Discrete Choice Modelling Final Assignment

Richard Öhrvall and Benjamin Jarvis

2024-03-14

Task

In the final assignment, you will, on your own, apply discrete choice methods to answer a social scientific research question (or questions) pertaining to a real-world data set. You ***MUST*** *apply at least two of the three* modeling approaches presented during the course:

1) Binary logistic regression (or linear probability models)
2) Multinomial logistic regression
3) Conditional logistic regression

HOWEVER*, those who manage to successfully apply all three methods will be given extra consideration when determining their overall assignment grade.*

You are limited to using one of two possible data sets, which have been uploaded in Lisam along with this assignment. The data sets are as follows:

- **Eurovision Contest Results**: These data describe songs and voting in the Eurovision competition. They list and characterize the songs that made it into the semi-finals (starting in 2009), which songs advanced to the finals, and (starting in 2016) which songs were awarded points–including how many points they were awarded–by both juries and polling in other countries. These data are spread across a number of files and may require some reshaping, joining, and appending to make them suitable for analysis. More information about these data can be found at **this Kaggle data page**.

- **American Community Survey**: These data describe the characteristics of a large sample of households in the United States in 2015, including basic demographic information, data about educational outcomes, recent migration, marriage and fertility, place of residence and work, and more. These data are described in more detail at **the IPUMS website**.

While you must use one of the above as your main dataset, you are allowed to integrate supplementary data to enrich these main data. For example, you might link in data about the demographic characteristics of countries, or the climatological features of geographic regions. If you do link in additional data, these should be adequately described in your report (see below). Note: you are not required or expected to integrate supplementary data, except to the extent that you need those data to answer your research questions.

To complete the assignment you must:

1. Identify your own social scientific research questions that can be answered using the above methods and data. You should have a distinct questions that can be addressed with the distinct methods, but these questions should be clearly linked. For example, in an analysis

of voting behavior, one might first ask how left-right ideology is related to the probability of voting (which one could model with a binary logit), and second, how left-right ideology relates to the choice of what party to vote for (which one could model with a multinomial logit).

2. Carry out analysis using the appropriate analytic tools and techniques in `R`. You should embed the code that you use to perform your analysis in a Quarto or R Markdown file. It should be possible for the instructors to run your code on their own computers with few or no modifications.
3. Write a report using Quarto or R Markdown, but compiled into a "clean" PDF that outlines our research questions, proposes hypotheses, presents results, and makes social scientific inferences related to your hypotheses. More on the structure of the report is below.

Submission and Deadline

You should submit your final assignment as both a Quarto (or R Markdown) file and a knitted PDF file in Lisam. If any supplemental files are needed in order to knit the pdf, those should also be submitted, e.g. bib files for references, external images, etc. The PDF should not include any R code or any console output.

The deadline for submitting the final assignment is Thursday, March 28 at 18.00. Do not wait until the last minute to submit. If you have problems with Lisam, e-mail the assignment before the deadline. **Late submissions will not be accepted**.

On Plagiarism and Large Language Models (LLMs)

Note that this is an individual assignment and LiU takes plagiarism very seriously. Plagiarism can include handing in an assignment that is, in whole or in part, identical to another person's work, not citing other peoples' texts and ideas properly, and more. Please consult LiU's guidelines at https://liu.se/en/article/plagiering-upphovsratt if you have any questions about what constitutes academically honest conduct.

We advise against using ChatGPT or other LLMs in this assignment, except perhaps to get advice with specific coding issues or to correct your drafts for grammar and clarity. In our experience, LLMs are still not capable of integrating together research questions, relevant social scientific literature, meta-data about tables and variables, and task-appropriate analytic techniques. However, if you use ChatGPT or another LLM to help you, you must provide **all relevant chat logs** along with your submission. These should make it clear what your contributions were to the final product. Please also indicate clearly in your report if and how you used an LLM, if you did so.

Report Content and Structure

Try to find a concise, relevant title for your research paper. The paper can follow any outline you prefer, but it should be structured in a way that makes it easy for the reader to follow. You could take inspiration from the articles we have discussed in the seminars. One potential structure is presented below, including suggestions about what kinds of content should go into each section:

- Introduction
  - Frame and motivate your study.
  - Enumerate your research question(s).
  - Provide an brief overview of your data and methods.

    - Preview the key findings.
- Background
    - Review (some) relevant empirical and/or theoretical literature.[1]
    - Provide essential information about your specific case (i.e., the social context captured in your data).
    - Develop hypotheses.
    - Discuss potential confounders and/or mediators.
- Data:
    - Identify the data source.
    - Describe how you generated your analytic sample from the data.
    - Explain how you treated missing values.
    - Clearly identify the dependent variable(s).
    - Identify and explain the operationalization of key independent variables.
    - Produce and discuss descriptive statistics.
- Methods:
    - Present your modeling strategy.
    - Indicate what statistical models you used and for which dependent variables.
    - Explain your model specification, including any (non-linear) variable transformations and/or interactions.
- Results:
    - Present your findings w/ reference to appropriate tables and figures.
    - Discuss fit statistics.
    - Interpret key coefficients in terms of odds, odds ratios, etc.
    - Cite key statistical hypothesis tests and related p-values.
    - Illustrate and discuss relevant predicted probabilities and/or marginal effects.
- Discussion/Conclusion:
    - Recapitulate your research questions and hypotheses.
    - Compare the relevant empirical results to your expectations.
    - Reflect on any theoretical implications of your findings.
    - Discuss potential limitations of your study and suggestions for future analyses.

Your research report should be no more than 3,000 words, but we strongly prefer shorter reports closer to 2,000 words. This word count does not include the the tables or bibliography. Please make sure that your paper is within the maximum limit – it could affect your grade.

In addition to your text, your paper should include an appropriate number of comprehensible, well-labelled tables and figures. These should be numbered and appropriately cited in the text. You should only use as many tables and/or figures as needed to describe your data and convey your results. Use a “less is more” approach here. The combined total number of graphs and tables should not exceed six (6). This includes any tables or graphs presented in an appendix at the end of the paper (if you have an appendix).

References and Bibliography

You should use references to show which previous works your own study is drawing on. You are free to choose the reference style you prefer, but you should be consistent throughout the paper. We

[1] Given the time limit for the assignment, we do not expect a full-fledged literature review, but you should be able to find some relevant references that could help you frame your study, develop theoretical expectations, present prior empirical findings, describe your study context, interpret your results, etc.

recommend, however, the Harvard system, i.e. author-year.

The paper should include a bibliography. You could freely choose the style of the bibliography, but it should be consistent. You could check other articles to get some ideas.

More information about references and bibliographies can also be found on LiU's website, https://liu.se/en/article/citeringsteknik.

Grading

The grading will to a large extent be based on the appropriate model choice, estimation, presentation and interpretation of the results. However, other aspects, such as the framing of the topic, positioning in relation to previous literature, description of the data, conclusions, clarity of the language, style, formatting, etc., will also be taken into account.

The final assignment will graded on a scale from A to F, in the following way.

A – Excellent (note: use of all three methods is required)

B – Very Good

C – Good

D – Satisfactory

E – Sufficient

Fx – Insufficient, some additional work required.

F – Insufficient, considerably more work required.

The grade Fx means that you can revise the final assignment and submit it again and possibly reach a sufficient grade, but you could also choose to submit a new final assignment as part of a re-examination instead. The grade F means that you have to do the re-examination in order to have an opportunity to pass.

Re-examination

There will be two opportunities for re-examination. The dates and timings for those re-examinations will be communicated at the conclusion of the course, but will most likely be timed for the beginning of the summer (just after the last week of the Spring Term) and the end of the summer (just before the first week of the Fall Term). Any subsequent re-examinations will be timed to align with final examinations in future offerings of this course. In the event of re-examination, the instructor(s) reserve the right to assign you the data set and research questions, meaning there may be no flexibility to determine your own topic. To a near certainty, re-exams will make use of entirely different data sets, so as not to provide an unfair advantage to those who do not pass or submit earlier exams.