Hertie School

Master of International Affairs
Master of Public Policy
Master of Data Science

Spring Semester 2023

Course Syllabus

GRAD-Xoooo: Statistics II: Statistical Modeling and Causal Inference with R

Area of Concentration(s):

1. General information

Course Format	Onsite
Instructor(s)	Prof. Dr. Simon Munzert
Instructor's e-mail	munzert@hertie-school.org
Assistants (if applicable)	– Sebastian Ramirez Ruiz, ramirez-ruiz@hertie-school.org – Francisca Castro, francisca.castro@hu-berlin.de – Carolina Diaz Suarez, A.Diaz-Suarez@students.hertie-school.org – Sahil Gill, s.gill@hertie-school.org
Instructor's Office Hours	Wednesdays, 10-12h Please email me in advance to make an appointment and to let me know what you would like to talk about.

Link to Study, Examination and Admission Rules and MIA, MDS and MPP Module Handbooks

For information on **course room, times and session dates,** please consult the Course Plan on *MyStudies*.

Instructor Information:
Simon Munzert is Professor of Data Science and Public Policy at the Hertie School and director of the Data Science Lab. His research interests include attitude formation in the digital age, public opinion, and the use of online data in social research. He received his Doctoral Degree in Political Science from the University of Konstanz.

2. Course Contents and Learning Objectives

Course contents:

This course continues the sequence in statistical modeling. Assuming prior knowledge in simple and multiple linear regression modelling, it introduces you to a new perspective on studying causes and effects in social science research. Based on a framework of causality, the course agenda covers various strategies to uncover causal relationships using statistical tools. We start with reflecting about causality, the ideal research design, and then learn to use a framework to study causal effects. Then, we revisit common regression estimators of causal effects and learn about their limits. Next, we will focus on matching, instrumental variables, difference-in-differences and fixed effects estimators, regression discontinuity designs, and techniques to explore moderated relationships. All classes divide time between theory and application.

Main learning objectives:

The goals are to (1) acquaint you with some of the most common statistical methods used to demonstrate causality, (2) enable you to implement these with statistical software, and (3) prepare you for Hertie's methods electives.

Software:

We will work with R, RStudio and RMarkdown to implement and practice the learned techniques. It is assumed that you have some basic knowledge in using R from Statistics I. If not, we strongly encourage you to familiarize yourself with R prior to the course so to be able to focus on the substance of the course. To that end, we will provide an introductory R webinar, for those who don't feel quite comfortable with R yet. The course will be held as two half-day online seminars in the week before class starts. The first part of the course covers the R, RStudio and R Markdown setup and will introduce you to the fundamentals of programming in R. The second part of the course will then introduce you to data wrangling with the Tidyverse, basic visualizations with ggplot2, and conclude with some advice on troubleshooting and best practice.

Teaching style:

Each session will consist of a pre-recorded lecture, which, alongside the course readings, serves to familiarize you with the session's topic. This is followed by interactive Q&A sessions with the lecturer during which core concepts are discussed and further clarified. To become acquainted with R and to learn how to implement various statistical techniques in practice, you will be enrolled in a small-group support lab session taught weekly by a teaching assistant.

Prerequisites:

Statistics I, basic command of R

Diversity Statement:

Science thrives on the diversity of ideas and opinions, so I try to represent this diversity in our course as well. In the choice of course readings, we strive to cover authors with a diversity of backgrounds as far as possible. During class sessions, I seek to create a trusting and inclusive atmosphere that allows students of all orientations and backgrounds to be comfortable and openly engage in discussions. The R community lives these values and I want you to become part of it. If you have any suggestions that contribute to this goal, I am always grateful for feedback.

3. Grading and Assignments

Grading is structured around 3 main groups of assignments, each roughly equally weighted.

Composition of Final Grade:

Assignment 1: Biweekly take-home assignments	Deadline: *About a week after publication*	Submit via Moodle	40%

Assignment 2: In-class (or online) final exam	Deadline: *Final exam week – exact date TBD*	Moodle Quiz (or onsite, TBD)	35%
Assignment 3: Final replication project	Deadline: *About three weeks after final exam week – exact date TBD*	Submit via Moodle	25%

Assignment Details

Assignment 1
This assignment comprises 5 take-home problem sets (the 4 best of which will be graded) that you will have to work on. The problem sets refer to 1-2 data sets that are shared by the instructor. The assignment will have to be submitted as a knitted R Markdown file containing both R code and your answers to the questions. Each of the four best assignments has an equal weight in the final grade for this component. You are generally encouraged to study and learn to use the software together. However, each student must hand in individual assignments, which will be graded separately.

Assignment 2
The in-class (or online) final exam will be comprised of a set of multiple-choice theoretical questions that draw on the class readings along with a set of questions on applied statistical problems based on analysis output.

Assignment 3
As a final assignment, course participants are asked to replicate, check, and amend the analyses of an existing academic paper. The instructor will provide a small set of replication data sets, and their accompanying code and manuscript. Participants will be asked to re-run the analyses reported in the manuscript, as well as go beyond them, by exploring additional questions with the data. This assignment will draw on skills built up during the weekly assignments. The replication project will be brief (6-8 pages). However, please keep in mind that replicating and understanding code written by someone else takes considerable time.

Late submission of assignments: For each day the assignment is turned in late, the grade will be reduced by 10% (e.g. submission two days after the deadline would result in 20% grade deduction).

Attendance: Students are expected to be present and prepared for every class session. Active participation during lectures and seminar discussions is essential. If unavoidable circumstances arise which prevent attendance or preparation, the instructor should be advised by email with as much advance notice as possible. Please note that students cannot miss more than two out of 12 course sessions. For further information please consult the Examination Rules §10.

Academic Integrity: The Hertie School is committed to the standards of good academic and ethical conduct. Any violation of these standards shall be subject to disciplinary action. Plagiarism, deceitful actions as well as free-riding in group work are not tolerated. See Examination Rules §16 and the Hertie Plagiarism Policy.

Compensation for Disadvantages: If a student furnishes evidence that he or she is not able to take an examination as required in whole or in part due to disability or permanent illness, the Examination Committee may upon written request approve learning accommodation(s). In this respect, the submission of adequate certificates may be required. See Examination Rules §14.

Extenuating circumstances: An extension can be granted due to extenuating circumstances (i.e., for reasons like illness, personal loss or hardship, or caring duties). In such cases, please contact the course instructors and the Examination Office *in advance* of the deadline.

4. General Readings

Throughout the class we rely heavily on the material in:

- **The Mixtape**

  Cunningham, S. (2021). Causal Inference: The Mixtape. New Haven: Yale University Press. Online version available at: https://mixtape.scunning.com/

- **The Effect**

  Huntington-Klein, N. (2021). *The Effect: An Introduction to Research Design and Causality*. CRC Press. Free online version available at: https://theeffectbook.net/

- **Impact Evaluation**

  Gertler, P.J., Martinez, S., Premand, P., Rawlings, L.B., Vermeersch, M.J. (2016). Impact Evaluation in Practice. Second Edition. World Bank Group. Preprint available at https://openknowledge.worldbank.org/handle/10986/25030.

In addition, you will encounter chapters from other books and journal articles to read; these primarily provide illustrations and more background. Articles that are designated as **Optional Readings** are not required, although references to them and techniques they use will be made during the lecture.

Articles listed under **Application Reading** serve as an example case during the lecture. It is mandatory to read them in advance.

This course has an exclusive focus on causal identification and inference. There are more books out there that people commonly consult to learn about these topics, including:

1. Angrist, Joshua David, and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press, 2009.

2. Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Basic Books.

3. Morgan, S. L., & Winship, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd ed.). New York: Cambridge University Press.

4. Hernán, M. A., & Robins, J. A. (2020). *Causal Inference: What If*. Boca Raton, FL: CRC Press. Preprint available at: https://cdn1.sph.harvard.edu/wp-content/uploads/sites/1268/2019/10/ci_hernanrobins_1oct19.pdf.

5. Angrist, J. D., & Pischke, J.-S. (2015). *Mastering 'Metrics: The Path from Cause to Effect*. Princeton: Princeton University Press.

Moreover, there are other books that focus more on statistical modeling and estimation. If you are interested in these topics, please consult the following:

1. Andrew G., & Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York: Cambridge University Press. [*a reference book for regression modeling and, particularly, mixed-effects models, from a Bayesian perspective*]

2. Enders, C. K. (2010). *Applied Missing Data Analysis*. New York: The Guilford Press. [*an accessible introduction into the problems of missing data for our analyses; includes chapters on maximum likelihood estimation*]

3. Fox, J. (2016). *Applied Regression Analysis and Generalized Linear Models* (3rd ed.). London: Sage Publications. [*theoretical treatment of OLS regression and GLMs, along with selected additional topics, e.g., multilevel models*]

4. Freedman, D. A. (2009). *Statistical Models: Theory and Practice*. Cambridge: Cambridge University Press.

5. Wooldridge, J. M. (2013). *Introductory Econometrics: A Modern Approach* (5th ed.). Mason, OH: Cengage Learning. [*an econometric classic*]

Finally, it's the 2020s and there are tons of video and audio resources around causal inference on the web. Here are some that I can recommend:

1. Causal Inference Bootcamp https://mattmasten.github.io/bootcamp/

2. Online Causal Inference Seminar https://sites.google.com/view/ocis/ , https://www.youtube.com/channel/UCiiOj5GSES6uw21kfXnxj3A/videos

3. Causal Inference Podcast, https://casualinfer.libsyn.com/

5. Session Overview

Course session times and dates can be found in the Course Plan on *MyStudies*.

Session	Session Title
1	**Counterfactual causality**
2	**The potential outcomes framework and experiments**
3	**Causal graphs***
4	**Regression estimators of causal effects**
5	**Matching***
6	**Instrumental variables**
Mid-term Exam Week: no class	
7	**Regression discontinuity designs***
8	**Difference-in-differences and synthetic controls**
9	**Panel data***

10	**Heterogeneity and mechanisms**
11	**Validity and generalizability***
12	**Discussion and conclusion**
Final Exam Week: no class	

Note: After sessions marked with an asterisk (*) a new assignment will go online.

6. Course Sessions and Readings

All mandatory readings that represent journal articles or select chapters from books will be available on the Moodle website.

Session 1: Counterfactual causality	
Learning Objective	(a) Understanding why causality matters in social science and policy; (b) Grasping the difference between associations, interventions, and counterfactuals.
Required Readings	**The Mixtape**: Chapter 1 (Introduction) **Impact Evaluation**: Chapters 1-2 (Why evaluate? Preparing for an evaluation) Also, watch any of the following videos (about 1 hour each): – Andrew Gelman: 100 Stories of Causal Inference https://youtu.be/jnI5KI843Lk – Interview with Judea Pearl https://youtu.be/hB9xDcumnHY – Interview with Donald Rubin https://youtu.be/mIjZc8Iy9KY – Interview with Esther Duflo https://youtu.be/WWW9q30MYxU – Judea Pearl: The New Science of Cause and Effect https://youtu.be/ZaPV1OSEpHw – Fireside chat with Susan Athey: https://youtu.be/ypHLWbtqepo
Optional Readings	Freedman, David A. "Statistical Models and Shoe Leather." *Sociological Methodology* 21 (1991): 291–313. Rubin, D. B. (2008). For Objective Causal Inference, Design Trumps Analysis. *The Annals of Applied Statistics*, 2(3), 808–840.

Session 2: The potential outcomes framework and experiments	
Learning Objective	(a) Understanding the potential outcomes framework (POF), and the perils of selection bias for inference; (b) Assumptions behind POF; (c) Average treatment effect; (d) Value of randomization and experiments.
Required Readings	**The Mixtape**: Chapter 4 (Potential outcomes causal model)
Optional Readings	**Impact Evaluation**, Chapters 3-4 (Causal inference and counterfactuals, Randomized assignment)

	The Effect: Chapter 10 Duflo, E., Glennerster, R., & Kremer, M. (2008). Using Randomization in Development Economics Research: A Toolkit. In T. P. Schultz & J. A. Strauss (Eds.), *Handbook of Development Economics* (Vol. 4, pp. 3895–3962). Amsterdam, NL: Elsevier. Morgan, S. L., & Winship, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd ed.). New York: Cambridge University Press. Chapter 2.

Session 3: Causal graphs	
Learning Objective	(a) Directed Acyclical Graphs (DAGs); (b) Uses and usefulness; (c) Principles of DAGs; (d) Identification criteria; (e) D-separation; (f) Confounding.
Required Readings	**The Mixtape**: Chapter 3 (Directed acyclic graphs) **The Effect:** Chapters 6 to 8
Optional Readings	Elwert, F. (2013). Graphical Causal Models. In S. L. Morgan (Ed.), *Handbook of Causal Analysis for Social Research* (pp. 245–273). Dordrecht: Springer Netherlands. Pearl, J., Glymour, M., & Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. Chichester, UK: Wiley. Chapter 1 [starting from 1.4] and Chapter 2. Morgan, S. L., & Winship, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd ed.). New York: Cambridge University Press. Chapter 3.

Session 4: Regression estimators of causal effects	
Learning Objective	(a) OLS mechanics and estimation. (b) Omitted variables. (c) Regression from a causal perspective. (d) Post-treatment bias.
Required Readings	**The Mixtape**: Chapter 2 (Probability and regression review; from Ordinary least squares onwards) **The Effect:** Chapter 13
Optional Readings	Gelman, Andrew, and Jennifer Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press, 2007. Chapter 9, esp. Section 9.5 onwards. Angrist, Joshua David, and Jörn-Steffen Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press, 2009. Section 3.2.
Application Reading	Egan, Patrick J., and Megan Mullin. "Turning Personal Experience into Political Attitudes: The Effect of Local Weather on Americans' Perceptions about Global Warming." *The Journal of Politics* 74, no. 3 (July 1, 2012): 796–809.

Session 5: Matching	
Learning Objective	(a) Matching compared to regression. (b) Matching methods. (c) Specification.
Required Readings	**The Mixtape**: Chapter 5 (Matching and subclassification) **The Effect:** Chapter 14
Optional Readings	**Impact Evaluation**: Chapter 8 (Matching) Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25, no. 1 (February 2010): 1–21. Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. "Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference." *Political Analysis* 15, no. 3 (June 20, 2007): 199–236. Morgan, S. L., & Winship, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd ed.). New York: Cambridge University Press. Chapter 5.
Application Reading	Gilligan, M.J., & Sergenti, E.J. (2008). Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference. *Quarterly Journal of Political Science* 3: 89—122.

Session 6: Instrumental variables	
Learning Objective	(a) IV definition. (b) Exclusion restriction. (c) Two-stage least squares.
Required Readings	**The Mixtape**: Chapter 7 (Instrumental variables) **The Effect:** Chapter 19
Optional Readings	**Impact Evaluation**: Chapter 5 (Instrumental variables) Sovey, Allison J., and Donald P. Green. "Instrumental Variables Estimation in Political Science: A Readers' Guide." *American Journal of Political Science* 55, no. 1 (2011): 188–200. Dunning, Thad. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge: Cambridge University Press, 2012. Chapter 4.
Application Reading	Kern, H.L. and J. Hainmueller. (2009). Opium for the Masses: How Foreign Media Can Stabilize Authoritarian Regimes. *Political Analysis* 17(4): 377-399.

Session 7: Regression discontinuity designs	
Learning Objective	(a) Connection with experimental assignment; (b) the *sharp* RD design; (c) choices in RD designs; (d) the *fuzzy* RD design
Required Readings	**The Mixtape**: Chapter 6 (Regression discontinuity) **The Effect:** Chapter 20

Optional Readings	**Impact Evaluation**: Chapter 6 (Regression discontinuity design) Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2019). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. New York: Cambridge University Press. Sekhon, J. S., & Titiunik, R. (2017). On Interpreting the Regression Discontinuity Design as a Local Experiment. In *Regression Discontinuity Designs* (Vol. 38, pp. 1–28). Bingley, UK: Emerald Publishing Limited. Hausman, C., & Rapson, D. S. (2018). Regression Discontinuity in Time: Considerations for Empirical Applications. *Annual Review of Resource Economics*, 10(1), 533–552. Lee, D. S., & Lemieux, T. (2015). Regression discontinuity designs in social sciences. In *The SAGE Handbook of Regression Analysis and Causal Inference* (pp. 301–326). London: SAGE Publications.
Application Reading	Carpenter, C. & Dobkin, C. (2009). The Effect of Alcohol Consumption on Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age. *American Economic Journal: Applied Economics* 1(1): 164—182.

Session 8: Difference-in-differences and synthetic controls	
Learning Objective	(a) DiD. (b) Synthetic control method. (c) Diagnostics.
Required Readings	**The Mixtape**: Chapters 9, 10 (Differences-in-differences, Synthetic control) **The Effect:** Chapters 17, 18
Optional Readings	**Impact Evaluation**: Chapter 7 (Difference-in-differences) Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105, no. 490 (2010): 493–505. Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. "Comparative Politics and the Synthetic Control Method." *American Journal of Political Science* 59, no. 2 (2015): 495–510.
Application Reading	Selb, P. and S. Munzert. (2018). Examining a Most Likely Case for Strong Campaign Effects: Hitler's Speeches and the Rise of the Nazi Party, 1927–1933. *American Political Science Review* 112(4): 1050–1066.

Session 9: Panel data	
Learning Objective	(a) Basics: interrupted time series models; (b) Features of panel data; (c) Adjustment strategies; (d) Fixed effects properties
Required Readings	**The Mixtape**: Chapter 8 (Panel data) **The Effect:** Chapter 16
Optional Readings	Bell, A., & Jones, K. (2015). Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data. *Political Science*

	Research and Methods, 3(01), 133–153. Imai, K., & Kim, I. S. (2019). When Should We Use Unit Fixed Effects Regression Models for Causal Inference with Longitudinal Data? *American Journal of Political Science*, 63(2), 467–490. Morgan, S. L., & Winship, C. (2015). *Counterfactuals and Causal Inference: Methods and Principles for Social Research* (2nd ed.). New York: Cambridge University Press. Chapter 11. Brüderl, J., & Ludwig, V. (2015). Fixed-effects panel regression. In *The SAGE Handbook of Regression Analysis and Causal Inference* (pp. 327–357). London: SAGE Publications. Clark, T. S., & Linzer, D. A. (2015). Should I Use Fixed or Random Effects? *Political Science Research and Methods*, 3(2), 399–408.
Application Reading	O'Grady, T. (2019). How do Economic Circumstances Determine Preferences? Evidence from Long-run Panel Data. *British Journal of Political Science*, 49(4), 1381–1406.

Session 10: Heterogeneity and mechanisms	
Learning Objective	(a) Moderation and mediation. (b) Heterogeneous effects. (c) Estimation and interpretation.
Required Readings	Brambor, Thomas, William Roberts Clark, and Matt Golder. "Understanding Interaction Models: Improving Empirical Analyses." *Political Analysis* 14, no. 1 (2006): 63–82. Nguyen, T. Q., Schmid, I., & Stuart, E. A. (2021). Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological Methods*, 26(2), 255.
Optional Readings	Baron, Reuben M., and David A. Kenny. "The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations." *Journal of Personality and Social Psychology* 51, no. 6 (1986): 1173–1182. Hainmüller, J., Mummolo, J., and Xu, Yiqing. (2019). How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice. *Political Analysis* 27(2): 163—192.
Application Reading	Gidengil, Elisabeth, Hanna Wass, and Maria Valaste. "Political Socialization and Voting: The Parent–Child Link in Turnout." *Political Research Quarterly* 69, no. 2 (June 1, 2016): 373–83.

Session 11: Validity and generalizability	
Learning Objective	(a) Construct validity, (b) Statistical conclusion validity, (c) Internal validity, (d) External validity, (e) Generalization of study results
Required Readings	**The Effect:** Chapter 22 McDermott, Rose. "Internal and external validity." *Cambridge Handbook of Experimental Political Science* (2011): 27-40.

Optional Readings	**Impact Evaluation**: Chapter 9 (Addressing methodological challenges) Hünermund, P. & Louw B. 2020. On the Nuisance of Control Variables in Regression Analysis. https://arxiv.org/abs/2005.10314. Kern, Holger L., Stuart, Elisabeth A., Hill, Jennifer, and Green, Donald P. 2016. Assessing methods for generalizing experimental impact estimates to target populations. *Journal of Research on Educational Effectiveness* 9: 103-127. Lesko, Catherine R., Buchanan, Ashley. L., Westreich, Daniel, Edwards, Jessie K., Hudgens, Michael G., and Cole, Stephen R. 2017. Generalizing study results: a potential outcomes perspective. *Epidemiology* 28: 553-561 Stuart, E. A., & Rhodes, A. (2016). Generalizing Treatment Effect Estimates from Sample to Population: A Case Study in the Difficulties of Finding Sufficient Data. *Evaluation Review*, 41(4), 357–388.

Session 12: Discussion and conclusion	
Learning Objective	(a) Getting the big picture, (b) Making policy analysis transparent and ethical, (c) How to broaden your data science toolkit
Required Readings	None
Optional Readings	None

Final Exam Week: no class