POLS0010 – 2019-2020

Data Analysis: Analyzing & Understanding Data

Lecturer: Stephen Jivraj (SJ), James Cheshire (NB) and Tom O'Grady (TO)

Office Hours:
NB: TBC
SJ: TBC
TO: 3-5pm, B.13 29-31 Tavistock Square

Teaching: 40 hours of lectures, 20 hours of computer tutorials

Credits: 30 credits/ 8 US Credits/ 15 ECTS Credits

Assessment: Term I: One 3,000 word essay (50% - submitted in two equally weighted parts)
Term II: One 3,000 word essay (50%)

Essay Deadline:
Term I essay Part 1:
Term I essay Part 2:
Term II essay: Tuesday 21st April 2020 by 2pm

Attendance: Attendance is compulsory at all lectures and seminars for which students are timetabled. Attendance will be monitored and no student will be entered for assessment unless they have attended and pursued the module to the satisfaction of the department.

USEFUL LINKS

Lecture and Seminar Times:
Online Timetable at www.ucl.ac.uk/timetable

Extenuating Circumstances
Link Pending

Penalties for Late Submission and Overlength Essays
Link Pending

Essay Submission Information
Link Pending

Essay Writing, Plagiarism and TurnItIn
http://www.ucl.ac.uk/current-students/guidelines/plagiarism

Content

This module aims to build skills in regression analysis using a variety of modelling techniques: linear, limited dependent, panel, time-series and longitudinal models. It also develops students' skills in spatial data analysis and practical skills in data analysis of sample social surveys. Students will be proficient users of RStudio by the end of the module. The module teaches skills that students can apply across a range of jobs—in the public, private and third sectors. Emphasis is placed on using real-world data, 'hands-on' lab sessions, analysis, interpretation and visualisation.

Lectures and tutorials
Each week there will be an introductory lecture followed by a computer tutorial. The lecture will last two hours and the tutorial will last one hour. The lectures will introduce students to many of the ideas and issues relating to the various topics. The computer tutorials will provide an opportunity to implement the techniques covered in the lectures.

Please check the online timetable for last minute changes to room bookings.

Term 1:
Lectures: TBC
Computer tutorial: TBC

Term 2:
Lectures: TBC
Computer tutorial: TBC

Assessment
The module is assessed through the completion of two essays due after each term. Each essay accounts for 50% of the total marks on the module. The essays must be a **maximum of 3,000 words**. The term 1 essay will be submitted in two equally weighted parts in terms of the word count and mark (i.e. 1,500 words each and 25% each of the total marks on the module). Please include the word count at the top of the essay and submit your essay using your candidate number as the filename. Please check the Department of Political Science essay submission checklist and penalties for late submission and exceeding word limits.

You will find useful guidance for writing and presenting essays on the Department of Political Science student website. These guidelines are designed to help you, and you should read them carefully and do your best to follow them. Good essays give clear and focused answers to the question asked, they have clear structures, and they will be adequately and appropriately referenced. They do not provide a vague and unstructured discussion of the topic. Plagiarism is taken extremely seriously and can disqualify you from the module (for details of what constitutes plagiarism see http://www.ucl.ac.uk/current-students/guidelines/plagiarism). If you are in doubt about any of this, ask the tutor.

Other non-assessed work
The computer tutorials will allow students to apply and test their knowledge of the material covered on the module and weekly exercises should be submitted for feedback from the module tutor. It is intended that students will complete weekly exercises outside of class.

Reading list
In order to gain a sufficient understanding of the concepts and techniques taught on this module, students will need to do background reading. No one book covers all of the content on the module, and it is worth reading as widely as possible. Note also there are useful online resources, with some examples included below.

Term 1:
Brunsdon, C. and Comber, L. 2015. An Introduction to R for Spatial Analysis and Mapping. London: Sage.
de Smith, M., Goodchild, M. and Longley, P. 2014. Geospatial Analysis. Available free online here: http://www.spatialanalysisonline.com/
Field, A., Miles, J. and Field, Z. 2012. Discovering Statistics Using R. London: Sage.
Gelman, A. and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.
Lumley, T. (2010) Complex Surveys: A Guide to Analysis Using R

Term 2:
- Christopher Dougherty. 2016. *Introduction to Econometrics.* Oxford University Press
- Andrew Gelman and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press.
- Kosuke Imai. 2017. *Quantitative Social Science: An Introduction*. Princeton University Press
- Gareth James, Daniela Whitten, Trevor Hastie and Robert Tibshirani. 2017. *An Introduction to Statistical Learning with Applications in R*
- Ted Kwartler. 2017. *Text Mining in Practice with R*. Wiley & Sons

Most of the reading is available in UCL library, although there are generally limited copies. Many items are also held in Senate House library.

Online resources

R resources
UCLA Statistical Consulting Group introduction to R
http://www.ats.ucla.edu/stat/r/seminars/intro.htm

Neat websites with basic data analysis commands described
http://www.statmethods.net/index.html
http://www.cookbook-r.com

The R Guide to UK Data Service key UK Surveys
http://ukdataservice.ac.uk/media/398726/usingr.pdf

Producing simple graphics with R
http://www.harding.edu/fmccown/r/

Term 1

Week 1 (NB)

Geographic Data and Methods

This week will mark the transition to working with geographic data in the rest of the module. It will revise some of the concepts briefly introduced last year as well as introduce others, such as the modifiable areal unit problem, in more depth.

Suggested Reading

Rogerson, P. 2010. Statistical Methods for Geography. Sage: London. Chapter 1.7

Longley, P., Goodchild, M., Maguire, D., Rhind, D. 2015. Geographic Information Science and Systems. Chapter 13.

Fotheringham, S. Brundson, C. and Charlton, M. 2007. Quantitative Geography. Perspectives on Spatial Data Analysis. Sage, London. **Chapter 4.**

Week 2 (NB)

Measures of Spatial Autocorrelation

This lecture and associated practical will introduce and discuss a fundamental set of statistics that are used to determine the significance of spatial patterns.

Suggested Reading

de Smith M J, Goodchild M F, Longley P A (2009) Geospatial Analysis, 3rd edition. **Chapter 2, Chapter 3.**
Brunsdon, C. and Comber, L. 2015. An Introduction to R for Spatial Analysis and Mapping. London: Sage. Chapter 8

Week 3 (NB)

Geographically Weighted Regression

Here we add the geographic dimension to the regression models you have become familiar with**.**

Suggested Reading

Brunsdon, C. and Comber, L. 2015. An Introduction to R for Spatial Analysis and Mapping. London: Sage. Chapter 7 & 8

de Smith, M., Goodchild, M. and Longley, P. 2014. Geospatial Analysis. Available free online here: http://www.spatialanalysisonline.com/ Chapter 5.6 "Spatial Regression".

Week 4 (NB)

Spatial Detection of Clusters

Many spatial datasets come in the form of point patterns – from disease outbreaks to Tweets – this weeks lecture and practical will outline the approaches used to determine if such points are spatially clustered or not.

Suggested Reading

Rogerson, P. 2010. Statistical Methods for Geography. Sage: London. Chapter 10

Brunsdon, C. and Comber, L. 2015. An Introduction to R for Spatial Analysis and Mapping. London: Sage. Chapter 6.5-6.5

Week 5 (NB)

Advanced R

This week will cover some more advanced elements of programming with R, such as writing functions and using R Markdown.

Suggested Reading
https://nicercode.github.io/guides/functions/

Week 6 (NB)

Revision and Assessment Workshop

This lecture concludes the first half of the module. It will provide an opportunity to answer any outstanding questions and the associated practical will serve as a help session to deal will any outstanding issues from previous weeks before transitioning to the next part of the module with Stephen.

Week 7: Multiple regression (SJ)
This lecture covers the theory and implementation of an extension to simple linear model by adding multiple explanatory variables (i.e. multiple regression), including dummy variables. Techniques of model selection and comparison are explored.

Core reading
Field et al. Chapter 7.

Supplementary reading
Gelman and Hill. Chapter 3.

Week 8: Further techniques in multiple regression (SJ)
The multiple regression model can be extended further by transforming variables to ensure linearity, polynomial effects, interaction terms, and variable centring.

Core reading
Gelman and Hill. Chapters 3-4.

Supplementary reading
Lumley. Chapter 5

Week 9: Sampling and non-response (SJ)
This lecture reintroduces key sampling concepts, types of probability samples, non-response types and bias. Weighting is introduced as a way to deal with survey non-response.

Core reading
Williams RL (2014). Survey sampling and weighting. In: Culyer AJ (ed), Encyclopedia of Health Economics, Oxford, UK: Elsevier, pp. 371–374.

Supplementary reading
Lumley. Chapters 1-3

Week 10: Item non-response and imputation methods (SJ)
This lecture introduces missing data mechanisms and approaches to impute data. These approaches include mean imputation, deterministic imputation, hot deck imputation and model-based imputation.

Core reading
Gelman and Hill. Chapter 25.

Supplementary reading
Lumley. Chapter 9.

TERM 2 (TO)

PART I: ADVANCED REGRESSION TECHNIQUES

Week 11: Core Skills in Regression

We begin the second half of term 2 with an overview of multiple regression, introducing some new concepts. We'll first remind ourselves how to use and interpret regression results, including prediction and marginal effects, using the mathematical tool of differentiation. We'll then learn the core statistical concepts of bias, consistency and efficiency, using them to explain why OLS regression may be optimal under certain conditions. Along the way, we'll introduce the idea of using simulation to study the properties of estimators and to conduct hypothesis tests.

Reading

- Gelman and Hill, Ch 7.1-7.2
- Imai, Ch. 7

Week 12: Binary Outcome Models and Maximum Likelihood Estimation

Many important outcomes in the social sciences are discrete rather than continuous, but standard linear regression assumes continuous dependent variables. In weeks 12 and 13 we'll look at logit models for binary outcomes (1 or 0) such as the decision to vote or not in an election. This week we focus on their use in predicting the probability of events, introducing some basic concepts from machine-learning to assess model performance. We'll also introduce the idea of maximum likelihood estimation, which is used to estimate these models in practice.

Core Reading

- James et al Ch. 4 [ignore pp. 138-143]

Supplementary Reading

- Dougherty Ch. 14 [useful if you want more mathematical details – ignore pp. 381-390]

Week 13: Further Issues in Binary Outcome Models – Interpretation, Prediction and Uncertainty

Logistic regressions differ in important ways from standard linear regression. Today we'll explain how to interpret these models and how to measure marginal effects and changes in predicted probabilities. We'll also cover uncertainty and hypothesis testing, including inference through simulation.

Reading

- Gelman and Hill Ch. 5

Week 14: Panel Data and Random Effects Models

Today we'll introduce data that varies across more than one level, starting this week with individual units observed at multiple points in time: panel data We'll cover the basics of panel data estimation, including the distinctions between full, partial and no pooling of observations across time. We'll end by introducing random effects models.

Reading

- Dougherty Ch. 14

Week 15: Multilevel/Hierarchical Models

We'll continue our exploration of multilevel data by looking at hierarchical data consisting of individuals observed in multiple geographic units such as cities or schools. We'll explain how to use multilevel modelling to understand relationships at both levels, avoiding the 'ecological fallacy', including models with intercepts and slopes that vary by geographical unit.

Reading

- Gelman and Hill Chs. 11-13

Week 16: Multilevel Logistic Regression and MRP (Multilevel Regression and Post-Stratification) Estimation of Public Opinion

One of the most important uses of multilevel modelling is for predicting responses to public opinion surveys (including voting intentions) within small geographical areas such as electoral constituencies, where survey responses are sparse. MRP combines logistic multilevel regression with demographic data to predict survey responses by area. One very notable use was in the 2017 General Election, where YouGov's constituency-by-constituency forecast using MRP was the only one to successfully predict seat losses by the Conservative Party. We'll learn how to construct MRP estimates, including the use of multilevel logistic regression, an extension of the techniques from week 15 to binary outcome data.

Core Reading

- Chris Hanretty, Benjamin Lauderdale and Nick Vivyan (2018). "Comparing Strategies for Estimating Constituency Opinion from National Survey Samples." *Political Science Research and Methods* 6 (3): 571-591

Supplementary Reading

- Gelman and Hill Ch. 14
- Benjamin Lauderdale, Delia Bailey, Jack Blumenau and Doug Rivers (2017). "Model-Based Pre-Election Polling for National and Sub-National Outcomes in the US and UK." *Working Paper,* available here and on Moodle

PART II: INTRODUCTION TO TEXT MINING IN THE SOCIAL SCIENCES

Week 17: Turning Text into Data and Describing it in R

The availability and use of text as a source of data in the social sciences has grown dramatically over the past couple of decades. We'll start with the basics of text analysis in R, including how to read in text data, the removal of stopwords and punctuation, and the use of stemming. We'll then learn how to turn texts into a 'bag of words' via document-term matrices and explore their features using word frequencies, wordclouds and other descriptive tools.

Core Reading

- Kwartler Chs. 2.1-2.3, 2.7-2.9, 3 and pp. 100-101

Supplementary Reading

- Imai Ch. 5.1

Week 18: Classifying and Scaling Documents I

We'll begin exploring how to classify documents using supervised dictionary methods, which are particularly useful in sentiment analysis and other well-defined classification tasks. We'll also begin exploring more open-ended text classification using tools from machine-learning that classify documents based on their similarity to a labelled test set, introducing the technique of Cross-Validation to assess model performance.

Reading

- Kwartler Chs. 6-7
- James et al Ch. 5.1

Week 19: Classifying and Scaling Documents II

We'll finish up our exploration of text classification by introducing the Support Vector Machines approach from machine learning, and then move on to the conceptually similar task of scaling, e.g. placing politicians on a left-right scale based on their rhetoric, introducing the widely-used WordScores scaling algorithm

Core Reading

- James et al Ch. 9
- Michael Laver, Kenneth Benoit and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* 97 (2), pp. 311-331

Week 20: Automated Data Collection and Web Scraping

Text data is often collected from the web using various automated methods such as calls to APIs, and a technique of computer-assisted downloading known as web scraping. We'll learn how to use R to collect, clean up and store the content of websites for future analysis, including the use of regular expressions.

Reading

- Kwartler Chs. 2.4-2.6 and 9