcome to use office hours to discuss questions related to these assignments.

Replication+Extension – 40%

As a final exercise, you will produce an single-authored paper that replicates and extends a work previously published in a top journal. The paper should have a maximum of 15 pages, and it should include a) the original theoretical argument, as well as the reasoning behind your proposed extension (which can be methodological or substantive in nature, or ideally both!); b) the replication of the original results, *which must not resort to any available replication code already available* for replication purposes (i.e., yours must be an independent replication of the original findings); and finally c) an empirical analysis that evaluates your proposed extension, including graphical interpretation tools. On February 22, you will submit the paper you wish to replicate to me, as well as the location of the data. The paper must have been published in the past 5 years, preferably in a high-impact journal. On the final day of classes, you must submit both the final paper manuscript and a replication archive to the UNC Dataverse. You will be assigned to replicate and review the work of one of your peers. By the date of the final exam, you will be required to submit set of written comments, emulating a journal review, to both your assigned peer and to the instructional team.

Grading Scale (in percentages)

Score	Grade	Score	Grade	Score	Grade	Score	Grade
> 94	H	≥ 75	P	≥ 60	L	< 60	F

Class Policies

Late work and Incompletes

Late work will not be accepted without *prior* (i.e. before the assignment is distributed) permission. No incompletes will be given for assignments or the course. Exceptions will be granted only under truly extraordinary circumstances. Prior arrangements should be made with the instructional team at least **two weeks** in advance.

Attendance

Lab attendance is mandatory. You will not be graded directly on your attendance to the class lecture. However, I strongly suggest students expecting to receive an H in this

course attend all lectures.

Technology in the classroom

You will frequently make use of computers in this course, during some lecture periods and during software training. Please be respectful to your instructor and your peers by using your computers only for class-related purposes. Put your phone away before class starts and don't bring it out. Please inform me at the beginning of the semester if you *don't* have a laptop computer you can bring to class for software training.

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 me indicating the need for accommodation and what type. This should be done during the first week of class.

Religious observances

Some students may wish to take part in religious observances that occur during this semester. If you have a religious observance that conflicts with your participation in the course, please meet with me or with the TA before the end of the second week of the semester to discuss appropriate accommodations.

Academic honesty

While problem-set assignments are designed to be collaborative, cheating and plagiarism will not be tolerated. This includes (but is not limited to) requesting or sharing work across assigned groups, as well as using someone else's code to conduct the analysis required for the final paper. More generally, I strongly encourage you to review the University's policies regarding academic honesty, which you can learn about at http://www.atschoolorientation.net/default.aspx?th=unc%2fhonor&. If you have any question regarding this issue, please feel free to ask any member of the instructional team.

Calendar with Topics, Required Readings and Assignments.

Week	Date	Lecture topic	Lab topic	Readings	Assignments
1	01/09	Introduction: Probability and regression models	No lab.	• Moore and Siegel (2015) Ch. 12, 13 • (A) Ch. 1	DataCamp: Linear Algebra

Continued on next page

Calendar – continued from previous page

Week	Date	Lecture topic	Lab topic	Readings	Assignments
2	01/14, 01/16	Linear Models I: Overview and Interpretation	R Review	• (G-H) Ch. 3.1-3.3 • Fox, Ch. 5	PS1
3	01/23	LM II: OLS and optimality conditions	LM in R	• (A) Ch. 2.1-2.3, 2.7.1 • Fox Ch 10	
4	01/28, 01/30	LM III: Uncertainty and Inference	Simulating assumption violations	• (A) Ch. 3 • (G-H) Ch. 3.4, 7.1, 7.2 • Fox Ch. 6	PS2 Due
5	02/04, 02/06	LM IV: Diagnostics and model fit	Simulation approaches to uncertainty	• (A) Ch. 2.4-2.6 • (G-H) Ch. 3.6	DataCamp: Intro. to the Tidyverse
6	02/11, 02/13	Generalized LM I: Likelihoods and MLE	LM diagnostics in R	• Eliason Ch. 1-3	PS3 Due
7	02/18, 02, 20	GLM II: The exponential family	Using optim to fit Normal linear model	• (A) Ch. 4 (not 4.3, 4.4, 4.6)	Replication Proposal due
8	02/25, 02/27	GLM III: Binomial outcomes & interpretation	Using optim to extend the linear model: heteroskedastic LM	• (G-H) Ch. 5, 6.3, 6.4 • (A) 5 (not 5.5)	PS4 Due
9	03/4, 03/06	GLM IV: Inference and measures of model fit	Binomial models in R	• (A) Ch. 4.3, 4.4, 4.6, 5.5 • (G-H) Ch 7.3, 7.4	DataCamp: Data Vis. w/. ggplot2 (Part 1)
10	03/18, 03/20	GLM V: Multinomial and ordered outcomes	Graphical interpretation of GLMs in R	• (A) Ch. 6 • (G-H) 6.5	PS5 Due

Continued on next page

Calendar – continued from previous page

Week	Date	Lecture topic	Lab topic	Readings	Assignments
11	03/25, 03/27	GLM VI: Counts outcomes	Multinomial and ordered probit in R	• (A) Ch. 7 • (G-H) 6.2	
12	04/01, 04/03	Correlated data I: MLM	Poisson Models in R	• (A) Ch. 9 (not 9.2.5) • (G-H) 11-13	PS6 Due
13	04/08, 04/10	Correlated data II: Multilevel GLM	Using lmer: estimation and interpretation	• (G-H) Ch. 12-13	
14	04/15, 04/17	Correlated data III: Panel Data	Using glmer: estimation and interpretation	• (A) 9.2.5	PS7 Due
15	04/22, 04/24	Correlated data IV: Other approaches to panel data	Using nlme: definition of working corr. Structure		