POL40950
Introduction to Statistics

Johan A. Elkink

School of Politics and International Relations
University College Dublin

jos.elkink@ucd.ie
Newman Building, Rm F304
http://www.joselkink.net/teaching

Autumn 2019

Introduction

This course is an introduction to quantitative data analysis in the social sciences, in particular political science, public policy, and sociology. Do you want to know whether more informed voters are more likely to have liberal values? Whether democracies are less likely to initiate a war? Whether high tax rates leads to higher levels of corruption? For many social science questions what we are really after is establishing whether there is a relationship between two variables of interest. And, we typically want to verify that there is no third variable explaining this relationship.

In statistical analysis, the first tool one usually reverts to to answer such questions is regression analysis. There are many other statistical tools available to the social scientist, but regression analysis is by far the most common and a thorough understanding of this method generally the key to being able to read or write quantitative social science papers and research reports. The course is therefore designed around introducing you to regression analysis—including model specification (which variables to include in a model?) and statistical inference (how do I know whether my findings hold for cases beyond my sample?).

Statistical concepts, such as plots, descriptive statistics, statistical inference, etcetera, will be introduced primarily in the context of regression analysis. This way, the course will form the background for basic statistical analysis in the social sciences, and a solid preparation for a more advanced course in advanced regression analysis or introductory econometrics.

Inter alia, we cover descriptive and inferential statistics, plotting numerical information, evaluating the distribution of variables, and inspecting correlations between variables. In statistics it is common to draw a random sample of individuals from a population to gather data on those individuals in order to derive insights about the population as a whole. The statistical inference component of the course will discuss the logic behind such inferences and some of the statistical tools available to do so, again primarily in the context of regression analysis.

The core textbook for the course is Ismay and Kim (2019), which is freely available online at `https://moderndive.com`. This book takes a modern, data science approach to regression analysis. The differences between data science and more typical quantitative social science will be discussed in class, in particular in the context of model specification. For basic statistical concepts such as variables, measures of central tendency, or basic linear regression, Blalock (1979) is a very good reference, while Moore, McCabe and Craig (2012) provides a more solid and mathematically coherent textbook for standard statistical analysis. Field (2009) and Kellstedt and Whitten (2009) are two further textbooks that might be helpful to refer to at times. In addition, there will be handouts with a description of the software commands you will need to learn to be able to do the homeworks and apply the theoretical material in practice.

You will learn how to use the statistical software package R, which is freely available at http://www.r-project.org (see also Verzani, 2014, Appendix A). You should download and install this at home, so you can get as much hands-on practice as possible. The labs take place in a regular classroom, so if you have a laptop, please do bring it along. It is also highly recommended that you install RStudio: `http://www.rstudio.com/`. **Make sure to install R and RStudio prior to class**, folowing the instructions in Ismay and Kim (2019, ch 2).

Classes

Classes take place once a week, with a lecture on Monday 1–2 PM in Q014 in the building of the Quinn School of Business and lab sessions Monday 2–3 PM in G5 of the Daedalus building, both at the Belfield campus of UCD.

Contact

I do not have fixed office hours, so if you want to make sure I am present, 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`).
Course materials will be uploaded to `http://www.joselkink.net/teaching`.

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/`

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

Homeworks

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 analysis—just the theory will never be sufficient. For that reason, there will be four practical homework assignments. The assignments will be available online. The standard penalty for late submission as outlined in the SPIRe Masters Handbook will apply, whereby it should be taken into account that a late submission might result in a delayed return of feedback to the entire class.

There will be four homework assignments during this module, whereby the third assignment is a preparation for the fourth. The first assignment, worth 25% of the grade, will be due **Monday 30/9, 10 am**, and the second, also worth 25% of the grade, **Monday 21/10, 10 am**. The third assignment will involve the regression analysis for the final assignment, worth 20% of the grade, and will be due **Wednesday 13/11, 5 pm**. The final assignment, worth 30% and building on the third assignment, will be due **Monday 9/12, 5 pm**. Instructions for each assignment will be uploaded to the course website in due course.

Assignments should be submitted electronically to `jos.elkink@ucd.ie`. The submission should contain either a PDF with the written-up answers and a separate file with all the commands used,[1] or a PDF with the two integrated. The latter, using Markdown, is the recommended approach.[2] Note that Microsoft Word files are *not* accepted and submitting in the wrong format first can cause your submission to be late. Lab sessions will be provided in R, so the default software package should be R Studio, but you can submit using alternative software, as long as all commands are provided.

Whichever software you use, command files need to be properly laid out, which includes:

- Include all commands that lead to the results used in the assignment, including opening files.
- Using only those commands that actually lead to the results used in the assignment, with no superfluous or additional commands.
- Using enough whitespace and indention to keep the code easy to follow.
- Inserting sufficient comments to clarify the commands. When the commands are integrated into the written submission, comments are only necessary where the purpose is not evident from the main text.

[1] E.g. `R` file in R, `SPS` file in SPSS, or `DO` file in Stata.

[2] Note that compiling ("knitting") Markdown to PDF is complicated, but it is straightforward to compile to HTML, then open in a web browser, and then save as (or print to) a PDF file—or compile to Word, open in Word, and then produce the PDF. Make sure you practice this in the lab session.

The grading will take into consideration the presentation of the results. Tables copy/pasted from computer output with some interpretation squeezed in will get lower grades than assignments that are properly presented and formatted, more akin to how you submit an essay . When a substantive interpretation is asked for, this means an interpretation in terms of the substantive content of the topic at hand: e.g. if the data is on attitudes and turnout in elections, the substantive interpretation is "what do these results tell us about voting behaviour?", not "is this relationship positive or negative, significant or not?". The translation back from the statistical results to the political science interpretation is crucial to a good grade in this course.

Course paper

The third and fourth homework will form the course paper of the module, whereby the third homework implements the statistical analysis that will be presented in detail in the fourth homework. A short list of potential topics will be provided in due course.

For the course paper you will be required to produce your own regression analysis and present the results and interpretation. This paper needs to be written as a proper course paper, with abstract, introduction, basic theoretical background and literature, description of methodology, data and model specification, and a discussion of the results. Have a look at published articles to get a good idea of the required structure, e.g. Ross (2004) is a good example of a layout for this assignment.

Plagiarism

Although this should be obvious, plagiarism—copying someone else's text without acknowledgement or beyond "fair use" quantities—is not allowed. Please carefully check the UCD policies concerning plagiarism[3] and its more extensive description of what is plagiarism and what is not[4].

9 September: Accessing & visualising data

What is quantitative political science? What are data? What is a variable? What are the different levels of measurement? How to describe your variables graphically, including pie charts, histograms. How to look at a distribution.

[3] http://www.ucd.ie/regist/documents/plagiarism_policy_and_procedures.pdf.

[4] http://www.ucd.ie/library/students/information_skills/plagiari.html

required	Ismay and Kim (2019, ch 1–3)
	Moore, McCabe and Craig (2012, ch 1)
recommended	Blalock (1979, ch 1-6)
optional	Pollock (2005, ch 1)
	Kellstedt and Whitten (2009, ch 5-6)
	Healey (2011, ch 1-4)
	Privitera (2011, ch 1-4)
	Argyrous (1997, ch 1-5)
	Fielding and Gilbert (2000, ch 1)
	Field (2009, §4.1-4.5)
	Diamond and Jefferies (2001, ch 2-5)
	Wright and London (2009, ch 1-3)
	Heiman (2001, ch 6-8)
R	Pollock (2014, ch 1, 3, 11)
	Field, Miles and Field (2012, ch 3)
	Verzani (2005, ch 1)
	Dalgaard (2002, ch 1)
	Maindonald and Braun (2007, ch 1)

16 September: Descriptive statistics

How to describe your variables numerically, including the mean, mode, median, variance, and standard deviation. How to describe relations between variables graphically, including bar charts, scatter plots, box plots. Discussion of covariance and correlation to look at numerical indicators of relationships.

required	Ismay and Kim (2019, ch 3–5)
	Moore, McCabe and Craig (2012, ch 1-2)
recommended	Blalock (1979, ch 3-6, 16-17)
optional	Pollock (2005, ch 3-4)
	Field (2009, §4.1-4.5)
	Privitera (2011, §2-4, ch 15)
	Diamond and Jefferies (2001, ch 2-5, 13)
	Healey (2011, ch 2-4, 14)
	Wright and London (2009, ch 1-3)
	Argyrous (1997, ch 2-4, 22)
	Kellstedt and Whitten (2009, ch 6)
	Heiman (2001, ch 6-8, 10)
R	Pollock (2014, ch 2, 4)
	Verzani (2005, §2.1-2.3, §3.1-3.3, ch 4)
	Dalgaard (2002, ch 3)
	Maindonald and Braun (2007, ch 2)
	Field, Miles and Field (2012, ch 4)

23 September: Simple regression

Descriptive univariate linear regression models—how to look at the relation between two continuous variables.

required	Ismay and Kim (2019, ch 6)
	Moore, McCabe and Craig (2012, §2.3-2.4)
recommended	Blalock (1979, ch 17)
optional	Pollock (2005, ch 7)
	Privitera (2011, §16.1-16.5)
	Diamond and Jefferies (2001, ch 13)
	Healey (2011, ch 14)
	Heiman (2001, ch 11)
	Argyrous (1997, ch 22)
	Miles and Shevlin (2001, ch 1)
	Wright and London (2009, ch 8)
R	Pollock (2014, ch 8)
	Field, Miles and Field (2012, ch 6-7)
	Verzani (2005, §3.4)
	Dalgaard (2002, ch 5)
	Maindonald and Braun (2007, ch 5)
further	Starnes, Yates and Moore (2010, §2.6)

30 September: Multiple regression

How to perform and interpret regression models with more than one independent variable. How to think about the difference between prediction and causal inference? Some discussion of model specification.

required	Ismay and Kim (2019, ch 6)
	Moore, McCabe and Craig (2012, ch 10-11)
recommended	Lewis-Beck (1980)
	Kellstedt and Whitten (2009, ch 9-10)
	Blalock (1979, ch 16, 18-19)
optional	Privitera (2011, ch 16)
	Argyrous (1997, ch 22)
	Hosker (2008, ch 10-11)
	Field (2009, ch 7)
	Healey (2011, ch 16)
	Miles and Shevlin (2001, ch 2)
R	Pollock (2014, ch 5, 8)
	Field, Miles and Field (2012, ch 7)
	Verzani (2005, ch 8, §10.3, ch 11)
	Dalgaard (2002, ch 9)
	Maindonald and Braun (2007, ch 6)
further	Moore, McCabe and Craig (2012, ch 9)
	Miles and Shevlin (2001, ch 4-5)
	Allison (1999)

7 October: Multiple regression—categorical independent variables

Categorical independent variables in multiple regression.

required	Kellstedt and Whitten (2009, ch 11-12)
recommended	Blalock (1979, ch 20)
	Hardy (1993)
alternatives	Miles and Shevlin (2001, ch 3)
R	Pollock (2014, ch 9)

14 October: Writing up regression results

How to present and interpret regression results. How to structure a quantitative research paper. How to convince the reader of the robustness of your results.

required	King (2006)

21 October: Multiple regression—interaction effects

Modeling interaction effects in multiple regression.

required	Kellstedt and Whitten (2009, ch 11-12)
recommended	Blalock (1979, ch 20)
	Hardy (1993)
alternatives	Miles and Shevlin (2001, ch 3)
R	Pollock (2014, ch 9)

4 November: Sampling distribution & Central Limit Theorem

What are probabilities and probability distributions? Introduction to the normal distribution. What is statistical inference? Introduction to sampling methods. What is the Central Limit Theorem?

required	Ismay and Kim (2019, ch 8)
	Moore, McCabe and Craig (2012, §1.3, §3.2-3.3, §5.1)
recommended	Kellstedt and Whitten (2009, ch 7)
	Blalock (1979, ch 7, 9)
	Hosker (2008, ch 3)
optional	Pollock (2005, ch 5)
	Hosker (2008, ch 4)
	Privitera (2011, ch 5-7)
	Diamond and Jefferies (2001, ch 7-8)
	Healey (2011, ch 5-6)
	Heiman (2001, ch 9, 12)
	Argyrous (1997, ch 6-7)
	Fielding and Gilbert (2000, ch 7, 10)
	Wright and London (2009, ch 4)
further	Blalock (1979, ch 21)
	Moore, McCabe and Craig (2012, ch 3-4)
R	Dalgaard (2002, ch 2)
	Verzani (2005, ch 5)
	Maindonald and Braun (2007, ch 3-4)

11 November: Hypothesis tests & confidence intervals

What are hypothesis tests and confidence intervals? How to think of statistical inference in multiple regression analysis.

required	Ismay and Kim (2019, ch 9–10)
	Moore, McCabe and Craig (2012, ch 6)
	Kellstedt and Whitten (2009, ch 8)
recommended	Blalock (1979, ch 8, 12)
	Senn (2012)
optional	Pollock (2005, ch 6)
	Hosker (2008, ch 5, 9-)
	Privitera (2011, ch 8, 11)
	Diamond and Jefferies (2001, ch 9-11)
	Healey (2011, ch 7)
	Heiman (2001, ch 13-14)
	Argyrous (1997, ch 8-9)
	Fielding and Gilbert (2000, ch 11)
	Wright and London (2009, ch 5)
SPSS	Norris et al. (2012, ch 10-12)
R	Pollock (2014, ch 6-7)
	Verzani (2005, ch 7)

18 November: Multiple regression—diagnostics & model fit

How to think about model fit in the contexts of prediction and causal inference. Statistical versus modelling considerations in model specification. Common problems in regression analysis (and hints at solutions).

required	Ismay and Kim (2019, ch 11–12)

25 November: Multiple regression—categorical dependent variables

Regression analysis when the dependent variable is binary—e.g. explaining turnout in elections. Introduction to logistic regression.

required	Moore, McCabe and Craig (2012, ch 14)
recommended	Pampel (2000)
optional	Pollock (2005, ch 8)
	Menard (2002)
R	Pollock (2014, ch 10)
	Field, Miles and Field (2012, ch 8)
	Dalgaard (2002, ch 11)
	Maindonald and Braun (2007, ch 8)

References

Allison, Paul D. 1999. *Multiple regression: a primer*. Thousand Oaks, CA: Pine Forge Press.

Argyrous, George. 1997. *Statistics for social research*. Basingstoke: MacMillan.

Blalock, Hubert M. 1979. *Social statistics*. 2nd ed. Tokyo: McGraw-Hill Kogakusha.

Dalgaard, Peter. 2002. *Introductory statistics with R*. New York, NY: Springer.

Diamond, Ian and Julie Jefferies. 2001. *Beginning statistics: an introduction for social scientists*. London: Sage.

Field, Andy. 2009. *Discovering statistics using SPSS*. 3rd ed. London: Sage.

Field, Andy, Jeremy Miles and Zoë Field. 2012. *Discovering statistics using R*. London: Sage.

Fielding, Jane and Nigel Gilbert. 2000. *Understanding social statistics*. London: Sage Publications.

Hardy, Melissa A. 1993. *Regression with dummy variables*. London: Sage.

Healey, Joseph F. 2011. *Statistics: a tool for social research*. 9th ed. Belmont CA: Wadsworth.

Heiman, Gary W. 2001. *Understanding research methods and statistics: an integrated introduction for psychology*. 2nd ed. Boston: Houghton Mifflin.

Hosker, Ian. 2008. *Starting statistics: Data handling for beginners*. Abergele: Studymates.

Ismay, Chester and Albert Y. Kim. 2019. "Statistical Inference via Data Science. A moderndive into R and the tidyverse.".
URL: *https://moderndive.com/*

Kellstedt, Paul M. and Guy D. Whitten. 2009. *The fundamentals of political science research*. Cambridge: Cambridge University Press.

King, Gary. 2006. "Publication, Publication." *PS: Political Science and Politics* 39(1):119–125.
URL: *https://gking.harvard.edu/files/gking/files/paperspub.pdf*

Lewis-Beck, Michael S. 1980. *Applied regression: an introduction*. Sage Publications.

Maindonald, John and John Braun. 2007. *Data analysis and graphics using R. An example-based approach*. 2nd ed. Cambridge: Cambridge University Press.

Menard, Scott. 2002. *Applied logistic regression analysis*. 2nd ed. London: Sage.

Miles, Jeremy and Mark Shevlin. 2001. *Applying regression & correlation: a guide for students and researchers*. London: Sage Publications.

Moore, David S., George P. McCabe and Bruce A. Craig. 2012. *Introduction to the practice of statistics*. 7th international edition ed. New York: W.H. Freeman.

Norris, Gareth, Faiza Qureshi, Dennis Howitt and Duncan Cramer. 2012. *Introduction to statistics with SPSS for social science*. Harlow: Pearson.

Pampel, Fred C. 2000. *Logistic regression: A primer*. London: Sage.

Pollock, Philip H. 2005. *The essentials of political analysis*. 2nd ed. Washington, DC: CQ Press.

Pollock, Philip H. 2014. *An R companion to political analysis*. Thousand Oaks, CA: CQ Press.

Privitera, Gregory J. 2011. *Statistics for the Behavioral Sciences*. Sage.

Ross, Michael L. 2004. "Does taxation lead to representation?" *British Journal of Political Science* 34:229–249.

Senn, Stephen. 2012. "Tea for three. Of infusions and inferences and milk in first." *Significance* (12):30–33.

Starnes, Daren S., Dan Yates and David S. Moore. 2010. *The practice of statistics*. 4th ed. New York: W.H. Freeman.

Verzani, John. 2005. *Using R for introductory statistics*. Boca Raton, FL: Chapman & Hall/CRC.

Verzani, John. 2014. *Using R for introductory statistics*. Boca Raton, FL: Chapman & Hall/CRC.

Wright, Daniel B. and Kamala London. 2009. *First (and second) steps in statistics*. 2nd ed. Los Angeles: Sage Publications.