Stefan Müller, PhD
Assistant Professor and Ad Astra Fellow
School of Politics and International Relations
University College Dublin
Belfield, Dublin 4, Ireland
✉ stefan.mueller@ucd.ie
https://muellerstefan.net

Autumn Trimester 2020

Introduction to Statistics (POL40950)

Last update: September 17, 2020

Note: Draft – subject to change

Latest version: https://muellerstefan.net/teaching/2020-autumn-introstats.pdf

---

Term: Autumn Trimester 2020
Time: Lecture (Mon, 13:00); Lab (Mon, 14:00)
Location: Lecture: QUI 006; Lab: F20 Newstead
ECTS: 10.0
Format: Lectures; lab work; blended learning

Convener: Stefan Müller
stefan.mueller@ucd.ie
https://muellerstefan.net
Office: Newman Building, G303
Virtual office hours: Tuesday, 10:00–12:00

Introduction

Welcome to Introduction to Statistics! In this course you will learn about concepts such as measurement, variables, statistical data and get equipped to answer a social science research question using linear statistical models and the statistical programming language R.

Do you want to know whether more informed voters are more likely to have liberal values? Are political parties responsive to the issue priorities by voters? Are democracies less likely to initiate a war? Do high tax rates lead to higher levels of corruption? Can voters accurately predict the government that will be formed after an election? Answering such questions usually requires the analysis of data – information about people, parties, communication, firms, or nations.

There are many other statistical tools available to the social scientist, but regression analysis is by far the most common. A thorough understanding of this method is required to read or write quantitative social science papers and research reports. The course therefore will mainly focus on 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?).

By the end of this module, you will have gained a basic understanding of statistics and the so called frequentist approach of hypothesis testing. The lab sessions and two homework assignments will make you familiar with the R statistical programming language and prepare you two write an original quantitative research paper.

The core textbook for the course is Ismay and Kim (2020), which is freely available online at https://moderndive.com.[1] 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.[2]

For the applied parts of this course, such as data import, data wrangling, and data visualisation, we will read parts of the following textbooks. Both books will also help you with your homework assignments and the technical elements of your course paper.

- Hadley Wickham and Garrett Grolemund (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol: O'Reilly. URL: https://r4ds.had.co.nz.
- Kieran Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press. URL: https://r4ds.had.co.nz.

We will work extensively with the R statistical programming language. The three books mentioned above (Ismay and Kim 2020; Wickham and Grolemund 2017; Healy 2019) provide detailed and intuitive examples and the corresponding R code (based on the `tidyverse` approach). In addition to these books, I recommend the following literature for introductions to statistical methods, regression, causal inference, and R:

- **Basic grasp of statistics**: David Spiegelhalter (2020). *The Art of Statistics: Learning from Data*. London/New York: Penguin Books.
- **Research design**:
  - Paul M. Kellstedt and Guy D. Whitten (2018). *The Fundamentals of Political Science Research*. 3rd edition. Cambridge: Cambridge University Press.
  - Gary King, Robert O. Keohane, and Sidney Verba (1994). *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.
- **R and regression analysis**:
  - Erik Gahner Larsen and Zoltán Fazekas (2019). *Quantitative Politics with R*.
  - Jennifer Bryan (2019). *Data Wrangling, Exploration, and Analysis with R*. URL: https://stat545.com.
  - Andrew Gelman, Jennifer Hill, and Aki Vehtari (2020). *Regression and Other Stories*. Cambridge: Cambridge University Press.
- **Data visualisation**: Claus O. Wilke (2019). *Fundamentals of Data Visualization: A Primer On Making Informative and Compelling Figures*. Sebastopol: O'Reilly.
- **Causal inference**:
  - Scott Cunningham (2020). *Causal Inference: The Mixtape (v. 1.8)*. Yale University Press (under contract).
- **RMarkdown**: Hadley Wickham and Garrett Grolemund (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol: O'Reilly: ch. 27, 29.

Learning Outcomes

1. basic understanding of working with R and RStudio

[1] Chester Ismay and Albert Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press. URL: https://moderndive.com.

[2] The structure of this module is similar to Introduction to Statistics taught at the School of Politics and International Relations in previous years. I thank Jos Elkink for allowing me to follow and adjust the structure of the previously taught module.

2. being able to wrangle, summarise, describe, and visualise statistical data
3. basic understanding of (frequentist) statistical inference
4. basic understanding of executing and interpreting multiple regression
5. preliminary understanding of logistic regression

Approaches to Teaching and Learning

The sessions consist of lectures and labs each week. Some of the lectures and labs will be online. The lectures focus on the fundamental aspects of statistical inference as well as the interpretation of these methods and examples.

In the lab sessions, students will be provided with clear instructions and solve problems related to data wrangling, visualisation and statistical methods. The homework assignments are structured so that they gradually lead up to a comprehensive regression analysis and associated social science paper, putting the technical material of the class in practice.

I will make extensive use of quizzes throughout the lectures to increase engagement in times of blended learning. In addition, I will distribute several short feedback surveys during the term in which you can indicate what you can provide feedback, ask questions, and make improvement suggestions.

Overview of Assessment

- Homework assignment (Week 3): 25%
- Homework assignment (Week 6): 25%
- Course paper (end of trimester): 50%

Expectations and Grading

Students submit two **homework assignments** during the term (after the end of Week 3 and Week 6). Each homework counts towards 25% of the final grade. The homeworks will be distributed via Brightspace 14 days before the submission deadline as an RMarkdown file.[3] Students fill in the answers and solutions in the same RMarkdown file, rename it to `hw_01/02_surname_firstname.Rmd`, knit it as an `html` file, and submit it via Brightspace. Only knitted `html` files will be accepted. More details on the homeworks will be provided in the first session(s) of the course.

Students also submit a **course paper** which counts towards 50% of the final grade. The research paper is a written analysis consisting of 4,000 words (including bibliography, captions, and footnotes). Students are required to answer a research question using quantitative methods and regression analysis. Students are free to answer questions from all fields of social science, but must justify their choice and the relevance of the question. The course paper must address the following aspects: research gap and relevance; theory and expectations (based on previous research); data and methodological approach; results; conclusion and outlook. The course paper must be submitted via Brightspace as a `pdf` document before **TBC**. Detailed instructions on the research paper, the presentation, and the in-class discussion will be provided in class and on Brightspace.

For information on academic writing, I recommend the following two sources:

- Stephen B. Heard (2016). *The Scientist's Guide to Writing: How to Write More Easily and Effectively Throughout Your Scientific Career*. Princeton: Princeton University Press.

[3] For a very short primer to RMarkdown see: https://rmarkdown.rstudio.com/articles_intro.html. We will discuss how to create and compile RMarkdown files in the first two weeks of the module.

- Patrick Dunleavy (2014). *How to Write Paragraphs in Research Texts (Articles, Books and PhDs)*. URL: https://medium.com/advice-and-help-in-authoring-a-phd-or-non-fiction/how-to-write-paragraphs-80781e2f3054.

If you require information on proper citation style, please refer to the guidelines of the American Political Science Association:

- APSA Committee on Publications (2018). *Style Manual for Political Science (Revised 2018 Version)*. URL: https://connect.apsanet.org/stylemanual/.

Student effort hours

Student effort type	Hours
Lectures	12
Computer Aided Lab	12
Autonomous Student Learning	200
Total	**224**

Plagarism

Although this should be obvious, plagiarism – copying someone else's text without acknowledgement or beyond 'fair use' quantities – is not allowed. Plagiarism is an issue we take very serious here in UCD. Please familiarize yourself with the definition of plagiarism on UCD's website[4] and make sure not to engage in it.

Late Submission Policy

All written work must be submitted on or before the due dates. Students will lose one point of a grade for work up to 5 working days late ($B-$ becomes $C+$). Students will lose two grade points for work between 5 and 10 working days late ($B-$ becomes C). When more than two weeks are necessary, the student will need to apply for extenuating circumstances application via the SPIRe Programme Office.

Questions and Problems

In this module, we will discuss concepts, methods, and software you might not have heard of before. I am aware that parts of this module could be challenging and I will assist you as best as I can. In addition to the lectures and lab sessions, I offer weekly office hours only for participants of this module. The office hours will take place via Zoom. I will share the link and password to the virtual room in the first lecture and post it on Brightspace.

If you struggle to solve problems relating to R or RStudio, please follow the steps outlined below before contacting me. It is very likely that at least one other person faced the same problem before or received the same error message.

1. Use the 'Search' function in the online books of the recommended textbooks (Ismay and Kim 2020; Wickham and Grolemund 2017; Healy 2019) and look up keywords that relate to your problem or the function that causes a problem. For questions about concepts, I recommend to consult the Glossary of Statistical Terms.

[4] https://libguides.ucd.ie/academicintegrity.

2. Try to summarise the problem in your own words and then google this summary. If the problem relates to R, add `rstats` to your search query. For example: `how to import csv file in rstats`. I am almost certain that you find a solution to most of your questions.

3. If your R code returns an error, I would advise you to Google the text the error message. For example: `Error: Can't subset columns that don't exist.`

⟶ If steps 1–3 still do not solve your problem or question, please get in touch with me. I am happy to help!

I will continuously update an FAQ page on Brightspace with questions that students have asked me and that might be relevant for everyone in the course.

Syllabus Modification Rights

I reserve the right to reasonably alter the elements of the syllabus at any time by adjusting the reading list to keep pace with the course schedule. Moreover, I may change the content of specific sessions, depending on the participants' prior knowledge and research interests. If I make adjustments, I will send an email to all seminar participants and upload the revised syllabus to Brightspace.

Course Structure

Week 1: Accessing and Visualising Data (21–25 September 2020)

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?

Important: Please install R[5] and RStudio[6] before the start of the first lecture. Ismay and Kim (2020: ch. 1) provide detailed instructions on how to install the required software.

Mandatory Readings

- Chester Ismay and Albert Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press: ch. 1–2.
- Jennifer Bryan (2019). *Data Wrangling, Exploration, and Analysis with R*. URL: https://stat545.com: ch. 2.

Optional

- David Spiegelhalter (2020). *The Art of Statistics: Learning from Data*. London/New York: Penguin Books: ch. 2.
- Erik Gahner Larsen and Zoltán Fazekas (2019). *Quantitative Politics with R*: ch. 7, 10.
- Jennifer Bryan (2018). "Excuse Me, Do You Have a Moment to Talk About Version Control?". *The American Statistician* 72 (1): 20–27.

Week 2: Descriptive Statistics (28 September – 2 October 2020)

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, scatterplots, and boxplots. Discussion of covariance and correlation.

Mandatory Readings

- Kieran Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press: ch. 1.
- Chester Ismay and Albert Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press: ch. 3–4.
- Erik Gahner Larsen and Zoltán Fazekas (2019). *Quantitative Politics with R*: ch. 3.

Optional

- Greg Wilson, Jennifer Bryan, Karen Cranston, Justin Kitzes, Lex Nederbragt, and Tracy K. Teal (2017). "Good Enough Practices in Scientific Computing". *PLoS Computational Biology* 13 (6): e1005510.

Week 3: Simple Regression (5–9 October 2020)

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

[5] Download R for Mac, Windows, or Linux at: https://cloud.r-project.org
[6] Download RStudio (Desktop Open Source License [free]) at: https://rstudio.com/products/rstudio/download

Mandatory Readings

- Chester Ismay and Albert Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press: ch. 5.

Optional

- Paul M. Kellstedt and Guy D. Whitten (2018). *The Fundamentals of Political Science Research*. 3rd edition. Cambridge: Cambridge University Press: ch. 9.

Week 4: Multiple Regression (12–16 October 2020)

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.

- Chester Ismay and Albert Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press: ch. 6.

Optional

- Paul M. Kellstedt and Guy D. Whitten (2018). *The Fundamentals of Political Science Research*. 3rd edition. Cambridge: Cambridge University Press: ch. 10.
- Erik Gahner Larsen and Zoltán Fazekas (2019). *Quantitative Politics with R*: ch. 11.
- Melissa A. Hardy (1993). *Regression with Dummy Variables*. Newbury Park, CA: SAGE Publications.

Week 5: Multiple Regression – Categorical Independent Variables and Interactions (19–22 October 2020)

Categorical independent variables in multiple regression. Modeling interaction effects in multiple regression.

Mandatory Readings

- John Fox (2015). *Applied Regresssion Analysis and Generalized Linear Models*. 3rd edition. Los Angeles: SAGE: ch. 7.
- Chester Ismay and Albert Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press: ch. 6.1.2.

Optional

- Jens Hainmueller, Jonathan Mummolo, and Yiqing Xu (2019). "How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice". *Political Analysis* 27 (2): 163–192.
- William Roberts Clark, Michael J. Gilligan, and Matt Golder (2006). "A Simple Multivariate Test for Asymmetric Hypotheses". *Political Analysis* 14 (3): 311–331.

Reading Week (26–30 October 2020)

Week 6: Sampling Distributions and Central Limit Theorem (2–6 November 2020)

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?

Mandatory Readings

- Chester Ismay and Albert Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press: ch. 8.
- Paul M. Kellstedt and Guy D. Whitten (2018). *The Fundamentals of Political Science Research*. 3rd edition. Cambridge: Cambridge University Press: ch. 7.

Week 7: Hypothesis Tests and Confidence Intervals (9–13 November 2020)

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

- Chester Ismay and Albert Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press: ch. 9–10.

Optional

- David Spiegelhalter (2020). *The Art of Statistics: Learning from Data*. London/New York: Penguin Books: ch. 7.

Week 8: Reporting (Regression) Results (16–20 November 2020)

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.

Mandatory Readings

- Gary King (2006). "Publication, Publication". *PS: Political Science & Politics* 39 (1): 119–125.
- Chester Ismay and Albert Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press: ch. 11.
- Glenn Firebaugh (2008). *Seven Rules for Social Research*. Princeton: Princeton University Press: ch. 1.

Optional

- David Spiegelhalter (2020). *The Art of Statistics: Learning from Data*. London/New York: Penguin Books: ch. 12.
- Erik Gahner Larsen and Zoltán Fazekas (2019). *Quantitative Politics with R*: ch. 13–14.

Week 9: Multiple Regression – Diagnostics and Model Fit (23–27 November 2020)

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).

- Paul M. Kellstedt and Guy D. Whitten (2018). *The Fundamentals of Political Science Research*. 3rd edition. Cambridge: Cambridge University Press: ch. 11.

Optional

- Erik Gahner Larsen and Zoltán Fazekas (2019). *Quantitative Politics with R*: ch. 11.4.

Week 10: Multiple Regression – Categorical Dependent Variables (30 November–4 December 2020)

Regression analysis when the dependent variable is binary – e.g. explaining whether or not a citizens turns out to vote on election day. Introduction to logistic regression.

- Paul M. Kellstedt and Guy D. Whitten (2018). *The Fundamentals of Political Science Research*. 3rd edition. Cambridge: Cambridge University Press: ch. 12.

Optional

- Andrew Gelman, Jennifer Hill, and Aki Vehtari (2020). *Regression and Other Stories*. Cambridge: Cambridge University Press: ch. 3.
- Jeff Gill and Michelle Torres (2019). *Generalized Linear Models: A Unified Approach*. Newbury Park, CA: SAGE Publications.

Week 11: TO BE DISCUSSED (7 December–11 December 2020)

Possible contents:

- Reproducible research
- Advanced data visualisation I

Optional

- Jeff Gill and Michelle Torres (2019). *Generalized Linear Models: A Unified Approach*. Newbury Park, CA: SAGE Publications.
- Kieran Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press: ch. 4–5.

Week 12: TO BE DISCUSSED (14 December–18 December 2020)

Possible contents:

- Advanced data visualisation II
- A primer on computational social science methods

Optional

- Kieran Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press: ch. 6, 8.
- Matthew J. Salganik (2017). *Bit by Bit: Social Research in the Digital Age*. Princeton: Princeton University Press.
- David Lazer, Alex Pentland, Duncan J. Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra González-Bailón, Gary King, Helen Margetts, Alondra Nelson, Matthew J. Salganik, Markus Strohmaier, Alessandro Vespignani, and Claudia Wagner (2020). "Computational Social Science: Obstacles and Opportunities". *Science* 369 (6507): 1060–1062.