E401/M518: Empirical Challenge

Review of Linear Regression - Part 1

Fall 2023

September 2023

Please work on this challenge with a partner. All challenges are based on real-world data and are similar to what you might encounter in your future career. The main purpose of this challenge is not on the application, but on getting you (and your class mates) more familiar with applying the techniques that we discussed in the lecture. Nevertheless, whatever you present should make economic sense and if it doesn't you should think about what could be going wrong with the data and/or your analysis. Please keep in mind that these are (potentially dirty) real-world data and I haven't checked every detail of it. Therefore, you are likely to run into a lot of problems. I strongly encourage you to come to my office hour to discuss any issues as well as your overall plan for your presentation a few days before the respective class. There is always a risk that there is not much interesting in your data set. As long as you are able to clearly document what you tried and have some conjecture/explanation for why you get the results you get, this is totally fine. It's very likely that you will be in similar situations regularly when taking a job as a data scientist. I designed this challenge to be pretty open-ended on purpose. When diving into the data you may find aspects that are totally different from what I had in mind. This is totally fine and another likely outcome in data science projects.

You are expected to give a presentation of roughly 25-30 minutes in class. Think of this presentation as one you would give to your boss or at a board meeting of a company or policy institution that hired you as a data scientist. Other students should think of themselves as board members who attend your presentation and are strongly encouraged to ask critical questions about your analysis and you should be prepared to answer them. Your presentation should contain the following elements: (1) a brief discussion of the data, i.e., where is it coming from, what are the most important variables, what is the unit of observation, what concerns do you have about the quality of the data etc., (2) the big picture business or policy question that you are trying to address with these data (other students have not necessarily read the questions in advance), (3) overview of the methodology that you used to answer the question, (4) your empirical results, (5) discussion of the results, policy implications, and potential caveats and suggestions for further steps. Lastly, this is not a presentation class, so don't invest in fancy PowerPoint slides! Having prepared a RScript in RStudio that generates all your results as we click through it is totally fine! However, I ask you to only work with code scripts. Avoid manual manipulation or loading of the data from a graphical interface at

all costs!

Main Techniques

In this challenge I will ask you to work mostly with linear regression models.

Data

In this challenge you will use data on a sample of Portuguese secondary school students (which we will use again in the challenge on cross validation). The data set includes demographic characteristics, test scores and alcohol consumption of each student. The data come from the UCI Machine Learning Repository and can be found on this website: https://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION .

The data consists of two files, one containing performance scores in math courses (`mat`) and one containing test scores in Portuguese language courses (`por`).

The variable definitions are as follows:

1. `school` - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2. `sex` - student's sex (binary: 'F' - female or 'M' - male)
3. `age` - student's age (numeric: from 15 to 22)
4. `address` - student's home address type (binary: 'U' - urban or 'R' - rural)
5. `famsize` - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6. `Pstatus` - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7. `Medu` - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
8. `Fedu` - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 – 5th to 9th grade, 3 – secondary education or 4 – higher education)
9. `Mjob` - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10. `Fjob` - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11. `reason` - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12. `guardian` - student's guardian (nominal: 'mother', 'father' or 'other')
13. `traveltime` - home to school travel time (numeric: 1 - 1 hour)
14. `studytime` - weekly study time (numeric: 1 - 10 hours)
15. `failures` - number of past class failures (numeric: n if 1<=n<3, else 4)
16. `schoolsup` - extra educational support (binary: yes or no)
17. `famsup` - family educational support (binary: yes or no)
18. `paid` - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19. `activities` - extra-curricular activities (binary: yes or no)
20. `nursery` - attended nursery school (binary: yes or no)

21. `higher` - wants to take higher education (binary: yes or no)
22. `internet` - Internet access at home (binary: yes or no)
23. `romantic` - with a romantic relationship (binary: yes or no)
24. `famrel` - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25. `freetime` - free time after school (numeric: from 1 - very low to 5 - very high)
26. `goout` - going out with friends (numeric: from 1 - very low to 5 - very high)
27. `Dalc` - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28. `Walc` - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29. `health` - current health status (numeric: from 1 - very bad to 5 - very good)
30. `absences` - number of school absences (numeric: from 0 to 93)

Several grades are recorded for each subject (Math or Portuguese):

1. `G1` - first period grade (numeric: from 0 to 20)
2. `G2` - second period grade (numeric: from 0 to 20)
3. `G3` - final grade (numeric: from 0 to 20, output target)

Before you run any analysis, make sure you familiarize yourself with the data and examine its quality. Briefly mention in your presentation, if some features look dubious to you.

Policy question

The Portuguese Ministry of Health has hired you to learn more about what determines students' success in school (measured by their test scores in Math and Portuguese). They are particularly worried about the recent rise in alcohol consumption among adolescents. The group of officials you are working for is very enthusiastic about using econometrics to answer important policy questions. However, they only know linear regression models. So that's why they ask you to use for your policy recommendation. Specifically, they are interested in the following:

- Which student characteristics have the biggest effects on test scores? Are there differences between Math and Portuguese grades?
- Most importantly, does alcohol consumption have a significant effect on student performance?
- Independently of students' grades, the Ministry is concerned with the health risks associated with an increase in alcohol consumption among adolescents. They would like to actively campaign against alcohol consumption, but they have only very limited resources. Can you develop a (linear regression) model to inform them which demographic groups exhibit the highest alcohol consumption? This could inform them which groups to target most.

After being positively surprised by the politicians' appreciation of formal econometrics and the quality of the data they provided you with, you get to work:

1. Try to answer the Ministry's questions using a series of linear regression models. Remember that you only have 30-40 minutes including questions; therefore, you may not get to answer all of their questions. Your presentation should be flexible enough to

accommodate running out of time, but you should also be able to say at least a little bit about each of their inquiries. You never know what they might ask about ...

2. For each question, start with a simple model that contains only one regressor. Think about and justify which variable you would choose for this. Then, work your way up to more complex models by including more regressors. Clearly explain your model specification and how you interpret the results. As usual, be very specific! Any time you talk about an econometric analysis you should be prepared to answer question about any of the numbers on the screen.
3. What would your final policy recommendation be? Is there a clear effect of alcohol consumption on grades? Which groups are particularly prone to high alcohol consumption? Explain clearly how you arrived at your conclusions and what some of its limitations are. What would be the most important you need to obtain or do in order to overcome some of the limitations?
4. Finally, think about what else the Ministry of Health might be able to learn from the data. What future steps would you suggest to take?