PUBLG088 Advanced Quantitative Methods

https://github.com/smikhaylov/PUBLG088

Slava Mikhaylov
s.mikhaylov@ucl.ac.uk
Office: Room 2.01 29/30 Tavistock Square
Office hours: Mondays 15:00-17:00

Version: December 11, 2015

Overview and goals

The course builds on the introductory level of statistics and probability theory and introduces students to concepts and techniques essential to the analysis of social science data. The goal of the course is to teach students to understand and confidently apply various statistical methods and research designs that are essential for modern day data analysis. Students will also learn data analytic skills using a statistical software package *R*. This combination provides students with the skillset that is increasingly required by employers in today?s highly competitive job market.

Prerequisites

This is an advanced course intended for students who've already had some training in quantitative methods for data analysis. The minimum prerequisite is a good knowledge of linear regression models and familiarity with a statistical software package (e.g., R, Stata, SPSS, SAS, MATLAB). This would roughly be equivalent to having already taken at least one module in introductory statistics and/or econometrics.

Before you take the course

You are strongly recommended before taking this course to complete the following:

- James, Witten, Hastie, and Tibshirani (2013) *Introduction to Statistical Learning with Applications in R*, Chapters 1-2.
- *An Introduction to R*, available from http://cran.r-project.org/doc/manuals/R-intro.pdf
- Downloading and installing RStudio, available from http://www.rstudio.com.
- A brief online introduction to R Markdown, which we will use for completing the exercises for the course, see http://bit.ly/R_markdown

What to expect

Reading

The handout lists the required readings for every week. This required reading should be completed prior to the lecture in a given week. Students are expected to read the material very carefully. You may even find it helpful to read it first, come to the lecture and then re-read it after the lecture.

Homework

This is a methodological course, developing skills in understanding and applying statistical methods. You can only learn statistics by doing statistics and, therefore, the homework for this course is extensive, including weekly homework assignments. The assignments consist of data analytic tasks that you will be asked to complete either on your own or in small groups. It is important to learn to work in small groups because that's how much of the private and public sector operates. You may also be asked to present the results of your group work at the subsequent seminar. This helps you develop public speaking skills and skills of presentation of the results of your analysis to a wider audience – something you'll have to do in your job after finishing the degree.

Homework and readings will keep you busy. But this is an intensive advanced level course where we cover a lot of ground in ten weeks. So if you take this course you shouldn't expect an easy ride. However, skills you learn in the course are transferrable well beyond your degree.

Assessment

The course is assessed with a 3,000-word research paper in the form of a replication assignment (100% of the mark for the course). The goal of the research paper is to help you develop a deeper understanding of data analytic methods in your field of study and ability to apply them to a concrete empirical problem.

1. Your paper should address a substantive problem in your field of interest.
2. We will discuss the basic principles of replication and your essay assignment before the Reading Week, giving you an opportunity to work on the project during the Reading Week.
3. The research paper should replicate an academic article published in the last 5-8 years in a top ranking political science or economics journal. The definition of top ranking is any of the top 20 journals from the Google Scholar rankings:

   - Political Science: http://bit.ly/GoogleScholar_PoliSci
   - Economics: http://bit.ly/GoogleScholar_Economics

   Higher quality articles tend to appear in higher ranked journals, and in your replication exercise you should be learning from the best. Also, the number of citations of an article on Google Scholar is an indicator of its importance to the discipline.
4. Your paper must use methods we have or will talk about in the course, or at about the same level of sophistication as the material we cover here.

5. You don't need to replicate every section in the original article. You can replicate only the part that you can justify as important (this may not be the part that the author considered important).

6. Do not choose an article unless you fully understand its argument, methods, and substance.

7. Your key task for the assignment is to improve and extend the analysis in one specific area. This should be reflected in your paper structure: replication of the original result should not take up more than a page or two of your paper, main emphasis should be on the substantive contribution and extension of the original results that you are making.

8. Improvements can include changing the way the original article dealt with missing data, selection bias, omitted variable bias, the model specification, the functional form, adding control variables or better measures, extending the time series and conducting out-of-sample tests, applying a better statistical model, etc.

9. By 10am in Week 10 upload a 2-page proposal outlining a clear replication and improvement plan to the corresponding folder on the Moodle page for the course. Your proposal should have a max 200-word abstract (just like any journal article that you read) that contains a sentence on how replicating this article will help you with your dissertation project. In our Week 10 seminar we'll discuss your proposals.

10. Your paper should be proofread. Follow Harvard referencing style, using the citation style of the American Political Science Review.

11. This assignment is a version of the replication paper exercise at Gary King's Gov2001 course at Harvard. For more details and guidance read Gary King "Publication, Publication," *PS: Political Science and Politics*, Vol. 39, No. 1 (January, 2006), 119-125. Updates are available here: http://gking.harvard.edu/papers/.

12. Follow Gary King's discussion in "Publication, Publication" for overall style, structure, and presentation of your paper.

13. Good examples of a replication exercise are (a) Gary King and Michael Laver "On Party Platforms, Mandates, and Government Spending" *American Political Science Review*; and (b) Bell, Mark and N. Miller "Questioning the Effect of Nuclear Weapons on Conflict" *Journal of Conflict Resolution*.

14. You should use R Markdown to write your replication project.

15. The final version of the paper is due on 14 January 2016. For the PUBLG088 course, only an online submission is required. Please ensure you read the documents in the essay information folder on Moodle page for the course for full details on how to complete the submission. In addition you must also upload your complete replication package (R Markdown .Rmd file and full replication set) to the corresponding folder on Moodle.

The mark for your assessment consists of two parts: replication (50%) and extension (50%). The marks for replication and extension reflect how well you address points in the assessment description above.

Important Specifics

Computer Software

Computer-based exercises will feature prominently in the course, especially in the lab sessions. The use of all software tools will be explained in the sessions, including how to download and install them. All of the class work will be done using R, using publicly available packages.

Main Texts

The primary texts are:

- James et al. (2013) *An Introduction to Statistical Learning: With applications in R*. Springer. The book is available from the authors' page: http://www-bcf.usc.edu/~gareth/ISL/
- Stock, James and Mark Watson. 2012. *Introduction to Econometrics*, 3rd edition. [We'll use this text for the revision of basic concepts in the first two weeks.]

The following are supplemental texts which you may also find useful:

- Zumel, N. and Mount, J. (2014). Practical Data Science with R. Manning Publications.
- Lantz, B. (2013). Machine Learning with R. Packt Publishing.
- Conway, D. and White, J. (2012) Machine Learning for Hackers . O'Reilly Media.
- Leskovec, J., Rajaraman, A. and Ullman, J. (2011). Mining of Massive Datasets . Cambridge University Press.
- Zafarani, R., Abbasi, M. A. and Liu, H. (2014) Social Media Mining: An introduction . Cambridge University Press.

Short Course Schedule

Below is a proximate schedule for the course. Some topics may need to be covered in more than one lecture. We will take as much time as needed on each topic, so we may not get to all the topics listed below.

Date	Topic	Details	Readings
Oct 6	Course overview and review of probability and statistics	We will use this session to get to know the range of interests and experience students bring to the class, as well as to survey the approaches to be covered. We will also review basic probability and statistics and demonstrate the R software.	James et al. Chapter 1; Stock&Watson Chapters 2-3.
Oct 13	Introduction to statistical learning	Statistical learning, supervised vs unsupervised learning, regression vs classification, prediction vs inference. We will also discuss and demonstrate R Markdown.	James et al. Chapter 2.
Oct 20	Linear regression	The basic linear regression model.	James et al. Chapter 3.
Oct 27	Classification	Logistic regression, LDA.	James et al. Chapter 4
Nov 3	Resampling methods	Cross-validation, bootstrap.	James et al. Chapter 5.
Nov 10	Reading week		
Nov 17	Model selection and regularization	Subset selection, shrinkage methods, lasso, ridge regression.	James et al. Chapter 6.
Nov 24	Non-linear models	Polynomial regression, splines, local regression.	James et al. Chapter 7.
Dec 1	Tree-based methods	Decision trees, bagging, random forests, boosting.	James et al Chapter 8.
Dec 8	Unsupervised learning and dimensional reduction	Principal components analysis, correspondence analysis, cluster analysis.	James et al Chapter 10.
Dec 15	Unstructured data analysis	Working with unstructured data (text) in R.	

Detailed Course Schedule

Tuesday, October 6: Course overview and review of probability and statistics

We will use this session to get to know the range of interests and experience students bring to the class, as well as to survey the approaches to be covered. We will also review basic probability and statistics and demonstrate the R software.

Required Reading:

- James et al. Chapter 1.

Recommended Reading:

- Stock&Watson Chapters 2-3.

Tuesday, October 13: Introduction to statistical learning

Statistical learning, supervised vs unsupervised learning, regression vs classification, prediction vs inference. We will also discuss and demonstrate R Markdown.

Required Reading:

- James et al. Chapter 2.

Tuesday, October 20: Linear regression

The basic linear regression model.

Required Reading:

- James et al. Chapter 3.

Tuesday, October 27: Classification

Logistic regression, Linear Discriminant Analysis

Required Reading:

- James et al. Chapter 4.

Tuesday, November 3: Resampling methods

Cross-validation, bootstrap.

Required Reading:

- James et al. Chapter 5.

Tuesday, November 17: Model selection and regularization

Subset selection, shrinkage methods, lasso, ridge regression.

Required Reading:

- James et al. Chapter 6.

Tuesday, November 24: Non-linear models

Polynomial regression, splines, local regression.

Required Reading:

- James et al. Chapter 7.

Tuesday, December 1: Tree-based methods

Decision trees, bagging, random forests, boosting.

Required Reading:

- James et al. Chapter 8.

Tuesday, December 8: Unsupervised learning and dimensional reduction

Principal components analysis, correspondence analysis, cluster analysis.

Required Reading:

- James et al. Chapter 10.

Tuesday, December 15: Unstructured data analysis

Principles of text mining. Working with unstructured data (text) in R.

Required Reading:

- Grimmer, J, and B M Stewart. 2013. "Text as Data: the Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis*.
- Benoit, Kenneth and Alexander Herzog. In press. "Text Analysis: Estimating Policy Preferences From Written and Spoken Words." In *Analytics, Policy and Governance*, eds. Jennifer Bachner, Kathyrn Wagner Hill, and Benjamin Ginsberg. 21(3): 267-297.
- Lucas et al. 2015. "Computer-Assisted Text Analysis for Comparative Politics." *Political Analysis*, 2: 254-277.