**3.4. Conclusion and Hypotheses**

**Hypothesis 1**

The highest level of education of the parents of a school dropout is lower than the highest level of education of the parents of an enrolled student.

**Hypothesis 1.1**

Lower levels of school dropouts can be related to higher levels of education of the parents.

**Hypothesis 2**

Structure of single-parent families have higher level of school dropouts.

**Hypothesis 3**

Households with higher socioeconomic status have lower level of school dropouts.

**CHAPTER 4**

**4.1. Data**

The data for this analysis come from the Colombian Longitudinal Survey of Los Andes University (ELCA) which follows around 10,000 Colombian households in rural and urban areas every three years collecting information about education, income, household composition, social capital, and health. The ELCA survey has been conducted in 2010, 2013, and 2016 and applied to 10,800 households (6,000 urban and 4,800 rural). The empirical analysis of the present research uses the three waves of the urban sample which is representative of households from stratum one to four at the national level and five regions of the country: Bogota, Central, Eastern, Atlantic, and Pacific. I concentrated in the questions regarding education, socioeconomic status, and structure of the household. Furthermore, it is important to mention that other data sources more suitable to analyze school dropouts like the SIMAT database of the Ministry of Education, which provides information of the whole education trajectory of children studying in public schools, and the SIMPADE database of the Secretary of Education of Bogota, which has information of all students enrolled in public schools in the city and the socioeconomic background of their parents, were not available to researchers as the databases contain sensible data of the children and observations are not anonymized.

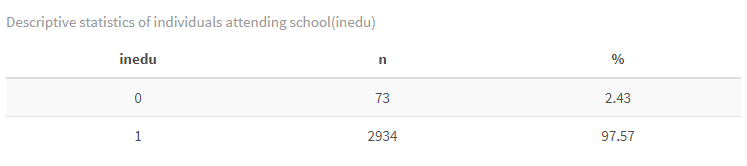
**4.2. Analytical sample**

The first wave in 2010 has an original sample of 22,179 observations, the age of the interviewees range between 4 to 60 years. The second wave in 2013 has an original sample of 20,574, the age of the interviewees range between 2 to 100 years. The third wave in 2016 has an original sample of 19,298, the age of the interviewees range between 5 to 96. To construct the analytical sample the three waves were joined, the respondents with missing values in the attendance to school and parent educational attainment variables were omitted, and the analysis was restricted to the age range between 5 and 16 years which corresponds to the compulsory education in Colombia. Thus, the analytical sample has a size of 3,007 observations and consists of boys and girls between 6 and 16 years of urban areas of five regions of Colombia.

**4.3. Analytical sample and sample statistics**

**4.3.1. Dependent variable**

The dependent variable is the school dropout rate. This study will define school dropout as the student that is not part of the education system during compulsory education years, between 5 and 16 years. It will be assumed that if the respondent belongs to this age range and at the time of the interview is not studying, then he or she has dropped out of school. The variable is operationalized through an education question that asks the children of the household if they are attending school at the moment of the survey. If the individual answered affirmatively, it is considered that the individual is enrolled in a school and the variable takes the value 1. In contrast, if the individual answered negatively, it is considered that the individual dropped out of school and the variable takes the value of 0.



The distribution of the variable shows that most children of the analytical sample are enrolled in school 97.5%, compared to 2.4% that dropped out.

Here I think the variation is too low so I don´t know if it is possible to do the analysis with this sample (I have a plan b for the definition of dropout that increases the sample, I develop this in the next section)

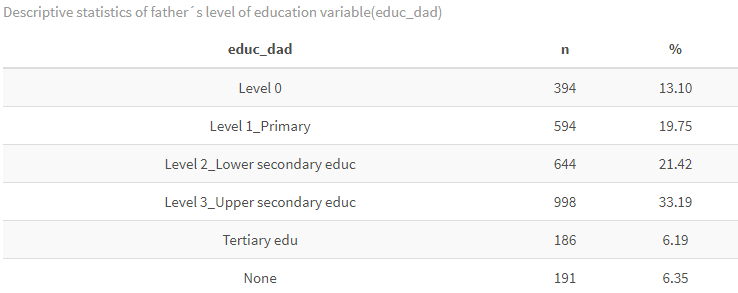
The sample is drastically reduced when I remove the NA´s from the father´s education variable, before this I have a sample of 10,358, and a variation for the inedu variable of 0 = 27,736, 1 = 13,462

**4.3.2. Independent variables**

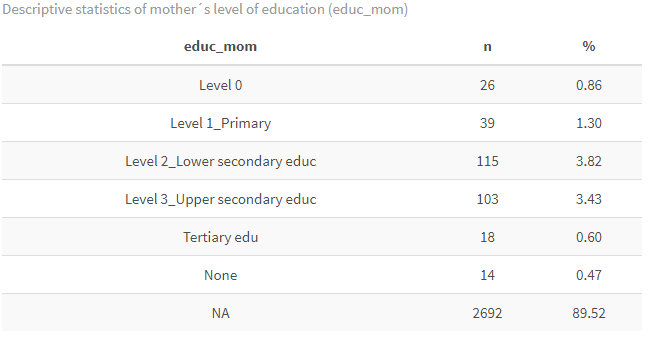
The main variables of interest are parent´s education level, family structure and socioeconomic status.

**Parent´s education**

The survey contains information of the education of mothers and the education of fathers. Individuals were asked what is the highest educational level attained by their mother and the father. This variable contains six categories: level 0, for some years of primary; level 1-Primary, for all primary school; level 2 -Lower secondary education, for some years of secondary education; level 3- Upper secondary education, for all secondary education; level 4- Post secondary non-tertiary education, for one or more years of technical and university without degree; and level 5- Tertiary education, for university with degree and postgrad. The levels of ELCA were classified in these categories based on the International Standard Classification of Education -ISCED (add SOURCE).



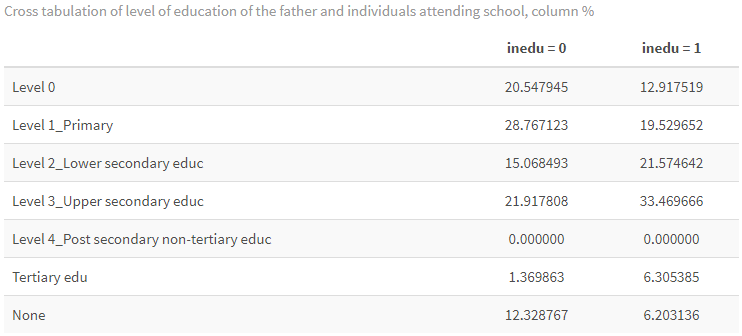
The highest level of education of fathers in the sample is upper secondary education with 33.2%, and lower secondary education with 21.4%. Only 6.2% of the fathers have achieved tertiary education.



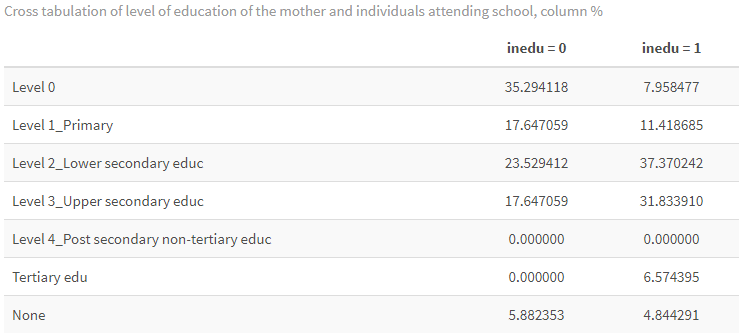
On the other hand, the education of the mothers does not provide enough information, it is possible to observe that the highest educational attainment for mothers is lower secondary education with 3.4%, but 89% of the observations are NA´s.

I kept the NA´s because the sample was drastically reduced if I remove them, so I don´t know if a should only consider the father´s information on this variable.

**Cross tables parent´s education and dependent variable (inedu)**



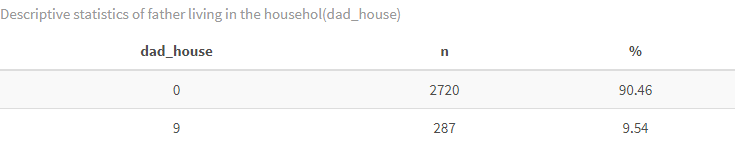
The table shows that 28% of students that dropped out of school (inedu = 0) have fathers in level 1- Primary, follow by 20% with fathers in level 0. In contrast, 33.5% of students enrolled in school have fathers in level 3- Upper secondary education and 21.6% have fathers in level 2-Lower secondary education. The share of students that dropped out of school is higher when the fathers have none education 12.3% in contrast with 6.2% for enrolled students, and when the fathers have level 0, 20.5% compared to 13%. Even though, the level of education of the father for both groups is not too high, it is possible to observe that for the students that dropped out the highest level of education of the fathers is lower than the level of education of the fathers for students that are enrolled in school.



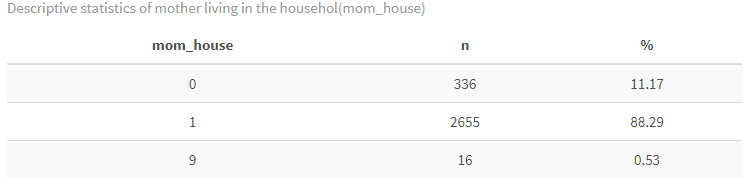
Regarding the mother´s education, 35% of students that dropped out of school have mothers in level 0 and 23% in level 2- Lower secondary education. For students enrolled 37.3% have mothers in level 2 and 32% in level 3. For students with mothers in level 0 the share of dropouts is considerably higher than for enrolled students, 35.3% to 8%.

**Family structure**

To analyze family structure, the survey asks the interviewees if the father and mother lives in the household. The variable has three levels, if the person lives in the household the variable takes the value 1, if the person does not live in the household the variable takes the value 0, and if the person died the variable takes the value 9.

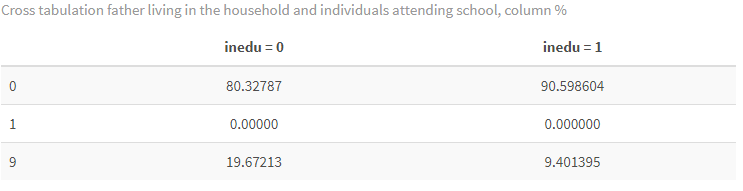


Observing the information of fathers there are no observations in the level 1, so apparently 90% of fathers in the sample do not live in the household and 9.5% died.



In contrast, 88.2% of the mothers live in the household, 11% do not live in the household and only 0.5% died.

**Cross tables family structure and dependent variable (inedu)**



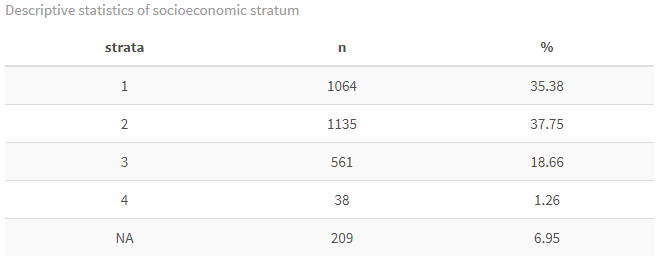
Contrasting the information of the fathers with attendance to school information on the level 0, 80% of the students that dropped do not live with their father and 90.5% of the students enrolled do live with their parent, but as almost all the variation is in this category the data is not informative. However, there is a significant difference for the level 9, 19% of the students that dropped out of school their father died, compared to 9.4% of the students enrolled.



In the case of the mother, it is possible to identify that when the mother does not live in the household the share of students that dropped out of school is higher, 29.5%, than the share of enrolled students, 11.3%. Additionally, when the mom lives in the house the share of dropouts is lower, 68.8%, than the share of enrolled students, 88%.

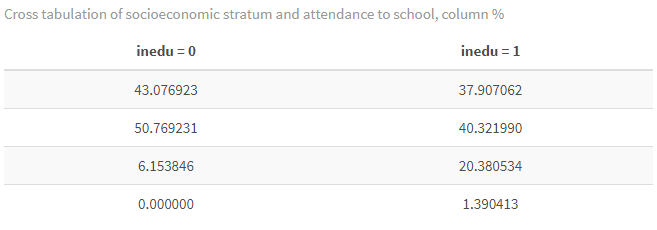
**Socioeconomic status**

In Colombia residential properties are classified in socioeconomic stratums, this classification is carried out mainly to charge for residential public utilities in a differential way by strata. According to the National Statistics Department -DANE- residential properties in urban areas are classified in six categories: 1-Low low; 2-low; 3-Medium-low;4-Medium;5-Medium-high; and 6-High (add SOURCE). It is important to mention that ELCA survey includes strata one to four and excludes strata five and six from the analysis. Thus, respondents were asked to which stratum their household belong to, this classification is used to operationalize the socioeconomic status variable.



From the analytical sample 37.7% of the respondents belong to the strata 2 and 35.4% to the strata 1. Even though, the sample mainly belongs to a low socioeconomic status it is possible to see the differences between the stratums.

**Cross table socioeconomic stratum and dependent variable (inedu)**



In strata 2 and 3 the share of dropouts is higher than the share of enrolled students, in strata 1, 43% compared to 37%, and in strata 2 50.8% and 40.3%. In strata the relation shifts and the share of enrolled students is higher than dropouts, 20.3% to 6%, and in strata 4 there are no dropouts.

**Covariates**

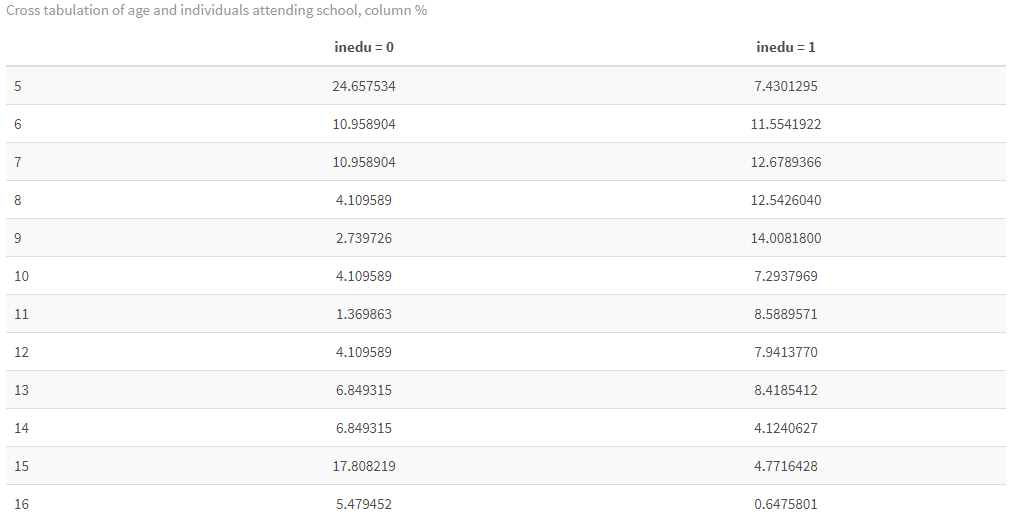
The study controls for the time-constant covariates age and gender (not sure about type of school yet)

**Age**

The highest percentage of individuals in the analytical sample are 9 years old with 13.7%, followed by individuals with 7 and 8 years old, 12.6% and 12.3% respectively. The lowest share of individuals is 16 and 14 years old with 0.76% and 4.2%.



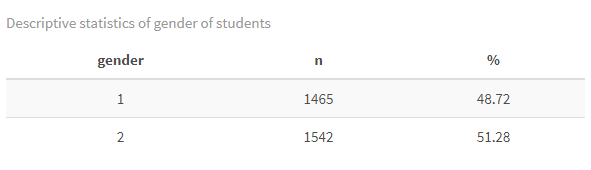
**Cross table age and dependent variable (inedu)**



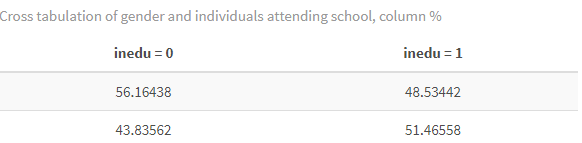
From the students that are not attending school 24.6% are 5 years old and 17.8% are 15 years old. In contrast, from the enrolled students 14% are 9-year-old and 12.6% 7 years old.

**Gender**

This variable takes the value of 1 for male and 2 for female. The analytical sample is balanced regarding gender, 48.7% are male and 51.2% are female.



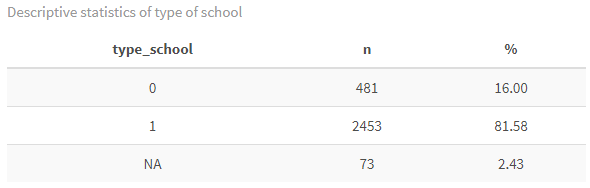
**Cross table gender and dependent variable (inedu)**

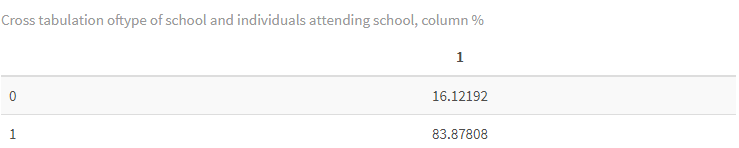


There is a higher percentage of male individuals that are not attending school 56.2%, compared to 43.8% of female. This relationship is inverted for students enrolled, there are more females enrolled that males 51.4% compared 48.5%.

**Type of school**

I think it is important to include the type of school, if it is public or private, but as the dropout variable (inedu) is operationalized by asking the individuals if they are studying, then for the dropouts there is no information about the institution.

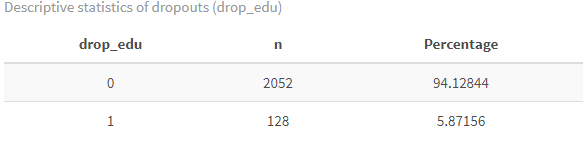




**APPROACH B OF DEPENDENT VARIABLE**

**Dropped education (drop\_edu)**

Other approach to operationalize dropouts using the ELCA survey is by using the question that asks the respondents if they stopped studying more than 2 months during the last three years. By using this approach the definition of school dropouts must change, to the assumption that if a student stopped studying more than 2 months then she or he dropped out of school. Moreover, the sample will take into account only wave 2 and 3, because this question is not asked in wave 1. After joining both waves, removing NA´s from the dropout variable and the education of the father variables and limiting the sample to the range of age between 5 and 16 years the analytical sample has 2,089 observations.

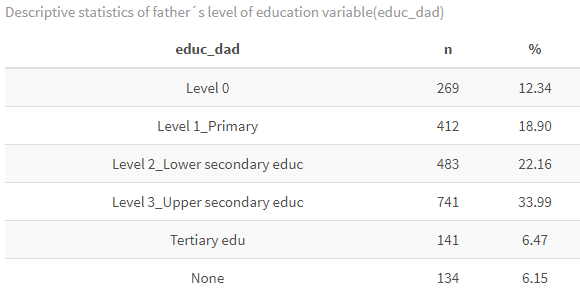


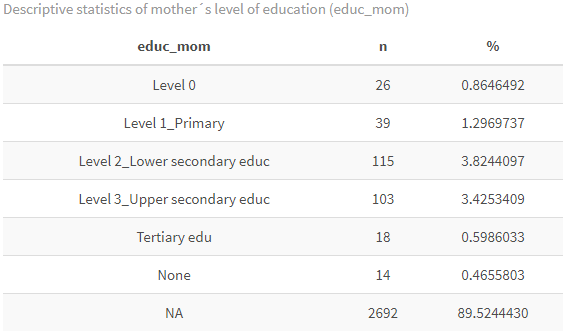
The sample of dropouts increases compared to the sample using the attending school variable. The analysis of the independent variables was done using this new sample.

**Independent variables**

**Parent´s education**

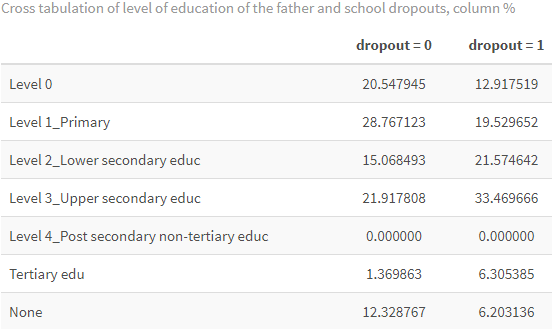
The highest level of education of the fathers is level 3- Upper secondary education with 34%, followed by level 2-Lower secondary education with 22.2%.



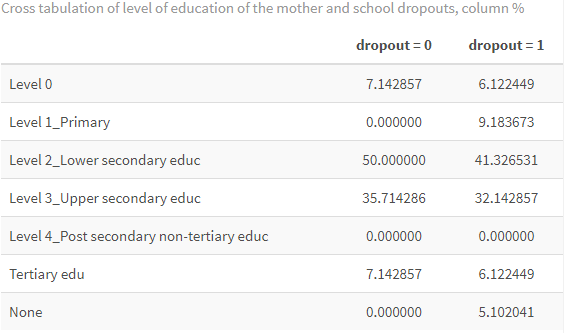


The mother´s education variable has the same problem as before, 89% of the observations are NA´s.

**Cross tables parent´s education and dependent variable (drop\_edu)**



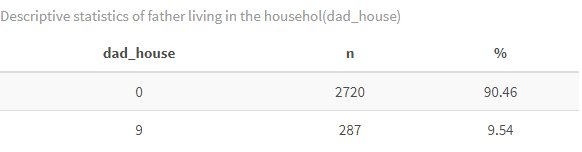
For dropouts, the highest level of education of the father is level 3-Upper secondary education and for enrolled students the highest level of education is level 1-Primary, this contradicts hypothesis 1.

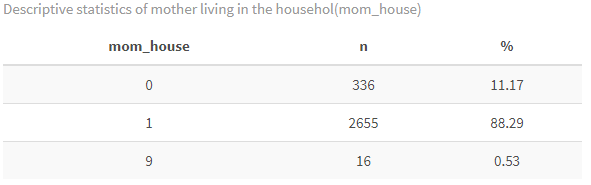


In the case of the mother´s education the same trend occurs, for the dropouts, the highest level of education of the mother is level 3-Upper secondary education and for the enrolled students level 2-Lower secondary education.

**Structure of family**

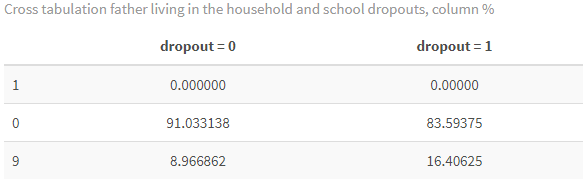
Th data for the presence of the father in the household shows that 90.5% of the fathers do not live in the household and 9.5% died.



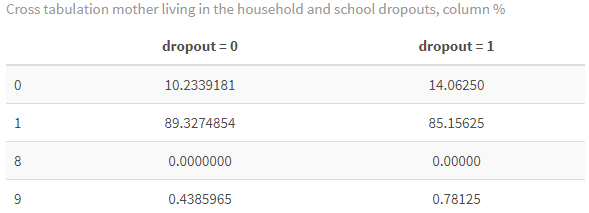


88.3% of the mothers live in the household, 11.2% do not live in the household and 0.5% died.

**Cross table family structure and dependent variable (drop\_edu)**

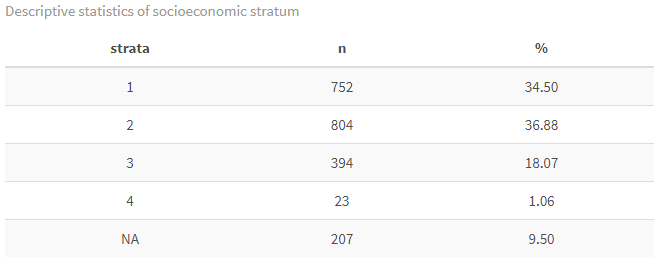


Apparently, when the father does not live in the household there is a higher percentage of enrolled students than dropouts, 91% compared to 83.5%, this contradicts the hypothesis 2. However, when the father died there is a higher share of dropouts 16.4% than enrolled students 9%.



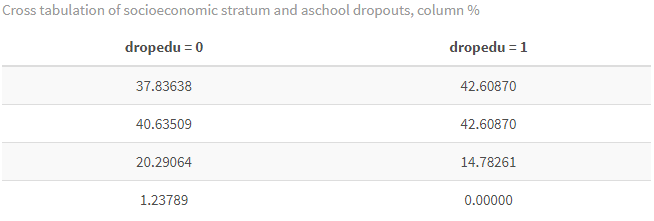
When the mother does not live in the household there is a higher percentage of dropouts 14% than enrolled students 10%. Moreover, when the mother lives in the household the share of enrolled students is higher than the dropouts, 89% and 85% respectively.

**Socioeconomic status**



37% of the analytical sample, using drop\_edu as the dependent variable, belongs to strata 2 and 34.5% to strata 1. The variation of the variable can allow a proper analysis and the NA´s are only 9%.

**Cross table socioeconomic stratum and dependent variable (drop\_edu)**

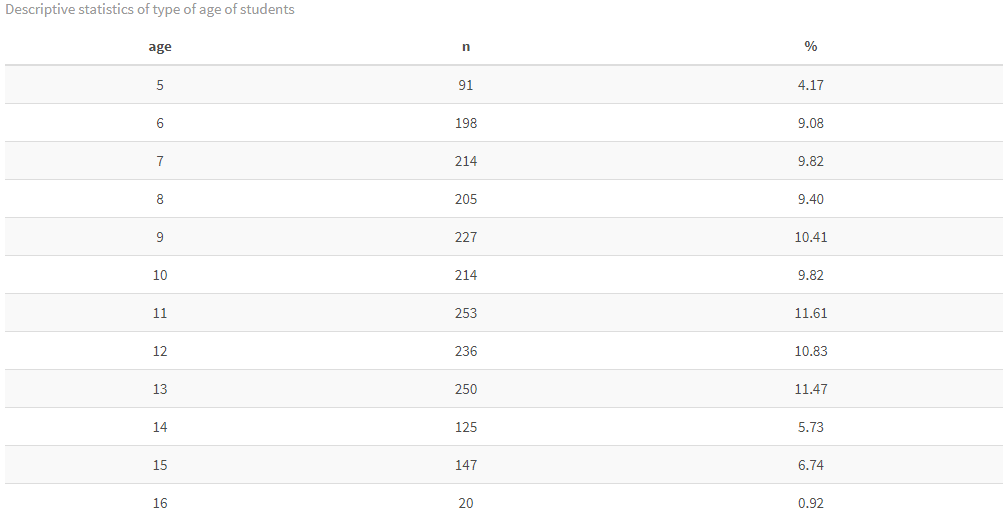


It is possible to observe that in strata 1 and 2 the share of dropouts is higher than the enrolled students, in strata 1, 43% compared to 38%, and in strata 2, 43% compared to 41%. In strata 3, the share of enrolled students is higher than dropouts, 20.2% and 15%. In strata 4 there are no dropouts.

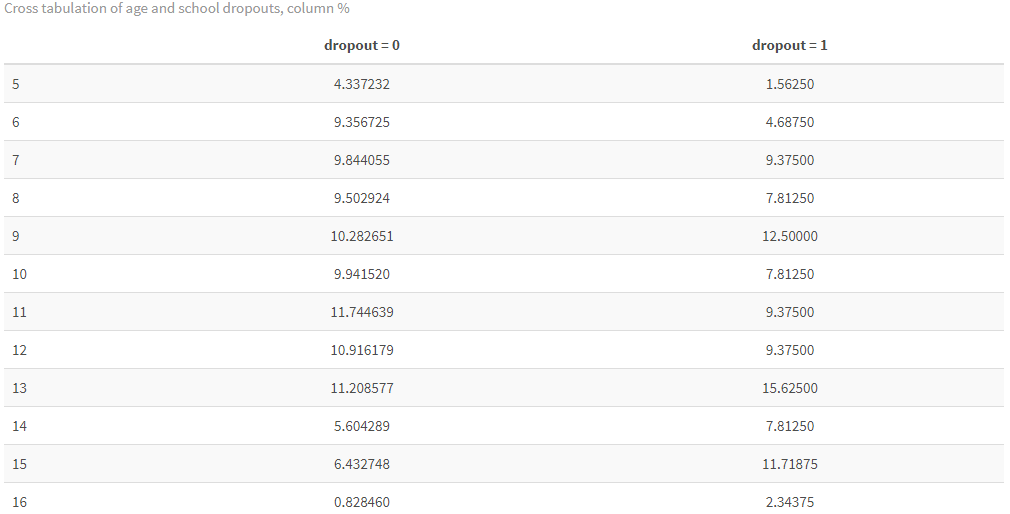
**Covariates**

**Age**

The highest percentage of individuals in the analytical sample are 11 years old with 11.6%, followed by individuals with 13 and 12 years old, 11.5% and 10.8% respectively. The lowest share of individuals is 16 and 5 years old with 0.92% and 4.2%.



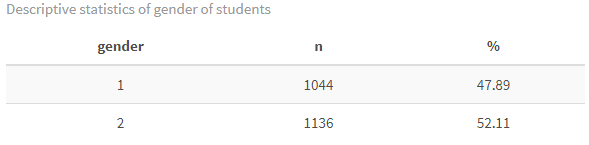
**Cross table age and dependent variable (drop\_edu)**



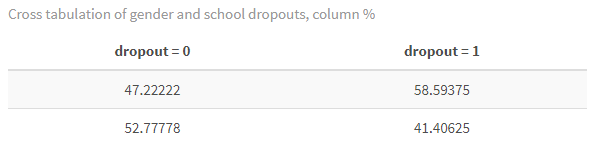
From the students that are not attending school 12.5% are 9 years old and 11.7% are 15 years old. In contrast, from the enrolled students 11.7% are 11-year-old and 11.2% 13 years old.

**Gender**

In the analytical sample 48% are male and 52% female.



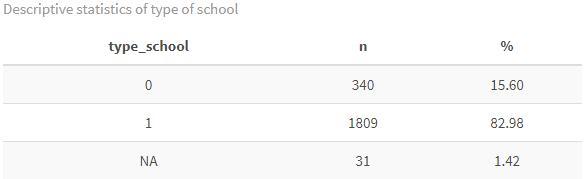
**Cross table gender and dependent variable (drop\_edu)**



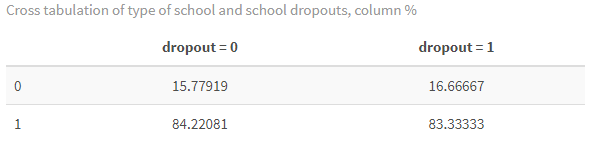
There is a higher share of males that dropout compared to females, 58.6% to 41.4%. In the same line, there are more enrolled females 52.7% than male 47.2%.

**Type of school**

83% of the schools in the sample are public, 16% private, and only 15 NA´s.



**Cross table gender and dependent variable (drop\_edu)**



83% of the dropouts studied in a public school and 16.6% in a private school. Also, 84% of the enrolled students study in a public school and 15.8% in a private school. Here I don´t see much variation between the two groups so should I keep this variable?

**Model**

Where should I explain that I´m not using longitudinal data but cross sectional and using the data of the 3 waves to run a pooled logit model?

This is the code I used to run the model:

#Shape the dataset as panel data for the dependent variable inedu

model01 <- pdata.frame(data01, index=c("id","wave"), drop.index=TRUE, row.names=TRUE)

#Running pglm model with inedu

model <- pglm(inedu ~ educ\_dad + strata + age + gender, model = "pooling", family = binomial("probit"), data = model01)

I don´t know if I´m on the right path in this part

**References**

Colombian Longitudinal Survey of the Universidad de los Andes - ELCA, public data. Produced and distributed by the Center for Economic Development Studies (CEDE), School of Economics, Universidad de los Andes. Bogotá, Colombia.

**Appendix**

Html attached with R code.