Student Paper

Young Lives
An International Study of Childhood Poverty

# Survival Analysis: Exploring the dropout motives in a panel of Peruvian children, using the Young Lives program dataset

Patricio Valdivieso

Paper submitted in part fulfilment of the requirements for the degree of MSc in Economics, University of Edinburgh.

The data used come from Young Lives, a longitudinal study of childhood poverty that is tracking the lives of 12,000 children in Ethiopia, India (Andhra Pradesh), Peru and Vietnam over a 15-year period. **www.younglives.org.uk**

Young Lives is funded by UK aid from the Department for International Development (DFID) and co-funded by the Netherlands Ministry of Foreign Affairs from 2010 to 2014 and by Irish Aid from 2014 to 2015.

THE UNIVERSITY
*of* EDINBURGH

MSc Dissertation

# Survival Analysis:
## *Exploring the dropout motives in a panel of Peruvian children, using the "Young Lives" program dataset*


*Exam Number:*
B063831

*Supervisor:*
Prof. Philippe
LeMay-Boucher


August $20^{th}$ 2015


**JEL Classification codes:** C4, C150, C230, I2, I250, I210.
**Key Words:** Microeconometrics, Survival Analysis, Development Economics, Education, High-School Dropout.

# Abstract

Our thesis explores the dropout motives of a panel of Peruvian children, provided by the Young Lives program. Boyden (April 2014). We used survival analysis with both, a non-parametric estimation (Kaplan-Meier survival estimates) and a semi-parametric estimation (Cox proportional hazards (PH) model). Our results suggest that the child's initial conditions are relevant determinants of the drop out decision. Also, we found evidence to support the importance of the wealth level, the location of the household and the sex of the caregiver, for the drop out decision. We believe that further research is required to refine the size of the effects and this can be achieved by the inclusion of the fourth round of data and the use of parametric models to explore the effect of the covariates that did not meet the proportionality assumption required by the Cox PH model.

# Contents

# Chapter 1

# Introduction

Education is a key variable when determining the level of inequality in a society, De Gregorio and Lee (2002). It has also shown to be a key determinant of inter-generational economic mobility, Iyigun (1999). Although many countries have made efforts to secure education supply for the majority of their population, opportunity inequality keeps affecting educational attainment, but through a different channel, that is, high-school dropout.

It is hard to argue about the importance of primary and secondary education in the human capital accumulation process, Heckman (1976). Intuitively, any person should want to finish high-school in their early years, but despite basic reasoning, runaways exist, and the reasons behind this attitudes must be understood.

High-school dropout in Peru represents around 8% of the Peruvian population between ages 13 and 19. This problem is not exclusive to Peru. In the United States, high-school dropout rates add up to 6.8%[1] for 2013 and

[1]U.S. Department of Education, National Center for Education Statistics. (2015).

in countries similar to Peru, like Chile or Ecuador, dropout rates are around 0.23% and 7.56%, respectively.[2]

**A dropout is defined as:**

> *"A person who was enrolled in the previous year but not in the current and have not concluded high-school", Franklin and Kochan (2000).*

As the definition states, since the person was already enrolled in school, our interest is solely based on the reasons that motivates the individual to leave school.

There are a few methods to measure dropout rates. One of the most popular ones, is the method that calculates the percentage of a country's population between 13 and 19 years old, which are not currently enrolled in school. This is the method used by most national statistical offices. In our case, since we will be using a specific panel, we consider as a drop out any person that was enrolled in school in round 1, but not in rounds 2 or 3, regardless of the possibility of returning to school in the future.

The first point, using only dropout values from rounds 2 and 3, responds to the fact that by the first round, none of the children from the "Old Cohort"[3] were in high-school, they were between 6 and 8 years old and that corresponds to $1^{st}$ and $3^{rd}$ grades of primary education. In the second round, these children were between 11 and 13 years old which corresponds to the

[2]Drop-out rate in lower secondary general education UNESCO. Figures correspond to year 2012 and are cumulative figures for the entire 5 years of secondary education.

[3]For a detailed explanation of the data and the correspondent composition of the "Old Cohort", please refer to section 3, methodology.

last year of primary school and the first 2 years of high-school. For the third round, all children between 13 and 16 years were enrolled in the last four years of secondary education. The second point, discarding the possibility of enrolment in the future, may seem as a limitation. Nevertheless, we believe that the act of not enrolling in a particular year, will respond to a similar motivation, regardless of the possibility of completing their studies further in their lives or not.

We have used survival analysis methodology, to explore this phenomena. Survival analysis is defined as:

- *"Generally, survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is time until an event occurs." Kleinbaum and Klein (2005)*

We think this is adequate, because graduating from high-school can be understood as a survival process, where people who graduate, survived the risk of drop out high-school and people who drop out, are considered as "failure events". These type of analysis will allow us to compare survival probabilities of a group affected by a covariate, against the group that was not affected by that covariate.

As an example, we can consider the dummy variable, "sex", where males were denoted by a 1 and females by a 0. When analyzing the survival probabilities of both groups we will be able to compare the instantaneous failure probability of men, against the instantaneous failure probability of women. If the female dropout probability is smaller than the male dropout probability, we will obtain a hazard ratio bigger than 1. If the male failure risk

is smaller than the female failure risk, then the result will be smaller than 1. Hence a hazard ratio of 1.5 would imply that the male group is 0.5 times more likely to fail (drop out), given that the individuals in the sample have survived until that point in time.[4]

In Peru, primary and secondary education is mandatory, it was established in the 1993 constitution. The current figures of the Education Ministry in Peru, show that dropout is stable around 8% and that differences across sexes are not evident, Figure (1.1). Although, 8% seems like a reasonable figure, there are differences amongst areas, where urban rates are consistently lower than the rural ones, Figure (1.2). In the urban areas, the rates are closer to the national average, 8%, while in the rural areas, the average is higher, despite the relevant improvement experienced by the year 2013.[5]

If we include effects such as socioeconomic status (SES), figures suggest a strong correlation between this variable and dropout rates. This is consistent with most authors findings about SES relevance, Figure(1.3). Alexander et al. (1997), Lavado et al. (2005), Rees and Mocan (1997), Rumberger and Thomas (2000), Woldehanna and Hagos (2012)

In the past 2 decades, Peru has experienced a positive evolution of their educational outcomes [6]. Nevertheless, this may be directly attributed to

[4]Further details on the estimation method and specific measures is provided in the methodology section, this explanation is meant to provide the reader with intuitive knowledge about the method.

[5]Rural dropout rates were close to 32% and Urban dropout rates were close to 23% by the year 2005. Ministry of Education Peru.

[6]Lavado et al. (2005) found that Peruvian urban rates were closer to 0.14 while the rural rates were around 0.35. Although this figures are not directly comparable, since their study considered the total dropout of the educational system, we can find similar figures in the Ministry of Education of Peru and results are quite similar for the urban areas (0.135), while in terms of results in rural areas, there has been substantial progress (0.182) in the last 8 years.

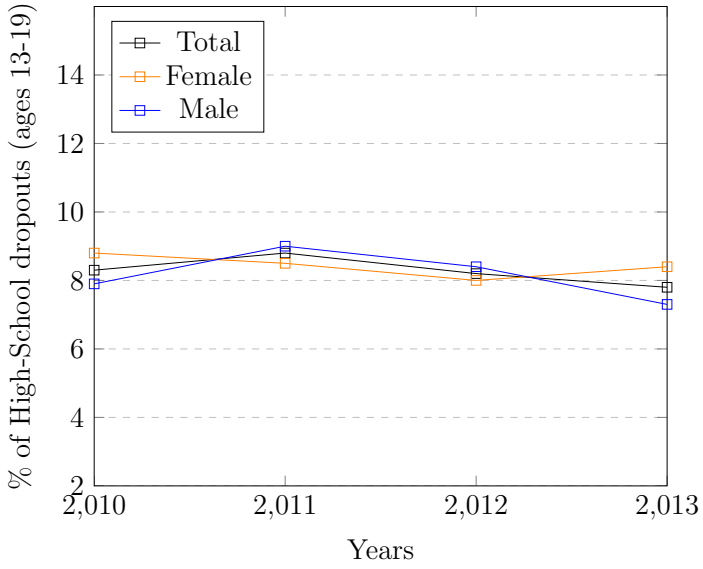


Figure 1.1: High School droput rates in Peru - Total and by Sex.
Source: Ministry of Education Peru

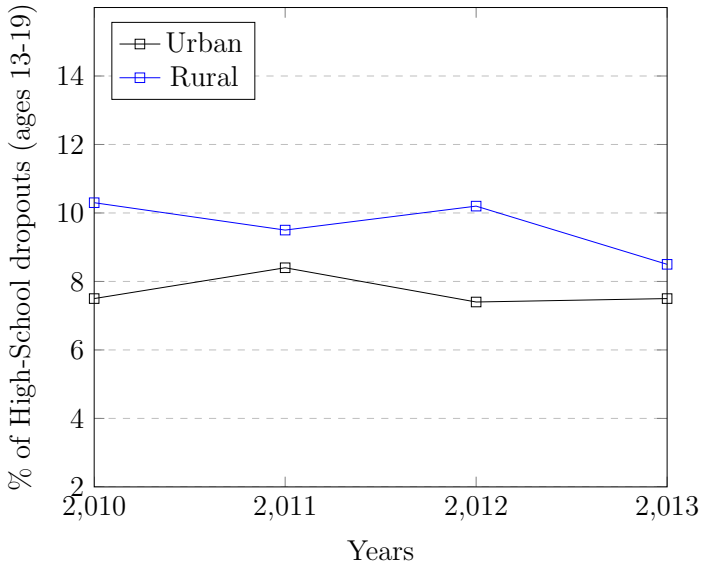


Figure 1.2: High School dropout rates in Peru - by Density
Source: Ministry of Education Peru

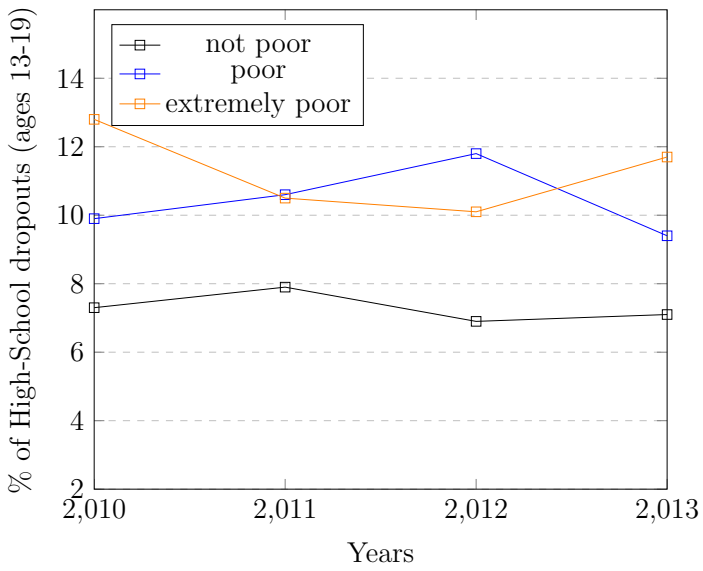


Figure 1.3: High School dropout rates in Peru - by SES
Source: Ministry of Education Peru

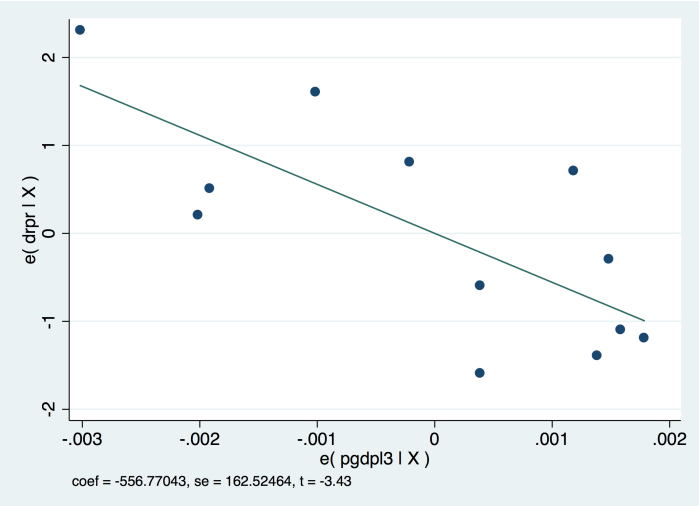


Figure 1.4: OLS regression using the "dropout rate" as the dependent variable and the "third lag of public investment as % of GDP in education" as the explanatory variable.

the strong and almost uninterrupted GDP growth that the country has experienced from 1991 until 2014, Figure(1.4)[7]. It is possible to argue that, since the country has experienced substantial growth and since educational indicators such as school dropout have evolved positively, there is not a real problem and it is only a question of time before this figures converge to their structural level.

Unfortunately, Peruvian growth relied heavily on commodity prices and although the country successfully reduced their poverty figures during the past two decades, structural changes didn't take place. Now that the commodity prices have stopped rising and that the country's growth have reduced significantly [8], it is reasonable to expect that this apparent "efficiency" that the educational sector was experiencing, will now end.

Education is one of the weakest sectors in Peru. The country has allocated the lowest budget as percentage of GDP for the sector with their peers in the region[9], and unsurprisingly, it shows the lowest results in the standardized test, PISA.[10]

[7]Figure 1.4 show the ols regression between the dependent variable "Dropout rate" and the explanatory variable, "public investment as a % of GDP". We decided to use lagged values of the explanatory variable, since any increment in the budget of the sector, would be implemented, and effectively affect students, a few years after being assigned.

[8]Peruvian average growth for the period 2010-2013 was 6.7 percent and in 2014 dropped to 2.4 percent. As shown in the "Marco Macroeconómico multianual 2016-2018" published in the Peruvian Central Bank website, GDP forecast for the year 2015 is 4.2 percent and for the next two years is 5.5 percent.

[9]Argentina 2012: 5.1, Brasil 2012: 6.3, Chile 2012: 4.6, Colombia 2013: 4.9, Ecuador 2012: 4.2, Peru 2013: 3.3, Uruguay 2011: 4.4. Source: World Bank Data.

[10]Is The Programme for International Student Assessment (PISA). This is a worldwide study by the OECD for 15-year-old students on the subjects of mathematics, science, and reading. Peru ranked poorly on 2014 and now that the 2015 results are out, Santiago Cueto from GRADE mention that: "We are achieving poorly as most of our students are placed at level 1 or below 1 in performance scales in mathematics(75 percent), science(69 percent) and reading(60 percent)".... "in short it could be said that they only manage to solve the most simple items, and sometimes not even that."

## 1.1 Research questions

Our main research question explores the possibility of influencing the dropout decision. In specific, we want to understand if it is possible to generate policy which provides the right incentives for children (or their parents) to pursue the high-school diploma. To explore this decision, we propose the following research questions:

**Our main research question is:**

1. Are the children's initial traits, determinants of drop out in Peru?

**Our secondary research questions are:**

1. Is the household socioeconomic status (SES) a determinant of high-school dropout in Peru?
2. Is the location of the households (region), a determinant of high-school dropout in Peru?
3. Does the gender of the caregiver affects the high-school dropout decision?

If we found evidence to support our main hypothesis, policy implementations should be directed towards influencing parents conditions even before the child is born. Otherwise, policy towards household conditions, labour market conditions or school quality conditions, will be valid.

The secondary research questions, are aiming to link the opportunity inequality issue with the dropout decision, where, if we found that the so-

cioeconomic status (SES) is significant, we can propose some kind of aid towards low-income families.

The next hypothesis, is aiming to check previous findings regarding the region where the household is located. In a country such as Peru, with diverse geography, we can find families living in the coast, highlands or jungle and with access to different weather, food, labour options and resources in general. Given this particular conditions, it is possible to expect an effect from the region where the household is located, where, given the level of economic concentration in the capital, Lima (which is located in the coast), we would expect to find a higher hazard for families living in the highlands and an even higher hazard for households located in the Jungle.

Finally, our fourth hypothesis addresses a common problematic to many traditional developing countries, that is, the role of women in the household. We would expect to find that households where the caregiver is a female have a lower hazard than those where the caregiver is a male. We will explore all these questions in further depth, in the literature review section.

The remaining of this document is structured in the following way: Chapter 2, explores the main and secondary hypothesis and links them with existing literature. Chapter 3, describes the data and the descriptive statistics from the sample while explaining and justifying the estimation models used. Chapter 4, presents the non-parametric results (Kaplan-Meier survival curves) and the semi-parametric section with the first-order correlation results and the complete-model results, both with the Cox PH model. Chapter 5 concludes, highlighting the main findings and limitations of this paper, and pointing out the direction that further research should take.

# Chapter 2

# Literature Review

Our paper explores 4 research questions that can be clustered into 4 main literature sections: child's initial conditions, household socioeconomic status, general households characteristics and the caregiver's characteristics.

## 2.1 Child's initial conditions

**Alexander et al. (1997)** proposed that high-school dropout is a process that begins in the early childhood and that is shaped by four specific types of predictors:

- Background characteristics
- Family context
- Children's personal resources
- School experiences

They referred to children's personal resources and denoted two clusters of variables.The first, attitudes towards self and school, **Gamoran and Nystrand (1992)**, highlighting that dropouts are more internally controlled, since most dropouts leave school in good academic standing, **Fine (1986)**, **Schneider et al. (1994)**. The second cluster involved engagement behaviours, like lateness, absences and time spent on the TV. This study proposed that, after controlling for SES and other relevant covariates, results suggest that boys are more affected by stressful home conditions, while initial academic differences alone, do not explain why youth moves along such different development paths. This proposition is remarkable, in the sense that if children's initial conditions are not as determinant as some might think, there is scope for adequate policies in order to address the dropout issue.

**Lavado et al. (2005)** found that there was little difference in dropout rates across genders, and they attributed these results to the application of the millennium development goals in Peru[1]. We do believe that this might be a slightly forced statement, given that other countries in the region showed similar figures regarding gender dropout rates and that it is also consistent with **Alexander et al. (1997)** results, where there is no significant difference across gender. Since the authors don't provide figures to backup that correlation, we believe it is questionable. We will explore this same question in our panel.

[1]As defined by the Millenium Project from the United Nations Development Program: "The Millennium Development Goals (MDGs) are the world's time-bound and quantified targets for addressing extreme poverty in its many dimensions-income poverty, hunger, disease, lack of adequate shelter, and exclusion-while promoting gender equality, education, and environmental sustainability."

**Eckstein and Wolpin (1999)** explored the relevance of initial traits and its magnitude and persistence over dropouts. Using the national longitudinal survey of youth (NLSY79) in the U.S., their estimation was based on the solution of the dynamic optimization problem with the maximization of a likelihood function that accounts jointly for annual observed work-schooling choices, wages, credits earned and grades. In order to explore the initial traits, they assumed 4 discrete types of youths who differ in the parameters that describe their preferences, **Eckstein and Wolpin (1990), Keane and Wolpin (1997), Heckman and Singer (1984)**. Their results suggest that working while attending high-school does reduce academic performance, but the effects are small. Their paper propose that dropping out of high-school is confined to youth with lower ability and motivation, lower value for a high-school diploma and consequently lower value of attending high-school and a higher value for leisure. Within the reasons they found to explain high-school dropout, they mentioned, disliking school, high value of leisure, low ability and motivation, good labour market opportunities and low expectation of the payoff to graduation. They explored questions like: What if the dropout type had the same initial traits as the other types? For example, if they had the same ability and motivation as graduates, or the same expected value of graduation. Their results propose that the dropout type would replicate the results of the other types, even when initial traits are different. This proposition is very relevant, in the sense that it supports our previously mentioned possibility of reducing the number of dropouts regardless of their initial conditions. Finally, they modelled some policy restrictions, such as forbidding youth to work and study simultaneously, or banning the youth to

work during the first four years of high-school. With this type of modelling restrictions, attendance rates fell in their model results. This is a surprising result, since it implies that policies that do not alter the traits with which youth come to high-school will have very limited effect upon dropout. In other words, forbidding youth to work, its not enough. As we proposed in our main hypothesis, it might be the case that policy needs to be directed to the parents conditions, even before the child is born.

**Montmarquette et al. (2007)** results, support the seminal **Angrist and Krueger (1990)** paper and their conclusions about compulsory school attendance laws. They also support the significance of minimum wage upon dropout and specify that this effect is 3 times larger in the G-type student than in the W-type student[2]. Although this might seem counter intuitive at first, with a 1% increase in minimum wage, the probability of dropping out will be reduced by 2.48 % for the G-type, while the reduction in the probability of dropping out for the W-type would be 0.97%. This is also consistent with an increase in the unemployment rate, which significantly decreases the probability of dropout, specially on the G-type. Specifically, they found that low unemployment rates have an increasing effect in dropout rates, specially for the W-type students, again supporting the idea that initial traits matter.

**Woldehanna and Hagos (2012)** and **Brown and Park (2002)**, both found significant gender bias, but in opposite directions. **Woldehanna and Hagos (2012)** found that Ethiopian boys were less likely to stay in school

[2]The G-type students are the student who's initial traits are driven to studying, while the W-type student are the students who's initial traits are driven to work, they value leisure more and school less.

while **Brown and Park (2002)** found that rural Chinese girls were more prone to dropout, where, gender bias in educational investment may be attributed to the lower returns from education, for girls.

## 2.2 Socioeconomic status

**Alexander et al. (1997)**, found nationwide evidence (U.S.) for propositions such as that half of the welfare families were headed by dropouts. They also mentioned that dropouts accounted for half of the prison population and that the average family income from the dropouts was equivalent to half of the graduates average family income, **McMillan and Whitener. (1994)**. Three very strong arguments supporting the relevance of SES. We believe that this study is revealing and sound, nevertheless, there is one aspect that we would like to point out and it is the way they constructed their SES. The authors used a composite of the parents educational level and occupational status and they complemented this information with the receipts of reduced price school meals, as a proxy of SES. In that sense we think that the last factor might be misleading and may account for significant bias, given that some families, even the ones with middle or high income, might want to save money in this type of expenditure or may complement their children's nutrition with other sources, hence, we do believe there is scope for improvement in the construction of the SES.[3]

**Lavado et al. (2005)**, Using data from the Peruvian Department of Education and national survey of homes 2002, the authors highlighted the

[3]We have used the Wealth index (WI) as the proxy of the SES. Our WI, accounts for this type of issues. The construction of our WI, is detailed in Section 3.

fact that children whom require more Education are the most prone to leave school, **Ravallion and Wodon (2000)**, which is consistent with the previously mentioned relevance of SES.

## 2.3 Households characteristics

**Alexander et al. (1997)** found that within the household context, variables such as family stressors (divorce, marriage, death, illness, moving to a new house), **Haveman et al. (1991)**, parents attitudes and values, **Seginer (1983)** and parents socialization practices (friends screening, extra curricular activities, after school care), **Posner and Vandell (1994)** were found to be significant.

**Lavado et al. (2005)** also mentioned factors that are relevant to the Peruvian case, such as living in a rural area or having a language different from Spanish as their native language, **Alarcón (1995)**. They also showed that by 2005, figures for the rural population were much more dramatic than those of today and the lack of education supply was still an issue. They found evidence of disadvantage for the rural boys from the amazon and the metropolitan Lima girls.

**Montmarquette et al. (2007)** found that male students who work more than 30 hours a week had a 14 % increase in their dropout probability. They also mentioned that for a female student with parental responsibilities and that worked more than 30 hours a week, there was a 20 % higher dropout probability. The case of female students with no parental responsibilities is different, since even those who worked more than 30 hours a week, had

lower dropout rates. They proposed that working over a threshold amount of time is detrimental, but since students spent a significant amount of their out-of-school time in unproductive leisure activities, working, at an extent, shouldn't be detrimental.

**Woldehanna and Hagos (2012)** introduced the effect of external shocks to the household, such as dead or illness or family members, livestock and many others. They found that most shocks occurred to the rural households and that rural children were less likely to stay in school than the urban children. Also, they found evidence of significance regarding the number of siblings below 7 years of age. This factor showed a negative impact upon the likelihood of dropout[4].

## 2.4 Caregiver's characteristics

**Lavado et al. (2005)** found that the parents education mattered, specifically they found that the level of education of the father was more important in the rural areas, while the level of education of the mother was more relevant in urban areas. The number of siblings coursing primary education also showed significant impact towards the increasing dropout probability, as in **Woldehanna and Hagos (2012)**.

**Montmarquette et al. (2007)** also supported that students with highly educated parents have a higher probability of working while in school. Evidently, this might be due to specific Canadian population characteristics, and it is highly unlikely to find this type of behaviour in Peru. Nevertheless,

[4]The bigger the number of children below 7 years the higher the likelihood of dropout of the older siblings.

**Brown and Park (2002)** found that the likelihood of dropout of primary school falls dramatically when women have a greater say in the enrollment decision, implying that women value education more than men. They also pointed out that an additional year of father's education reduced the likelihood of dropping out by 12% to 14%.

# Chapter 3

# Methodology

## 3.1 The Dataset

Our database is composed by 3 rounds of collection, built by The Young Lives Program. We are only using the "Household and Child Survey" and within it, the Peruvian section of the data. The Peruvian portion of the dataset corresponds to 2,052 children who were born in 2001 and 2002 and 714 children who were born in 1994 and 1995, with 3 interventions, in September 2002, March 2007 and August 2009. We will refer to the first group as the "young cohort" and to the second group as "Old cohort". Given that we are studying the high-school dropout phenomena, we will only use the information from the "Old Cohort", since the children in that group will be 8 years old in the first round, 12 on the second round and 15 on the third round. This corresponds to $3^{rd}$ grade, $7^{th}$ grade and $10^{th}$ grade, respectively (Out of a total 11 years required for the high-school diploma). In Peru, children begin their school life at the age of 3, this stage is called pre-school

and lasts 3 years. By the age of 6, children begin the first year of primary education, which lasts 6 years. By the age of 12, children begin secondary education, which lasts for 5 years. An average Peruvian child is expected to finish school by the age of 16.

Within these 714 Children, we have a total number of 2,122 records, given a minimal attrition rate, between the 3 rounds. The sample design considered 20 sentinel sites, where poor families were over-sampled and the sampling of clusters was randomized, **Sanchéz and Melendez (2015)**. At the time of the design, a poverty map provided by (FONCODES 2001)[1] was used to select the 20 sentinel sites. One of the objectives of the sampling method, was to over-sample the poor areas. For this reason, the highest ranked 5% of the districts were excluded (All districts were located in Lima), approximately 75% of the sites were considered to be poor and 25% were considered non-poor. Each district had a probability of being selected proportional to its population size. Although the sample wasn't intended to be nationally representative, on average, The Young Lives sample includes households with more education, better access to services and a bigger amount of assets than the average. However, after adjusting the sample for the size proportional selection probability, many of the differences between Young Lives and the DHS 2000[2] are not significant. Attrition rate for the first 3 rounds was 5.7%.

[1]FONCODES is a program directed by the Development and social integration ministry. The main objective of the fund, is to promote economic independence and integration of the rural families in extreme poverty situation.

[2]Demographic and Health Survey from the United States International Development Agency

## 3.2 Descriptive statistics

Table 3.1: Sample by Gender: Old Cohort (Born 1994/1995)

| | Men | | | Women | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | R1 | R2 | R3 | R1 | R2 | R3 |
| Rural | 24.6% | 24.5% | 22.1% | 27.1% | 25.2% | 24.7% |
| Urban | 75.4% | 75.5% | 77.9% | 72.9% | 74.8% | 75.3% |
| Total Obs. | 386 | 368 | 362 | 328 | 317 | 316 |

Source: Own estimates, based on Young Lives program data.
Note: *a*. Figures in rows 1 and 2 show the distribution of gender by round between rural and urban children.

The sample is balanced regarding gender distribution, where approximately there is a 75% of the total sample that lives in urban sites, independent of sex, Table (3.1). Also, in both genders, there is an increasing concentration in urban sites throughout the rounds. In the case of men, urban men represent a 78% by the third round, while in the case of women, the increase adds-up to a total of 75% by the third round.

Table 3.2: Parents education: Old Cohort (Born 1994/1995)

| | Dad with Secondary Ed. | | | Mom with Secondary Ed. | | |
|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) |
| | R1 | R2 | R3 | R1 | R2 | R3 |
| Rural | 11.3% | 10.7% | 8.9% | 6.3% | 4.8% | 4.2% |
| Urban | 88.7% | 89.3% | 91.1% | 93.7% | 95.2% | 95.8% |
| Total Obs. | 230 | 234 | 213 | 223 | 227 | 216 |

Source: Own estimates, based on Young Lives program data.
Note: *a*. Figures in rows 1 and 2 show the distribution of parents education between rural and urban children.

Regarding the parents education, of the total number of parents with secondary or higher education, only between 11.3% and 9% of the sample's fathers lived in rural areas, while around 90% of them lived in urban sites, Table (3.2). This results are not surprising, since economic concentration is one of the big unsolved issues in Peru (9.8 million citizens live in Lima, the capital city, out of a total 31.1 million[3].). In the mothers case, the situation is even more skewed towards urban sites, where around 95% of the old cohort's mothers with secondary or higher education lived in urban sites, and about a 5% of the mothers with secondary or higher education, lived in rural sites. Since the level of education of the parents is a common control variable, the low presence of parents with secondary education in rural locations, would suggest a positive correlation between living in a rural site and higher high-school dropout rates.

Another widely used control variable is the Wealth Index (WI). In our case, it was constructed as a composite of 3 sub-indexes:

- Housing quality index (hq)
- Access to services index (sv)
- Ownership of consumer durables (cd)

All of these indexes have equal weights in the estimation of the wealth index. Then, we can define the wealth index (WI) as:

$$wi_i = \frac{hq_i + sv_i + cd_i}{3} \tag{3.1}$$

[3]Estimations form Peruvian national institute for statistics (INEI) for the year 2015

The composition of the housing quality index is a simple average of the following variables:

- Crowding (number of people sleeping in a room)
- Main material of walls (which will be represented by a dummy variable, that will take the value of 1, if the main material of the walls satisfy the basic construction quality norms, otherwise, it will take the value of 0. This same principle apply for the other dummies)
- Main material of the roof
- Main material of the floor

The Access to services index includes the following variables:

- Access to electricity
- Access to safe drinking water
- Access to sanitation
- Access to adequate fuels for cooking

Finally, the consumer durables index, is a simple average of a set of dummy variables that take the value of 1 if a household member owns at least one of each consumer durables. Only those consumer durables who were available across all 3 rounds were included.

From Table (3.3), we can observe that rural households show an average WI close to 0.3, while in the urban households, the average is above 0.5. Evidently, the closer the index gets to 1, the higher the wealth level of the

Table 3.3: Summary Statistics: Old Cohort (Born 1994/1995) - By Typesite

| | Avg.WI | | | | | | |
|---|---|---|---|---|---|---|---|
| | (1) Round 1 | (2) Round 2 | (3) Round 3 | (4) 2 v 1 | (5) 3 v 2 | (6) Abs | (7) Rel |
| Rural | 0.284 | 0.266 | 0.378 | 7.47% | 42.00% | 512 | 24.7% |
| Urban | 0.553 | 0.582 | 0.645 | 5.27% | 10.95% | 1565 | 75.3% |
| Total | 0.474 | 0.503 | 0.583 | 6.13% | 15.88% | 2077 | 100.0% |
| Total Obs. | 708 | 684 | 675 | | | | |

Source: Own estimates, based on Young Lives program information. Note: *a.* The total number of observations for each round (1),(2),(3), represents the observations used for the calculation of the total average wealth index. *b.*(4),(5),(7), represents the percentage changes. *c.* (6) Represents the absolute number of observations from each site considering the 3 rounds together.

household. Also, there is an absolute increase in the average WI for both types of sites. While the increase in the rural case is more steep than that of the urban sites, both types of sites present a higher increase from rounds 2 to 3, which is consistent with the country's GDP growth throughout that period.[4]

In terms of the distribution by geographical region, the coastal households resemble the urban behaviour from Table(3.4). Evidently, given the amount of households in Lima (which is in the coastal region and it is mainly urban), the average values for urban sites should be similar to the coastal region values. Also, the households from the jungle, present a significantly lower wealth index during the first two rounds, but this situation is addressed in round 3, where differences between households in the highlands and households in the jungle are unnoticeable. This is due to the impressive 35% increase in

[4]Peru GDP growth rate 2006: 7.5%; 2007: 8.5%; 2008: 9.1%, Source: Peruvian central bank.

Table 3.4: Summary Statistics: Old Cohort (Born 1994/1995) - By Region

| | Avg.WI | | | | | | |
|---|---|---|---|---|---|---|---|
| | (1) Round 1 | (2) Round 2 | (3) Round 3 | (4) 2 v 1 | (5) 3 v 2 | (6) Abs | (7) Rel |
| Coast | 0.574 | 0.620 | 0.671 | 7.98% | 8.20% | 865 | 41.6% |
| Highland | 0.415 | 0.436 | 0.515 | 5.14% | 18.16% | 899 | 43.3% |
| Jungle | 0.382 | 0.382 | 0.516 | 0.06% | 35.02% | 313 | 15.1% |
| Total | 0.474 | 0.503 | 0.583 | 6.13% | 15.88% | 2077 | 100.0% |
| Total Obs. | 708 | 684 | 675 | | | | |

Source: Own estimates, based on Young Lives program information. Note: *a.* The total number of observations for each round (1),(2),(3), represents the observations used for the calculation of the total average wealth index. *b.*(4),(5),(7), represents the percentage changes. *c.* (6) Represents the absolute number of observations from each site considering the 3 rounds together.

the average wealth index of the households in the jungle.

Table 3.5: Distribution of shocks: Old Cohort (Born 1994/1995)

| | Death/Ill. of HH mem | | | Death Livestock | | | Drought/crop fail | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) R1 | (2) R2 | (3) R3 | (4) R1 | (5) R2 | (6) R3 | (7) R1 | (8) R2 | (9) R3 |
| Rural | 0.0% | 23.0% | 21.0% | 41.0% | 64.0% | 50.0% | 57.0% | 57.0% | 65.0% |
| Urban | 0.0% | 77.0% | 79.0% | 59.0% | 36.0% | 50.0% | 43.0% | 43.0% | 35.0% |
| Total Obs. | 0 | 148 | 136 | 27 | 14 | 12 | 53 | 88 | 82 |

Source: Own estimates, based on Young Lives program information.
Note: *a.* Figures in rows 1 and 2 show the distribution of shocks between rural and urban children.

From the distribution of the shocks experienced by the households, Table (3.5), we can observe that the first shock, death or illness of a household member, is distributed as expected (Similar to the total sample distribution by typesite), where around 22% of the rural households experienced a shock

of this type, while 78% of the Urban households had a similar experience. This would imply that there is no evidence for a higher incidence of death or illness of a household member, related to living in a rural or urban site.

The situation is quite different for the shocks concerning death of the livestock, crop failure or droughts, where the rural households are much more prone to suffer one of these effects, Table (3.5). Although, in some cases the number of observations is below 50, we believe that logic supports these results, since rural households are much more dependent of their crops, livestock and the weather. This is also consistent with previous studies, **Woldehanna and Hagos (2012)**.

Table 3.6: Distribution of Dropout: Old Cohort (Born 1994/1995)

| | Men Dropout | | | Women Dropout | | | Total Dropout | | |
|---|---|---|---|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) |
| | R1 | R2 | R3 | R1 | R2 | R3 | R1 | R2 | R3 |
| Rural | 60.0% | 0.0% | 46.9% | 50.0% | 0.0% | 41.2% | 57.1% | 0.0% | 44.9% |
| Urban | 40.0% | 100% | 53.1% | 50.0% | 100% | 58.8% | 42.9% | 100% | 55.1% |
| Total Obs. | 5 | 4 | 32 | 2 | 2 | 17 | 7 | 6 | 49 |
| % of Total. | 12.2% | 9.8% | 78.0% | 9.5% | 9.5% | 81.0% | 11.3% | 9.7% | 79.0% |

Source: Own estimates, based on Young Lives program information.
Note: *a*. Figures in rows 1 and 2 show the distribution of dropouts between rural and urban children per segment. *b*. Figures in row 4 show the proportion of dropouts form each segment, e.g., the figure 12.2% corresponds to the 5 men that dropped out in round 1 divided by the total men that dropped out in the 3 rounds, that is 41. The result form that fraction (5/41) is 12.2%.

Finally, in terms of the dropout distribution, we can observe a few interesting things. First, about 80% of the dropouts occur in the third round, Table (3.6), which suggests an increasing probability of dropout or a decreasing survival function, through time. This finding was taken into considera-

tion when choosing survival analysis as the estimation methodology and the possibility of using a semi-parametric method such as the Cox proportional hazards model, that approximates the correct distribution[5]. We can also observe that the dropout contribution is almost equivalent for rural and urban sites. Nevertheless, this could imply a higher probability of dropout for the rural cohort, given that the initial sample distribution is biased towards urban sites.

## 3.3 Estimation Method

The model we will use is specified as follows:

$$\mathbf{D_{it}} = \alpha + \beta \mathbf{X_{it}} + \gamma \mathbf{H_{it}} + \theta \mathbf{W_{it}} + \lambda \mathbf{L_{it}} + \epsilon_{\mathbf{i}} \tag{3.2}$$

Where, $\mathbf{D_{it}}$ is the enrollment status, which will be 1 if the student is still enrolled or 0 if the person is not enrolled. This dependent variable is conditioned on the previous period enrollment.

$$\mathbf{D^*_{it}} = \beta \mathbf{X_i} \mid \mathbf{D_{t-1}} = \mathbf{1} \tag{3.3}$$

$\mathbf{D_{t-1}} = \mathbf{1}$, if the individual enrolled for that period. $\mathbf{D_{t-1}} = \mathbf{0}$, if the individual did not enrolled for that period. Equation 3.3 represents the basic modelling idea behind our dependent variable. $\mathbf{X_{it}}$ represents a vector of

[5]Although we won't be using it, it would be possible to use a parametric Accelerated time failure (AFT) model with a Weibull distribution, but we have opted for the Cox proportional hazards model, since it fairly approximates the correct distribution. This is explained in detail in the next section

child characteristics, $\mathbf{H_{it}}$ is the vector of the household characteristics, $\mathbf{W_{it}}$ is the wealth index, $\mathbf{L_{it}}$ is the dummy for location or region or residence, $\epsilon_i$ is the vector of residuals and $\alpha$ is a constant term.

For our results section, we will use 2 approaches, the Cox proportional hazards (PH) model and Kaplan-Meier survival curves. We have opted for this two estimation methods for various reasons. First, because of the ideal structure for the type of data we are using, that is, survival data. Second, the Kaplan-Meier survival curves are a non-parametric method which is easy to understand, implement and analyze, while the Cox proportional hazards (PH) model, is a semi-parametric method and posses the feature of approximating accurately the correct distribution of the process, Kleinbaum and Klein (2005). Third, two of the most relevant papers for our analysis use some of these methods. Both papers analyze dropout phenomena, while one, uses the same dataset,[6] Woldehanna and Hagos (2012), and the other, the country of study, Lavado et al. (2005). The use of the Cox PH model incorporates survival times into consideration and uses the censored data, while in an alternative logistic model, we would ignore survival times and censoring. In short, the Cox model uses more of the available information than the logistic model.

To understand the relevance of survival times and censoring, let's consider our data. We are using enrolment variation and the variation of a group of covariates to explore the decision of abandoning high-school. This presents us with two issues/features. First, we have censored data. Censoring means that the value of our measurement (enrolled or not enrolled) is partially

[6]Young lives dataset, but for Ethiopia.

known. In other words, we don't know the survival time exactly. This happens because when the value of our dependent variable is $\mathbf{D^*_{it}} = \mathbf{1}$, we know that the person has remained until that period, but that does not means that the person will finish high-school or that she wouldn't dropout in the near future. So what we do know, is that until that point in time, when $\mathbf{D^*_{it}} = \mathbf{1}$ at either $\mathbf{T} = \mathbf{1}, \mathbf{2}, \mathbf{3}$, the person has survived the failure event until that point in time, but that does not means that the risk of failure has disappeared. Also, even for individuals present in the 3 rounds of collection $\mathbf{D^*_{it}} = \mathbf{1}$ and $\mathbf{T} = \mathbf{1}, \mathbf{2}, \mathbf{3},$, we can't confirm that they will finish high-school, since our $3^{rd}$ round corresponds to the $10^{th}$ year of schooling (out of a total 11 years required to finish high-school). The second feature is the possibility of using survival time as the outcome of the analysis. This is a key feature of survival analysis, where T = Survival time (T≥0), t = specific value for T and $\delta$ = (0,1) this coefficient is the dummy variable which determines if the individual abandoned school or was censored. $\delta = 1$ indicates failure (in our case, that is, not enrolling in school for the current period) and $\delta = 0$ indicates censorship (attrition or still enrolled, but either case censored since we don't know if the individual actually finished high-school).

Two of the main outputs of survival analysis are S(t), the survivor function and h(t), the hazard function. The survivor function indicates the probability that a person survives more than the specified time t.

$$S(t) = P(T > t) \tag{3.4}$$

The hazard function, h(t), represents the instantaneous potential per unit

of time for the event to occur, given that the individual has survived up to time t.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t} \tag{3.5}$$

We must keep in mind that this hazard function expresses the instantaneous potential, in other words the hazard function express the probability of dropout, given that the person has enrolled until the previous round. But this does not imply that in the next period, the probability will remain unchanged, since the available information will increase.

The general formulae for S(t) and h(t) are:

$$S(t) = exp[-\int_0^t h(u)du)] \tag{3.6}$$

$$h(t) = -[\frac{\frac{\mathrm{d}S(t)}{\mathrm{d}t}}{S(t)}] \tag{3.7}$$

This set of equations allow us to grasp the interdependence of both functions, where equation (3.6) expresses the survival function in terms of an integral involving the hazard function. Equation (3.7) describes the hazard function, in terms of a derivative that includes the survivor function, Kleinbaum and Klein (2005).

### 3.3.1 Kaplan-Meier survival curves

We will employ Kaplan-Meier (KM) survival curves to represent survival probabilities. These curves are basically, the previously mentioned survival function, but ordered by failure time. The failure times are ordered from the

smallest to the largest and this method also makes use of the censored data. The Kaplan-Meier formula is the product of each conditional probability until the specified period. The KM formula is limited to the product terms up to the required survival week, that is why it is also known as the product-limit formula.

$$\hat{S}(t_{(j)}) = \prod_{i=1}^{j} \hat{P}r[T > t_{(i)} \mid T \geq t_{(i)}] \tag{3.8}$$

Alternatively, equation (3.8) can be expressed as the product of the survival estimate for the previous period, times the conditional probability of surviving past the present failure time. Kleinbaum and Klein (2005)

$$\hat{S}(t_{(j)}) = \hat{S}(t_{j-1}) * \hat{P}r[T > t_{(i)} \mid T \geq t_{(i)}] \tag{3.9}$$

### 3.3.2 Cox proportional hazards (PH) model

The Cox proportional hazards (PH) model is the commonly used mathematical model when working with survival data.

$$h(t, X) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i} \tag{3.10}$$

The model express the hazard at time t, as the product of the baseline hazard $h_0$(t) and e elevated to the linear sum of $\beta_i X_i$ over p explanatory variables. From the formula, we can observe that the first term, the **baseline hazard**, is a function of t, but not of X. Inversely, the second term, the exponential expression, is a function of X but not of time. It is important to mention that if the X's were time-dependent, the model wouldn't be a proportional Cox model, the proportionality assumption would not be satis-

fied and would require an extended Cox model. It is important to mention that the baseline hazard, $h_0$(t), is an unspecified function, making this a semi-parametric model. It is possible to use baseline hazard functions, with known functional forms, that is the case of the Weibull distribution used by Lavado et al. (2005). Despite the semi-parametric condition of the Cox PH model, the model has proven to be "robust" with results that constantly approximate the results of the correct parametric option.

The estimates of the parameters of this model are obtained via maximum likelihood (ML) estimation. This estimates will be called $\hat{\beta}_i$. The ML estimation is usually based on the specified outcome distribution, in this case, the distribution has not been specified, hence the Cox likelihood is based on the order of the events rather than their distribution. For this reason, these approach is called "partial likelihood" and it is considered a semi-parametric approach.

$$L = L_1 * L_2 * L_3 * ... * L_k = \prod_{j=1}^{k} L_j \tag{3.11}$$

Then, setting the partial derivatives of the natural log of L to zero and solving the system of equations we obtain the correspondent scores.

$$\frac{\partial lnL}{\partial \beta_i} = 0 \tag{3.12}$$

Where i=1,2,3,...,p and p = # of parameters.

### The Hazard ratio

Once we obtain the estimates, we will compute hazard ratios to make statistical inference. The hazard ratio is defined as the ratio of two different individuals hazard functions.

$$\widehat{HR} = \frac{\hat{h}_0(t)e^{\sum_{i=1}^{p}\hat{\beta}_i X_i^*}}{\hat{h}_0(t)e^{\sum_{i=1}^{p}\hat{\beta}_i X_i}} \tag{3.13}$$

Which after simple algebraic manipulation, can be expressed as:

$$\widehat{HR} = e^{\sum_{i=1}^{p}\hat{\beta}_i(X_i^* - X_i)} = exp[\sum_{i=1}^{p}\beta_i(X_i^* - X_i)] \tag{3.14}$$

This expression allows us to grasp two important features, first, the baseline hazard, cancels out. Second, the only difference between the numerator and the denominator is generated by the two different sets of X's, $X_i^*$ and $X_i$.

### Adjusted survival curves using the Cox PH model

Another desired outcome from survival analysis is obtaining estimated survival curves, using the Cox model. Given that these survival curves are constructed with estimations from a Cox model, they are called, adjusted survival curves. Where the adjustment refers to the inclusion of explanatory variables used as predictors. Lets remember that if we don't use a model to fit survival data, we can still construct a survival curve using the aforementioned Kaplan-Meier method.

We can obtain adjusted survival curves, for two different levels of an exposure variable, where $X_1 = 1$ and $X_0 = 0$.

$$\hat{S}(t, X_1) = [\hat{S}_0(t)]^{exp[\hat{\beta}_1(1) + \sum_{i \neq 1} \hat{\beta}_i \overline{X}_i]} \tag{3.15}$$

$$\hat{S}(t, X_0) = [\hat{S}_0(t)]^{exp[\hat{\beta}_1(0) + \sum_{i \neq 1} \hat{\beta}_i \overline{X}_i]} \tag{3.16}$$

Where, equation 3.15 refers to exposed subjects and equation 3.16 to unexposed subjects. It is also possible to obtain an adjusted survival curve which considers all the covariates in the model, using equation 3.17.

$$\hat{S}(t, \overline{X}) = [\hat{S}_0(t)]^{exp[\sum \hat{\beta}_i \overline{X}_i]} \tag{3.17}$$

This would give us a single adjusted survival curve, rather than different curves for each type of group.

### The PH assumption

Finally, we will explain the proportional hazards (PH) assumption, which is one of the key assumptions of this model and we will need to test this assumption upon our data before proceeding with estimation. The PH assumption, proposes that the hazard ratio between any two specifications should be a constant ratio over time.

$$\widehat{HR} = \hat{\theta} = exp[\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)] \tag{3.18}$$

Equation (3.18) shows that the entire expression is independent of time,

hence, we can claim it as a constant. An easy way to check the compliance of this assumption is to graph each group, if the functions cross at any point, we can claim that the PH assumption is not met. If this is the case, the Cox PH model should not be used, Kleinbaum and Klein (2005). This is the basic theoretical framework required to understand the results section. Although, we cover most of the key concepts, we will explain any other technical tool used in the results section. Most of the additional notes, will be included in the appendix section or as footnotes.

# Chapter 4

# Empirical Results

## 4.1 Kaplan-Meier survival estimates (Non-parametric estimation)

We have obtained the Kaplan-Meier survival estimates for a set of covariates that affected the "Old Cohort" during the period of study. Woldehanna and Hagos (2012) used a similar fashion in their paper, where they explored the Kaplan-Meier survival estimates for factors such as: death or illness of family members, death of livestock, drought, crop failure, pest infestation and diseases, gender and location. We have obtained the Kaplan-Meier survival estimates for the covariates that help us explore our hypothesis of interest, that is: Child initial traits (gender of the child), socioeconomic status (wealth index), location of the household (typesite or region), caregiver gender (sex). We have also included an additional set of Kaplan-Meier survival estimates to explore the effect of idiosyncratic shocks.

Before running the estimates, we applied the Stata command "**stsplit**" to split the data across the years of schooling achieved by the children. To understand the effect of this transformation, lets consider an arbitrary individual from our panel, individual id:18044. Originally we have 3 entries of data for this id, rounds 1, 2 and 3, of collection. Nevertheless, in round 1, id:18044 reported to be enrolled in school and currently studying $2^{nd}$ grade of primary school; in round 2, the same id reported to be enrolled in $6^{th}$ grade of primary school; and in round 3 he reported to drop out and presented a missing value for the grade he was currently in, since he wasn't enrolled. What the "**stsplit**" command will do is create values for grades $3^{rd}$,$4^{th}$ and $5^{th}$, using the same values as the ones obtained in the $1^{st}$ round of collection, and since in the $3^{rd}$ round of collection he responded not to be enrolled, the last value for this individual will be $6^{th}$ grade. It is obvious that there is a missing information problem, since, the individual may have enrolled in $7^{th}$ grade and dropped during the $8^{th}$ grade or any other combination. Since not all the children will be in the same grade in each round of collection, we have been able to use the "stsplit" command to construct a 10-grade (from $1^{st}$ to $10^{th}$) process and analyze the dropout probabilities in that time-frame, allowing us to observe any grade-specific effects. Our initial sample was of 2,142 (714 individuals and 3 round of observations.) observations, now we have created 3,227 observations which accounts for a total of 5,369 observations. It is also important to consider the distribution of the "time" variable, child grade, "**chgrade**", which approximates a normal distribution throughout the 10 years of spread, with flat tails and right-skewness.[1]

[1]Appendix Figure A.1.

**Gender of the Child:** We begin exploring our main hypothesis, that is, child's initial conditions. Since gender is a non-time-varying variable, the effects are easy to interpret. Results show that, consistent with Woldehanna and Hagos (2012) and contrary to Lavado et al. (2005), boys present a smaller survival rate, consistently throughout their entire school-life, Figure (4.1) and this rate decreases throughout high-school, where in grades, $7^{th}$, $8^{th}$ and $9^{th}$, survival rates show the highest gap between boys and girls. Also, consistent with most literature, the probability of failure (dropout) monotonically increases for both genders during the entire 10 years of the study, which is an interesting finding, policy wise. If we revise Lavado et al. (2005) findings, which correspond to almost a decade ago, the situation was quite different. The authors found that the probability of dropout was slightly higher for women. It seems that this is not the case anymore and that now the shape of the problematic has changed. It is possible to speculate that the multiple efforts executed by the Peruvian government, during the past 20 years, in behalf of woman's rights protection, have worked, at an extent. Although there is still an embarrassing amount of gender related incidents, Peru has a government agency specifically devoted to protect women and women's related issues, like husband abuse or work discrimination. Other notorious examples are the inclusion of women in the military forces and their participation in the traffic police. Also, most of the traditional Peruvian families still rely on their maternal figures for house labour and children care.

**Socioeconomic status (SES):** Our second hypothesis is aiming to confirm the relevance of socioeconomic status. This has been proven in most

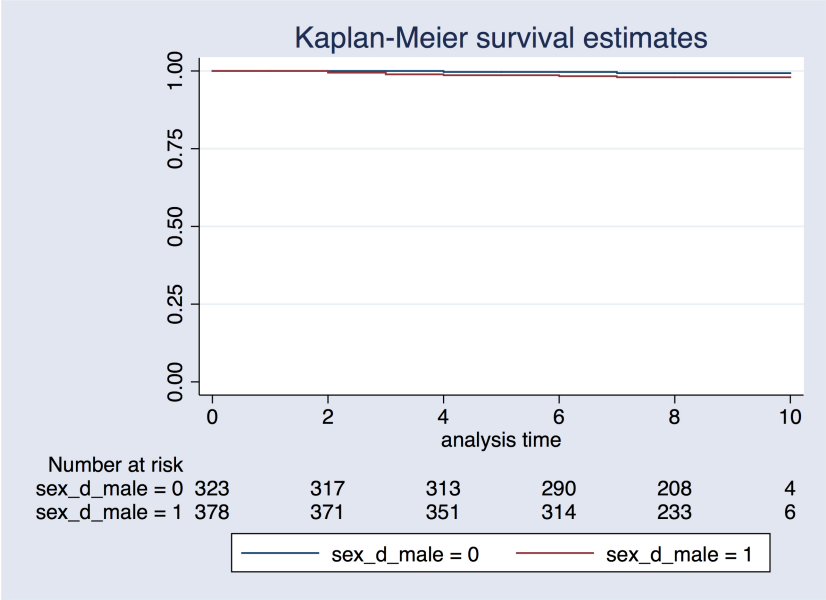


Figure 4.1: Kaplan-Meier survival estimates by gender in all rounds.

cases and ours is not an exception. We represent SES by our previously explained wealth index (WI), equation (3.1), which is a composite of 3 sub-indexes. Since it is a continuous variable, that changes over time, we created a dummy for those who had a WI above the average of the total sample. This proved to have a remarkable effect, where, from grades $6^{th}$ onwards, the survival probability diminishes substantially for those with a WI below average. The largest drop is experienced in the $7^{th}$ grade, where the gap between those with an above average WI and those below average, reaches its maximum separation, Figure (4.2). Presumably, from the evolution of the "at risk" populations, we can infer that the sample of those with a WI above average, remain fairly constant around 300, while those with a WI below average, experience substantial drop outs, from grades $6^{th}$, onwards.

**Location of the household (Urban vs. Rural):** Location proved to

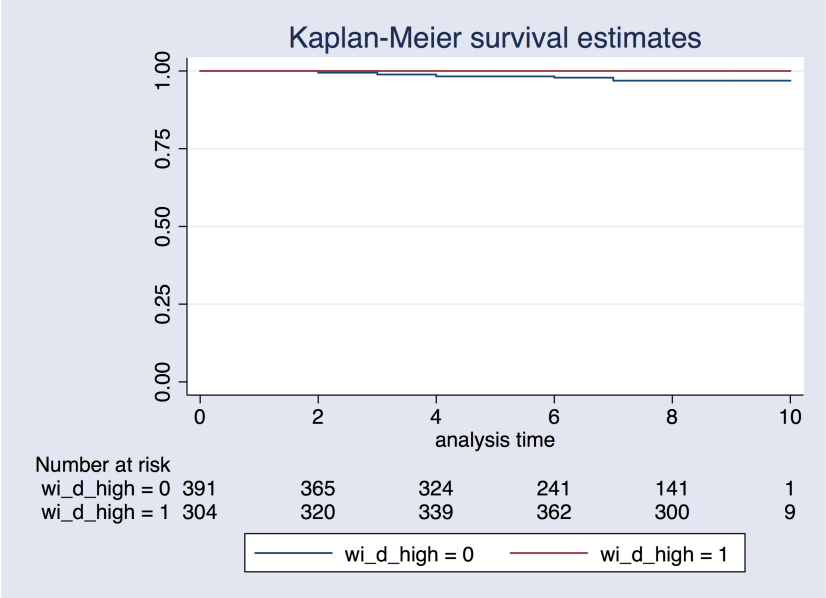


Figure 4.2: Kaplan-Meier survival estimates by wealth index in all rounds.

be a very relevant issue. Both, **Lavado et al. (2005)** and **Woldehanna and Hagos (2012)** found significantly higher hazard rates for rural groups, nevertheless, we have found that this is not the case for our sample, in fact, there is a slightly higher survival probability for the rural children from grades $2^{nd}$ to $6^{th}$ (Figure 4.3), which can be related to the strong policy towards education supply during the past 2 decades in Peru. This is also an important finding since there is a persistent tradition in Peru, of underestimating the rural outcomes, nevertheless, this stereotype doesn't seem to apply anymore, where most of the problems are now emerging from the poverty related to the big urban centers.[2].

**Caregiver gender:** Our third hypothesis, examined if the gender of the caregiver was relevant in terms of the decision to drop out. We had some

[2]Lima is the capital of Peru and is a city with an estimated population of 9,838,251 up to 2015. Source:Peruvian National Institute of Estatistics (http://www.inei.gob.pe).

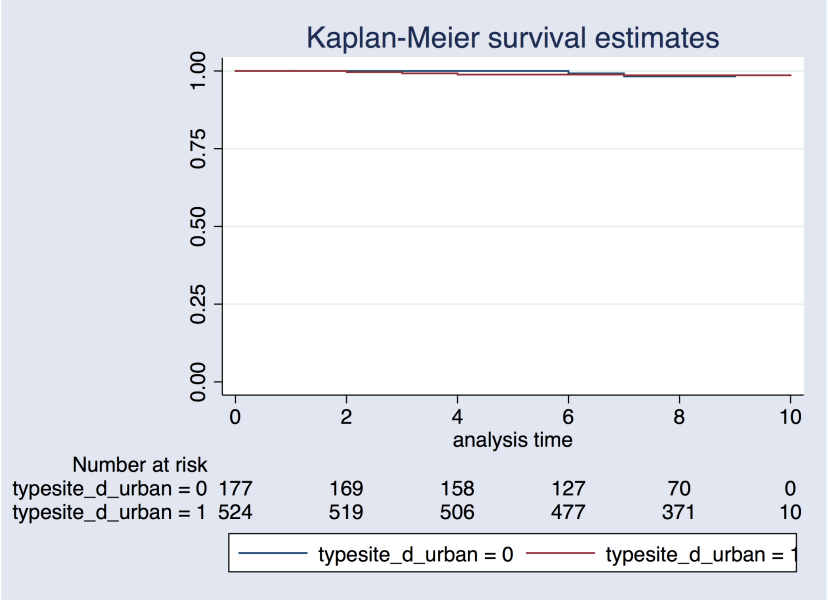


Figure 4.3: Kaplan-Meier survival estimates by Location of the household (Urban vs. Rural) in all rounds.

ex-ante expectations upon the relevance of the presence of the mother in the household and the effect of the person in charge of the children's care, within the household. According to (Figure 4.4), children with a male caregiver present higher failure probabilities than those with female caregivers. This probability is specially high for grades $6^{th}$ to $10^{th}$, reaching it's maximum distance in the $7^{th}$ grade, where presumably, children are becoming older and have more of the required skills to get a paid job. We can begin to suspect that the $7^{th}$ grade is one of the natural cut-offs and since it is the grade of transition form primary to secondary school (first year of high-school) it is reasonable to observe a higher dropout probability in that grade. Unfortunately, the sample size for those with a male caregiver is very small and making inference upon such small figures, is risky, specially, in a non-parametric approach. We will revise this hypothesis again, in the next

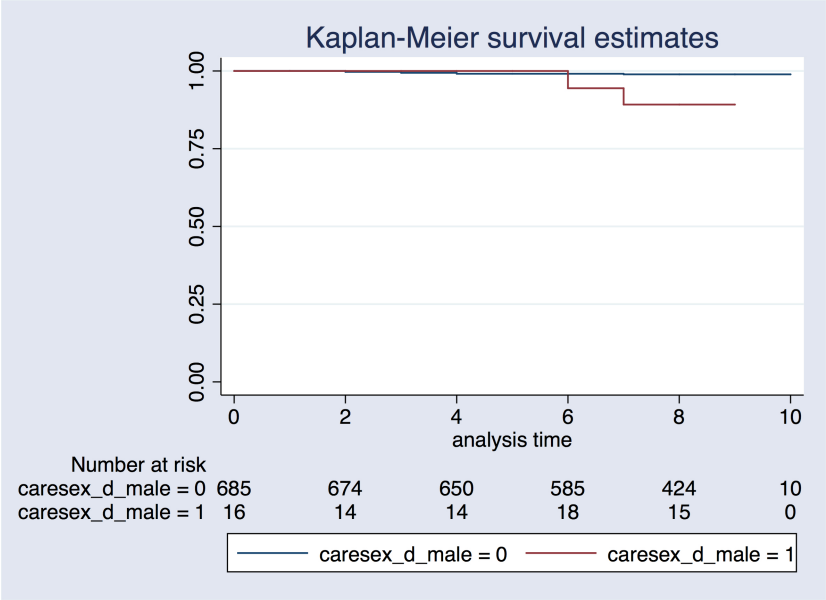


Figure 4.4: Kaplan-Meier survival estimates by Caregiver gender in all rounds.

section, when we will run the semi-parametric estimations and compare.

**School type (Private vs. Public):** Another interesting question, although it wasn't included in the hypothesis, is that of the school characteristics. Our dataset is not focused in such type of covariates, nevertheless, it includes the private school vs. public school dummy, which address such a question. The outcome confirms our ex-ante expectations, since there is a higher survival probability for the children in private schools, where the results are close to 100% survival, while public schools show the aforementioned increasing-risk during high-school. This is also interesting, since it could point out a few things. First, children in private schools might be receiving "better education", if so, they probably value this "better education" accordingly, and hence decide not to proceed with the drop out. Second, it seems that once a child manages to get into private education in Peru, the

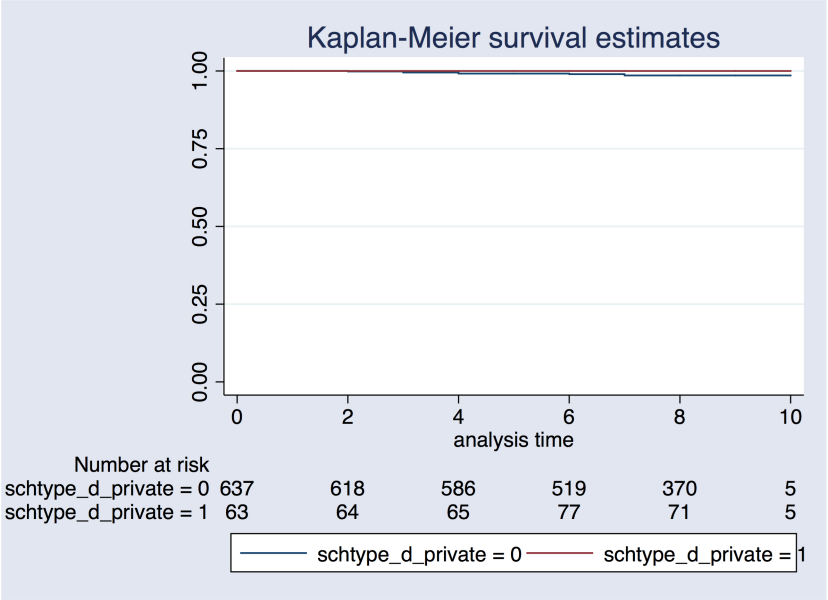


Figure 4.5: Kaplan-Meier survival estimates by School type (Private vs. Public) in all rounds.

probability of dropout diminishes significantly. Third, since there is a higher monetary cost associated to private education, parents with a higher valuation of education will be willing to invest a higher amount of capital in their children's education, providing them with private education. It is possible to argue then, that the value perceived from private education in Peru, by both, parents and children, is high enough, to persuade them from leaving high-school, while this is not the case with public education, Figure (4.5).

**Idiosyncratic Shocks:** We have included 3 sets of shocks that are known to affect rural and sometimes even urban populations. Shocks such as the death or illness of a family member, The death or illness of livestock and natural effects such as, Drought, crop failure, pest infestation and diseases. These are the 3 conglomerates of idiosyncratic shocks that may affect some of the households in the study. We found that for the first group, death or illness

of a family member, there is no difference in the survival estimates for both groups, Figure (A.2), nevertheless we also found that the highest dropout risk is experienced in the $7^{th}$ grade, which also happens in the next two cases. The second group, death or illness of livestock presents a slightly higher survival probability for those who experienced the shock, unfortunately, the sample for that group is smaller than 25 in every grade, hence we won't perform any between-group inference. Finally, the third group, drought, crop failure and pest infestation and diseases, showed a slightly higher survival probability for those who experienced the shock, this, until $7^{th}$ grade, where again both groups experience a similar survival probability. This effect might be explained by the idea that children who experience any of these shocks might not be asked to help in agricultural labour anymore and may be allowed to attend school more regularly. The effect is too small to support this proposition, nevertheless, other papers have pointed similar explanations, **Woldehanna and Hagos (2012)**.[3]

[3]Figures for these Idiosyncratic shocks are presented in the appendix section, A.2, A.3, A.4.

## 4.2 Cox proportional hazards (PH) estimates (Semi-Parametric estimation)

Before providing the results for the complete model, we will present the table of first-order correlations, to get a sense of which covariates show significance, before controlling for any type of characteristic.

In the first group, the child characteristics group, there were 10 characteristics that where evaluated, Table(4.1). Most of these covariates, can be considered initial conditions of the child. First, the gender of the child **"sexdmale"**, which has a hazard ratio of 3.120 but is not significant at the 10% level. Since this is a dummy value which indicates if the child is male or not, the hazard rate of 3.120 implies that the probability of failure (drop out) of the boys divided by the probability of failure of the girls is 3.120, hence it is preferable to be a girl under these circumstances. This value confirms the proposition of the previously found disadvantage for boys, in the non-parametric section, (Figure 4.4). Nevertheless, since it is not significant at the 10% level, we can only use the direction of the effect as an indication. The proportionality test conducted confirmed that the proportionality conditions where met for this covariate. It is important to remember that if the value of the $\chi^2$ is high enough and is associated with a P-Value smaller than 0.05 then we could reject the null hypothesis of proportionality at the 5% confidence level and shouldn't use that covariate with a Cox proportional hazards model. Since this is not the case with this covariate, we can't reject the proportionality assumption, and hence, we can use the Cox PH model.

We also explore the variable "bmi" or Body mass index, which is a contin-

uous variable, hence its interpretation is different from the previous variable which was a dummy. In this case, the hazard ratio of 0.990 indicates that a 1% increase in the bmi index would reduce the probability of dropout in 0.01%. Again, the P-Value associated to the Z-statistic is 0.908, hence, we lack of significance at the 10% level.[4]

Then we used the "ppvtraw" indicator, which is the Peabody picture vocabulary test, raw score. This is used to asses our main hypothesis about child's initial conditions. We found that a 1 point increase in the raw score[5] has a positive effect of 6.2% on the instantaneous probability of survival. This value is significant at the 1% level. This would be the first strong finding supporting our main hypothesis, but we will assess this again in the complete model specification.

We also evaluated the dummy variable "stunt", which indicates those children whom presented 2 standard deviations below the median in the height for age score (HFA). This variable can also be considered as a child's initial conditions indicator and the results show a hazard rate of 5.297, than means that the instantaneous probability of the group of those children who where found to be stunted, where 5.2 times higher than that of those who were not found to be stunted, this at the 5% significance level. This results reinforce the idea of the importance of the child's initial conditions which fits into the idea of the Heckman 2-period model, and the importance of the investment if the early stages of human capital development, Heckman (1976).

[4]We only showed these first two variables to guide the reader through the interpretation methodology.

[5]The minimum score in the test is 10 and the maximum is 125.

A Categorical variable for ethnicity was also included "*chethnic*", finding the "mestizo" ethnicity to present a higher hazard than the others. The 3 ethnicity's presented were compared against the "white" ethnicity, but the sample size is heavily biased towards "mestizo" ethnicity and we believe that the size of the effect is just a reflection of the 4,960 "mestizo" observations out of a total of 5,360 observations. Although it is significant, it is hard to use such a figure to make inference.

The next categorical variable is child religion, "*chldrel*" where the benchmark is catholic religion. Again the sample is biased towards catholic religion, although the sample for evangelists is also large enough (693 observations) to allow for inference. The hazard rate found is 3.322 which implies that the evangelists have an instantaneous failure probability 3.3 times higher than the catholics, at the 10% significance level. It can be argued that this is not a child's "initial condition", but it is certainly a defining characteristic which affect the child's behaviour and their household conditions.

Finally, within the child characteristics, we found that a 1 unit (cm) increase in the child height[6], "*chheight*", showed a positive effect of 6% in the instantaneous survival probability of the child, this at the 1% significance level. Curiously, the hazard rate for child weight, "*chweight*", is very similar, but is not signifcant at the 10% level.

Within the household characteristics $HH_{it}$, there are quite a few covariates of interest, Table (4.2). First, the size of the household, "*hhsize*". Some studies, mentioned the importance of this variable and specially of the number of children living in the house. This idea comes from the fact that in

[6]Child height had a minimum of 93.9cm and a maximum of 180cm.

some developing countries, when both, father and mother must get a paid job, the older brother or sister is required to take care of the younger siblings and these activities may require them to use part of their school time. As we mentioned earlier, missing school is a strong determinant of dropout. In our sample, we found a hazard rate of 1.306 for this continuous variable. This implies that for a 1 unit increase in the number of members of the household[7], the instantaneous risk of failure (drop out) increases by 30.6%, this at the 1% confidence level.

Second, we explored the categorical variable, caregivers education, "*caredu*". Within the 14 possible categories, two were found significant. First, $11^{th}$ grade or complete secondary education and $13^{th}$ grade or complete technical education, where the hazard rates where very close to zero in both cases. We believe that, although only this two levels where found statistically significant, there are some other interesting findings. First, grades $2^{nd}$, $3^{rd}$ and $5^{th}$ presented the higher hazard rates. This is consistent with the idea that the higher the level of education of the caregiver, the lower the probability of failure of the child. Also, we believe that once we get access to the fourth round of data collection and we have a larger sample for grades $10^{th}$, and $11^{th}$, we will be able to have more precise estimates in terms of the size of the effect.

Finally, within the household characteristics, we explored the relevance of the caregivers sex, which is in line with our fourth hypothesis. The dummy "*caresexdmale*" which indicates if the caregiver is a male, presented a hazard rate of 11.375 at the 1% significance level. This implies that children with

[7]Which had a minimum of 1 and a maximum of 17 members in our sample.

male caregivers have an instantaneous failure probability that is 11.4 times higher than that of those with women caregivers.

Our second hypothesis explored the relevance of socioeconomic status, represented by our constructed wealth index, "*wi*", where we found a hazard rate of 0.001, significant at the 1% level. This would imply that a 1% increase in the wealth index, have a 0.99% positive effect on the probability of survival, confirming previous findings on the relevance and direction of the wealth effect in the determination of dropout. This is almost a 1 to 1 effect.

Our third hypothesis, explored the importance of the household location,"$L_{it}$", although, none of the covariates were found significant at the 10% level, there seems to be a higher hazard associated with living in the jungle or in the highlands, while urban locations presented a hazard rate slightly above 1. The urban hazard rate, 1.028 is in line with the non-parametric estimation, where we found that the survival probability was slightly lower for the urban children, up to the $7^{th}$ grade.

Now that we have explored the first order correlations results and contrasted them with our main and secondary hypothesis, we will analyze the results from the complete specification.

$$\mathbf{D_{it}} = \alpha + \beta \mathbf{X_{it}} + \gamma \mathbf{H_{it}} + \theta \mathbf{W_{it}} + \lambda \mathbf{L_{it}} + \epsilon_{\mathbf{i}} \tag{3.2}$$

We have included the following covariates in the complete specification: $X_{it}$: Child sex dummy (male = 1), bmi, ppvtraw, stunt, cladder, agemon, ethnicity "mestizo", ethnicity "amazon native", ethnicity "negro", religion

"evangelist", religion "mormon", religion "none", child weight, child height. $HH_{it}$: Household size, caregiver education (categorical variable with 14 possible levels of achievement), caregiver sex dummy (male = 1). $WI_{it}$: Wealth index (0 to 1). $L_{it}$: Location dummy (urban = 1), region of the household (categorical variable which includes the comparison of the highland region and the jungle against the benchmark region, which is the coast), Tables (4.3) and (4.4).

All these covariates where include in a single Cox proportional hazards (PH) regression. In regards to our main hypothesis, child's initial conditions, 4 factors were found to be significant, after controlling for sex, bmi, age, religion and cladder[8]. Those significant factors are, ppvtraw, stunt, ethnicity and child height

The first significant covariate, that supports our main hypothesis, is the "ppvtraw" score, which was also significant in the previous regressions, and which maintains a coefficient very close to the one found on the first order correlations Table (4.1), but with a 1% significance level. In this case, a 1 point increase in the score would represent a 9.4% increase in the probability of survival of the subject, which is consistent with the size and direction of the previously found effect.

The second significant covariate is "stunt", which indicates those children whom presented 2 standard deviations below the median in the height for age score (HFA). This variable can also be considered as a child's initial conditions indicator and was found significant at the 5% level. This is consistent

[8] "cladder" is an indicator of the child's ladder or life satisfaction, which goes from 1 to 9.

with our previous results from the first-order correlations.

Also, regarding initial conditions, child weight was found significant with a hazard ratio of 0.127 which means that a 1 kilogram increase would represent an instantaneous survival probability increase of 88%, this at the 5% significance level. In the same fashion, child height was found significant but in the opposite direction, where a one unit increase in height, would increase the instantaneous failure (drop out) probability in 151% at the 5% significance level. This result is surprising, since in the first order correlations we found that child height was also significant but in the opposite direction. We believe that the direction of the coefficient obtained from the complete model should be more accurate, since it incorporates all the required controls.

Given the results from these covariates, we are inclined to believe that the answer to our main hypothesis: Are the child's initial traits, determinants of dropout in Peru? Is a probable, yes. The size and direction of the effects for 3 of the 4 significant covariates are relatively consistent and point to the same direction as the results, before including the controls.

Regarding our secondary hypothesis, the child's socioeconomic status and its relevance, we used the wealth index and found that it is again, significant at the 1% level, with a coefficient of 6.13e-07, which is smaller than the previously found 0.001, but the direction and interpretation is fairly similar with consistent statistical significance. Since we found consistent results in the 3 stages of our estimation, that is, the Kaplan-Meier survival estimates, the Cox PH first order correlations and the Cox PH complete model estimation, we are inclined to conclude that there is a positive effect regarding SES and that the lower the SES the higher the instantaneous probability of failure.

This is consistent with most literature.

Our third hypothesis is about the household characteristics, in specific, the location of the household. In this respect, we did not found evidence in the Kaplan-Meier survival estimates, neither in the first order correlations (Although the direction and size of the coefficients made sense), nevertheless, in the complete model Cox estimates, we found the dummy "location" (urban = 1) to be significant at the 5% significance level, while the hazard ratio proposes an increase of 6,773.5 in the instantaneous probability of failure of those who lived in a urban site. Again, we are suspicious about the magnitude of the effect, we believe that once the fourth round of collection is included, the size effects will be much more accurate, nevertheless, we do believe there is an increased risk of failure for those living in urban sites as compared to those living in rural sites.

We also found a significant coefficient for those living in the highlands and associated hazard ratio of 0.059, significant at the 5% level, which would imply that children living in the highlands have a 94% lower instantaneous failure probability compared to those living in the coast.

Finally, our last hypothesis, questioned the relevance of the caregivers gender. In regards to this, we found the covariate "caresedmale", to be significant at every stage, in the Kaplan-Meier estimates, in the Cox first-order estimates and in the Cox complete-model estimates. In all the cases, a male caregiver represented an increase in the instantaneous failure risk, which accounts for a hazard rate of 1,509.2 at the 10% level. This would imply a much higher risk of failure for those children with male caregivers. Again, we believe that the significance and sign of this effect is quite clear,

while the magnitude should be clarified once we include the fourth round of collection. Regardless of the size of the effect, we can fairly answer our fourth hypothesis: Does the gender of the caregiver affects the high-school dropout decision? All of our estimates seem to indicate that it does, at least in the Peruvian case.

Table 4.1: Cox proportional hazards model (semi-parametric estimation, first-order correlations)

| | Cox PH Estimation | | | PH test | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| | HR | SE | P-Val | $\chi^2$ | P-Val |
| $X_{it}$ | | | | | |
| sexdmale | 3.120 | 2.484 | 0.153 | 1.08 | 0.297 |
| bmi | 0.990 | 0.084 | 0.908 | 0.01 | 0.942 |
| ppvtraw | 0.938*** | 0.017 | 0.001 | 0.60 | 0.437 |
| stunt | 5.297** | 3.909 | 0.024 | 2.80 | 0.094 |
| cladder | 0.838 | 0.167 | 0.379 | 3.34 | 0.067 |
| agemon | 1.002 | 0.021 | 0.905 | 0.14 | 0.704 |
| *chethnic* | | | | | |
| mestizo | 1.79e+15*** | 1.22e+15 | 0.000 | 0.00 | 1.000 |
| amazon native | 0.367 | 0.234 | 0.116 | 0.00 | 1.000 |
| negro | 0.367 | 0.305 | 0.229 | 0.00 | 1.000 |
| *chldrel* | | | | | |
| evangelist | 3.322 | 2.344 | 0.089 | 4.41 | 0.035 |
| mormon | 3.82e-18 | . | . | 4.41 | 0.035 |
| none | 5.17e-19 | . | . | 4.41 | 0.035 |
| chweight | 0.950 | 0.036 | 0.188 | 0.15 | 0.699 |
| chheight | 0.947*** | 0.017 | 0.004 | 0.15 | 0.693 |

Source: Own estimates, based on Young Lives program data.
Note: *a*. Figures in column 1 represent the hazard ratio obtained with the cox proportional hazard model. Figures in column 2 are the robust standard errors and the corresponding P-value are presented in column 3. *b*. Columns 4 and 5 correspond to the proportional hazards test, where the $\chi^2$ is presented in column 4 and its corresponding P-value in column 5. *c*. Values with (*) correspond to the 10% significance level, values with (**) correspond to the 5% significance level and coefficients with(***) correspond to the 1% significance level.

Table 4.2: Cox proportional hazards model (semi-parametric estimation) - First order correlations

| | Cox PH Estimation | | | PH test | |
|---|---|---|---|---|---|
| | (1) | (2) | (3) | (4) | (5) |
| | HR | SE | P-Val | $\chi^2$ | P-Val |
| $HH_{it}$ | | | | | |
| hhsize | 1.306*** | 0.095 | 0.000 | 0.00 | 0.954 |
| **caredu** | | | | 4.80 | 0.569 |
| 1 | 1.53e-19 | . | . | | |
| 2 | 3.393 | 4.760 | 0.384 | | |
| 3 | 6.548 | 7.565 | 0.104 | | |
| 4 | 1.53e-19 | . | . | | |
| 5 | 5.089 | 6.219 | 0.183 | | |
| 6 | 1.520 | 1.858 | 0.732 | | |
| 7 | 1.52e-19 | . | . | | |
| 8 | 1.52e-19 | . | . | | |
| 9 | 1.52e-19 | . | . | | |
| 10 | 1.52e-19 | . | . | | |
| 11 | 1.52e-19*** | 1.51e-19 | 0.000 | | |
| Inc.Tech.College | 1.52e-19 | . | . | | |
| Com.Tech.College | 1.52e-19*** | 1.52e-19 | 0.000 | | |
| Inc.University | 4.16e-19 | . | . | | |
| Com.University | 4.14e-19 | . | . | | |
| caresexdmale | 11.375*** | 8.477 | 0.001 | 2.52 | 0.112 |
| $WI_{it}$ | | | | | |
| wi | 0.001*** | 0.001 | 0.001 | 3.3 | 0.069 |
| $L_{it}$ | | | | | |
| typesitedurban | 1.028 | 0.853 | 0.973 | 4.77 | 0.029 |
| regiondjungle | 1.904 | 1.758 | 0.485 | 1.54 | 0.464 |
| regiondhighland | 1.344 | 1.035 | 0.700 | 1.54 | 0.464 |

Source: Own estimates, based on Young Lives program data.
Note: *a*. Figures in column 1 represent the hazard ratio obtained with the cox proportional hazard model. Figures in column 2 are the robust standard errors and the corresponding P-value are presented in column 3. *b*. Columns 4 and 5 correspond to the proportional hazards test, where the $\chi^2$ is presented in column 4 and its corresponding P-value in column 5. *c*. Values with (*) correspond to the 10% significance level, values with (**) correspond to the 5% significance level and coefficients with(***) correspond to the 1% significance level.

Table 4.3: Cox proportional hazards estimation (semi-parametric estimation of the complete model.)

$$\mathbf{D_{it}} = \alpha + \beta\mathbf{X_{it}} + \gamma\mathbf{H_{it}} + \theta\mathbf{W_{it}} + \lambda\mathbf{L_{it}} + \epsilon_{\mathbf{i}} \tag{3.2}$$

| | Haz. Ratio (1) | Std.Err. (2) | $P > \|Z\|$ (3) | [95% Conf. (4) | Interval] (5) |
|---|---|---|---|---|---|
| $X_{it}$ | | | | | |
| Childsex Male | 0.452 | 0.617 | 0.561 | 0.031 | 6.582 |
| bmi | 88.427** | 162.616 | 0.015 | 2.405 | 3250.316 |
| ppvtraw | 0.905 | 0.056 | 0.104 | 0.802 | 1.021 |
| stunt | 8.333* | 10.412 | 0.090 | 0.719 | 96.481 |
| cladder | 0.732 | 0.238 | 0.339 | 0.387 | 1.385 |
| agemon | 1.086 | 0.064 | 0.161 | 0.968 | 1.219 |
| **Ethnicity** | | | | | |
| mestizo | 5.57e+13 | . | . | . | . |
| amazon native | 2.94e-08 | . | . | . | . |
| Negro | 2.99e+22 | . | . | . | . |
| **Religion** | | | | | |
| evangelist | 2.013 | 2.831 | 0.619 | 0.128 | 31.679 |
| mormon | 172756.4 | . | . | . | . |
| none | 4.02e-18 | . | . | . | . |
| chweight | 0.127** | 0.115 | 0.022 | 0.022 | 0.744 |
| chheight | 2.508** | 1.051 | 0.028 | 1.104 | 5.699 |
| Wald $\chi^2_{(19)}$ | 182.40 | | | | |
| N | 3,991 | | | | |

Source: Own estimates, based on Young Lives program data.
Note: *a*. Column (1) expresses the results for the Hazard rates from the Cox PH estimates. *b*. Values with (*) correspond to the 10% significance level, values with (**) correspond to the 5% significance level and coefficients with(***) correspond to the 1% significance level.

Table 4.4: Cox proportional hazards estimation (semi-parametric estimation of the complete model.) (Continuation of Table 4.3)

| | Haz. Ratio (1) | Std.Err. (2) | $P > \|Z\|$ (3) | [95% Conf. (4) | Interval] (5) |
|---|---|---|---|---|---|
| $H_{it}$ | | | | | |
| hhsize | 2.645 | 2.244 | 0.252 | 0.502 | 13.953 |
| **caredu** | | | | | |
| 1 | 3.66e-21 | . | . | . | . |
| 2 | 0.263 | 0.924 | 0.703 | 0.001 | 252.249 |
| 3 | 5.287 | 6.932 | 0.204 | 0.405 | 69.061 |
| 4 | 2.95e-28 | . | . | . | . |
| 5 | 8.126 | 21.178 | 0.421 | 0.049 | 1343.68 |
| 6 | 0.845 | 1.269 | 0.911 | 0.045 | 16.011 |
| 7 | 5.26e-29 | . | . | . | . |
| 8 | 1.68e-19 | . | . | . | . |
| 9 | 1.76e-22 | . | . | . | . |
| 10 | 2.62e-23 | . | . | . | . |
| 11 | 4.84e-34 | . | . | . | . |
| Inc.Tech.College | 1.49e-18 | . | . | . | . |
| Com.Tech.College | 2.01e-19 | . | . | . | . |
| Inc.University | 4.15e-19 | . | . | . | . |
| Com.University | 2.16e-20 | . | . | . | . |
| caresexdmale | 1509.249* | 6446.974 | 0.087 | 0.348 | 6528161 |
| $WI_{it}$ | | | | | |
| wi | 6.13e-07*** | 3.16e-06 | 0.005 | 2.53e-11 | 0.015 |
| $L_{it}$ | | | | | |
| urban | 6773.544** | 30508.79 | 0.050 | 0.993 | 4.62e+07 |
| **region** | | | | | |
| highland | 0.059** | 0.072 | 0.019 | 0.006 | 0.632 |
| jungle | 1.016 | 1.951 | 0.993 | 0.024 | 43.757 |
| Wald $\chi^2_{(19)}$ | 182.40 | | | | |
| N | 3,991 | | | | |

Source: Own estimates, based on Young Lives program data.
Note: *a*. Column (1) expresses the results for the Hazard rates from the Cox PH estimates. *b*. Values with (*) correspond to the 10% significance level, values with (**) correspond to the 5% significance level and coefficients with(***) correspond to the 1% significance level.

# Chapter 5

# Conclusions

Our research shows that the dropout process has some specific noticeable particularities and that policy can be refined within those specifics. First, we found that the $7^{th}$ grade is one of the cutoff points for dropouts, which is consistent with previous studies. We don't have a large enough sample in the latest grades of high-school in order to check for a secondary cutoff such as $10^{th}$ or $11^{th}$ grades, which other studies found as secondary cutoffs. Nevertheless, we believe this will be overcome with the next round of information. Regarding our main hypothesis, evaluating the relevance of children's initial conditions, we showed that at least 3 out of 4 significant covariates presented consistent results and effects, even after including the required controls. These results incline us to believe that a child's initial traits are indeed determinants of dropouts in Peru and that policy considering this fact should be evaluated. Our secondary hypothesis, evaluating the relevance of socioeconomic status, measured by the wealth index, was found to be significant at the 1% level and with a hazard ratio very close to zero.

This would imply that, for a 1% increase in the wealth index, there is a increase in the instantaneous survival probability of a factor smaller than 1%. Our third hypothesis explored the relevance of the location of the household. In this sense, we found a consistent increased risk for those living in urban sites. Nevertheless, we can not confirm the magnitude of our location coefficients, given the high concentration in urban sites. Also, regarding location, we found a decreasing risk for those living in the highlands, but only in the semi-parametric results. Finally, our fourth hypothesis asked if the gender of the caregiver mattered, in terms of dropout probabilities. Our evidence suggests that it does. Specifically, having a male caregiver increases the probability of dropout in a statistically significant way, but the magnitude of the effect is questionable, again because of sample issues. Although we found consistency across the non-parametric and semi-parametric estimates, some of the coefficients require the fourth round of data in order to get a more precise magnitude of the coefficients. Finally, a few covariates didn't meet the proportionality assumption required by the Cox PH model, hence using a fully parametric model, such as an Accelerated Failure Time (AFT) model with a Weibull distribution, would be a recommendable direction for further research.

# Appendix A

# Appendix

- **Stata Code**

  For the stata do file, please refer to the electronic submission files. The name of the file is: **stperu3.do**

- **Data file**

  For the data file containing the data used for this document, please refer to the electronic submission files. The name of the file is: **peru constructed.dta**

- **Main regression code**

  stcox typesitedurban i.region hhsize i.caredu caresexdmale sexdmale bmi ppvtraw stunt cladder agemon i.chethnics i.chldrel chweight chheight, vce(robust)

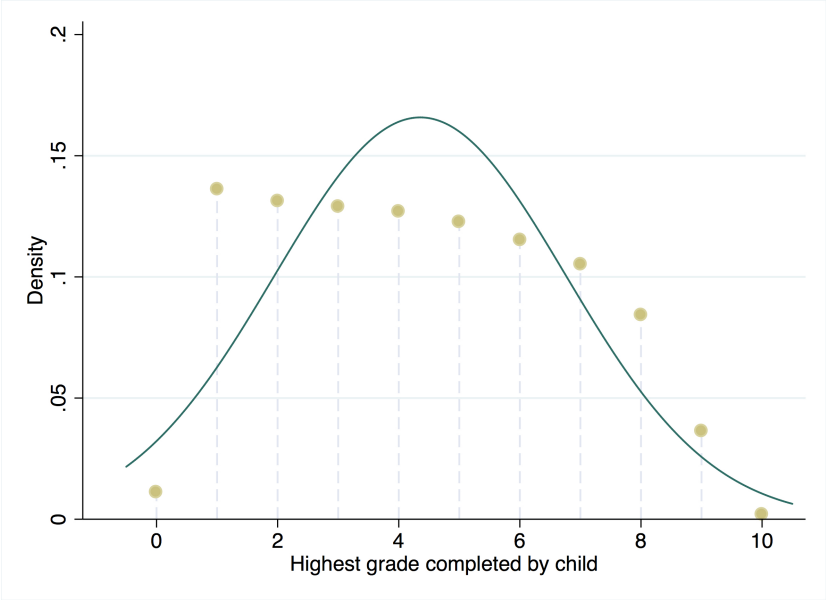


Figure A.1: Distribution of the sample through the years of schooling

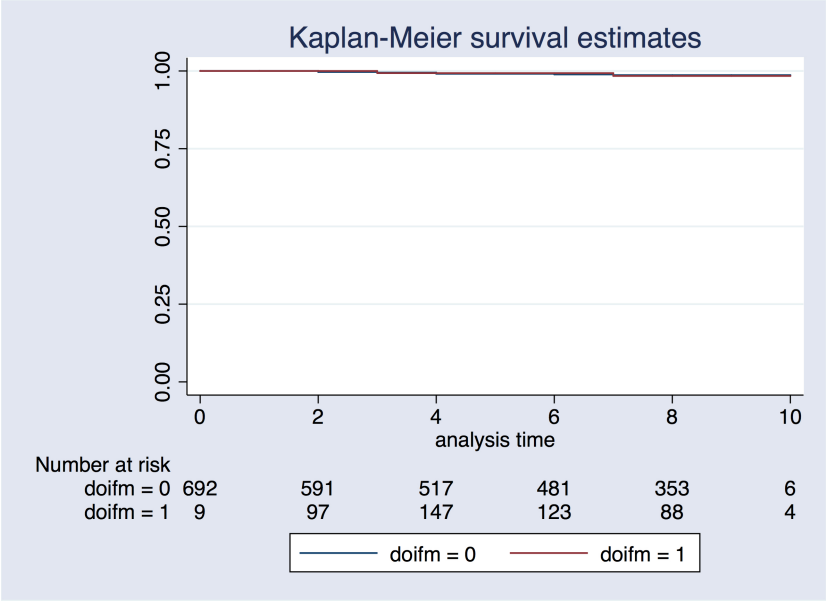


Figure A.2: Kaplan-Meier survival estimates by death or illness of family member in all rounds.

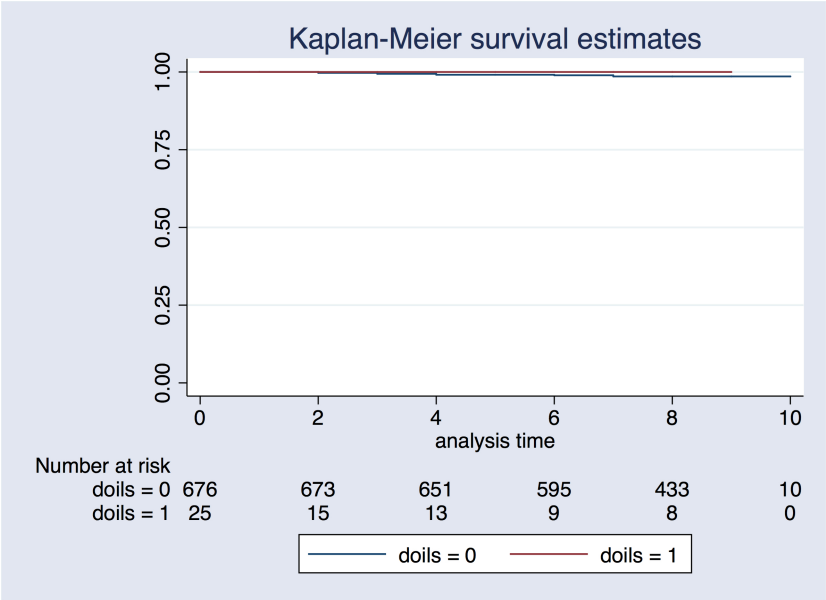


Figure A.3: Kaplan-Meier survival estimates by death or illness of livestock in all rounds.

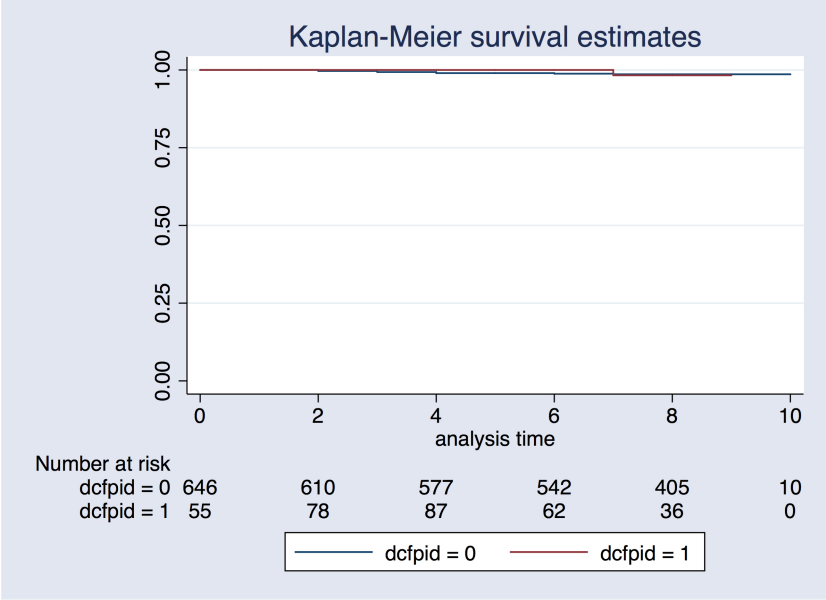


Figure A.4: Kaplan-Meier survival estimates by Drought, crop failure, pest infestation and diseases in all rounds.

# Bibliography

Alarcón, Walter (1995), *Atraso y deserción escolar en niños y adolescentes*. INEI. Dirección Técnica de Demografía y Estudios Sociales.

Alexander, Karl L, Doris R Entwisle, and Carrie S Horsey (1997), "From first grade forward: Early foundations of high school dropout." *Sociology of education*, 87–107.

Angrist, Joshua D and Alan B Krueger (1990), "Does compulsory school attendance affect schooling and earnings?" Technical report, National Bureau of Economic Research.

Boyden, J. (April 2014), "Young lives: an international study of childhood poverty: Rounds 1-3 constructed files, 2002-2009 [computer file]." *Colchester, Essex: UK Data Archive [distributor]*, SN: 7483 , http://dx.doi.org/10.5255/UKDA-SN-7483-1.

Brown, Philip H and Albert Park (2002), "Education and poverty in rural china." *Economics of Education Review*, 21, 523–541.

De Gregorio, José and Jong-Wha Lee (2002), "Education and income in-

equality: New evidence from cross-country data." *Review of income and wealth*, 48, 395–416.

Eckstein, Zvi and Kenneth I Wolpin (1990), "Estimating a market equilibrium search model from panel data on individuals." *Econometrica: Journal of the Econometric Society*, 783–808.

Eckstein, Zvi and Kenneth I Wolpin (1999), "Why youths drop out of high school: The impact of preferences, opportunities, and abilities." *Econometrica*, 1295–1339.

Fine, Michelle (1986), "Why urban adolescents drop into and out of public high school." *The Teachers College Record*, 87, 393–409.

Franklin, Bobby J and Susan Kochan (2000), "Collecting and reporting dropout data in louisiana."

Gamoran, Adam and Martin Nystrand (1992), "Taking students seriously." *Student engagement and achievement in American secondary schools*, 40–61.

Haveman, Robert, Barbara Wolfe, and James Spaulding (1991), "Childhood events and circumstances influencing high school completion." *Demography*, 28, 133–157.

Heckman, James and Burton Singer (1984), "A method for minimizing the impact of distributional assumptions in econometric models for duration data." *Econometrica: Journal of the Econometric Society*, 271–320.

Heckman, James J (1976), "A life-cycle model of earnings, learning, and consumption." *Journal of political economy*, S9–S44.

Iyigun, Murat F (1999), "Public education and intergenerational economic mobility." *International Economic Review*, 40, 697–710.

Keane, Michael P and Kenneth I Wolpin (1997), "The career decisions of young men." *Journal of political Economy*, 105, 473–522.

Kleinbaum, DG and M Klein (2005), "Survival analysis: a self-learning text." *Statistics*.

Lavado, Pablo, José Gallegos, et al. (2005), "The dynamics of the schooling dropout in peru: a framework using duration models." Technical report.

McMillan, Phillip Kaufman, Marilyn M. and Summer D. Whitener. (1994), "Dropout rates in the united states." *U.S. Department of Education, Office of Educational Research and Improvement*.

Montmarquette, Claude, Nathalie Viennot-Briot, and Marcel Dagenais (2007), "Dropout, school performance, and working while in school." *The Review of Economics and Statistics*, 89, 752–760.

Posner, Jill K and Deborah Lowe Vandell (1994), "Low-income children's after-school care: Are there beneficial effects of after-school programs?" *Child development*, 65, 440–456.

Ravallion, Martin and Quentin Wodon (2000), "Does child labour displace schooling? evidence on behavioural responses to an enrollment subsidy." *The Economic Journal*, 110, 158–175.

Rees, Daniel I and H Naci Mocan (1997), "Labor market conditions and the high school dropout rate: Evidence from new york state." *Economics of Education Review*, 16, 103–109.

Rumberger, Russell W and Scott L Thomas (2000), "The distribution of dropout and turnover rates among urban and suburban high schools." *Sociology of Education*, 39–67.

Sanchéz, Santiago Penny Mary Miranda Alejandra, Alan Cueto and Guido Melendez (2015), "Young lives survey design and sampling in peru." *Young Lives*.

Schneider, Barbara, David Stevenson, and Jeffrey Link (1994), "Social and cultural capital: Differences between students who leave school at different periods in their school careers." In *annual meeting. American Educational Research Association, New Orleans*.

Seginer, Rachel (1983), "Parents' educational expectations and children's academic achievements: A literature review." *Merrill-Palmer Quarterly (1982-)*, 1–23.

Woldehanna, Tassew and Adiam Hagos (2012), "Shocks and primary school drop-out rates."