POL30430
Data Analytics for Social Science

Johan A. Elkink

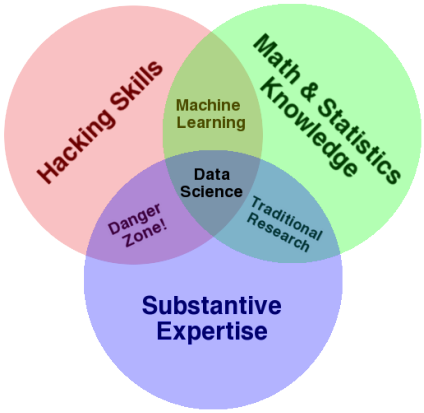
School of Politics and International Relations
University College Dublin

http://www.joselkink.net

Spring 2020

Introduction

There is currently a hype around "big data" and "data science" as new applications of statistics and computer science, with companies regularly hiring specialists in the manipulation of (large) data sets and visualising and analysing patterns in the data. Rather than just a combination of computer science ("hacking skills") and statistics, the field of data science is really at the intersection of three disciplines, including "substantive expertise". Most of data science is concerned with the predictive modelling of human behaviour, i.e. the core specialism of a social scientist. The most important contrast with regular applications of statistics to social science is that data scientists are typically interested in *predicting* behaviour, while social scientists are interested in *understanding* human behaviour. This leads to different modelling decisions.



This course provides an overview of common statistical methods applied to the social sciences, but also a number of techniques more common to data science and machine learning. The applications will primarily come from political science, or neighbouring disciplines such as sociology, public policy and development. It starts with a brief recap of the basic principles of statistical analysis, then discusses how to access, manipulate, and summarise data, and then moves on to a range of different methods—regression analysis, logistic regression, dimension reduction techniques, quantitative text analysis, etcetera.

All material is discussed using real world examples of data analysis, with both micro- and macro-level data, that also form the basis for the continuous assessments. Rather than delving deeply into the mathematical properties of various techniques, this module focuses on the application and the types of problems where particular techniques can be applied. All analyses will be performed in R and RStudio (see also Altman et al., 2011), which is freely available software and you are encouraged to install this on your personal computer prior to the class from `http://www.r-project.org/` and `http://www.rstudio.com/`.[1] R is quickly becoming the most popular statistical software in political science and is already the predominant tool in statistics.[2] R is a statistical programming language and it will take some time to get used to programming, instead of using a graphical user interface. Not to worry, however: we will do the R programming together, in class, and assignments will be based on the in-class analyses.

The learning outcomes associated with this twelve-week course are aimed at students being able to:

- Basic understanding of statistical analysis in the social science;
- Ability to manipulate data sets to prepare for statistical analysis;
- Ability to select the appropriate statistical technique for a range of different types of empirical questions;
- Ability to execute a range of standard techniques;
- Ability to describe, interpret, and present statistical analysis to a wider audience;
- Develop a greater familiarity with a range of techniques and methods through a diverse set of theoretical and applied readings;
- Know where to go to learn more about the techniques in this class and other that were not covered in this class.

Prerequisites

There are no prerequisites for this course, but a basic introductory course to statistics will help. Note that different students in the class will have different backgrounds, some including reasonably advanced econometrics training and others with barely any familiarity with statistics. We will attempt to keep it exciting for all students, but do keep in mind that not all students in the class might be at the same level. This also means that you should never hesitate to ask a question—you are likely to overestimate how many students in the class are following the lecture when you are confused.

[1] Install R first, as RStudio will detect the installed R version. After that, you will only use RStudio, not R directly.

[2] Data scientists also make extensive use of Python, but those two packages are nearly exchangeable in terms of functionality.

Texts

This course will assign a variety of reading materials, some required and some supplementary. A large proportion of the material is covered in James et al. (2013), which is a relatively accessible and very well-written introductory textbook to data science. It would be a good idea to purchase this book.

A free accessible book online is Grolemund and Wickham (2016), which is useful because it is very applied, providing clear sample code, and because it is available for free online at `http://r4ds.had.co.nz/`, but it is not as good in coverage and quality as the James et al. (2013) volume.

For all readings, the focus should be on the basic intuition behind each technique, the typical applications, and the limitations. You can skip most of the technical parts.

Classes

Classes take place once a week, Thursday 9–11 in lab G5 of the Daedalus building at UCD and are a mixture between lectures and labs.

Grading

The only way to properly learn statistics is by hands-on training. You will need to work with actual data and produce your own statistical analyses—just the theory will never be sufficient. Furthermore, probably the most difficult part is to properly describe your analysis and interpret the results. For that reason, the grading will be based on three essays, each of which will present and interpret a different analysis. The analyses themselves will be part of the lab exercises, so will be largely done during class, but the write-up and interpretation is done in the essays. While the lab exercises should largely suffice for the analysis, you are of course encouraged to add similar analyses, as you see fit.

For late submissions the standard policies apply.[3] It should also be taken into consideration that a late submission might result in a delayed return of feedback to the entire class. Exemptions will be granted only on the basis of illness or bereavement, documented in all cases.[4]

All assignments should be submitted electronically to `jos.elkink@ucd.ie`, consisting of either: a PDF file containing the essay and an R-file with all the commands necessary for the analysis, or: a PDF file generated using RMarkdown with the essay and all necessary commands and results. Note that the R-file or the RMarkdown output should contain:

[3] `http://www.ucd.ie/governance/resources/policypage-latesubmissionofcoursework/`.

[4] `https://www.ucd.ie/science/study/currentundergraduatesciencestudents/extenuatingcircumstances/`.

- all commands used for the analysis;
- comments to explain what the code is doing, especially to explain what variable names represent;
- enough white space (empty lines, spaces between elements of a command, etc.) to keep the file readable.

Each essay should consist of a short introduction, a description and motivation of the data and methods used (approximately 25% of the essay), the analysis including necessary graphs and tables (approximately 35%), and an interpretation and conclusion (approximately 40%). Everything needs to be properly referenced. For the theoretical section, you can simply refer to the article the analysis is based on—focus the essay on the analysis and methodology.

1. **Project I 20%**. The first analysis will concern survey data and make use primarily of graphical and descriptive statistics. Deadline: **24 February, 1 pm** (1000–1250 words).
2. **Project II 40%**. The second analysis will focus on the use of regression analysis and classification methods. Deadline: **6 April, 1 pm** (2000–2500 words).
3. **Project III 40%**. The third analysis will focus on the statistical analysis of text. Deadline: **5 May, 1 pm** (2000–2500 words)—if you have exams in May, make sure to plan your time!

Plagiarism

Although this should be obvious, plagiarism—copying someone else's text without acknowledgement or beyond "fair use" quantities, or that of your own in another submission or publication—is not allowed. UCD policies concerning plagiarism can be found online.[5]

Contact

I do not have fixed office hours, so you can make an appointment by email. If a personal visit is not necessary, the easiest way to reach me is by email (`jos.elkink@ucd.ie`). I will not have my own office this term, so we will have to arrange a meeting location and time—the easiest is typically right before or after class.
Course materials will be uploaded to `http://www.jos.elkink.net/teaching.php`.

To stay up to date with developments in the UCD School of Politics and International Relations, also keep an eye on the following social media:
Web: `http://www.ucd.ie/politics/`

[5] http://www.ucd.ie/governance/resources/policypage-plagiarismpolicy/

Blog: http://politicalscience.ie/
Twitter: http://twitter.com/ucdpolitics
Facebook: http://www.facebook.com/ucdspire

Schedule details

PROJECT I: SURVEY DATA

23 January: Introduction

Basic introduction to R and accessing data in R. On using comments and markdown files. Basic introduction to conversion between Excel and R and to downloading and accessing social science data sets. Introduction to graphs and tables to inspect the data.

required	James et al. (2013, ch 1–2)
	Chang (2013, ch 2)
recommended	Bradley (2015)
	Zumel and Mount (2014, ch 2)
	Grolemund and Wickham (2016, ch 1–2, 11, 22, 26–27)
further	Peng (2016, ch 4–6)

30 January: Distributions and descriptive statistics

Basic statistical descriptions, such as mean, median, variance, covariance. How to think about and visualise distributions.

required	Chang (2013, ch 6)
	Grolemund and Wickham (2016, ch 7)
recommended	Zumel and Mount (2014, ch 3)
	Grolemund and Wickham (2016, ch 3)
	"Misleading axes on graphs"
	http://callingbullshit.org/tools/tools_misleading_axes.html
further	Wickham (2010)

6 February: Comparing through visualisation

Using graphs to visualise relationships between variables. More advanced use of the ggplot2 library.

required	Grolemund and Wickham (2016, ch 28)
	Chang (2013, ch 10)
recommended	`http://www.cookbook-r.com/Graphs/`
further	Chang (2013)
	Wickham (2015)

PROJECT II: COUNTRY-LEVEL DATA

13 February: Linear regression and data wrangling

Introduction to multiple linear regression for cross-sectional data. How to execute a regression, how to present the results, and how to interpret the findings. How to interpret coefficients for binary independent variables. Also a discussion on accessing unprepared raw data. Tools for manipulation and transformation of data. Dealing with missing data categories.

required	Bartholomew et al. (2008, ch 6)
	Chang (2013, ch 5)
	James et al. (2013, ch 3, 6–7)
recommended	Zumel and Mount (2014, 4, §7.1)
further	Lantz (2013, ch 6)
	Grolemund and Wickham (2016, ch 4–5, 9–12, 14–16, 23–25)
	Peng (2016)

20 February: Logistic regression

Using multiple regression when the dependent variable is binary.

required	James et al. (2013, ch 4)
	Bartholomew et al. (2008, ch 4)
	`http://www.ats.ucla.edu/stat/r/dae/logit.htm`
recommended	Zumel and Mount (2014, §7.2)

5 March: Trees and forests

Using multiple regression when the dependent variable contains multiple categories, ordered or not. Using tree-based machine learning algorithms.

required	James et al. (2013, ch 8–9)
recommended	Lantz (2013, 5)
	Zumel and Mount (2014, ch 5–6)
further	Zumel and Mount (2014, ch 8–9)

26 March: Geography and networks

How to use geocoded data in R, including visualising maps and calculating distance matrices. How space can be interpreted as a network. A very crude and quick overview of different network and spatial regression methods.

required	
recommended	
further	Anselin (2002)

PROJECT III: TEXT DATA

2 April: Cluster analysis

Using data to group observations in clusters based on similarity. Introduction to using text as data.

required	Bartholomew et al. (2008, ch 2)
	James et al. (2013, §10.3)

9 April: Principal components and multidimensional scaling

A few basic models to understand underlying dimensional structure of data, including factor analysis, principal component analysis and MDS. How to execute and present results?

required	Bartholomew et al. (2008, ch 3, 5, 7)
	James et al. (2013, ch 10)
further	Shashua (2008, ch 5–6)

16 April: Wordscores

Introduction to data cleaning for statistical text analysis. Performing dimensional analysis of text using Wordscores.

required	Grimmer and Stewart (2013)
	Laver, Benoit and Garry (2003)
recommended	Baturo and Mikhaylov (2014)
further	Slapin and Proksch (2008)

23 April: Topic models

Using topic models to interpret large quantities of text data.

required | Greene and Cross (2016)

References

Altman, Micah, John Fox, Simon Jackman and Achim Zeileis. 2011. "An introduction to the special volume on "political methodology"." *Journal of Statistical Software* 42.

Anselin, Luc. 2002. "Under the hood. Issues in the specification and interpretation of spatial regression models." *Agricultural Economics* 27:247–267.

Bartholomew, David J., Fiona Steele, Irini Moustaki and Jane I. Galbraith. 2008. *Analysis of multivariate social science data*. Boca Raton: CRC Press.

Baturo, Alexander and Slava Mikhaylov. 2014. "Reading the Tea Leaves: Medvedev's Presidency through Political Rhetoric of Federal and Sub-National Actors." *Europe-Asia Studies* 66(6):969–992.

Bradley, Theresa. 2015. "The Unplumbed Depths of Government Data." *Motherboard* . 4 February 2015, accessed 30 January 2017.
URL: *http://motherboard.vice.com/en_uk/read/the-unplumbed-depths-of-government-data*

Chang, Winston. 2013. *R Graphics Cookbook*. O'Reilly.
URL: *http://ase.tufts.edu/bugs/guide/assets/R%20Graphics%20Cookbook.pdf*

Greene, Derek and James P. Cross. 2016. "Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach." *CoRR* abs/1607.03055.
URL: *http://arxiv.org/abs/1607.03055*

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: The promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* 21(3):267–297.

Grolemund, Garrett and Hadley Wickham. 2016. *R for Data Science*. O'Reilly.
URL: *http://r4ds.had.co.nz/*

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With applications in R*. Springer.

Lantz, Brett. 2013. *Machine Learning with R*. Birmingham: Packt Publishing.
URL: *https://archive.org/details/LantzMachineLearningWithR2013*

Laver, Michael, Kenneth Benoit and John Garry. 2003. "Extracting policy positions from political texts using words as data." *American Political Science Review* 97(2):311–331.

Peng, Roger D. 2016. *R Programming for Data Science*. Leanpub.
URL: *https://leanpub.com/rprogramming*

Shashua, Amnon. 2008. *Introduction to Machine Learning*.
URL: *https://arxiv.org/pdf/0904.3664.pdf*

Slapin, Jonathan B. and Sven-Oliver Proksch. 2008. "A scaling model for estimating time-series party positions from texts." *American Journal of Political Science* 52(3):705–722.

Wickham, Hadley. 2010. "A Layered Grammar of Graphics." *Journal of Computational and Graphical Statistics* 19(1):3–28.
URL: *http://vita.had.co.nz/papers/layered-grammar.pdf*

Wickham, Hadley. 2015. *ggplot2: Elegant Graphics for Data Analysis*. Springer.
URL: *http://ms.mcmaster.ca/∼bolker/misc/ggplot2-book.pdf*

Zumel, Nina and John Mount. 2014. *Practical Data Science with R*. Manning.