Statistical models beyond linear regression

Syllabus (Spring 2024)

Silje Synnøve Lyder Hermansen

silje.hermansen@ifs.ku.dk

Abstract

Political scientists are in high demand as analysts due to their ability to both understand societal questions and answer them using empirical data. This course puts you in that category of job seekers. You acquire a set of transferable skills that are valuable, regardless of whether you contemplate a career as an academic or a policy analyst for a governmental body, NGO or in a consulting firm. The course also provides a toolkit for students looking use quantitative methods for their master's thesis.

This is an applied methods course. I explain the statistical theory behind models, but the emphasis is on understanding when different models are useful, how to employ them and interpret the results. The course helps students 1) identify appropriate statistical models that describe different data types and 2) interpret these models in a meaningful way.

Many of the phenomena political scientists take interest in can be classified into categories or events that often are not independent from each other. This includes outcome variables like voters' choice of party, number of social media posts in a time span, or time between violent events. The course introduces students to a number of models that are specifically designed to describe the underlying phenomenon that generates such data while possibly leveraging their nested structure. My focus is on observational data. The purpose is to help students gear the statistical analysis towards a realistic yet analytical description of the data.

Topics include

- introduction to R as a statistical software
- binary outcomes (binomial models)
- categorical outcomes (models of choice: multinomial and ordered regression)
- count outcomes (poisson, negative binomial and hurdle models)
- event history data (duration/survival models)
- hierarchical (nested) data (fixed effects and hierarchical/multilevel models)
- missing data (imputations)

Learning outcomes

Students will acquire an overview of typical data structures in political science and form a mental map over models designed for those structures. They will have an intuition of the statistical process that underpins these models, their assumptions (limitations) and what problem each model seeks to address.

Students will obtain a mastery of R as a state-of-the-art software for data analysis. Exiting this course, they can boast hands-on experience with statistical analysis. They will also be able to understand and communicate their findings to policy makers and the wider public in an intuitive way.

Students will be able to devise sound strategies for analyzing observational data for which linear models (OLS) is not appropriate. By reading, discussing and replicating research articles, they further have practical knowledge of how researchers use statistical methods to answer substantive political science questions.

The course has three learning objectives that the exam and class activities are geared to help you acquire:

1. Provide you with a mental map over different data and outcome types as well as the models that are fit for analyzing them.
2. The estimation of these models (in R) and the assumptions underlying these models.
3. The interpretation of regression results.

Class activities

This is a work-intensive class insofar as students are expected to do the readings for each class and work through examples in R. The best way to do so, is to work in groups and exchange insights.

Time and place

We meet twice per week. As a general rule, lectures with theory focus will be on Mondays followed by data labs/seminars with focus on practical implementation in R.

- *Monday (week 7-13, 15-20) 0800-1000 in CSS 5-0-28*
- *Wednesday (week 7-13, 15-20) 1300-1500 in CSS 5-0-28*

You will learn the most if you have already had a first stab at the readings before the class. The readings contain examples with R codes. You will get an intuitive understanding of the material if you work through these examples; especially if the statistical language intimidates you. Be prepared to share your problems and solutions in class!

Class activities and compulsory contributions

Students will contribute to the teaching through *two student-led activities*. At the beginning of the semester, you will therefore sign up for two out of three possible presentation types. You can work in groups for these presentations.

Theory recap At the beginning of the first data lab after we have started a new topic, students will give a recap to jolt our memory about what we have learned. Please aim at 10 minutes. Afterwards, you will share slide deck on Absalon. At the end of the semester, the presentations will help you draft the executive summary that heads the portfolio exam.

- Linear regression and curvilinear effects
- Hierarchical models
- Binary outcomes/Binomial regression
- Models of choice/descrete outcomes
- Count outcomes
- Event history/duration models
- Imputation

R-tips As part of the data labs, we will go through a few tips and tricks in R that may come in handy when you work in R. None of these techniques are strictly necessary to pass the exam, but they are undeniable time-savers when you work with data. For your presentation, aim at 15-20 minutes mixing your presentation with student activities. Afterwards, you will share your notes on Absalon.

- Regular expressions for recoding and data extraction
- Visualization of regression effects using the `ggeffects`-package
- Loops: How to do the same operation many times.
- Functions: How to standardize operations in order to re-use them.
- Dynamic reporting (rmarkdown/LaTeX in R)
- `shiny`-apps
- Other ideas? Feel free to contact me with a proposal!

Replication Your two portfolio items will involve a replication of extant research. You may sign up to present your first part of the replication study to the rest of the class: The descriptive statistics and initial replication. Aim at a 10 minutes presentation + a student round-table where we may discuss potential challenges and solutions. After your presentation, you will share your replication codes on Absalon.

- Portfolio item 1 (April 8th)
- Portfolio item 2 (May 6th)

Exam

The evaluation form for the class is a "portfolio exam". As per usual, you may coauthor the portfolio and submit it as a group exam. *The portfolio exam is due on Absalon on June 1st 2024.* I will apply the criteria specified in the following.

Examination form

The workflow, feedback and form of this portfolio are somewhat different from what you are used to. Make sure to go online and check the list of requirements before handing in the exam. The portfolio will contain the following elements.

1. Preface: A course summary (ca. 2000 characters)

Your portfolio is prefaced by an executive summary of the class. Its aim is to summarize the decision map of when the different models are used. The text should contain references to the relevant readings. You may inspire yourself from the presentations in class and this decision tree for GLMs.

2. Two replication studies/short analyses

The core of your portfolio are two short replication studies assigned in the second half of the semester. Rather than an essay, each assignment consists in a set of questions that you will answer. Each paper is maximum 19.200 characters (24.000 for two students; 28.800 for three students). However, tables, graphics, R-codes and references are not included.

Preliminary feedback You can get *oral* feedback from me and fellow students for each of the portfolio replications during the "workshop/help-desk" sessions. If you also want *written* feedback on your first assignment before the final submission, you may hand in a draft on Absalon at before the set deadline (April 14th). I will not grade your paper at that point, so feel free to submit work in progress.

3. Data and replication codes

We follow standard scientific procedure, meaning that all your analytical work – including your graphical displays – is documented through R-codes. To do so, you will have to do two things:

a. Share your R-codes

You may share your R-codes in one of two ways. You can either create a separate, self-contained and commented R-script that you attach to your portfolio *or* you can show the code embedded in your analytical text. To do the latter, you may use rmarkdown to generate a pdf containing your text, images and codes intertwined.

b. Share your data

To run the codes, I will have to access the data set on which you ran your codes when you made your analysis. You share the data as a hyperlink to a shared folder (e.g. googledocs) or on Absalon.

3. Potential link to a shiny app

A core learning objective for the class is to communicate the implications of statistical models in an intuitive way. There will be an emphasis on graphical display. You may do this as graphics displayed in your pdf.

However, the best students will also illustrate their replication results as an interactive graphic where the reader may play around with different scenarios and get help to understand the predicted outcome of the model. These graphics are generated in RStudio using Shiny. The application/interactive graphic is hosted on a server (e.g. shinyapps.io). To share the application, you create a link to your graphic under your answer to the relevant exam question.

4. Documentation of class activities

Everyone has something to contribute with to help the group learn. To validate the class – and thus get a passing grade – you will have to participate in the compulsory class activities. Your portfolio allows you to document that participation.

a. Links to your two group works/presentations on Absalon
b. Links to your other course participation, if relevant (e.g. sharing of notes and solutions on Padlet).

Assessment

The elements in your portfolio will weigh differently in the final assessment. They are designed to test the learning outcomes of the course.

As a minimum, you will have to demonstrate an adequate understanding of the basics of the course. This includes providing an overview of the model approaches we go through, participation in class activities and the replication of selected results in the research articles assigned for your two papers.

Your grade will primarily be based on your two replication studies. They give you the opportunity to demonstrate the degree to which you have assimilated the learning outcomes. This includes

1) an ability to critically identify and assess the assumptions and limitations of the model. There are often several solutions to a single problem, none of which may be perfect. Spelling some of these solutions out, applying them and discussing their adequacy requires a very good command of the material.

2) Once the model assessment is done, I am interested in your ability to take the model results seriously and apply them in an intuitive interpretation. This means that your

communication of the results in words, numbers and graphics (TTT; "tekst, tal og tegning") will be given substantial attention. The two above-mentioned criteria have equal weight. The truly excellent portfolios will also include an interactive graphic that demonstrates your mastery of statistical interpretation.

The work provided in your two portfolio papers may leave some uncertainty as to what grade your work qualify for. In these cases, I will look to the quality of the accessory documentation to get a better idea about your command of the material. In particular, I will look to the R-codes for clarification if your text leaves me in doubt about what you intend to communicate or what your statistical analysis in fact does.

You can find information about the general examination form and grading criteria at the Department of Political Sciences' home pages.

Online resources

Course webpage You will find slides and other information on my webpage. I will update the page as we advance throughout the semester. This is also where I publish my slides prior to the class.

chatGPT (openAI) is a sophisticated language model/Artificial Intellegence with which you can chat online. It may help you understand the R-codes you will work with this semester. Be aware that this is a machine, so you will have to check the answers you get, just as you do when you google.

Fora for R codes and statistics Some popular fora where users help each other out with statistical questions and R codes are https://stats.stackexchange.com and https://stackoverflow.com/.

Dynamic reports/RMarkdown You can write up your notes and portfolio exam using RMarkdown. It allows you to intertwine R-codes, graphics, tables and your text in one document. You work in RMarkdown in RStudio and can generate html, pdf and word documents with it. The project has its own webpage with tutorials.

Course plan

Week	Topic	Date	Reading
1	Introduction to R as a statistics software	05.02; 07.02	Hermansen (2023), ch. 1-4, p. 19-70
2	Descriptive statistics and graphical display	12.02; 14.02	Hermansen (2023), ch. 5-6, p. 73-119
3	Linear regression	19.02; 21.02	Hermansen (2023), ch. 7-9, p. 123-194

Week	Topic	Date	Reading
			Gelman and Hill (2007), ch 3-4, p. 29-79 Berry, Golder, and Milton (2012) King, Tomz, and Wittenberg (2000) (supplementary reading)
4-5	Hierarchical data structures	26.02; 28.03 04.03; 06.03	Gelman and Hill (2007), ch 11-12, p. 235-278 Gelman and Hill (2007), ch 13, p. 279-300 Gelman and Hill (2007), ch 14-15, p. 301-342
6	Binary outcomes (logistic regression)	11.03; 13.03	Ward and Ahlquist (2018), ch. 3, p. 43-78 Ward and Ahlquist (2018), ch. 6, p. 119-132 Gelman and Hill (2007), ch. 6, p. 109-134 (supplementary reading)
7-8	Categorical outcomes (multinomial and ordered logistic regression)	18.03; 20.03; 03.04	Ward and Ahlquist (2018), ch. 8-9, p. 141-189
	Assignment 1 is given	03.04	
9	Workshop week	08.04; 10.04	Assignment 1 presentation, Assignment helpdesk, Dynamic reporting
	Assignment due (optional)	14.04	
10-11	Count outcomes (poisson, negative binomial and hurdle models)	15.04; 17.04	Linear Digressions: podcast on poisson distribution Ward and Ahlquist (2018), ch. 10, p. 190-216 Gelman and Hill (2007), ch. 6, p. 109-134 (supplementary reading)
11-12	Event history data (survival models)	22.04; 24.04; 29.04	Ward and Ahlquist (2018), ch. 11, p. 190-216

Week	Topic	Date	Reading
	Assignment 2 is given	29.04	
13	Workshop week	06.05; 08.05	Assignment 2 presentation, Assignment helpdesk, `shiny`-app
14	Missing data	06.05; 13.05	Ward and Ahlquist (2018), ch 12, p. 249-270 Gelman and Hill (2007), ch 25, p. 529-545
	Deadline portfolio exam	01.06	

Literature

Berry, William D., Matt Golder, and Daniel Milton. 2012. "Improving Tests of Theories Positing Interaction." *The Journal of Politics*, July. https://doi.org/10.1017/S0022381612000199.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge ; New York: Cambridge University Press.

Hermansen, Silje Synnøve Lyder. 2023. *R i praksis - en introduktion for samfundsvidenskaberne*. 1st ed. Copenhagen: DJØF Forlag.

King, Gary, Michael Tomz, and Jason Wittenberg. 2000. "Making the Most of Statistical Analyses: Improving Interpretation and Presentation." *American Journal of Political Science* 44 (2): 341–55.

Ward, Michael D., and John S. Ahlquist. 2018. *Maximum Likelihood for Social Science: Strategies for Analysis*. Analytical Methods for Social Research. Cambridge: Cambridge University Press.