# Content

# Study guide for Statistics and Data Science II, Fall 2023

*Last updated: October 15th, 2023*

## Teachers


Course director and teacher: Maria Brandén, PhD in Sociological Demography, Associate Professor in Sociology and in Analytical Sociology maria.branden@liu.se

Teacher for seminars: Cheng Lin, PhD student in Analytical Sociology, MSc in Computational Social Science cheng.lin@liu.se

Teacher for labs: Martin Arvidsson, Postdoc in Analytical Sociology (background in Statistics / Statistics and Machine Learning) martin.arvidsson@liu.se


## Course description

This course intends to familiarize the student with the underlying principles of linear regression modelling and its applications in social sciences, with a special emphasis on interpretation, causal inference and methods to address this using observational data. The aim is to develop the student's understanding of what the methods are all about and how they work in practice.

The course moves between (1) discussing the methods and discuss research applying them in social scientific settings, and (2) applying the methods on empirical data, using the statistical software R.

Focus is on the application and not the theoretical parts of quantitative modelling. While we do cover some of the basics, we will spend limited time on the underlying maths. Students will learn the important fundamentals of applied statistical analyses, including how to assess a quantitative research design, execute and interpret standard models as well as more advanced models considering causal inference.

The course consists of lectures, seminars and labs. The lectures are given by Associate professor Maria Brandén who is also the course coordinator. The seminars are held by PhD student Cheng Lin and the labs are held by Postdoc Martin Arvidsson.

All seminars are mandatory. In the case of missing a seminar, the student shall hand in one-page of answers to the questions for the seminar. This shall be handed in before the next meeting.

Students are strongly encouraged to contribute to a positive and active learning environment. Please feel free to ask questions during class, or let the teacher know if something is unclear or strange. There are no stupid questions! Please show respect to your classmates when they ask questions, this is a diverse group of students from different backgrounds and what is obvious for one of you may be completely new for someone else.

### Structure of lectures, seminars and labs

Most themes will be covered by three meetings. The first part is a lecture where a method/concept is introduced and discussed. The second part is a seminar. Between the lecture and the seminar, the students are expected to have read the listed scientific paper and be prepared to discuss it in smaller groups and in class. Questions will be provided by the instructor before each seminar.

Labs will follow after each lecture+seminar, and will let the students familiarize themselves with the methods from the lecture, using R.

After each lab session (except the lab on Monte Carlo simulations), students are expected to hand in a lab report which will be graded with a pass or fail. Students are encouraged to help each other while doing labs, however, **handing in identical lab assignments is not allowed.**

All lab reports are to be submitted in Lisam before the next lab (see deadlines below). At this point in time, we upload the correct solutions to Lisam. If a student fails to upload a lab report before the correct solutions are posted, s/he will need to hand in an additional assignment. No detailed feedback will be provided to the lab reports.

## Intended learning outcomes

After completion of the course, the student should be able to:

1. use statistical software to estimate appropriate linear regression models for cross sectional and panel data and explain the statistical principles underlying these estimates;
2. use statistical software to assemble appropriate data structures for estimating regression models and implementing robustness checks;
3. interpret the parameters of linear regression models, produce predictions, and evaluate goodness of fit;
4. describe the logic of causal inference and how it applies to regression models, distinguishing between causality and correlation;
5. identify common threats to causal interpretation of linear models, and assess and justify modeling approaches for solving these threats;
6. evaluate the validity and robustness of causal inferences under a variety of assumptions about how the data was generated.

## Examination and grades

*All examination is individual. Students are encouraged to help their peers and ask their peers for help but are not allowed to hand in identical assignments.*

Grades range from A to F/Fx and are based on how well the student has achieved the intended learning outcomes. The learning outcomes are assessed as follows:

Active participation in seminars

- 1 ECTS, OBLx (pass/fail)

Lab reports

- 2 ECTS. ASSx (pass/fail)

Final project

- 4 ECTS. PROx (E-A, or F/Fx)

Oral comments on another student's final project, and a short (~1 page) written summary of comments

- 0.5 ECTS. OPPO (E-A, or F/Fx)

Active participation in seminars and the lab reports are graded by pass or fail.

The final project and the oral comments on another student's final project are graded (E-A, or F/Fx). The following dimensions are considered:

1. The ability to present methods for analyzing cross sectional and/or panel data
2. The ability to discuss and critically evaluate methods for analyzing cross sectional and/or panel data
3. The ability to show an understanding of the logic of causal inference and how it applies to regression models, distinguishing between causality and correlation
4. The ability to identify common threats to causal interpretation of linear models, and assess and justify modeling approaches for solving these threats. Not only in a generic sense, but also how it applies creatively to your problem.
5. The ability to refer to existing literature properly (given the availability of ChatGPT, it is very important that you show that you base your reasoning on the literature)
6. The ability to show an understanding of the importance of transparency in how choices have been made when deciding on a research design, including (but not limited to) choosing a method, operationalization of variables, and choosing a data set

For the grades on the final project, all the dimensions are initially graded on a 3-level scale (good, acceptable and fail). These grades are then translated to a final grade according to the following:

| | |
|---|---|
| A | Good on all dimensions |
| B | Good on a minimum of 5 dimensions (fail on none) |
| C | Good on a minimum of 4 dimensions (fail on none) |
| D | Good on a minimum of 3 dimensions (fail on none) |
| E | Acceptable on all dimensions |
| Fx | Fail on at most 2 dimensions. Possibility to re-examine those dimensions to receive an E. |
| F | Fail on more than 2 dimensions. Full re-examination needed. |

## Deadlines

All labs are to be handed in before the following lab. (Lab1 12/11; Lab2 23/11; Lab3 6/12; Lab4 14/12)

14/12: Deadline for e-mailing research question for final project to maria.branden@liu.se

12/1: E-mail first version of final project to maria.branden@liu.se and peer according to list. You will only receive comments by your peer on this first draft, i.e. not by the instructor.

19/1: Final version of final project submitted in Lisam (it will be tested for plagiarism through Urkund). Also submit the 1 page of comments you prepared for your peer for the final seminar at Lisam. This is also the deadline for the terminology-assignment.

**Please make sure you meet the deadlines. If a student fails to meet a deadline, s/he will need to hand in an additional assignment.**

## Final project

In the final project, I want you to demonstrate what you have learned throughout the first part of the course (until December 15th) by discussing an own (fictive or real) project where you could apply one (or preferably more than one) of the methods that we have learned. Make sure you have read the grading criteria in the course guide, so that you fulfil the criteria for the grade you aim for.

Note that the hands-on data work is evaluated through the labs – no actual analysis is required for the final project. But if you want to, you are welcome to include analysis also in the final project.

The final assignment should be around 2000-2500 words and include:

- A social scientific research question of causal nature (i.e. how does X affect Y?). Draw graphs to explain the link between X and Y, and make sure you carefully discuss potential confounders, mediators, moderators etc. (i.e. how does X affect Y, what are the potential Z's ?). Make sure you use the literature when discussing causality related to your research question.
- A discussion of what kind of data you would need to answer the question. Think of what kind of data you ideally would like to have, and what kind of data you think is possible to get access to. Please consult online codebooks for potential data sources or think of own questionnaires or equivalent that you could use. Also describe how you would operationalize at least 3 relevant concepts in your analysis.
- Present at least one of the methods we have learned on 1-2 pages (make sure to refer to the literature) and discuss how it can be used to answer your research question. If you aim for a higher grade than D, it is not enough to only focus on OLS regression. What are the drawbacks of the method, and what considerations do you need to make? Do you think any results you would get could be said to be of causal nature? Why? Why not?

## Final seminar

The final assignment is to be discussed at the final seminar which takes place from 8.15 to 17 on January 16th. At this seminar, each student is the designated peer on a fellow student's assignment. All students are to prepare a 10-minute presentation presenting the other student's project and to list at least 3 good things and 3 things that can be improved. After the seminar, these points should be summarized in a word document and given to the student. The word document should also be submitted at Lisam.

## Re-examination

F or Fx on *active participation in seminars* is re-examined by handing in a 1000 words reading response discussing the literature for the seminar

F or Fx on the *lab reports* are re-examined by handing in a new version of the lab report by the date of the re-examination.

Fx on the *final project* is re-examined by adapting the project according to comments by the instructor. F on the *final project* is re-examined by handing in a completely new final project, according to comments by the instructor.

F or Fx on *the peer review* (oral and written comments on another student's final project) is re-examined by reading and commenting on another student's final project, as instructed by the instructor.

## Plagiarism

Plagiarism is a serious offense against good academic practice and can if worse comes to worst result in temporary suspension from studies by decision of The Disciplinary Board at Linköping University. We have had these problems previously, and take them seriously. We expect all students to know about the rules regarding plagiarism at LiU. Here you can re-familiarize yourself with these rules: https://liu.se/en/article/plagiering-upphovsratt

We also view it as plagiarism to copy-paste from ChatGPT. You can use ChatGPT for isolated tasks, and we encourage you to use it if you get stuck on labs and as a tool for learning. Ultimately, your work should be based on your own pool of knowledge, stemming from labs, seminars, course literature, and online sources, including chat GPT.

## Course literature

*Books*

Gelman, A & Hill, J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press.

Angrist, J D & Pischke, J-S. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press. Possibly available at https://www.researchgate.net/publication/51992844_Mostly_Harmless_Econometrics_An_Empiricist's_Companion/download

*Scientific articles and other sources*

See the **Reading instructions** below for each lecture

## Reading instructions

* indicates non-mandatory reading

All literature is supposed to be read prior to the lecture/seminar/lab.

*Note: Some minor additional readings may be added and some of the literature to the seminar may be replaced (in yellow)*

**Lecture 1, 2/11**

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... & Carlsson, R. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337-356. Available at https://journals.sagepub.com/doi/pdf/10.1177/2515245917747646

Engel, R. J. & Schutt, R. K. (Eds.). (2014). Conceptualization and measurement. Chapter 4 in *Fundamentals of social work research*. Sage Publications. Available at https://us.sagepub.com/sites/default/files/upm-binaries/61666_Chapter_4.pdf

**Lab1, 3/11**

No readings

**Lecture 2:1, 7/11**

Gelman, A & Hill, J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press. Chapters 3 and 4.

Frost, J. (2017). Understanding Interaction Effects in Statistics. Blogpost available at http://statisticsbyjim.com/regression/interaction-effects/

* Angrist, J D & Pischke, J-S. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press. Chapters 1, (2), 3.1 and 3.2 (no necessity to understand the math). Possibly available at https://www.researchgate.net/publication/51992844_Mostly_Harmless_Econometrics_An_Empiricist's_Companion/download

**Lecture 2:2, 10/11**

*To be discussed at seminar. It is crucial to have read prior to seminar, reading instructions will be available on course page.*

Magnusson, C. (2010). Why Is There a Gender Wage Gap According to Occupational Prestige?: An Analysis of the Gender Wage Gap by Occupational Prestige and Family Obligations in Sweden. Acta Sociologica, 53(2), 99–117. Available at https://journals.sagepub.com/doi/abs/10.1177/0001699310365627

**Lab 2, 13/11**

Gelman, A & Hill, J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press. Chapters 3 and 4.

**Lecture 3:1, 16/11**

Gangl, M. (2010). Causal inference in sociological research. Annual review of sociology, 36, 21-47.

Gelman, A & Hill, J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press. Chapters 9 and 10.3

**Lecture 3:2, 21/11**

*To be discussed at seminar. It is crucial to have read prior to seminar, reading instructions will be available on course page.*

Massoglia, M. (2008). Incarceration, health, and racial disparities in health. Law & Society Review, 42(2), 275-306. Available at https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-5893.2008.00342.x

**Lab 3, 24/11**

Gelman, A & Hill, J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press. Pages 206-209.

Olmos, A., & Govindasamy, P. (2015). Propensity scores: a practical introduction using R. Journal of MultiDisciplinary Evaluation, 11(25), 68-88. Available at http://journals.sfu.ca/jmde/index.php/jmde_1/article/view/431/414

**Lecture 4:1, 28/11**

Allison, P. D. 2009. Fixed effects regression models: SAGE publications (chapter 2).

Angrist, J D & Pischke, J-S. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press. Chapter 5. Possibly available at https://www.researchgate.net/publication/51992844_Mostly_Harmless_Econometrics_An_Empiricist's_Companion/download

Gelman, A & Hill, J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press. Chapter 10.7

Collischon, M., & Eberl, A. (2020). Let's talk about fixed effects: Let's talk about all the good things and the bad things. *KZfSS Kölner Zeitschrift für Soziologie und Sozialpsychologie*, *72*(2), 289-299. https://link.springer.com/article/10.1007/s11577-020-00699-8

**Lecture 4:2, 30/11**

*To be discussed at seminar. It is crucial to have prepared prior to the seminar, reading instructions will be available on course page. Note that all groups will be assigned to read one paper each, to present and discuss in class.*

**Group 1:** Gangl, M., & Ziefle, A. (2009). Motherhood, labor force behavior, and women's careers: An empirical assessment of the wage penalty for motherhood in Britain, Germany, and the United States. Demography, 46(2), 341-369. Available at https://link.springer.com/content/pdf/10.1353%2Fdem.0.0056.pdf

**Group 2:** De Neve, J. E., & Oswald, A. J. (2012). Estimating the influence of life satisfaction and positive affect on later income using sibling fixed effects. Proceedings of the National Academy of Sciences, 109(49), 19953-19958. Available at http://www.pnas.org/content/pnas/109/49/19953.full.pdf

**Group 3:** Amato, P. R., & Anthony, C. J. (2014). Estimating the effects of parental divorce and death with fixed effects models. Journal of Marriage and Family, 76(2), 370-386. Available at https://onlinelibrary.wiley.com/doi/abs/10.1111/jomf.12100

**Group 4:** Bygren, M., & Szulkin, R. (2010). Ethnic environment during childhood and the educational attainment of immigrant children in Sweden. Social Forces, 88(3), 1305-1329. Available at https://academic.oup.com/sf/article-abstract/88/3/1305/1936392

**Lab 4, 7/12**

Gelman, A & Hill, J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press. Chapter 10.7

Colonescu, C. 2016 . Principles of Econometrics with R. Chapter 15. Available at https://bookdown.org/ccolonescu/RPoE4/panel-data-models.html (html-version) and at https://bookdown.org/ccolonescu/RPoE4/RPoE.pdf (pdf-version).

**Lecture 5:1, 8/12**

Gelman, A & Hill, J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press. 10.5 and 10.6, and p.228-229

Angrist, J D & Pischke, J-S. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press. Chapter 4.1 and 5.2. Possibly available at https://www.researchgate.net/publication/51992844_Mostly_Harmless_Econometrics_An_Empiricist's_Companion/download

* Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program. Journal of the American statistical Association, 105(490), 493-505.

* Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. Journal of Economic perspectives, 15(4), 69-85.

**Lecture 5:2, 12/12**

*To be discussed at seminar. It is crucial to have read prior to seminar, reading instructions will be available on course page.*

Hjalmarsson, R & Lindquist, M. J. The causal effect of military conscription on crime. The economic journal. 129, 2522-2562.

**Lab 5, 15/12**

Angrist, J D & Pischke, J-S. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press. Chapter 5.2. Possibly available at https://www.researchgate.net/publication/51992844_Mostly_Harmless_Econometrics_An_Empiricist's_Companion/download

**Remaining meetings:** No required readings

## Schedule (Rooms subject to change – please confirm with TimeEdit)

https://cloud.timeedit.net/liu/web/schema/s/s.html?i=x7w_AuZu1nQnkOy1QZwZZ6lQCwc0QYZh710jycQ7sY7u6nfv0XQoZb4QQlZyQQ0ZgoAQbobybZ806y4QZZqu0u

| DAY | DATE | START | END | TYPE OF MEETING | ROOM | TEACHER | CONTENT |
|---|---|---|---|---|---|---|---|
| **Thursday** | 2023-11-02 | 10:15 | 12:00 | Lecture 1. Roll call | K22 | MB | Introduction. Choosing method, data and operationalization to answer a research question. |
| **Friday** | 2023-11-03 | 10:15 | 12:00 | Lab 1 | K25 | MA | Data management in R |
| **Friday** | 2023-11-03 | 13:15 | 15:00 | Lab 1 | K25 | MA | Data management in R |
| **Tuesday** | 2023-11-07 | 10:15 | 12:00 | Lecture 2 | K22 | MB | OLS-regression. |
| **Friday** | 2023-11-10 | 10:15 | 12:00 | Seminar for Lecture 2 | K22 | CL | Seminar on OLS |
| **Monday** | 2023-11-13 | 10:15 | 12:00 | Lab 2 | K22 | MA | Lab on OLS-regression |
| **Monday** | 2023-11-13 | 13:15 | 15:00 | Lab 2 | K22 | MA | Lab on OLS-regression |
| **Thursday** | 2023-11-16 | 10:15 | 12:00 | Lecture 3 | K22 | MB | Causality 1: Potential outcomes and matching techniques |
| **Tuesday** | 2023-11-21 | 10:15 | 12:00 | Seminar for lecture 3 | K22 | CL | Seminar on matching techniques. |
| **Friday** | 2023-11-24 | 10:15 | 12:00 | Lab 3 | K25 | MA | Lab on matching techniques |
| **Friday** | 2023-11-24 | 13:15 | 15:00 | Lab 3 | K22 | MA | Lab on matching techniques |
| **Tuesday** | 2023-11-28 | 10:15 | 12:00 | Lecture 4 | K22 | MB | Causality 2: Using panel or clustered data to assess causality |
| **Thursday** | 2023-11-30 | 10:15 | 12:00 | Seminar for lecture 4 | K22 | CL | Seminar on fixed effects methods |
| **Thursday** | 2023-12-07 | 13:15 | 15:00 | Lab 4 | K24 | MA | Lab on fixed effects methods |
| **Thursday** | 2023-12-07 | 15:15 | 17:00 | Lab 4 | K24 | MA | Lab on fixed effects methods |
| **Friday** | 2023-12-08 | 10:15 | 12:00 | Lecture 5 | K24 | MB | Causality 3: Instrumental variables and difference-in-difference. Intro to final project. |
| **Tuesday** | 2023-12-12 | 10:15 | 12:00 | Seminar for lecture 5 | K22 | CL | Seminar on instrumental variables. |
| **Thursday** | 2023-12-14 | | | | | | **DEADLINE 1: Email research question to Maria** |
| **Friday** | 2023-12-15 | 10:15 | 12:00 | Lab 5 | K22 | MA | Lab on difference-in-difference |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Friday** | 2023-12-15 | 13:15 | 15:00 | Lab 5 | K22 | MA | Lab on difference-in-difference |
| **Monday** | 2023-12-18 | 13:15 | 15:00 | Lecture 6 | K22 | MB | Non-mandatory: What can go wrong? Statistical fallacies and common pitfalls. Opportunity to discuss final project and other issues related to course. |
| **CHRISTMAS BREAK** | | | | | | | |
| **Tuesday** | 2024-01-09 | 10:15 | 12:00 | Lecture 7 | K22 | MA | Monte Carlo simulations |
| **Wednesday** | 2024-01-10 | 13:15 | 15:00 | Lab 7 | K22 | MA | Monte Carlo simulations |
| **Wednesday** | 2024-01-10 | 15:15 | 17:00 | Lab 7 | K22 | MA | Monte Carlo simulations |
| **Friday** | 2024-01-12 | | | | | | **DEADLINE 2: First version of final project** |
| **Tuesday** | 2024-01-16 | 08:15 | 17:00 | Final seminar | K22 | MB | Read and comment on designated final project. |
| **Friday** | 2024-01-19 | | | | | | **DEADLINE 3: All remaining submissions** |