



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 2021

Introduction to Statistics (POL40950)

Version: September 14, 2021

Latest version at: <https://muellerstefan.net/teaching/2021-autumn-introstats.pdf>

Term: Autumn Trimester 2021
Time: Lecture & Lab: Mon, 13:00–14:45
Location: Q279
Credits: 10.0
Format: Lectures; lab work; blended learning

Module Coordinator: Stefan Müller, PhD
stefan.mueller@ucd.ie
<https://muellerstefan.net>
Office: Newman Building, G312
Office hours: Tue, 13:00–15:00 ([sign up here](#))

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 to 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.

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

The differences between data science and more typical quantitative social science will be discussed in class, in particular in the context of model specification. In addition, the textbook by Llaudet and Imai (Forthcoming)² provides intuitive examples and explanations for most of the concepts discussed in this module.

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.

- H. Wickham and G. Grolemund (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol: O’Reilly. URL: <https://r4ds.had.co.nz>.
- K. Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press. URL: <https://socviz.co>.

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 and quantitative methods:**
 - D. Spiegelhalter (2020). *The Art of Statistics: Learning from Data*. London/New York: Penguin Books.
- **Research design and causal inference:**
 - N. Huntington-Klein (2021). *The Effect: An Introduction to Research Design and Causality*. Boca Raton: CRC Press. URL: <https://theeffectbook.net>.
 - S. Cunningham (2021). *Causal Inference: The Mixtape*. New Haven: Yale University Press. URL: <https://mixtape.scunning.com>.
- **R and regression analysis:**
 - E. G. Larsen and Z. Fazekas (2021). *Quantitative Politics with R*. URL: <http://qpplr.com>.
 - A. Gelman, J. Hill, and A. Vehtari (2020). *Regression and Other Stories*. Cambridge: Cambridge University Press.
- **Data visualisation:** C. O. Wilke (2019). *Fundamentals of Data Visualization: A Primer On Making Informative and Compelling Figures*. Sebastopol: O’Reilly. URL: <https://clauswilke.com/dataviz/>.
- **RMarkdown:** H. Wickham and G. Grolemund (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. Sebastopol: O’Reilly: ch. 27, 29. URL: <https://r4ds.had.co.nz>.

Learning Outcomes

1. basic understanding of working with R and RStudio
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

²E. Llaudet and K. Imai (Forthcoming). *Data Analysis for Social Science: A Friendly Introduction*. Princeton: Princeton University Press.

5. preliminary understanding of logistic regression

Approaches to Teaching and Learning

The sessions consist of lectures and labs each week. 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 distribute several short feedback surveys during the term in which you can indicate what you can provide feedback, ask questions, and make improvement suggestions. We will communicate via [Slack](#), a business communication platform, to clarify questions about the course content, homeworks, code, and examples. I will share the link to the Slack workspace on Brightspace. Students should check the Slack workspace at least three times per week. Questions should be posted in the `#questions` channel. This ensures that you can support each other and that everyone will have access to the solutions.

Overview of Assessment

- Homework assignment (Week 4): 25%
- Homework assignment (Week 7): 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. Table 1 shows the grade conversion scheme for that will be applied to the homework assignments.

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 **Friday, 10 December 2021 (8pm)**. 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:

³For 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.

Table 1: Grade conversion scheme for homeworks

Homeworks	UCD Grade
97–100%	A+
94–96%	A
91–93%	A–
88–90%	B+
85–87%	B
83–84%	B
80–82%	C+
77–79%	C
74–76%	C–
71–73%	D+
68–70%	D
65–67%	E+
54–64%	E
44–53%	E–
0–32%	F

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

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/>.

Table 2: Student effort hours

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

Plagiarism

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 website⁴ 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

⁴<https://libguides.ucd.ie/academicintegrity>.

necessary, the student will need to apply for extenuating circumstances application via the SPIRE Programme Office.

Office Hours

We can meet either in person (Room G312, Newman Building) or online during my office hour on Tuesday from 13:00–15:00. Please sign up for a meeting at <https://calendly.com/mueller-ucd/office-hours>.

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 in person or 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](#).
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 of 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!

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.

Additional Covid-19 Guidelines

Covid-19 continues to pose a threat to our well-being and health. We all need to follow UCD’s guidelines, which involves wearing masks in the lecture rooms. I will also wear a mask at all times. If you come to my office hours in person, please make sure to wear a mask. If you are unwilling or unable to wear a mask, we can meet virtually. If you are not feeling well, stay home! I try to make all relevant materials available to everyone: I live-will record all lectures, share the slides, and upload all readings. Protecting everyone’s health is most important. Should you be sick or need a longer period of absence, please get in touch and I happily work with you to ensure your success in

this module. We are in this together – let’s try our very best in the months to come and support each other.

Course Structure

Week 1: Accessing and Visualising Data (13–17 September 2021)	6
Week 2: Descriptive Statistics (20–24 September 2021)	7
Week 3: Simple Regression (27 September–1 October 2021)	7
Week 4: Multiple Regression (4–8 October 2021)	7
Week 5: Multiple Regression – Categorical Independent Variables and Interactions (11–15 October 2021)	8
Week 6: Sampling Distributions and Central Limit Theorem (18–22 October 2021)	8
Week 7: Bank Holiday Monday: Virtual Q&A Session (25–29 October 2021)	8
Week 8: Hypothesis Tests and Confidence Intervals (1–5 November 2021)	9
Week 9: Reporting Data, Methods, and Results (8–12 November 2021)	9
Week 10: Multiple Regression – Diagnostics and Model Fit (15–19 November 2021)	9
Week 11: Logistic Regression (22–26 November 2021)	9
Week 12: Improving the Interpretation and Visualisation of Regression Results (29 November–3 December 2021)	10

Week 1: Accessing and Visualising Data (13–17 September 2021)

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 R⁵ and RStudio⁶ 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

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

Optional

- D. Spiegelhalter (2020). *The Art of Statistics: Learning from Data*. London/New York: Penguin Books: ch. 2.
- E. G. Larsen and Z. Fazekas (2021). *Quantitative Politics with R*: ch. 7, 10.

⁵Download R for Mac, Windows, or Linux at: <https://cloud.r-project.org>

⁶Download RStudio (Desktop Open Source License [free]) at: <https://rstudio.com/products/rstudio/download>

- J. 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 (20–24 September 2021)

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

- K. Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press: ch. 1.
- C. Ismay and A. Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press: ch. 3–4.
- E. G. Larsen and Z. Fazekas (2021). *Quantitative Politics with R*: ch. 3.

Optional

- G. Wilson, J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal (2017). “Good Enough Practices in Scientific Computing”. *PLoS Computational Biology* 13 (6): e1005510.

Week 3: Simple Regression (27 September–1 October 2021)

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

Mandatory Readings

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

Optional

- P. M. Kellstedt and G. D. Whitten (2018). *The Fundamentals of Political Science Research*. 3rd edition. Cambridge: Cambridge University Press: ch. 9.
- A. McCauley and A. Ruggeri (2020). “From Questions and Puzzles to Research Project”. *The SAGE Handbook of Research Methods in Political Science and International Relations*. Ed. by L. Curini and R. Franzese. Thousand Oaks: Sage: 26–43.

Week 4: Multiple Regression (4–8 October 2021)

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.

- C. Ismay and A. Y. Kim (2020). *Statistical Inference via Data Science: A ModernDive into R and the tidyverse*. Boca Raton: CRC Press: ch. 6.
- P. M. Kellstedt and G. D. Whitten (2018). *The Fundamentals of Political Science Research*. 3rd edition. Cambridge: Cambridge University Press: ch. 10.

Optional

- E. G. Larsen and Z. Fazekas (2021). *Quantitative Politics with R*: ch. 11.
- M. A. Hardy (1993). *Regression with Dummy Variables*. Newbury Park, CA: SAGE Publications.

Week 5: Multiple Regression – Categorical Independent Variables and Interactions (11–15 October 2021)

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

Mandatory Readings

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

Optional

- J. Hainmueller, J. Mummolo, and Y. Xu (2019). “How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice”. *Political Analysis* 27 (2): 163–192.
- W. R. Clark, M. J. Gilligan, and M. Golder (2006). “A Simple Multivariate Test for Asymmetric Hypotheses”. *Political Analysis* 14 (3): 311–331.

Week 6: Sampling Distributions and Central Limit Theorem (18–22 October 2021)

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

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

Optional

- R. Hoekstra, R. D. Morey, J. N. Rouder, and E.-J. Wagenmakers (2014). “Robust Misinterpretation of Confidence Intervals”. *Psychonomic Bulletin & Review* 21 (5): 1157–1164.

Week 7: Bank Holiday Monday: Virtual Q&A Session (25–29 October 2021)

The lecture and lab session are cancelled due to the Bank Holiday on 25 October. However, I will facilitate a virtual Q&A session, dealing with questions focusing on questions relating to theory, coding, or your final research projects. This session will take place on Zoom. We will discuss a date and time for this meeting in Week 2.

Week 8: Hypothesis Tests and Confidence Intervals (1–5 November 2021)

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

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

Optional

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

Week 9: Reporting Data, Methods, and Results (8–12 November 2021)

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

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

Optional

- D. Spiegelhalter (2020). *The Art of Statistics: Learning from Data*. London/New York: Penguin Books: ch. 12.
- K. Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press: ch. 4–5.
- E. G. Larsen and Z. Fazekas (2021). *Quantitative Politics with R*: ch. 13–14.

Week 10: Multiple Regression – Diagnostics and Model Fit (15–19 November 2021)

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).

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

Optional

- E. G. Larsen and Z. Fazekas (2021). *Quantitative Politics with R*: ch. 11.4.

Week 11: Logistic Regression (22–26 November 2021)

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

- P. M. Kellstedt and G. D. Whitten (2018). *The Fundamentals of Political Science Research*. 3rd edition. Cambridge: Cambridge University Press: ch. 12.
- P. Roback and J. Legler (2020). *Beyond Multiple Linear Regression: Applied Generalized Linear Models and Multilevel Models in R*. Boca Raton: CRC Press: ch. 6.

Optional

- J. Gill and M. Torres (2019). *Generalized Linear Models: A Unified Approach*. Newbury Park, CA: SAGE Publications.

Week 12: Improving the Interpretation and Visualisation of Regression Results (29 November–3 December 2021)

How to improve the presentation of substantive effects and interactions given that regression tables and coefficients may be difficult to interpret.

- G. King, M. Tomz, and J. Wittenberg (2000). “Making the Most of Statistical Analyses: Improving Interpretation and Presentation”. *American Journal of Political Science* 44 (2): 341–355.

Optional

- A. Gelman, J. Hill, and A. Vehtari (2020). *Regression and Other Stories*. Cambridge: Cambridge University Press: ch. 3.
- K. Healy (2019). *Data Visualization: A Practical Introduction*. Princeton: Princeton University Press: ch. 6, 8.