

POL606: Advanced Quantitative Methods for Social Sciences

Purdue University, Department of Political Science

Fall 2017, 09/07/2017

Professor: Viridiana Ríos, Ph.D.

Email: viririos@purdue.edu

Time commitment: 3-5 hours per week

Peer Meeting: TBD by students (weekly, optional)

Problem Solving Meeting: Tuesdays, 3:00pm-4:00pm (EST) (one per module, optional)

Data analysis has become increasingly salient for social scientists. Policy outcomes, news contents, political interactions, geolocations, and even dating preferences are now data points that can be used to expand our understanding of the world.

As a result, technology has dramatically transformed how we conceive our discipline. It allows social scientists to effectively develop and use a large variety of empirical tools to order, filter, and analyze substantive information that just a decade ago would have been inaccessible.

This transformation has turned social science into a powerful tool that systematically impacts decision makers, media, and policy. But it has also left social scientists in a unique position of responsibility. Our analyses can now be held accountable as predictions, traced, and replicated.

The goal of this course is to equip students with the tools, and thus the responsibility, to conduct accountable quantitative social science. Students will learn advanced data analysis skills to implement and replicate methods that are now considered basic staples of quantitative science. Specifically, we will study statistical learning techniques like linear regression, classification, sampling, subset selection, and unsupervised learning. By the end of this course, students will be empowered to conduct better, faster, and more efficient quantitative research, and to replicate or evaluate the research conducted by others.

What To Expect

This course is targeted education. It has been designed to extract the best of both traditional and distance learning. Unlike traditional class settings where students are required to advance at a similar pace, learn from a one-time explanation given by the professor, and devote time to commute to be physically present, this course allows students to learn at their own unique pace, using many complementary sources of information, and with complete travel flexibility. Yet, unlike distance learning courses, where students lack a structured pre-established classroom environment to work over issues that may be difficult to solve on their own, this course provides structured meetings with the professor to go over the most challenging material. In addition, this course encourages and rewards students that work together in small learning meetings.

The course is divided into six modules and an introductory lesson. There is a total of 16 weeks, each one requiring a minimum of 120 minutes of learning, a weekly 60 minute long meeting with peers, and additional time for problem solving meetings with the professor.

- **Module 0.** Introduction (1 week).
- **Module 1.** Linear Regression and R (2 weeks).
- **Module 2.** Classification: Binary and Ordinary Outcomes (3 weeks).
- **Module 3.** Resampling Methods: Cross-validation and Bootstrap (2 weeks).
- **Module 4.** Subset Selection: Forward and Backward Stepwise (2 weeks).
- **Module 5.** Unsupervised Learning (2 weeks).
- **Module 6.** Unsupervised Learning (2 weeks).
- **Final.** (1 week).

The two most important ingredients of success in this course are reading all the class materials, and working with your peers (see *peer meeting* section). Learning will be team work. Assignments will be solved in groups. Make sure you answer the questions of others (there will be bonus points every time you do). If you still need some help, request an appointment during office hours.

Teaching Team

– *Viridiana Ríos* (Professor) (viririos@purdue.edu) is a visiting assistant professor in the Department of Political Science at Purdue University. She holds a Ph.D. in Government from Harvard University where she was a member of the Institute for Quantitative Social Sciences. Her research uses quantitative methods to study how government structures may induce illegality, either by securing unequal privileges, facilitating criminal activities, or by triggering conflict and violence.

– *Mylene Cano* (Teaching Assistant) (mylenecano9@gmail.com) is an economics Master’s student at ITAM (Instituto Tecnológico Autónomo de México). She holds a B.A. in Economics and Political Science from ITAM.

Course Requirements

To be successful in this course, you need to acquaint yourselves with using R before starting (see software requirements). Also, students who have previous knowledge of linear regression will get much more out of this course.

– **Assignments** (80%): The professor will provide a published academic paper and its dataset. Students will use the tools learned in class to replicate the paper. Assignments must be delivered as HTMLs or PDFs created in RStudio (see software requirements). Data sets and RStudio code used to replicate the original code must also be delivered. Assignments can be done in teams but must be uploaded as individual zip files into Blackboard on the due date. Late assignments won’t be accepted. The time at which the zip file was uploaded will be considered the official time of delivery.

– **Final exam** (20%): A take-home examination to be solved individually.

– **Peer meeting** (up to 15% bonus): In my experience, group learning meetings between students are the best way to learn the material and to solve questions about the assignments. I strongly encourage you to meet at least once per week for an hour at the Schaffer Space (BRNG 2243) to present your doubts or solve the doubts of others about the code, the reading materials or the assignments. Attendance is optional. Please, set up a time that works the best for all of you. I will make sure the room is available. At the end of the semester, I will provide additional points to the students that were the most helpful in resolving questions at these meetings. If you are not in town, you may attend the meeting via Zoom (see software requirements).

– **Problem Solving meeting**: Students will hold regular meetings with Prof. Rios via Zoom (see software requirements) to go over the most challenging materials. Meeting ID: 553 268 0407. We will hold one meeting per module on Tuesdays from 3:00pm to 4:00pm (See Google calendar for details). Please send questions in advance. Attendance is not required.

Software Requirements

– **Computing**: Code labs are done in R. Download R and RStudio. To be successful in this course you need to be familiar with both. If you are not, take the time to go over an online tutorial like the Pre-fresher for Political Scientists that I taught at Harvard, or the Future Learning Webinar taught at the statistics department in Purdue. You need to be capable of uploading packages and data, conducting descriptive statistics, vector/matrix operations (including sub-setting), basic plots, understanding how a simple function and loop is created, sampling from a distribution, and calculating p-values. If you prefer to learn from videos, take a look at the R-tutorial YouTube channel. Our textbook, James (2015) also has a basic intro to R in section 2.3. Take our “Refresh Test” (on Blackboard) to check if you have enough R knowledge to take this course. Take some

time to feel comfortable with R-Studio. Here is a good tutorial.

– *Apps*: We will use Blackboard, Google Calendar, Perusall and Zoom as our basic tools. Take some time to become familiar with them. Blackboard is the course website. Google Calendar will allow you to know when assignments are due and schedule office hours. Perusall is a tool for group reading. Zoom is a tool for group conversations. Whenever meeting the professor at Zoom, use the following meeting ID: 553 268 0407.

Learning Tools

Each module contains four materials:

– *Video*: Brief lecture. Professor explains the tools covered, their relevance, and provides examples of implementation. Available at Blackboard.

– *Textbooks*: Assigned to provide detailed math behind each method/approach. Create an account at Perusall to access them. Use the code RIOS-9120. Perusall will be our depository of readings. Add comments and resolve the comments of your peers to get up to a 15% grade bonus.

– *Slides*: Developed to cover the most important contents that you need to grasp to be successful at this course. They will be available at Perusall. Students are encouraged to go over the slides and the readings simultaneously. Add comments and resolve the comments of your peers at Perusall to get up to a 15% grade bonus.

– *Code*: Code to implement and interpret each learned tool in R (see computing below). Available at Blackboard.

Reading materials

Two textbooks:

– James, G., Witten, D., Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning, with Applications in R*. Springer, 2013 (edition 2015).

– Scott Long, J. *Regression Models for Categorical and Limited Dependent Variables*. Sage Publications, 1997.

One paper per module:

– Dreher, A., and F. Schneider. “Corruption and the Shadow Economy: An Empirical Analysis.” *Public Choice* 144.1-2 (2010): 215-238.

– Holland, B. E., and V. Rios. “Informally Governing Information How Criminal Rivalry Leads to Violence against the Press in Mexico.” *Journal of Conflict Resolution* (2015): 1-25.

– Rios, V., “The Role of Drug-related Violence and Extortion in Promoting Mexican Migration.” *Latin American Research Review* (2014); 49(3): 199-217.

– The World Bank. “Haiti’s Post-earthquake dataset.”

– Rios, V., and M. Coscia. “Knowing Where and How Criminal Organizations Operate Using Google.” *CIKM* (2012); 12: 1412-1421.

– Blair, R., Blattman, C., and Hartman, A. “Predicting local violence: Evidence from a panel survey in Liberia.” *Journal of Peace Research*, 54(2), 298-312.

And support material:

– Fox, J. *Applied Regression Analysis, Linear Models, and Related Methods*. Sage Publications, Inc, 1997.

– Greene, W. H. *Econometric Analysis*. Pearson Education, 2008.

– Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. Springer Series in Statistics, 2009.

– Wooldridge, J. M. *Introductory Econometrics: A Modern Approach*. Nelson Education, 2000.

Calendar

Each student is responsible for following the Google Calendar of the class. Make sure you can view and send invites to the calendar of the class. You will receive an invite to the calendar a couple of

weeks before the class starts. Students are responsible for going over the materials of each module.

Office hours

To schedule an appointment send a Google Calendar invite to viridianarios@gmail.com. Verify that the time slot you are requesting is not taken by another student, falls between office hours (Tuesday 4:00-5:00pm EST), and is not longer than 20 minutes. Students may not request more than two slots in a given day. Most issues can be resolved in 15 minutes if you come prepared with concrete questions. We will meet using Zoom.

Course Agenda

Module 0 Introduction

Video: What is this course about?

Readings: James et.al. 2015, Chapter 2 (sections 2.1, 2.2).

Lecture Contents:

- What is statistical learning?
- Installation of R and RStudio.

Module 1. Linear Regression and R.

Video: Refreshing linear regression and learning the basics of R.

Readings:

- *Textbook:* Scott Long, Chapter 1 (Intro, section 1.1), Chapter 2.
- *Replication analysis:* Dreher, A., and F. Schneider. “Corruption and the Shadow Economy: An Empirical Analysis.” *Public Choice* 144.1-2 (2010): 215-238.

Slides Contents:

- Basic linear equations.
- Linear regression models.
- Assumptions.
- Omitted variable bias.
- Estimating coefficients.

Assignment 1: Run your own regression.

Module 2. Classification: Binary and Ordinary Outcomes.

Video: What is classification?

Readings:

- *Textbook:* Scott Long, Chapter 3, Chapter 5.
- *Replication analysis:* Holland, B. E., and V. Rios. “Informally Governing Information How Criminal Rivalry Leads to Violence against the Press in Mexico.” *Journal of Conflict Resolution* (2015): 1-25.

Slides Contents:

- Binary estimation: Logit and probit.
- Binary interpretation.
- How good is our binary model?
- Ordinal estimation: Logit and probit.

Assignment 2: Run your own probit or logit .

Module 3. Resampling Methods: Cross-Validation and Bootstrap.

Video: What is resampling?

Readings:

- *Textbook*: James et.al. 2015, Chapter 2 (section 2.2, except 2.2.3), Chapter 5.
- *Replication analysis*: Rios, V. “The Role of Drug-related Violence and Extortion in Promoting Mexican Migration.” *Latin American Research Review* 2014; 49:(3): 199-217.

Slides Contents:

- MSE and TMSE.
- Cross-Validation.
- The bootstrap.
- Are resampling methods good?

Assignment 3: Run your own cross-validation and bootstrap.

Module 4. Subset Selection: Forward and Backward Stepwise.

Video: What is subset selection?

Readings:

- *Textbook*: James et.al. 2015, Chapter 6 (section 6.1).
- *Replication analysis*: The World Bank. “Haiti’s Post-earthquake dataset.”

Slides Contents:

- Choosing the optimal model.
- Best subset selection.
- Forward Stepwise Regression.
- Backward Stepwise Regression.

Assignment 4: Run your own stepwise.

Module 5. Unsupervised Learning.

Video: What is a shrinkage method?

Readings:

- *Textbook*: James et.al. 2015, Chapter 6 (section 6.3), Chapter 10 (all sections).
- *Replication analysis*: Rios, V., and M. Coscia. “Knowing Where and How Criminal Organizations Operate Using Google.” *CIKM*, 2012; 12: 1412-1421.

Slides Contents:

- Dimension reduction methods.
- Principal components analysis.
- Clustering methods.

Assignment 5: Run your own cluster analysis.

Module 6. Tree-Based Methods.

Video: What are tree-based methods?

Readings:

- *Textbook*: James et.al. 2015, Chapter 8 (all sections).
- *Replication analysis*: Blair, R., Blattman, C., and Hartman, A. “Predicting local violence: Evidence from a panel survey in Liberia.” *Journal of Peace Research*, 54(2), 298-312.

Slides Contents:

- The basics of decision trees
- Bagging, Random forests, boosting.

Assignment 6: Run your own tree-based method.

Final examination.**Acknowledgments and Credits.**

This course stands on the shoulders of many professors who directly (or indirectly) taught me statistical methods. Contents and exercises developed by Hung Chen (National Taiwan University), Adam Glynn (Harvard University), Jens Hainmueller (Stanford University), Trevor Hastie (Stanford University), Gareth James (USC), Curtis Kephart (UCSC), Gary King (Harvard University), Clayton Nall (Stanford University), Ian Pardue (PSU), John Scott Long (Indiana University), Laura Simon (PSU), Robert Tibshirani (Stanford University), Daniela Witten (University of Washington) and Derek Young (PSU) were used as base material for parts of this course. I am also thankful to John Bishop (Google), Rosalee Clawson (Purdue University), Jay McCann (Purdue University), and MaryShannon Williams (Purdue University) for support, mentor ship and helpful comments.

Please, read the sections below (extracted from Purdue sample syllabus).

Purdue Honors Pledge.

“As a boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together - we are Purdue.”

Academic Dishonesty.

Purdue prohibits dishonesty in connection with any University activity. Cheating, plagiarism, or knowingly furnishing false information to the University are examples of dishonesty. [Part 5, Section III-B-2-a, Student Regulations] Furthermore, the University Senate has stipulated that the commitment of acts of cheating, lying, and deceit in any of their diverse forms (such as the use of substitutes for taking examinations, the use of illegal cribs, plagiarism, and copying during examinations) is dishonest and must not be tolerated. Moreover, knowingly aiding and abetting, directly or indirectly, other parties in committing dishonest acts is in itself dishonest. [University Senate Document 72-18, December 15, 1972]. See this. Academic integrity is one of the highest values that Purdue University holds. Individuals are encouraged to alert university officials to potential breeches of this value by either emailing (integrity@purdue.edu) or by calling 765-494-8778. While information may be submitted anonymously, the more information that is submitted provides the greatest opportunity for the university to investigate the concern.

Use of Copyrighted Materials.

Students are expected, within the context of the Regulations Governing Student Conduct and other applicable University policies, to act responsibly and ethically by applying the appropriate exception under the Copyright Act for the use of copyrighted works in their activities and studies. The University does not assume legal responsibility for violations of copyright law by students who are not employees of the University.

A Copyrightable Work created by any person subject to this policy primarily to express and preserve scholarship as evidence of academic advancement or academic accomplishment. Such works may include, but are not limited to, scholarly publications, journal articles, research bulletins, monographs, books, plays, poems, musical compositions and other works of artistic imagination, and works of students created in the course of their education, such as exams, projects, theses or dissertations, papers, and articles. See this.

Grief Absence Policy for Students.

Purdue University recognizes that a time of bereavement is very difficult for a student. Students will be excused for funeral leave and given the opportunity to earn equivalent credit and to demonstrate evidence of meeting the learning outcomes for missing assignments, or other assessments, in the event of the death of a member of the student's family. See this.

Emergencies.

Students with health or other emergencies are responsible for submitting assignments, but extensions

will be provided. Each case will be analyzed independently. Policies will be defined by contacting the instructors or TA via email. See this.

Note that the Purdue University Student Health Center (PUSH) does not provide students with “excuse” notes. Unless the student is acutely ill, there is nothing for PUSH to verify. Instead, encourage students to communicate with the instructor or TA soon as possible, in the event of an illness, so that you can work together for a positive solution.

CAPS Information.

Purdue University is committed to advancing the mental health and well-being of its students. If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, such individuals should contact Counseling and Psychological Services (CAPS) at (765)4946995 or here during and after hours, on weekends and holidays, or through its counselors physically located in the Purdue University Student Health Center (PUSH) during business hours.

Accessibility and Accommodations.

Purdue University strives to make learning experiences as accessible as possible. If you anticipate or experience physical or academic barriers based on disability contact the professor, and the Disability Resource Center (DRC). Students may present a Letter of Accommodation to you at any point in the semester. Should you have questions about accommodations, please contact the DRC (4941247, drc@purdue.edu).

Nondiscrimination.

Purdue University is committed to maintaining a community which recognizes and values the inherent worth and dignity of every person; fosters tolerance, sensitivity, understanding, and mutual respect among its members; and encourages each individual to strive to reach his or her own potential. In pursuit of its goal of academic excellence, the University seeks to develop and nurture diversity. The University believes that diversity among its many members strengthens the institution, stimulates creativity, promotes the exchange of ideas, and enriches campus life.

Purdue University prohibits discrimination against any member of the University community based on race, religion, color, sex, age, national origin or ancestry, genetic information, marital status, parental status, sexual orientation, gender identity and expression, disability, or status as a veteran. The University will conduct its programs, services, and activities consistent with applicable federal, state and local laws, regulations, and orders and in conformance with the procedures and limitations as set forth in Executive Memorandum No. D-1, which provides specific contractual rights and remedies. Any student who believes they have been discriminated against may visit here to submit a complaint to the Office of Institutional Equity. Information may be reported anonymously. See this.