Statistical models beyond linear regression

Syllabus (Spring 2023)

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 (multinomial and ordered regression)
- count outcomes (poisson, negative binomial and hurdle models)
- event history data (survival models)
- hierarchical (nested) data (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 students with a mental map over different types of data and outcome 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

Time and place

The class runs twice per week. We also share the same Absalon page. You are in principle welcome to attend any of the two classes regardless of where you are enrolled. However, given that we will have group work and presentations, you need to stick to the class where you have enrolled for the activities.

- Thursday (week 6-13, 15-20) 0800-1000 in CSS 2-1-24
- Thursday (week 6-13, 15-20) 1000-1200 in CSS 7-0-34

Feedback is also given throughout the semester. Students will thus have the opportunity to work on the portfolio on which their grade is based during the semester. Specifically, students will provide and receive tips and comments that they can later incorporate in the own papers.

Class activities and compulsory contributions

The class is designed for group learning. The class activities are therefore designed to help you learn from each other while simultaneously preparing you for your own exam. 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.

We will divide our time between lectures (every other week) and data labs/seminars in which students get to apply what they have learned theoretically. At the beginning of the semester you will be divided up in groups.

Each group is responsible for communicating on behalf of the class the main insights about the topic we are working with that week. *The group will be the main contact point for the other students that need help.* I will also, of course, be available for assistance if the groups wants to touch base and/or need help. The communication will take place on the class' twitter account (@beyond_LM) as well as Absalon and in class. When you sign up for the groups, be aware that you will have a work-intensive week with tight deadlines. Note this in your calendar and set aside the time required.

The theoretical weeks will revolve around a plenary lecture. 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.

The group responsible this week (the "theory group"), will a) work together to make a twitter thread where you summarize the main insights. This thread may constitute a part of the executive summary for the portfolio exam. b) At the beginning of the seminar the next week, the group will make a short presentation where they also summarize what we learned last week. You will make your slides available on Absalon.

The lab weeks will revolve around the implementation of the models in R. At the end of the seminar, I will publish and explain the portfolio topic for the week. The students interested in getting feedback on their portfolio will hand in their work on Absalon before the next lecture.

The group responsible for the week (the "R group") will in the days following the seminar work through the replication part of the portfolio assignment and a) present their findings on Twitter (together with the graphics they produced) as well as b) sharing their code with the rest of the class on Absalon. Once again, the work provided will help the class prepare their exam, but also give the group members a flying start.

I expect the groups responsible for a topic (the theory and R groups) to collaborate and exchange to help each other out. The are also the main go-to for fellow students.

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 June* 1^{st} *2023.* I will apply the criteria specified in the following.

Examination form

The workflow, feedback and form of this portfolio are very specific and somewhat different from what you are used. *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 tweets and presentations in class.

2. Two replication studies/short analyses

The core of your portfolio are two short replication studies assigned from the topics covered over the semester. I assign several such homeworks over the semester, and you will choose to hand in *two* of those topics. 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. The questions for each study are assigned at the end of the relevant seminar. If you want feedback on your work before the final submission, you may hand in a draft on Absalon at midnight before the next lecture.

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 folder with your name on Absalon (files/Student files/Your name).

3. 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. A screenshot of your tweet(s)

You will document your Twitter activity on behalf of the class by providing a screenshot of the tweets you and your group were responsible for.

b. A link to group work on Absalon

If you are in a group that presents the theoretical intuition of the model, you will share a link to your slides on Absalon. If you are in a group that presents the empirical application of models, you will share a link to your R-codes used for the tweets (on Absalon).

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, apply them and discuss 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	09.02	Hermansen (2023), ch. 1-4, p. 19-70

Week	Topic	Date	Reading
2	Descriptive statistics and graphical display	15.02	Hermansen (2023), ch. 5-6, p. 73-119
3	Linear regression	23.02	Hermansen (2023), ch. 7-9, p. 123-194 Gelman and Hill (2007), ch 3-4, p. 29-79
4-5	Binary outcomes (logistic regression)	02.03; 09.03	Ward and Ahlquist (2018), ch. 3, p. 43-78 Ward and Ahlquist (2018), ch. 6, p. 119-132
6-7	Categorical outcomes (multinomial and ordered logistic regression)	09.03; 16.03	Ward and Ahlquist (2018), ch. 8-9, p. 141-189
	Assignment due	22.03	
8-9	Count outcomes (poisson, negative binomial and hurdle models)	23.03; 30.03	Linear Digressions: podcast on poisson distribution Ward and Ahlquist (2018), ch. 10, p. 190-216
	Assignment due	12.04	
10-11	Event history data (survival models)	13.04; 20.04	Ward and Ahlquist (2018), ch. 11, p. 190-216
	Assignment due	26.04	
12-13	Hierarchical data structures	27.04; 4.05	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
	Assignment due	10.05	
14-15	Missing data	11.05; 18.05	Ward and Ahlquist (2018), ch 12, p. 249-270 Gelman and Hill (2007), ch 25, p. 529-545
	Assignment due	24.05	

Week	Topic	Date	Reading
	Deadline portfolio exam	01.06	

Literature

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. https://www.djoef-forlag.dk/book-info/r-i-praksis.

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. https://doi.org/10.1017/9781316888544.