Impact of Financial Access on School Attendance

Group No. 5

February 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

Education has long been considered important to economic growth. The general idea behind this is that people with higher education have higher income, as evidenced by Ashenfelter (1997). There are many factors that could affect why a child attends or does not attend school. In particular, financial access could affect schooling through access to credit and savings with better conditions. By limiting the incentives and capacity to invest in human capital, credit constraints play an important role in determining aggregate productivity, national income distributions, social mobility, and economic growth and development Becker (1975). According to Aportela (1998) it was found that access to financial services increases savings on low-income people. What this could imply is that with better access to financial instruments that allow people to save money, a household head may better prepare for the expenses that will have to be made when the child goes to school.

The question we seek to answer is if the mass introduction in 2002 of a bank (Banco Azteca) which offered new financial services to communities in Mexico which previously had no access to these particular banking services, affected the attendance to school of students in those communities. The intention of the project is to establish the relationship between attendance to school by children under 15 years of age and access to a banking service in Mexico. To establish the relationship we will make use of the appropriate econometric tools and empirical strategy. The relevance of this question is given by the likelihood of productivity being affected as well. Even if having access to banking services does not affect school attendance, the question would still be relevant to policy making since it could be a dimension for the decision of the policy maker.

Data

Our empirical strategy will be a difference in differences regression analysis for which we need panel data containing characteristics of mexican families in 2002, previous to the launch of Banco Azteca, and in 2005 after the launch of Banco Azteca; we will further need data detailing the municipalities where Banco Azteca opened branches. For this purpose, We will use two sources for our data. The first is the ENNVIH panel data which contains several characteristics of mexican families and is publicly available in the Mexican Family Life Survey website [1]. The second dataset which contains the list of bank branches in different municipalities in Mexico at different points in time, including Banco Azteca, and is publicly available from the Comision Nacional Bancaria y de Valores (CBNV) website. [2]

The data gathered from the Mexican Family Life Survey (MXFLS) website details several aspects of a Mexican Family displayed throughout different categorized books. Among these details of mexican families, several variables related to school attendance can be found. The Mexican Family Life Survey consists of several books of panel data which are divided into two major categories: household characteristics and individual characteristics. These categories are consequently subdivided into several books detailing different aspects, for example, control, economy, consumption, reproductive health among others. Within these particular books, there are different questionnaires fitting into each category.

Since Banco Azteca started operations on several locations on December 2002, the basis for choosing the year of the dataset has to consider that a period before that is needed. The next dataset will be for the year 2005-2006 and the reason 2005- 2006 is chosen is because it is the next immediate dataset available. If datasets from later years were chosen, more things would likely need to be controlled for because other policies could have come into play.

Potential Issues

The data from the Mexican Family Life Survey is in .dta form, however this should not pose a significant problem to the overall project; some data tidying and arranging will have to be done but the information is easily accesible. On the other hand, data from the Comision Nacional Bancaria y de Valores is in .xls form and requires some editing and cleansing to tidy the data set. This should not pose a significant problem either since the data is easily accessible as well and once the datasets have been properly arranged, analysis should be relatively easy to perform..

[1] http://www.ennvih-mxfls.org/english/ennvih-1.html

[2] http://portafoliodeinformacion.cnbv.gob.mx/bm1/Paginas/infhistoriaoperativa.aspx

References

- Aportela, Fernando, *Effects of Financial Access on Savings by Low-Income People*, Mimeo, MIT, 1998.
- Ashenfelter, Orley and Cecilia Rouse, *Income, Schooling, and Ability: Evidence from a New Sample of Identical Twins*, Quarterly Journal of Economics, Vol. 113, p. 253-284, 1997.
- Becker, Gary, *Human Capital*, 2nd ed., Columbia University Press, New York, NY, 1975.

Loading the Household Level Data for 2002

We start by loading the control book for 2002 (M X F L S - Book C). This book contains several control characteristics at an individual level and at a household level. Within the book there are four pertinent data sets to be used:

1. `c_portad.dta`: contains the location of the households by State, Municipality and Locality. It further contains a variable indicating number of inhabitants in Locality. This is a Household level data set.
2. `c_cv.dta`: contains physical characteristics of the household (i.e. whether it has a cooking room, a telephone, etc.). This is a Household level data set.
3. `c_cvo.dta`: this data set extends the previous one with more physical characteristics of the households (i.e. the construcction materials, access to electricity, etc.). This is a household level data set.
4. `c_ls.dta`: it contains control characteristics such as income, level of education, gender, among others. This is an individual level data set.

Removing non-informative variables in Book C

Once the data sets were loaded we proceeded to eliminate several non-informative variables. Some of them were redundant variables such as more identification characteristics for the household members, other variables seemed to ask the same thing but with a slightly different wording. For more details please check the Codebook for book C annexed in the project files.

Recoding variables in Household Level Data for 2002

The next step consisted of recoding the values of certain variables to reflect the different categories as the codebook instructed. The codebook provided numeric values to categorical variables which could be slightly confusing, we therefore recoded the values to reflect in a string what they actually meant. Additionally, we recoded dummy variables to be equal to one if the condition was true and zero otherwise (they were previously coded as threes as opposed to zeros); lastly, we created a resumed categorical variable for the size of the locality, if population was bigger thatn 100 thousand we categorized as `urban` and `rural` otherwise.

Dropping variables with excessive missing values

In many of the books many missing values for certain variables were observed. We created a data frame per data se to show us how many missing values each variable had and kept only the variables which didn't have an excessive amount of missing values. The first data set we noticed a missing values problem was `c_cv` as the following results show:

The second data set where we applied the same logic was `c_cv`:

We remove the variables which have over forty percent of missing values in these data sets.

Merging Household level data, recoding and transforming book C data

Once we kept the variables with no issues in all the data sets previously mentioned we merged the data sets into a master data set. It's important to mention that the merging was done by the variable `folio` which is the individual household id; by doing so, there were several household level characteristics which were attributed to an individual level. Observations of the households contruction materials for example will be the same for each individual member of the household. Finally, some recoding was done to as before to reflect the true categoreis of a variable with strings and dummies were appropriately transformed.

It is worth mentioning the transformation of three individual level variables to household level variables. The first variable measured the highest education grade the household member had attended; since the focus will be children between 6 and 15 years of age we transformed the variable to show the average **adult** household member education. It seems likely that children whose accompanying adult household memberhave a higher education level will be more likely to have more education.

The second variable we transformed was income. Since income only reflected the individual level of income and since many children do not have income themselves we transformed the variable to show the average income per household. Intuitively, if the aggregated or average income per household is bigger then perhaps children have more opportunities to attend school. It would have seemed naive to impute missing values of income for children with things other than a household level transformation of the individual level variable income.

The third variable we transformed was relationship status. Since it is safe to assume that most children under 15 are single we decided to create a household level variable which reflected the relationship status of the household head. The intuition behind this is that perhaps there is some correlation between having married figures as parents or household heads.

Once we created all the household level transformations to the mentioned variables, we filtered the data to only include children older than 5 and younger than 15.

Child Individual Data

We then proceeded to load individual level data for children by loading the book V for 2002 (M X F L S - Book V). This book contains several characteristics at an individual level for children under 15 yearsof age. Within the book there are six pertinent data sets to be used:

1. `v_edna.dta`: contains information regarding children's education.
2. `v_emn.dta`: contains information regarding child labor.
3. `v_cen.dta`: contains information regarding outpatient utilization for children.
4. `v_atn.dta`: contains information regarding time allocation for children.
5. `v_esn.dta`: contains information regarding overall children's health.
6. `v_hsn.dta`: contains information regarding inpatient utilization for children.

Exploring missing values

These data sets contained many varibles with excessive amounts of missing values. Once again we went through all data sets to identify variables which had excesssive amounts of missing values.

Recoding and merging to prepare master data set for children

Once we identified the variables with many missing values we then proceeded to discard redundant variables and to merge all of the data sets into a master data set. Categorical variables were also transformed to show the actual string category they represent as opposed to numeric value.

Recoding dummies

The last step consisted of recoding dummy variables to have a value of one if condition was true and zero otherwise.

Merging Book V and Book C master data sets

To have the working master data set we merge both of the master data sets from book C and book V.

THE FOLLOWING IS NOT READY, WE HAVE TO SEE IF IT'S WORTH IT