STAT230A: Final Project Proposal

Irina Degtiareva, Dmitrii Timoshenko

2023-04-10

Introduction

While significant attention has been given to understanding the determinants of educational outcomes in recent years, it remains an important research area for social science policymakers.

In the context of our final project, we are conducting a comprehensive investigation into the causal impact of students' intentions to pursue higher education on their academic performance. Our research aims to explore the complex interplay between various demographic, social, and economic factors and their interactions with the aforementioned causal effect.

The current body of literature on this subject presents conflicting hypotheses. One perspective suggests that students' intentions on getting higher education are merely a reflection of their current academic performance, without any tangible impact on their future efforts or opportunities. In contrast, an alternative perspective suggests that these expectations may serve as a cognitive driver, directing individual behavior towards the attainment of educational goals through preparatory commitment and dedicated effort.

We anticipate that our findings will contribute to a deeper understanding of the mechanisms that underlie academic success and inform policies and interventions aimed at improving educational outcomes.

To interpret causal effects, we make the crucial assumption that all relevant factors that may influence the outcome, in this case, students' final grades, have been fully accounted for in our analysis. Additionally, to assess the causal impact of this factor on academic performance, we must assume that there has been no external intervention that may influence the outcome, such as changes in teaching methods or curriculum, that are unrelated to students' higher education plans. This assumption is necessary to ensure that any observed effects are solely attributable to the factor of interest, rather than other external factors that may have influenced the outcome.

Data

We will work with Student Performance Dataset from UC Irvine Machine learning Repositiory [1]. The data attributes include student grades, demographic, social and school related features and it was collected by using school reports and questionnaires. Two datasets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). We have different number of observations in two subjects – 395 for mat and 649 for por –, because not every student took Math exam. A full table of features is available in the Appendix 1.

The following code helps to import data to R:

```
data_mat <- read.csv('data/student-mat.csv', sep=";")
head(data_mat, 1)
```

```
##   school sex age address famsize Pstatus Medu Fedu    Mjob    Fjob reason
## 1     GP   F  18       U     GT3       A    4    4 at_home teacher course
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
```

[1] Data can be found **here**

```
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
##   absences G1 G2 G3
## 1        6  5  6  6
```

```
data_por <- read.csv('data/student-por.csv', sep=";")
head(data_por, 1)
```

```
##   school sex age address famsize Pstatus Medu Fedu    Mjob    Fjob reason
## 1     GP   F  18       U     GT3       A    4    4 at_home teacher course
##   guardian traveltime studytime failures schoolsup famsup paid activities
## 1   mother          2         2        0       yes     no   no         no
##   nursery higher internet romantic famrel freetime goout Dalc Walc health
## 1     yes    yes       no       no      4        3     4    1    1      3
##   absences G1 G2 G3
## 1        4  0 11 11
```

In the data we have a lot of categorical features which we will encode using one-hot encoder. The following dataset contains everything what we need to answer the initial research question.

Methodology

First, we plan to run a regression analysis with an outcome variable `G3` - Mathematics final grade. There are two additional variables, `G1` and `G2`, that correspond to the first and second semester grades, respectively. These variables are strongly correlated with `G3`. We will discard these variables and focus on social and economic factors.

Models

To answer the initial question, we will evaluate several models:

- OLS regression
- LASSO regression
- Ridge regression
- Grouped LASSO regression

We make an assumption that we have enough data points to make a statistical inference on our coefficients, obtained by models. Due to the large number of covariates, we can expect better robustness of coefficients with penalized models. We expect that the Grouped LASSO model will demonstrate better interpretability, as we want to see the same penalties for different levels of some factors.

We also plan to run a Chow test to see if there are any differences in coefficients among the subjects. Probably, we don't need to fit two different models on subjects and can combine data from both.

Finally, we want to evaluate different interactions between covariates. Based on significance of coefficients we will give an interpretation of the interaction terms. This analysis will help us to determine what additional effort students need to attain to succeed in school.

Model Selection

To select our final model, we propose the following:

- Choose RSS, R^2-adjusted, AIC, BIC as a target metrics
- Forward/backward/mixed selection methods
- Select the final model using mentioned metrics and methods

The final decision on the model will depend on the best metric among the fitted models and its interpretability. For LASSO and Ridge regression, we will use K-fold cross-validation technique.

If we have enough time, we also plan to fit a logistic regression model to predict the probability of passing the exams (`G3` > 10).

Appendix

Variable	Meaning
sex	student's sex (binary: 'F' - female or 'M' - male)
age	student's age (numeric: from 15 to 22)
address	student's home address type (binary: 'U' - urban or 'R' - rural)
famsize	family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
Medu	mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
Fedu	father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
reason	reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
traveltime	home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
studytime	weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
failures	number of past class failures (numeric: n if 1<=n<3, else 4)
schoolsup	extra educational support (binary: yes or no)
famsup	family educational support (binary: yes or no)
paid	extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
activities	extra-curricular activities (binary: yes or no)
nursery	attended nursery school (binary: yes or no)
higher	wants to take higher education (binary: yes or no)
internet	Internet access at home (binary: yes or no)
romantic	with a romantic relationship (binary: yes or no)
famrel	quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
freetime	free time after school (numeric: from 1 - very low to 5 - very high)
goout	going out with friends (numeric: from 1 - very low to 5 - very high)
Dalc	workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
Walc	weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
health	current health status (numeric: from 1 - very bad to 5 - very good)
absences	number of school absences (numeric: from 0 to 93)
G1	first period grade (numeric: from 0 to 20)
G2	second period grade (numeric: from 0 to 20)
G3	**final grade** (numeric: from 0 to 20, output target)

Citations

1. Stephan Dochow, Sebastian Neumeyer. An investigation of the causal effect of educational expectations on school performance. Behavioral consequences, time-stable confounding, or reciprocal causality?. Research in Social Stratification and Mobility (2021), 100579, ISSN 0276-5624, https://doi.org/10.1016/j.rssm.2020.100579.

2. P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

3. Stinebrickner Ralph & Stinebrickner Todd R., 2008. "The Causal Effect of Studying on Academic

Performance," The B.E. Journal of Economic Analysis & Policy, De Gruyter, vol. 8(1), pages 1-55