School Attendance in Mexico

Group No. 5

March 11, 2019

Members:

- Hector Hernandez Heimpel (NetID: hh744) - email: hh744@georgetown.edu
- Tingjie Meng (NetID: tm1305) - email: tm1305@georgetown.edu
- Liumin Chen (NetID: lc1077) - email: lc1077@georgetown.edu

Github Repository: https://github.com/hheimpel/Group-No.-5

Problem Statement

We will undertake two different but related questions in this project. First, we will make use of machine learning techniques to try to establish which covariates predict school attendance for children under 15 years of age in Mexico. A good model could be useful to policy decision makers because in the abscence of information regarding school attendance, other variables could be used to try an predict school attendance. The second question we are interested in is establishing the effect of access to low-income financial services on school attendance. For this, we will make use of an event which happened in Mexico in 2002 where a bank called Banco Azteca opened up a plethora of bank branches throughout Mexico. The services Banco Azteca provides include low-income financial services (which were the first of it's kind in Mexico) and we assume that proximity to Banco Azteca branches is equivalent to having access to these services. The effect of access to these services could be important to policy decision makers because a proper subsidy or tax incentive could be given to banks which offer these services in communities with low attendance to school.

Data Analysis and Manipulation

There are multiple data sets which will comprise the source of our analysis for the prediction problem (Question 1). The first group of data sets is contained in the Control Book for 2002 in the Mexican Family Life Survey (ENNVIH). The data sets contained within this book contain household level characteristics and they are divided into two categories. The first category contains characteristics observed in the household by the interviewer (i.e. proximity to water source, type of floor, etc.) and the second category contains answers to individual level survey questions (i.e. marital status, maximum education grade, etc.). The second group of data sets are contained in Book V for 2002 in the Mexican Family Life Survey. These group of datasets contain information about characteristics of children under 15 years of age, they include but are not limited to, child employment, child health characteristics and more. The third group of data sets is contained in the Bank Operational Information from the

Comision Nacional Bancaria y de Valores (CNBV). This data set contains information about the locations of bank branches in 2002 for Mexico (including Banco Azteca).

The data sets previously described contain a lot of problems which means we are going to have to make use of several tools to prepare them for analysis; for example, missing data seems to be a substantial problem in our data sets, we have to be careful in dealing with missing values because they might represent different things. In addition to R's `base` package, we will have manipulate the data extensively with the `tidyverse` and the `recipes` package in order to prepare it for analysis and the machine learning process. The data sets also contain problems with strings, in order to be able to join the different data sets we will have to manipulate strings. Furthermore, plenty of exploratory analysis will also be required to check for patterns of missing data, outliers, variance and covariance amongst different variables. Our main tool for this purpose will be the package `ggplot2` which will provide graphics to easily and rapidly identify problems and/or patterns in the data. For the second research question we will use the same data sets for year 2005.

Testable Hypothesis

Using machine learning techniques we will try to find the best model for prediction of school attendance using the previously described data sets. Our goodness of fit measure will be the Mean Squared Predition Error (MSPE) and we will choose the model which produces the lowest one. To estimate the impact of low-income financial services on school attendance we will mainly be interested in the coefficient in the difference in differences regression model which measures said impact. Our main hypothesis in this case is that access to low-income financial services will increase school attendance; in our Difference in Differences strategy municipalities with no Banco Azteca branches will be our control group and municipalities with Banco Azteca will be our treatment group.

For both of these questions it is important to state that the outcome variable has not yet been chosen. Our data sets contain variables related to school attendance such as hours spent doing school work, days missed from school in the school year, whether the child is actively subscribed to school in that year, amongst others. We will run different models for both questions and assess which outcome variable provides the best outcomes in terms of goodness of fit for question 1 and statistical significance and non-bias for question 2.

Main Products

The main output from our analysis will be for research Question 1 the table containing the model, its coefficients and corresponding MSPE. We will also use `ggplot2` graphics that help us identify the most important systematic patterns in the data; these include a density map to describe the distribution of banks geographically, a bar chart showing the number of banks in each municipality, a stacked bar chart presenting the percentage of students who chose to attend schools and who did not in each municipality, among others. The main output for Question 2 will be the values of the coefficients for the covariates in the models, their statistical significance and other important statistics.

References

Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Max Kuhn and Hadley Wickham (2018). recipes: Preprocessing Tools to Create Design Matrices. R package version 0.1.4. https://CRAN.R-project.org/package=recipes