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: September 28, 2016

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 modern 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. An introduction to statistics or econometrics at undergraduate level would serve as a very useful foundation for this course, although no formal prerequisites are required. Familiarity with computer programming or database structures is a benefit, but not formally required. If you are unsure whether your prior statistical training is sufficient to take this module please contact the module tutor for confirmation.

If you don't have the required prerequisites but still wish to take the module, you should work through any good textbook for introductory level statistics with R. For example,

- Gelman, Andrew, and Jennifer Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Before you take the course

You should complete the following:

- Data Camp R tutorials https://www.datacamp.com/courses/free-introduction-to-r
- Data Camp R Markdown tutorials https://www.datacamp.com/courses/reporting-with-r-markdow
- *An Introduction to R*, available from http://cran.r-project.org/doc/manuals/R-intro.pdf
- Kellstedt, Paul M. and Guy D. Whitten (2013). *The Fundamentals of Political Science Research*, 2nd edition, Cambridge University Press, Chapters 1-4. *These sections on research design are important for your weekly assignment and final assessment*.

You should also download and install the latest version of RStudio (http://www.rstudio.com) and R (https://cran.r-project.org) on your computer.

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 subsequent seminars. 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.

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 and RStudio as its integrated development environment (IDE): https://www.rstudio.com
- We are also using R Markdown as a format for all document creation (including your final assessment) from R: http://rmarkdown.rstudio.com
- GitHub is used to distribute materials for the module: https://github.com/smikhaylov/PUBLG088
- We will use Piazza to discuss any issues relating to the module: http://bit.ly/PUBLG088-2016

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/
- Hastie et al. (2009) *The Elements of Statistical Learning: Data mining, inference, and prediction*, Springer. The book is available from the authors' page: http://statweb.stanford.edu/~tibs/ElemStatLearn/

Additional recommended texts that we'll be using in the module:

- Kuhn, M. and K. Johnson (2013). *Applied Predictive Modeling*. Springer.
- Lesmeister, C. (2015). *Mastering Machine Learning with R*. Packt Publishing.
- Zumel, N. and Mount, J. (2014). *Practical Data Science with R*. Manning Publications.
- Leskovec, J., Rajaraman, A. and Ullman, J. (2014). *Mining of Massive Datasets*. 2nd edition. Cambridge University Press.

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

- Kromer, P. and R. Jurney (2015). *Big Data for Chimps*. O'Reilly.
- Lantz, B. (2013). *Machine Learning with R*. Packt Publishing.
- Conway, D. and White, J. (2012). *Machine Learning for Hackers*. O'Reilly Media.

Weekly drop-in R surgeries

Additional R session/surgery is available for SPP students taking Intro or Advanced methods modules in Term 1. This drop-in surgery is on Thursdays 2-5pm at Gordon House 106, led by methods Teaching Fellows. You can use this slot to work on your assignments and/or ask specific R-related questions.

Assessment

The assessment follows a model of a data hackathon, but with a longer time frame. A data hackathon is an intensive exercise that asks researchers to do their best to turn information into knowledge. Data hackathons (datathons) use research questions and datasets to advance knowledge.

For the final assessment you will frame a research question, create and implement a research design, mobilize data resources and present your findings in the form of 3,000-word research paper. This research paper forms 100% of the mark for the module.

We will hold several mini-datathons during the term as formative assessments. These are designed to help you develop necessary expertise for the final, summative assessment.

For the datathon you will be provided with a dataset of textual data. Using this dataset and at least one other dataset of your choice you are asked to address any issue of interest to you that involves extracting insights from textual data. All papers must have a strong social theory, political theory or policy perspective and a strong methodology perspective. Methodology used in your paper should be at least at the level of methods covered in the module.

Research Paper structure

1. Introduction

    - What social/policy question was asked or challenge addressed? Why is this question important or the challenge critical?
    - What datasets were used?
    - What is the novel contribution?
    - What is the key methodology or methodologies used?
    - What are the key findings?

2. Literature Review

    - Review the relevant literature
    - Use the literature review to point to a gap in the literature that your study aims to fill.
    - You can also use the review to highlight the motivation for the study – an unanswered question, an ongoing debate – or to show that a particular methodological approach is commonplace or novel and, thus, warranted.
    - Synthesize and summarize the evidence from the relevant literature (e.g. a table of main effects and most significant findings).

3. Theoretical Framework

- You want to use this section to build upon the Literature Review. In this section you look to demonstrate a logical link between predictors and outcome variables.
- You may well begin by detailing the assumptions that are required for your logic to hold. For example, in international relations realists would typically assume that the international system is anarchic and that states are the primary actors of the system. These assumptions enable these theorists to subsequently make logical arguments about the effect of weak international institutions upon the likelihood of conflict between states.
- You may also want to be offering key conceptual definitions in this section to clarify what your variables look like. Importantly, this is not operationalization (measurement) but rather a discussion of what the state is and what an anarchic international system really means (just following up on the examples from above).
- Ideally, this section would conclude with the statement of some testable research hypotheses. These are propositions about the relationship between predictors and outcome.

4. Methodology/Research Design

    - This section serves the purpose of detailing how you will go about testing the hypotheses laid out above.
    - Begin by detailing the overview of your research methodology – data collection (for additional datasets you use) and analysis.
    - You want, also, to identify your unit-of-analysis and your spatial and temporal domains.
    - You then describe the operationalization of your variables: outcome variable, predictors, and controls.
    - You should also provide summary statistics for your variables, possibly also a visualization of main spatio-temporal trends in your variables.

5. Results

    - Simply detail findings from most important to least. Begin, therefore, by discussing the results as they pertain to your hypotheses – can you reject your hypotheses? Then move to discussing any additionally interesting findings on the control variables.
    - You should present the findings in summary form – a table of results or a graph.
    - You should think about substantive interpretation of your results. What do your findings mean for the policy problem at hand?

6. Conclusion and discussion

    - Provide an interesting overview of the study. What have we learned?
    - Next, discuss what are the limitations of your study, also what could be done to improve/build upon the study. You can demonstrate some self-criticism.
    - Finally and perhaps most importantly, discuss policy implications of your study. Ideally you can also provide policy recommendations.

Submission

- Follow submission guidelines on Moodle page for the module (Essay Information folder).
- You will need to submit a PDF file compiled from R Markdown document to Turnitin.

- In addition, you will need to submit your source R Markdown file with the full replication set (as one Zip archive) to a corresponding folder on Moodle page for the module.
- By 10am on Monday in Week 10 upload to a corresponding folder on Moodle page for the module a max 2-page proposal outlining the Introduction (see Research Paper Structure section above) of your research paper. This may be in bullet points. In our Week 10 seminar we'll discuss your proposals.
- Your paper is due on 10 January 2017, 2pm. For the PUBLG088 course, only an online submission is required.

Marking Guidelines

A strong paper will have the following components:

- Follow the structure of the paper outlined above.
- Employ multiple data sets, including the provided textual dataset.
- Include a high quality visualization.
- Develop an empirical model that would support answering a key social or policy issue that you formulate.
- Generate a new empirical finding that challenges or provides novel support for existing social or political theory.
- In addition, a strong paper should be well-written and provide some level of creativity in its use of or combination of data.

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	Core Readings
Oct 4	Course overview and introduction to statistical learning	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 discuss the concept of statistical learning	James et al. Chapters 1-2.
Oct 11	Linear regression	The linear regression model.	James et al. Chapter 3.
Oct 18	Classification	Logistic regression, LDA.	James et al. Chapter 4
Oct 25	Resampling Methods and Model Selection and Regularization	Cross-validation, bootstrap, shrinkage methods, ridge and lasso.	James et al. Chapters 5-6.
Nov 1	Non-linear Models and Tree-based Methods	GAMs, local regression, decision trees, random forests, boosting.	James et al. Chapters 7-8.
Nov 8	Reading week		
Nov 15	Neural Networks and Deep Learning	Neural Networks, fitting, deep belief networks, deep Bolzmann machines	Hastie et al. Chapter 11.
Nov 22	Support Vector Machines and Unsupervised Learning	SVM, principal components analysis, cluster analysis.	James et al Chapters 9-10.
Nov 29	Vector Space Model	Vector Space Model for text data, tf-idf, Naive Bayes classifier, correspondence analysis, text clustering	Manning, Raghavan and Shutze Chapters 2, 6, 13-17.
Dec 6	Word embedding models	word2vec, doc2vec	Mikolov et al. (2013)
Dec 13	Topic models	Latent Dirichlet Allocation, probabilistic topic models, dynamic topic models	Blei and Lafferty (2009)

Detailed Course Schedule

Tuesday, October 4: Course overview and introduction to statistical learning

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 discuss the concept of statistical learning

Required Reading:

- James et al. Chapters 1-2.
- Hastie et al. Chapter 2.

Recommended Reading:

- Zumel and Mount Chapter 10.
- Kuhn and Johnson Chapter 1.

Tuesday, October 11: Linear regression

The linear regression model.

Required Reading:

- James et al. Chapter 3.
- Hastie et al. Chapter 3.2.

Recommended Reading:

- Lesmeister Chapter 2.
- Zumel and Mount Chapter 7.1.
- Kuhn and Johnson Chapters 5, 6.1-6.2.

Tuesday, October 18: Classification

Logistic regression, LDA.

Required Reading:

- James et al. Chapter 4.
- Hastie et al. Chapter 4.1-4.4.

Recommended Reading:

- Lesmeister Chapter 3.
- Zumel and Mount Chapter 7.2.
- Kuhn and Johnson Chapters 11, 12.1-12.3, 13.5-13.6.

Tuesday, October 25: Resampling methods and model selection and regularization

Cross-validation, bootstrap, shrinkage methods, ridge and lasso.

Required Reading:

- James et al. Chapter 5-6.
- Hastie et al. Chapter 3.3-3.5, 7.10-7.11.

Recommended Reading:

- Lesmeister Chapter 4.
- Kuhn and Johnson Chapters 4, 6.3-6.4, 12.4-12.5.
- Martin Wainwright (2014). "Structured Regularizers for High-Dimensional Problems: Statistical and Computational Issues." *Annual Review of Statistics and Its Application*, 1: 233-253.
- Buhlmann, P., M. Kalisch, L. Meier (2014). "High-Dimensional Statistics with a View Toward Applications in Biology." *Annual Review of Statistics and Its Application*, 1: 255-278.
- Lange, K., J. Papp, J. Sinsheimer, E. Sobel (2014). "Next-Generation Statistical Genetics: Modeling, Penalization, and Optimization in High-Dimensional Data." *Annual Review of Statistics and Its Application*, 1: 279-300.

Tuesday, November 1: Non-linear models and tree-based methods

GAMs, local regression, decision trees, random forests, boosting.

Required Reading:

- James et al. Chapter 7-8.
- Hastie et al. Chapter 9.1-9.4, 10.1.

Recommended Reading:

- Lesmeister Chapter 6.
- Zumel and Mount Chapter 9.1-9.3.
- Kuhn and Johnson Chapters 7.2, 8, 14.
- Muchlinksi, D., Siroky, D., Jingrui, H., Kocher, M., (2016) "Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data." *Political Analysis*, 24(1): 87-103.

Tuesday, November 15: Neural Networks and Deep Learning

Required Reading:

- Hastie et al. Chapter 11.

Recommended Reading:

- Ruslan Salakhutdinov (2015). "Learning Deep Generative Models." *Annual Review of Statistics and Its Application*, 2: 361-385.
- Goodfellow et al. *Deep Learning*, MIT Press. Book is available online from authors' page: https://www.deeplearningbook.org.

Tuesday, November 22: Support Vector Machines and unsupervised learning

SVM, principal components analysis, correspondence analysis, cluster analysis.

Required Reading:

- James et al. Chapter 9-10.
- Hastie et al. Chapter 12.1-12.3, 14.3, 14.5.
- Leskovec et al. Chapter 11.

Recommended Reading:

- Lesmeister Chapter 5, 8-9.
- Zumel and Mount Chapters 8.1, 9.4.
- Kuhn and Johnson Chapters 7.3, 13.4.

Tuesday, March 1: Vector Space Model

Required Reading:

- Manning C., Raghavan P., and H. Shutze (2009). *An Introduction to Information Retrieval*. Chapters 2.2, 6.2-6.4, 13.1-13.4, 14-17. The book is available online from the authors' page: http://nlp.stanford.edu/IR-book/
- Jurafsky and Martin (2009) *Speech and Language Processing*, 2nd edition. Chapter 2.
- Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97: 311-331.
- Lowe, William. 2008. "Understanding Wordscores." Political Analysis 16(4): 356-371.
- Greenacre, M. (2007). *Correspondence Analysis in Practice*, 2nd edition. Appendix A & B.

Recommended 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*.
- Spirling, A. (2012), "U.S. Treaty Making with American Indians: Institutional Change and Relative Power, 1784âĂŞ1911." *American Journal of Political Science*, 56: 84–97.
- Herzog, A. and K. Benoit (2015), "The most unkindest cuts: Speaker selection and expressed government dissent during economic crisis." *Journal of Politics*, 77(4):1157–1175.
- Borcard, D., F. Gillet, P. Legendre (2011). *Numerical Ecology with R*. Springer.

Tuesday, December 6: Word embedding models

Required Reading:

- Mikolov, Tomas et al. "Efficient Estimation of Word Representations in Vector Space."
- Goldberg, Yoav and Omer Levy "word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method."

Recommended Reading:

- Mikolov, Tomas; Sutskever, Ilya; Chen, Kai; Corrado, Greg S.; Dean, Jeff (2013). "Distributed representations of words and phrases and their compositionality." Advances in Neural Information Processing Systems.
- Levy, Omer; Goldberg, Yoav; Dagan, Ido (2015). "Improving Distributional Similarity with Lessons Learned from Word Embeddings." Transactions of the Association for Computational Linguistics.
- Pennington et al. "GloVe: Global Vectors for Word Representation."
- Huang et al. "Improving Word Representations via Global Context and Multiple Word Prototypes."

Tuesday, December 13: Topic models

Required Reading:

- David Blei (2012). "Probabilistic topic models." *Communications of the ACM*, 55(4): 77-84.
- Blei, David, Andrew Y. Ng, and Michael I. Jordan (2003). "Latent dirichlet allocation." *Journal of Machine Learning Research* 3: 993-1022.
- Blei, David (2014) "Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models." *Annual Review of Statistics and Its Application*, 1: 203-232.
- Roberts, Stewart, Tingley, Lucas, Leder-Luis, Gadarian, Albertson, and Rand (2014). "Structural topic models for open-ended survey responses." *American Journal of Political Science*, 58(4): 1064-1082.

Recommended Reading:

- Blei, D. and J. Lafferty "Topic Models." In *Text Mining: Classification, clustering, and applications,* A. Srivastava and M. Sahami (eds.), pp 71-94, 2009. Chapter available here: http://www.cs.princeton.edu/~blei/papers/BleiLafferty2009.pdf.
- Blei, David M., and John D. Lafferty. "Dynamic topic models." In *Proceedings of the 23rd international conference on machine learning*, pp. 113-120. ACM, 2006.
- Mimno, D. (April 2012). "Computational Historiography: Data Mining in a Century of Classics Journals." *Journal on Computing and Cultural Heritage* 5 (1).
- Lesmeister Chapter 12.