Heterogeneous effects of dropout on labor market outcomes : the French higher education case

Gaspard Tissandier - Université Paris 1 Panthéon Sorbonne

20/01/2022

Abstract:

In the sense of Spence (1973), education sends a signal to the labor market which helps to reveal the ability of former students. Dropout is usually perceived as a negative signal, leading to lower wages and employment rates. This paper tests for the presence of heterogeneous effects of dropout in french higher education (post-high school diploma) on labor market outcomes from 2010 to 2013. These effects are measured on the time in employment and average wages at an individual's entry into the labor market. I analyze if the heterogeneous structure of the effects is conditional on individual characteristics such as diploma, social origin, or gender. I use the Generalized Random Forest algorithm with the distance to the closest higher education institution at 6th grade as an instrument to estimate individual Conditional Average Treatment Effects. In line with the literature, I find a negative effect of dropping out on the time in employment. However, this effect is heterogeneous across individuals, ranging from -59% to -41% pp. I find two subgroups with a significant and negative effect of dropping out on the average wages, with effects ranging from -1300€ per month to a -600€ on average monthly wage. The gender, study field, study duration, and the social origin of parents, especially the mother's one, have an active role in shaping the heterogeneity of the dropout effect.

JEL code : J01, J24, I2, I24

Keywords : dropout, higher education, grf, instrumental variable, labor market outcomes

1 Introduction

For a large part of the workforce, higher education is a crucial phase for accumulating competencies, knowledge, and skills that will be later valued in the labor market. While it is decisive for students to acquire diplomas to testify to their abilities, dropping out is a recurrent event in the French higher education system, whether voluntary or involuntary (Aina et al. 2018). Dropping out has a strong effect on labor market outcomes and can penalize new entrants in the long run (Schnepf 2014). Dropouts are numerous in the French higher education system: in 2018, 23.9% of students enrolled in their first year of higher education dropped out. This phenomenon is persistent trough time as 4.1% of students who began their study in 2014 dropped out at the end of the second year, and 10.4% at the end of the third year[1].

Following Spence (1973), dropping out send a negative signal to the labor market about the ability of the individual. Ability is defined as an underlying variable reflecting the capacity of a worker to perform well in a task, or a set of tasks. Indeed, the negative signal of dropping out on the ability competes with the positive signal brought by the diploma, affecting the wage, the probability of being employed, or the path of the worker in the labor market. Dropout has a negative effect on labor market outcomes, either on salary, opportunity or rate of employment (Flores-Lagunes and Light 2007, Matkovic and Kogan 2012 ,Reisel 2013, Schnepf 2014) . It has been observed that the negative effect of dropout can be heterogeneous conditional on the motivation of the dropout (Bjerk 2012) or even on the diploma or the age of the individual (Brodaty, Gary-Bobo, and Prieto 2008, Navarro, Fruehwirth, and Takahashi 2016, Scholten and Tieben 2017).

It has been shown that the social origin and individual characteristics of students are highly determinant in the dropout process (Aina et al. 2018, Vignoles and Powdthavee 2009), and that specific variables such as the academic path or gender have an impact on the structure of the effect of dropout. However, no study tried to analyze the heterogeneity of the dropout effect on labor market outcomes concerning a large set of individual characteristics. Acknowledging that the consequences of dropout are not the same depending on individuals is crucial to understand educational choices from students, and to design better policies against student dropout.

This paper asks whether the effect of dropping out on labor market outcomes, such as wages and probabilities of being employed, follows a heterogeneous distribution. Then, the distribution of this effect is studied to test if it is conditional on individual characteristics and more specifically if exogenous variables such as social origin or gender play a more important role than endogenous variables such as diplomas.

According to both fundamental models of education (Becker 1993 or Spence 1973), acquiring more year of education bring higher earnings, as detailed in the analysis of Fang (2006) or the review of Psacharopoulos and Patrinos (2018). The role of higher education as a signaling

[1]Repères et références statistiques 2019 - Direction de l'évaluation de la prospective et de la performance

method is widely explored in the literature. The study of Arcidiacono, Bayer, and Hizmo (2010) shows that college diplomas act as an ability signaling method, as college graduates get wages matching their measured ability quicker in their carrier than high school graduates. They use an external individual measure of ability (an ability test matched with the students' database) to prove that the revealed ability is not only by diplomas levels but that education conveys a very precise signal of ability allowing the companies to find better fitting individuals more quickly.

However, dropout raises less consensus about its effects on labor market performance. As highlighted by Schnepf (2014), the literature about labor market performance of dropout is scarce. Bjerk (2012) studies the effect of dropping out on criminal activity and labor market outcomes. The author finds that dropping out has a strong negative effect on both indicators. However, one of the main findings lies in the heterogeneity of the dropout effect: students who drop out for "passive reasons" have lowest performances than those who drop out with plans, or on purpose. In Schnepf (2014), the author finds that in most of the European countries, dropouts are benefiting from their study time, compared to students who didn't enter higher education. In this paper, Schnepf uses a propensity score matching model on data from the 2011 Programme for the Internationale Assessment of Adult Competencies to pursue the study on many different European countries. This conclusion is similar to Reisel (2013), where the author shows that in the United States, it is beneficial to integrate higher education even without graduating, compared to individuals without any higher education experience. Similarly, Matkovic and Kogan (2012) compares the effect of dropping out on labor market performances in Croatia and Serbia and corroborates the finding of the overall negative effect. They also find that the longer a student stays in higher education, the smoother the transition in the labor market is, especially in Serbia. This result is similar to the one of Flores-Lagunes and Light (2007) in the United States, where the sheepskin effect (the premium of having graduated from a diploma) is highly conditional on the number of years of schooling. In France, the study from Brodaty, Gary-Bobo, and Prieto 2008 covers the effect of delayed graduation (of which dropout is a special case) on labor market performance. The authors find a negative effect of delayed graduation, with significant differences between the effects conditional on the highest diploma. In Norway and the United States from 1989 to 1999, Reisel (2013) finds heterogeneity in the return to education due to the distribution of women and minorities across the income distribution, while Scholten and Tieben 2017 finds that in Germany, for individuals born between 1944 and 1986, the dropout effect is mostly conditional on the previous diploma, which acts as a "safety net".

The main issue in estimating the effect of dropout or delayed graduation is the endogeneity of the event with the underlying ability of the student. As seen before, the propensity score matching is used to solve this issue, as in Schnepf (2014). On the other hand, recent papers like Mahjoub 2017 use the period of birth as an instrument, inspired by Angrist and Krueger (1991). An alternative instrument is a distance to the closest higher education institution, as proposed by Card (1993). In Brodaty, Gary-Bobo, and Prieto (2008), the authors use a dense system of geographical IV with the distance to the closest university in 6th grade, and the

number of openings of higher education institutions in the geographical area during secondary education.

To allow the estimation of heterogeneous treatment effect, I apply the Generalized Random Forest (GRF) methods, developed by Athey, Tibshirani, and Wager (2019), on a French database of 18000 young workers who finished their education in 2010. Their work records are surveyed from 2010 to 2013, which helps us to construct two variables indicating the average of wages and time in employment for every individual. The GRF algorithm, based on the Random Forest structure (Breiman 2001), allows us to estimate individual Conditional Average Treatment Effect (CATE). Individuals are then gathered in subgroups following their CATE magnitude to compute the Average (Conditional) Local Average Treatment Effect (A(C)LATE) on these subgroups (Imbens and Angrist 1994, Athey and Wager 2020). The asymptotic property of the A(C)LATE allows to do inference on the quality of the estimators and to test for the presence of heterogeneity in the CATE distribution. To assess if the shape of the CATE distribution is conditional on individual characteristics, I estimate the parameters of a logistic model for being less penalized by dropout. By estimating the likelihood of belonging in the top 50% as the dependent variable, we can study which individual characteristics are the most important in shaping the effect of dropout.

The endogeneity of dropout is tackled with an instrumental variable setting adapted to the Random Forest structure of the GRF. I use the square of the distance to the closest Higher Education institution at 6th grade as an instrument. Paired with a vector of controls to estimate the predicted probabilities of dropout, I obtain an efficient instrumental variable setting allowing me to identify heterogeneous causal effects of dropout on labor market outcomes. The distance is measured at 6th grade to avoid the endogeneity due to the use of the distance between higher education institutions and the high school diploma city (Brodaty, Gary-Bobo, and Prieto 2008).

I succeed to find two groups with heterogeneous treatment effects for both indicators. After estimating the individual Conditional Average Treatment Effect (CATE - the individual effect), the sample is split around the CATE median, and the Average (Conditional) Local Average Treatment Effect (A(C)LATE) is estimated on both subsamples. This methodology, proposed by Athey and Wager 2020, allows estimating a doubly robust treatment effect on the treated compatible with the Instrumental Forest method. According to this estimation, dropping out has a negative effect of -59% of the observed period for the most penalized group, and of -41% for the less penalized group. On average wages, the effect goes from -1350€ per monthly wage for the most penalized group to -612€ for the less penalized group. The difference between the A(C)LATE of these subgroups is tested with a Student test, and these two tests confirm the heterogeneous distribution of dropout effect on both labor market outcomes. To understand which individual characteristics are the most important in shaping the individual's CATE distributions, I estimate a logit model for the likelihood of having a dropout effect above the median of CATE.

For the CATE distribution of dropping out on the time in employment, the most important

variables are the mother's diploma and occupation. Having a mother who acquired more years of education or work in a higher position (and blue-collar position) will increase a student's likelihood of being in the less penalized part of the CATE distribution. The student's diploma is less significant than the mother's diploma and occupation. Compared to fathers having a High School diploma or an employee and blue-collar occupation, all others education and occupation are decreasing the likelihood of being above the median. For this labor market indicator, the social origin is the most important element for reducing the penalty of dropping out, especially that of the mother.

When studying the dropout CATE on average wages, the parents' occupation and education play a far less active role in shaping the CATE distribution. The mother's diploma and father's occupation are the most preponderant variables in shaping the heterogeneity of the dropout effect. The effect of the mother's diploma is still similar while more heterogeneous depending on the diploma. The highest diploma tried is still less important than social origin to explain the heterogeneity of the dropout effect.

Generally, social origin and especially the mother's one are more important than the tried diploma or academic achievements to determine the magnitude of the dropout effect. However, the prevalence for exogenous variables is less pronounced for the average wages than for the time in employment.

This paper sheds new light on the integration of french tertiary education dropouts in the labor market. The main contribution is the application of machine learning techniques that helps to account for individuals' characteristics and unfold the heterogeneous structure of the dropout effect on labor market outcomes. I showed that social origin acts as a safety net for dropout on the time in employment, while the acquired diplomas are always less important than social origin. These results will help to understand better the path of higher education dropouts and to design a policy that prioritized students who could benefit from it the most.

2 Data

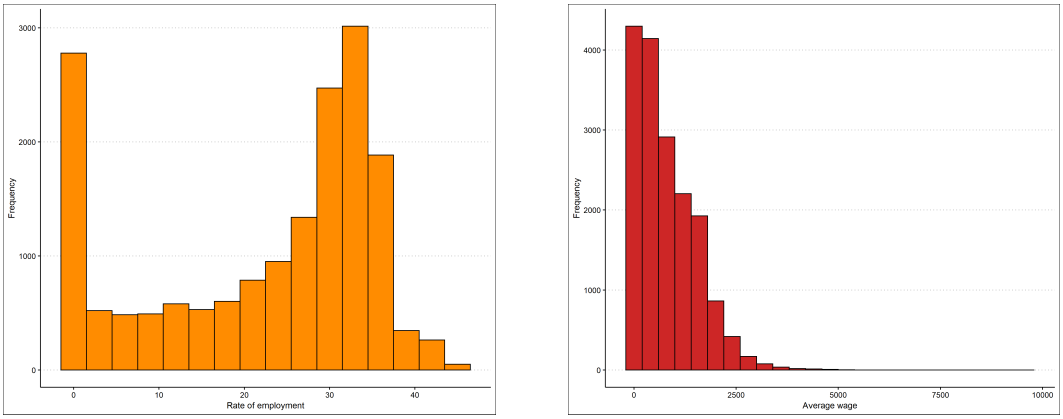
To identify the effect of dropping out on former students' labor market outcomes, I use "Génération 2010", a longitudinal survey provided by the CEREQ (Centre d'Etudes et de Recherches sur les Qualifications) [2]. This survey is conducted on individuals who have finished their education in 2010 (between October 2009 and October 2010), without any interruption before. Individuals are surveyed in 2013, three years after they left the educational system. The resulting database consists of a panel gathering information about former students' background, education, and a detailed schedule of employment from 2010 to 2013. The survey covers 33547 individuals with a wide range of education, social background variables, and professional records. I restrain this data set to individuals who at least, tried to obtain a higher education diploma. This array goes from high school diploma holders who tried one year of higher education to Ph.D. graduates. This represents a data set of 17094 individuals.

[2]Génération 2010 – Interrogation à 3 ans – 2013 (2013, CEREQ)

I create indicators variables for dropout, the number of months worked, and the average wages. Dropping out is defined here as not having validated a diploma in 2010, or exiting the educational system before the last year of said diploma. For example, if a student didn't graduate of her Master 2 because she didn't pass the exams, she will be considered as a dropout. A student who interrupted her study in the second year of undergraduate, out of the three required years will also be considered as a dropout. According to this definition, the database consists of 4309 individuals who dropped out, and 12785 who didn't (25% of dropout)'.

Each individual's employment curriculum is entered in a side database where employment and unemployment periods are filled in. For each working sequence, the beginning and ending salaries are specified precised, as the duration in months. This setting allows us to create two variables in order to test our hypothesis.

The first variable, *Time in employment*, consists in the number of months worked full time over 36 months (3 years), expressed as months (to ease the understanding, I will often express this indicator in percentage). The second variable, *Average Wages*, is an average of the wages on the whole period (36 months). Then, if an individual works 12 months with a salary of 1200€, *Average Wages* will equal 400€.



(a) Time in Employment (b) Average wages

Figure 1: Distribution of dependent variables

The following variables are kept : highest diploma tried on 32 levels and 12 levels, if foreign study travel or internship have been done by the student, geographical location in 6th grade, high school, and in 2010, when the individual left the education system. I also keep the gender of the individual, the professional occupation, professional sector (public or private) and diplomas of both parents, the number of siblings, and information about past education such as an indicator of presenting delay in 6th grade, at the high school diploma and the discretized grade of the high school diploma. Finally, I keep information on the type of high school diploma acquired (general or professional, and also scientific, economics, literature). The descriptive

statistics are presented in table 1. For commodity reasons, the diploma is presented on 12 levels and the social origin is presented only for the mother.
The time in employment and average wages conditional on dropping out and individual characteristics are presented in Appendix.

	No dropout	Dropout	No dropout (in %)	Dropout (in %)	Total	Percentage
Gender						
Male	5472	2239	70.96 %	29.04 %	7711	45.11%
Female	7313	2070	77.94%	22.06%	9383	54.89%
Highest diploma tried						
HSD +2 Health	3006	48	98.43%	1.57%	3054	17.87%
HSD +2 Industrial	827	416	66.53%	33.47%	1243	7.27%
HSD +2 Tertiary	1290	2339	35.55%	64.45%	3629	21.23%
Vocational Degree	952	60	94.07%	5.93%	1012	5.92%
HSD +3/4 Soc sci., Econ Law	955	690	58.05%	41.95%	1645	9.62%
HSD +3/4 STEM	396	287	57.98%	42.02%	683	4.00%
HSD +5 Soc sci., Econ Law	1408	112	92.63%	7.37%	1520	8.89%
HSD +5 STEM	865	43	95.26%	4.74%	908	5.31%
Business School	343	12	96.62%	3.38%	355	2.08%
Ingeniering School	935	29	96.99%	3.01%	964	5.64%
PhD	1808	273	86.88%	13.12%	2081	12.17%
Mother's highest diploma						
N.A	1328	532	71.40%	28.60%	1860	10.88%
No diploma	2428	1112	68.59%	31.41%	3540	20.71%
Below HSD	2013	685	74.61%	25.39%	2698	15.78%
HSD	2435	852	74.08%	25.92%	3287	19.23%
HSD +2 years	1618	407	79.90%	20.10%	2025	11.85%
HSD +3/4 years	1775	433	80.39%	19.61%	2208	12.92%
HSD 5+ years	1188	288	80.49%	19.51%	1476	8.63%
Mother's occupation						
N.A	1314	596	68.80%	31.20%	1910	11.17%
Blue collar	1017	443	69.66%	30.34%	1460	8.54%
Employee	5735	1992	74.22%	25.78%	7727	45.20%
Intermediary	1031	299	77.52%	22.48%	1330	7.78%
White collar	3016	778	79.49%	20.51%	3794	22.19%
Independent	463	155	74.92%	25.08%	618	3.62%
Farmer	209	46	81.96%	18.04%	255	1.49%

Table 1: Summary statistics by dropout status

HSD corresponds to the High School Diploma (or Baccalaureate). The time needed to acquire the diploma is counted as "+y" : HSD +2 corresponds to two years or study after the HSD. "Licence Pro" corresponds to vocational degrees, needing 3 years to complete. The short degrees (two and three years) lead to a precise field, and are considered as "professional degrees". The rest of the notations are self-explanatory.

Concerning parents' occupation, the levels are defined using the type of occupation. It is actually entirely possible to have a Blue collar earning more than an Employee. The independent level groups craftsman and company owners. The farmer level is broad and includes agronomic workers.

The unequal distribution of males and females is common and well documented in France, as female are more represented in higher education. The distribution of the professional situation

of the mother shows that they are mostly employed in employee positions, while it is more balanced for the fathers.

In table 1, we can also observe the distribution of dropout by highest diploma tried. The percentage presented for the dropout columns is the distribution of dropout for the considered diploma, while it is the percentage of the considered diploma among the whole population in the total column. For short degrees holders, Health majors are very less likely to drop out than tertiary or industrial graduates. For three or four years degrees, the percentage of dropouts is the same with around 40% of the students who are dropping out. For all others degrees, the dropout rate is noticeably lower, with a maximum of 13% for Ph.D.

Regarding social origin, the dropout rates conditional on the mother's education or professional occupation do not exhibit very different values.

Finally, I base my instrumental variable setting on the distance to the closest higher education institution from the student's 6th-grade city. This distance is computed by using GPS coordinates and the distance between both points on the geodesic[3]. The geographical unit is the *zone d'emploi*, dividing France into around 310 areas. If there is a university or a school in the *zone d'emploi* of the 6th-grade city, the distance is then 0. The density function of this variable is presented in figure 2. I use the square of the distance as an instrument.

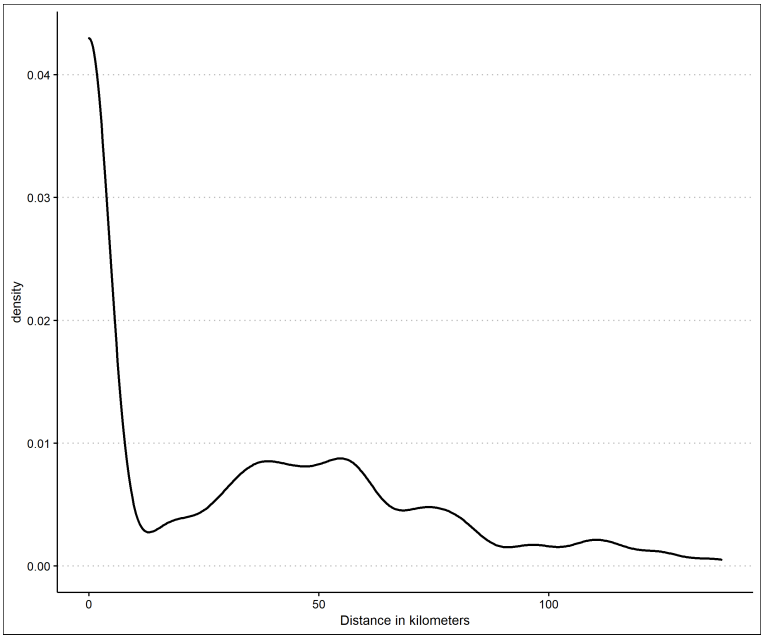


Figure 2: Distribution of the distance from 6th grade home to the closest university

While widely used as an instrument for educational attainment, the distance to the closest university is also a valid instrument for dropout or delay in graduation .With a hypothesis of rational individuals seeing education as a sequential process, the distance to the closest

[3]For computation methodology, see : C.F.F. Karney, 2013. Algorithms for geodesics, J. Geodesy 87: 43-55. doi: 10.1007/s00190-012-0578-z. Addenda: https://geographiclib.sourceforge.io/geod-addenda.html. Also, see https://geographiclib.sourceforge.io/

university will have a negative effect on educational attainment : for two similar students except that one is living farther than the other from the local university, the first one will have a higher cost of education, thus have a higher probability of not engage in the next year.

The distance to the closest university also affects negatively the dropout probability in two main ways. First, the instrument affects negatively the years spent to acquire education. Thus, if the distance reduces the time in education, it also reduces the occasion of dropping out. Second, dropping out of a degree leads to sunk costs. If a student quits a two-year degree, even if this experience can be valued in a certain manner, some benefits of having acquired the diploma will be lost. For two students who attained the same degree, the one which had to move from her parents' place will have a higher sunk cost, and then will be less likely to drop out.

3 Methodology

The objective of this analysis is to identify subgroups with different treatment effects of dropout, conditional on a vector of covariates X. If we want to test for every interaction that this vector allows, the number of interaction terms could be gigantic and will obviously detect spurious correlation. To avoid this pitfall, I rely on the Generalized Random Forest developed by Athey, Tibshirani, and Wager (2019), and use the data structure to identify heterogeneous treatment effects. This method allows us to compute Conditional Average Treatment Effect (CATE), the individual treatment effect, and the corresponding standard error, and then to average these effects on selected partitions of the population as Average (Conditional) Local Average Treatment Effects (A(C)LATE). This method relies on regression trees to estimate the CATE, and average the estimated CATE across all trees. This methodology is called the Random Forest (Breiman 2001) and allows to account for large possibilities of interactions between covariates, without risking over-fitting. To avoid spurious correlations due to using similar data to construct the trees and estimating the treatment effect, the authors rely on the "honest methodology". Finally, the endogeneity of dropout forces the use of an adapted instrumental setting. In this section, we will develop the Generalized Random Forest algorithm, the instrumental variable setting, and then the Average (Conditional) Local Average Treatment Effect estimation.

The objective of our paper is to evaluate the causal effect of dropout. As defined by Rubin (1974), we want to compute the individual difference in potential outcome $\tau_i = Y_i(1) - Y_i(0)$ with $Y_i(W_i)$ the outcome depending on the treatment status W_i. Since we do not observe both $Y_i(1)$ and $Y_i(0)$, alternative estimators are needed. Thus, we focus on the estimation of the Conditional Average Treatment Effects (CATE) defined as $\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0) \mid X_i = x]$. This estimator is used as a subsample average treatment effect on the individuals sharing $X_i = x$. Thus, for a combination of the vector $X_i = x$, we will be able to compute the treatment effect on this combination x, corresponding to individuals showing similar characteristics with i. In order to estimate the individual CATE, I rely on the Generalized Random Forest Algorithm

(GRF).

In this section, we will avoid to use too much technical explanations and try to focus on the general idea of the methodology. See Athey, Tibshirani, and Wager 2019 for all the technical details of the Generalized Random Forest.

3.1 The Generalized Random Forest algorithm

(All the notation are taken from either Hastie, Tibshirani, and Friedman (2009), Athey and Imbens (2016), Wager and Athey (2018) or Athey, Tibshirani, and Wager (2019)).

The GRF is based on the regression tree algorithm developed by Breiman et al. (1983) (called the CART for Classification and Regression Trees) and adapted as causal honest tree by Athey and Imbens (2016). I will proceed by first describe the honest causal tree, and then the adaptation to the Generalized Random Forest.

The objective of an honest causal tree is to create subgroups in the population on which the Conditional Average Treatment Effects (CATE) are evaluated. For a given dataset, we observe (Y_i, W_i, X_i), for $i = 1, ..., N$, with Y_i the outcome, X_i a vector of covariates and W_i the treatment status. In our example, if the considered individual has dropped out, she shows $W_i = 1$ and $W_i = 0$ if she didn't.

In the initial paper by Breiman et al. (1983), the regression tree use a training sample $\mathcal{S}^{tr}$ for which we know (Y_i, X_i) and a target sample for which we know only X_i. By fitting a tree model on $\mathcal{S}^{tr}$, the objective is to predict correctly the outcomes for the target sample. To do so, the algorithm first search for a splitting point s on a splitting variable X_j in order to create two subsample $R_1(j, s) = [X \mid X_j \leq s]$ and $R_1(j, s) = [X \mid X_j > s]$. In this setting, s is found by minimizing the mean squared error defined as :

$$MSE = \left[\sum_{x_i \in R_1} (y_i - \bar{y_1}(j, s))^2 + \sum_{x_i \in R_2} (y_i - \bar{y_2}(j, s))^2 \right] \tag{1}$$

Finally, the algorithm repeat this method until a stopping point (minimum number of individuals in subsamples, maximum number of subsamples, for example). To compute prediction for another sample, the algorithm fit new observations into its corresponding subsample, and then assign the mean of this subsample as the predicted outcome $\hat{Y}_i$. Compared to linear regression or similar methods, the CART allow us to account for high dimensional interactions between all covariates in X_i and help to build strong predictive models.

If the CART is efficient to produce prediction on a target sample, it is not yet suitable to estimate CATE. For this aim, we need two modifications of the original algorithm : introduce an "honest" design and use an modified splitting rule.

The honest design, firstly applied to regression tree by Athey and Imbens (2016), help to solve the over fitting problem. Over fitting arise when a model match too closely the data and then present no generalization power. Indeed, if we use the same sample to build the regression tree and to estimate the CATE in every created subsamples, wee will obtain completely biased estimators. In the honest design, we use two different and randomly drawn subsamples to build the tree with the first one, and then to estimate effect in the subsamples build by the regression tree in the second one.

In order to account for the second stage estimations, we need to adapt the objective function. We will focus on the Expected Mean Square Error, an adapted estimator of the Mean Squared Error.

We introduce here the estimated Conditional Average Treatment Effect, the estimated expression of the CATE presented below. With $\hat{\mu}^2$ the conditional mean of a subsample, it is defined as :

$$\hat{\tau}(x; \mathcal{S}) = \hat{\mu}(w_i = 1, x, \mathcal{S}) - \hat{\mu}(w_i = 0, x, \mathcal{S})$$

This expression estimate the CATE on individuals with $X_i = x$ as the difference between the both treated and non treated conditional mean on the given subsample. With an adapted estimate of the CATE, it is possible to design an objective function which suit our need. With N^{tr} the size of the training sample (made equal to the size of the estimation sample), l a subsample, $S^2_{\mathcal{S}^{tr}}(l)$ the subsample estimated variance of $\hat{\tau}$ and p the probability of being treated , the adapted expected Mean Squared Error is defined as :

$$\widehat{EMSE}_\tau(\mathcal{S}^{tr}) = \frac{1}{N^{tr}} \sum_{i \in \mathcal{S}^{tr}} \hat{\tau}^2(X_i; \mathcal{S}^{tr}) - \frac{2}{N^{tr}} \sum_l \left(\frac{S^2_{\mathcal{S}^{tr}_{treated}}(l)}{p} + \frac{S^2_{\mathcal{S}^{tr}_{control}}(l)}{1-p} \right) \tag{2}$$

This estimator of the Expected Mean Squared Error is almost composed as the MSE, but add a negative effect of within subsample variance of the CATE. This allow the algorithm to take into account that finer partition generate greater variances. Then, with this objective function, the algorithm will search for split that maximize treatment heterogeneity in treatment effect while avoid generating too much in-partition variance. For more details on the construction of this objective function, please refer to Appendix.

Since we have a efficient splitting criterion, one problem remain : due to the initial honest design, the built tree will greatly depend of the initial random splitting. To solve this issue, we apply the Random Forest algorithm first developed by Breiman (2001) and applied to causal inference by Wager and Athey (2018). The objective of the causal Random Forest is to create causal honest trees on subsamples of the whole population. For example, we draw a partition α of the initial population, and build the honest causal tree on this partition as described below. Then, the algorithm average all the individual CATE given by all trees to compute the individual CATE. This method provide estimates of individuals treatment effects with the associated standard error. One of the main assumption of this model is the unconfoundedness

i.e $W_i \perp (Y_i(0)), Y_i(1), X_i)$. This assumption is satisfied in a random treatment assignment setting such as Random Control Trials. Since it is almost impossible to randomize the dropout, I have to include a instrumental variable setting in the framework.

The Generalized Random Forest developed by Athey, Tibshirani, and Wager (2019) propose a general framework to estimate CATE with methods such as causal Random Forest and Instrumental Forest. The main divergence from the initial causal Random Forest come from the usage of a gradient-based loss criterion rather than the exact loss criterion (2). The gradient-base criterion is an approximation of (2) build with gradient-based approximations of $\hat{\tau}$ for each subsamples. This method, designed as a general framework for estimation in non-linear setting, help to use IV and is less costly in computation.

In this paper, I use the GRF to build individual CATE by using the following variables : the highest diploma tried on 12 levels, if the student made internship or international travel, the higher education institution region, the high school region, the type of high school diploma (general, technical or professional) on three variables, a categorical variable for the grade at the high school diploma, the delay at the high school diploma, the professional occupation and diploma of both parents, the number of siblings and the gender.

3.2 The instrumental variable setting

In this setting, I want to identify the effect of dropout on wages and time in employment. Unfortunately , these two variables have (at least) a common generating variable usually defined as the ability. If I believe our treatment and outcome variable to be link by the model $Y_i = \mu(X_i) + \tau(X_i)W_i + \epsilon_i$, a clear endogeneity problem arise as the treatment is correlated to the error term via the individual ability. To solve this issue, I need to find an instrument Z_i, correlated with the treatment W_i (having dropped out) but not with the error term i.e the ability.

In the case of the GRF algorithm, estimating a Instrumental Forest is equivalent to apply the Wald formula for individuals with $X_i = x$. The interactions terms generated by the GRF, change for every tree and then help us to account for high dimension heterogeneity. Since I have at our disposal an instrument Z_i satisfying all the IV assumptions, I can estimate the treatment effect as :

$$\tau(x) = \frac{Cov[Y_i, Z_i \mid X_i = x]}{Cov[W_i, Z_i \mid X_i = x]} \tag{3}$$

In this setting, I can implement an IV setting with a binary or continuous instrumental variable. However, since the A(C)LATE is used to estimate treatment effect on group of former students, we need a dichotomous instrumental variable. We need to compute the CATE as a doubly robust score as proposed by Athey and Wager (2020), and average them over subsamples to get unbiased results for every A(C)LATE. As precised in Athey and Wager (2020), we need a

binary instrument to compute the doubly robust score for the LATE.

To propose a binary instrumental variable with an high correlation, I adapt the Procedure 18.1 from Wooldridge (2010) with a logit model. This procedure, established for endogenous treatment, use a logit model to predict the probability of treatment, including all exogenous control and instrumental variable in the model.

In first step I estimate the following logit model :

$$P(W = 1 \mid X, Z) \tag{4}$$

With w the treatment status, x a vector of control variables and z a vector of instrumental variables described in section 3. The fitted probabilities $\hat{W}_i$ can be used as instrument in the GRF setting. However, since the LATE estimator need a dichotomous instrumental variable, I transform $\hat{W}_i$ as follows :

$$\begin{cases} z_{GRF} = 1 & \text{if} \quad \hat{w}_i > p(\alpha) \\ z_{GRF} = 0 & \text{if} \quad \hat{w}_i \leq p(\alpha) \end{cases}$$

With $p(\alpha)$ the value corresponding to the α^{th} percentile and $\hat{w}_i$ the estimated probability of dropping out. My choice of α is motivated by the Local Average Treatment Effect estimation step. The LATE is the average treatment effect on the compliers i.e individual who respond positively to the instrument. Since the LATE is computed by averaging the treatment effect times a weighting function which is divided by product of compliance scores, we need to keep the compliance scores as high as possible. The compliance score is defined as the individual propensity to dropout conditional on (x, z). The threshold which maximize the product of the scores is $(\alpha) = 0.80$.

$$\begin{gathered} X = (\text{geographical location in 6th grade}, \text{professional situation of parents}, \\ \text{diplomas of parents}, \text{gender}, \text{delay in 6th grade} \\ Z = (\text{distance to the closest university in 6th grade})^2 \end{gathered}$$

With the described setting, I can estimate CATE for every individuals conditional on their characteristics. Since I build an instrumental setting, the obtained $\hat{\tau}(x)$ are the causal effect of dropout, conditional on individuals' covariates.

3.3 Doubly robust estimation and Average Conditional Local Average Treatment Effect

The instrumental forest described previously generate individual Conditional Average Treatment Effect (A(C)LATE), formally $\tau(X) = \frac{Cov[Y,Z|X=x]}{Cov[W,Z|X=x]}$. In their paper Athey and Wager

2020, the authors propose a method inspired from Chernozhukov et al. 2016 to estimate doubly robust score of $\tau(X)$. To assess potential heterogeneity in the estimated treatment effects, we average the doubly robust scores to obtain the Average Conditional Local Average Treatment Effect. The A(C)LATE is asymptotically normally distributed, thus we can interpret it as an estimator of the doubly robust treatment effect on the compliers for a chosen subgroup.

The method chosen to assess CATE heterogeneity is to used the estimated treatment effect value generated by the Instrumental Forest built with the GRF methodology, to split the sample around the median of estimated CATE and to compute the A(C)LATE on each subsample. Since the A(C)LATE is a asymptotically normal, we can test if each subsample groups individuals with a significantly different from 0 treatment effect, and if the difference between the both groups A(C)LATE is significant.

The doubly robust score is computed as the sum of the estimate CATE by the Instrumental Forest and the multiplication of the Y residuals multiplied by a debiasing weight :

$$\Gamma = \tau(X) + g(X, Z)\left(Y - \mathbb{E}[Y|X] - (W - \mathbb{P}[W = 1|X])\tau(X)\right) \tag{5}$$

With $g(X, Z)$ the vector of debiasing weight :

$$g(X, Z) = \frac{1}{\Delta(X)} \frac{Z - \mathbb{P}[Z = 1|X]}{\mathbb{P}[Z = 1|X](1 - \mathbb{P}[Z = 1|X])} \tag{6}$$

In (6), $\Delta(X)$ is the vector of compliance score : $\mathbb{P}[W|Z = 1, X]$. It represent the propensity of an individual to dropout if the instrument is positive. The compliance score are computed using a causal forest (see Aronow and Carnegie 2013 for detailed explanation around the compliance score). For the practical way of estimating the doubly robust score, see Athey and Wager (2020).

Finally, the A(C)LATE is estimated as the average of all doubly robust scores. The A(C)LATE are computed on each subsample divided around the median of the CATE.

4 Results

4.1 First stage regression

As described before, dropping out is not randomly distributed in the student population and thus in the used sample. We need to rely on an instrument that is correlated with the dropout indicator, but excluded from the outcome equation. To assess the quality of our instrument, the first linear stage is presented in table 2. The first stage regression is run in the first model without controls, then with controls i.e presenting delay before 6th grade, diplomas, and professional position and work sector (public, private) of both parents, gender, and geographical area of birth (on 12 levels for the whole world).

Dropout	(1)	(2)
(min^2 /100)	-0.00403	-0.00458
	(0.00096)	(0.00095)
Constant	0.26052	0.69521
	(0.00388)	(0.14470)
Control variables	No	Yes
R Squared	0.001	0.036
F-Stat for the instrument	17.68	23.36

Table 2: Linear first stage

The initial correlation of the instrument with dropout is strong, and even better after controlling. With a final F-Statistics of the instrument of 23.36, I can rule out the weak instrument bias. As described in the Data section, the minimum distance to the closest higher education establishment is negatively correlated with the probability of dropout.

4.2 Assessing treatment heterogeneity

As described in the methodology part, the constructed distributions of CATE are the results of a data-mining process used to discover heterogeneity in the effect of a treatment. Since the whole process is made to maximize the heterogeneity in $\hat{\tau}$, it is entirely possible that the distributions are the results of noise in the data. Then, we need to rely on statistical estimators and their properties to test for the presence of heterogeneity in the distribution.

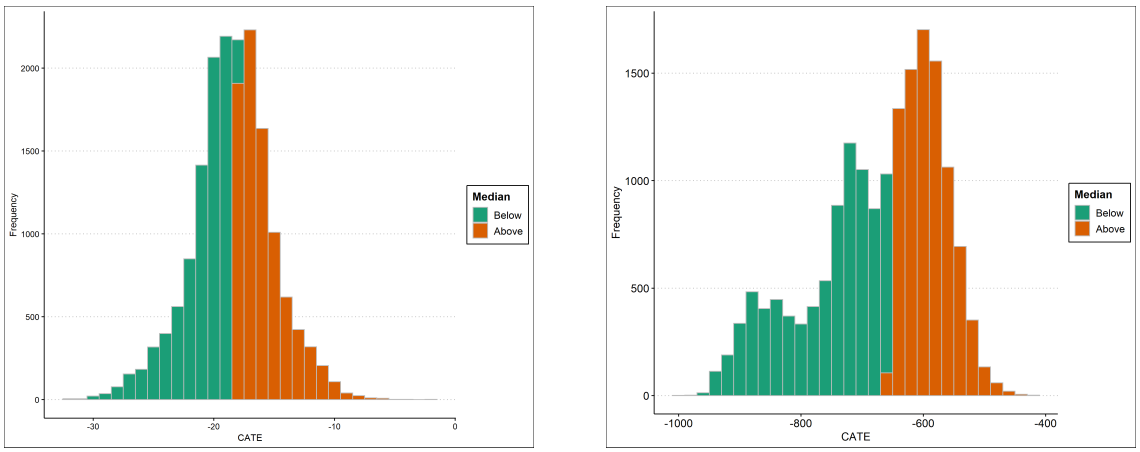
The CATE distributions for the time in employment and average wages are presented in figure 3. Simple statistics are presented in table 3.

In this section, I split the dropout effect distributions on the time in employment and the average wages around the median, compute the A(C)LATE on both subsamples and test for the statistical significance of their differences.

The estimated CATE follow the distributions showed in figure 3 and table 3.

	Min	1st Quartile	Median	Mean	3rd Quartile	Max
Time in Employment	-32.39	-20.37	-18.37	-18.45	-16.43	-2.125
	(-89.98%)	(-56.59%)	(-51.0%)	(-51.23%)	(-45.63%)	(-5.90%)
Average Wages	-991.9	-735.0	-651.9	-674.7	-594.6	-417.9

Table 3: Distribution of CATE of dropout on the Tile in Employment and the Average Wages



(a) Time in Employment (b) Average wages

Figure 3: Conditional Average Treatment Effect of dropout

The CATE on the time of employment range from -90% to -5.90% of the potential employment duration (36 months), with the median equal to -56.59% and the mean equal to -51%. This range indicates that individuals show CATE with a slight or null effect, to very important individuals effects with more than two years and a half of unemployment. This wide gap on individual effect needs to be validated through a A(C)LATE groups difference student test. Half of the individuals show a negative effect of -51.01 % or less, which means that around 50% of the sample could undergo a negative effect of more than a year and a half of unemployment, out of three years.

The CATE on average wages range from -992€ to -420€, with the median at -652€ and mean at -675€ in monthly wage. This first step results seriously denote with precedent findings, as the distribution show individuals effects going from deeply negative (first quartile at -809€ per month) to slighter negative effect for the less disadvantaged individual, almost from simple to double.

As explained in the Methodology section, even if the CATE distribution can be indicative of the dropout effect, we need to rely on the doubly robust score average presented in the methodology part. As proposed by Athey and Wager 2019, the individuals are split following their individual CATE, around the Median. I define $\hat{\tau}_1 \leq Median(\hat{\tau})$ the group of individuals showing a CATE below the median of estimated CATE and $\hat{\tau}_2 > Median(\hat{\tau})$ the group of individuals showing a CATE equal or above the median of estimated CATE. The $\hat{\tau}_1$ corresponds to a group showing a strong negative effect of dropout, while the $\hat{\tau}_2$ correspond to a group showing a slighter negative effect.

The A(C)LATE on each subgroups are presented in Table 4 for A(C)LATE on time in employment, and in Table 5 for A(C)LATE on average wages.

	A(C)LATE	Std.Error
$\hat{\tau}_1^{roe} \leq Median(\hat{\tau}^{roe})$	-21.23	2.26
$\hat{\tau}_2^{roe} > Median(\hat{\tau}^{roe})$	-14.69	2.34
T-Statistic	-2.01 (p.val = 4.44%)	

Table 4: Average Local Average Treatment Effect of dropping out on the Time in Employment

For the effect of dropping out on the time in employment, we observe one subsample with a strong negative effect on the compliers of -21.23 months, so a negative difference of 59 % for dropout compared to those who didn't. The subsample grouping individual with a CATE above the median shows an A(C)LATE of -14.69 months out of 36, so a decrease of 41% in employment for dropout. The student test of the difference of A(C)LATE between these two groups indicates that this difference is statistically significant at less than 5% level.

	A(C)LATE	Std.Error
$\hat{\tau}_1^{aw} \leq Median(\hat{\tau}^{aw})$	-1365.63	199.05
$\hat{\tau}_2^{aw} > Median(\hat{\tau}^{aw})$	-612.11	153.48
T-Statistic	-2.99 (p.val = 0.27%)	

Table 5: Local Average Treatment Effect of droping out on the Average Wages

The A(C)LATE of dropping out on the average wages indicates that 50% of the population of actual dropout is highly penalized with an effect of -1365€ per month on average wages. The top 50% of the distribution shows a A(C)LATE of -612.11. These two A(C)LATE are significantly different from 0 and highly significant. We can conclude that we correctly identify two subgroups of dropouts with heterogeneous effects of the dropout. While the first group presents a strong negative effect, the second one presents almost half of the effect.

In this section, we correctly assessed the presence of heterogeneity in the estimated CATE for both interest indicators. However, we need to study the social composition of the CATE distribution to understand the social dynamic behind the effect of dropout on labor market outcomes.

(For the distribution of actual dropout around the CATE median, see Appendix 1).

4.3 Composition of subsamples

In this section, we study the social and academic composition of the subsamples built around the median for both indicators. The objective is to understand if the CATE distribution is conditional on individual characteristics and how they are influencing the effect of dropout. To do so, a Logit model of being in the top 50% of the CATE distribution is estimated.

I include all the variables used to build the Generalized Random Forest. The main independent variables of interest are gender, the diplomas, if the student did a foreign trip during her study, or an internship, the education, and the occupation of both parents. The objective of this analysis is, first, to understand how these variables are influencing the CATE distribution i.e if having parents with a certain occupation can increase the chance to be in the less penalized group. Second, to understand which variables are the most important between exogenous (parent's characteristics) and endogenous (diploma and academic characteristics) variables to shape the distribution of the CATE.

4.3.1 Time in employment

The logarithm of the odds ratios for the gender, highest diploma tried and other academic variables are reported in table 6. The other odds ratio are reported in the Appendix. I perform the GVIF analysis in order to detect potential multicolinearity in the variables, however no measure goes above the recommended level. This analysis can be found table 9. The results for every variables of interest are presented in the Appendix, section 6.2.

	Dependent variable:
	$\hat{\tau} > Median(\hat{\tau})$
Female	0.047
	(0.051)
Highest diploma tried	
Bac+2 Industrial	0.077
	(0.146)
Bac+2 Tertiary	-0.759^{***}
	(0.100)
Licence Pro	-0.123
	(0.127)
Bac+3/4 HS Econ Law	-0.409^{***}
	(0.112)
Bac+3/4 STEM	-0.137
	(0.138)
Bac+5 HS Econ Law	-0.217^{*}
	(0.111)
Bac+5 STEM	-0.039
	(0.127)
Business School	-0.229
	(0.179)
Ingeniering School	-0.396^{***}
	(0.131)
PhD	-0.204^{*}
	(0.112)
Other academic characteristics	
Foreign trip : One	0.056
	(0.073)
Foreign trip : Many	0.189^{*}
	(0.102)
Internship : Yes	-0.083
	(0.052)
Observations	17,094
Log Likelihood	-5,213.640
Akaike Inf. Crit.	10,683.280
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 6: Logit model for being in the top 50% (ROE)

Following the estimation of the Logit model, we observe that females have a higher likelihood of ending in the top 50% of the distribution. a female represents an advantage to mitigate the negative effect of dropping out.

The effect of the highest diploma tried is mixed and varies depending on the field and duration of the considered diploma. The reference is Bac +2 in Health. This category doesn't

have a high dropout rate. Compared to Bac+2 Health, Bac+2 Industrial don't show any significant effect, while short degrees in the tertiary field are highly penalized by dropout. This result can be explained either by the high demand for Health and Industrial short degree graduates, or because the share of dropout in Tertiary is higher with also more graduate from this field.

For longer degrees, from Bac+3 to Bac+5, students who dropped out from a Bac+3/4 in economics, law, and social science or from Engineering school (equivalent to Bac +5), the likelihood of ending in the top half of the distribution is significantly reduced compared to Bac+2 in Health. For the rest of the degrees, the effect of dropping out is not significantly different from that of Bac +2 in Health degree.

Among every degree, the most penalized students are those who dropped out from a Bac+2 Tertiary, then Bac +3/4 Econ and Law and Engineering School. The penalty goes from simple to double. While dropouts are common among the first two, it is rarer for engineering students, and thus are far more concerning for the first two degrees dropout.

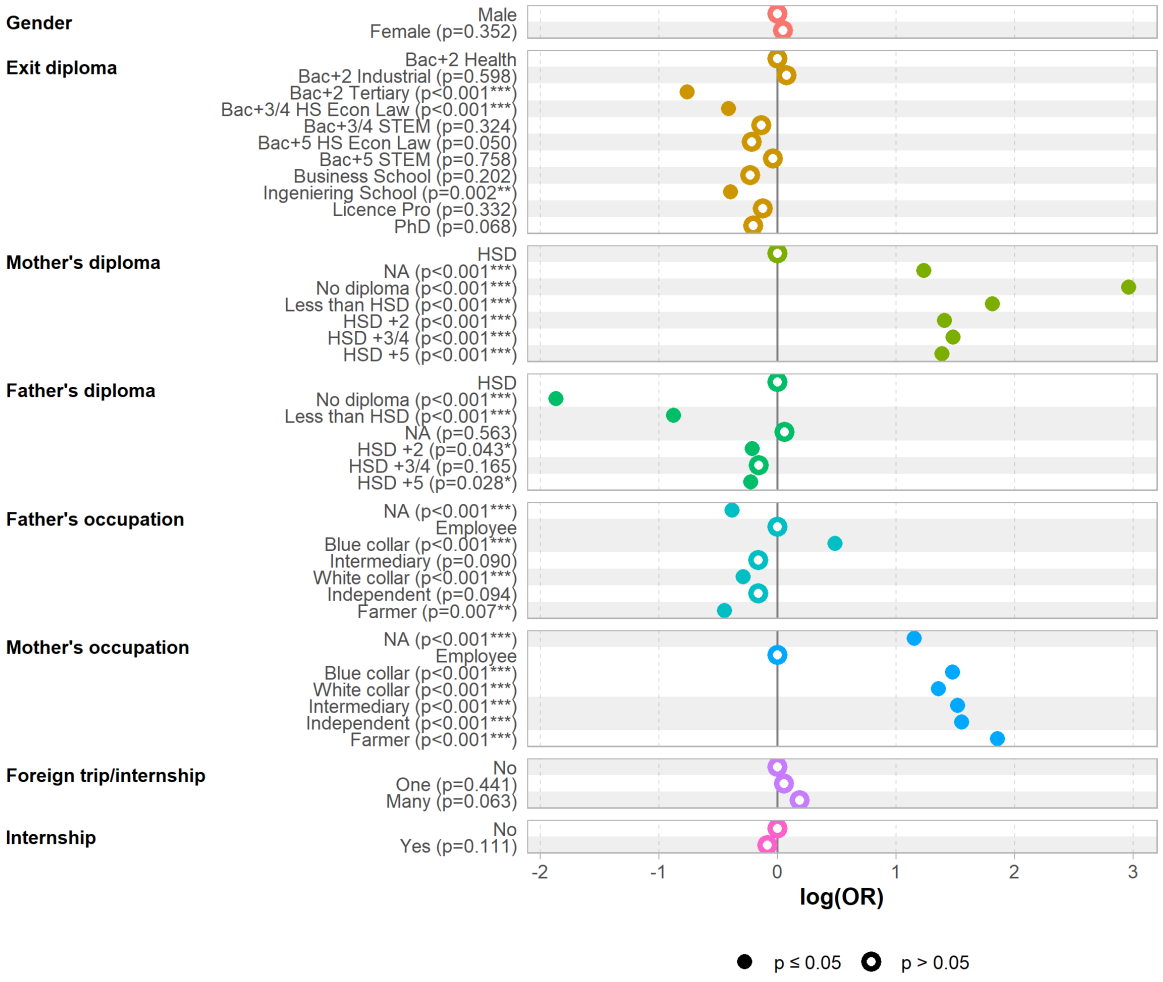


Figure 4: Log(odd ratios) for being in the top 50% of the CATE distribution (Time in employment)

In this figure, the reference levels are also reported. The reference level for mother and

father's highest diploma is the high school diploma. All the other mother's diplomas have a positive and significant effect compared to having a mother with an HSD. Having a mother with no diploma increase a lot the likelihood of being above the median of the CATE. All the other levels show a positive effect compared to the HSD with a smaller effect for having a mother with less than a diploma. Overall, the mother's education has a positive impact on the likelihood of being highly penalized by dropping out. The worst possibility is to have a mother with an HSD, which increases the most the likelihood of ending in the bottom 50%. On the contrary, having a father with an HSD or HSD+2 is the more advantaging social origin. All the others levels exhibit negative or null effects for being in the top 50%, with a decrease of the magnitude of this effect when the level of the diploma increase. The positive effect of the mother's diploma is higher than the father's diploma, indicating that it is the mother's characteristics that can help the most to mitigate the negative effect of dropout.

Generally speaking, having a mother with less than HSD can reduce the dropout penalty, while it is highly increasing this penalty if it is the father which has less than an HSD. Having educated parents is also mixed: while it advantaging if it is the mother, it is penalizing when the student's father is educated.

Regarding the parent's occupation, the reference level is Employee, which is also the biggest group in terms of the social original for the parents. Compared to having an employee mother, every other social origin has a positive impact on the likelihood of being the less penalized student, with a slightly smaller effect for the White-collar mother. On the contrary, every other occupation than an employee have a negative effect on the father, except for Blue collars. Having a Blue collar father is highly advantaging while having a farmer or White collar father doesn't help to end in the top half of the distribution.

I included in the analysis two secondary academic characteristics: if the student made a foreign trip during her study, or if she did an internship. The objective is to understand if the academic investment of students past the diploma is significant to reduce the dropout penalty. However, these two variables don't have any significant effect. We can conclude that, while the highest diploma can affect the magnitude of the dropout effect, secondary academic characteristics such as these don't play an active role in shaping the structure of the CATE, and can't act as a "security net".

By analyzing the conditional effect of dropout on the time in employment, it appears that having an educated and well-employed mother affects positively a student's probability of being less penalized by dropout, while the effect is completely reversed regarding the father's occupation and diploma. Overall, the social situation of both parents is playing a far more active role in shaping the CATE distribution than the highest diploma tried or secondary academic characteristics.

4.3.2 Average wage

The logarithm of the odd ratios are reported in table 7. The complete table is reported in Appendix.

	Dependent variable:
	$\hat{\tau} > Median(\hat{\tau})$
Female	4.930***
	(0.107)
Highest diploma tried	
Bac+2 Industrial	0.060
	(0.177)
Bac+2 Tertiary	1.324***
	(0.130)
Licence Pro	0.438***
	(0.163)
Bac+3/4 HS Econ Law	0.012
	(0.141)
Bac+3/4 STEM	0.236
	(0.183)
Bac+5 HS Econ Law	0.137
	(0.140)
Bac+5 STEM	0.252
	(0.165)
Business School	0.171
	(0.228)
Ingeniering School	−0.029
	(0.189)
PhD	0.506***
	(0.145)
Observations	17,094
Log Likelihood	-6,444.328
Akaike Inf. Crit.	13,006.660
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 7: Logit model for being in the top 50% (Average wage)

While the gender variable was not significant for the dropout effect, it is highly significant once we analyze the dropout effect on the average wages. Being a female is now highly significant in reducing the likelihood of being highly penalized for dropping out. It could mean that, while being equally employed in terms of time, females engage in work with fewer wages penalty for dropouts.

As seen in the previous section, students who dropped out of Bac+2 Tertiary were the most

penalized by dropout in terms of time in employment. Once we weigh the time in employment by the wages, these students are now less penalized by dropout. It can mean that, while they are highly disadvantaged in finding employment compared to Bac+2 Health, once they find one their wages are far more competitive. This result can be explained by the employment structure of the health and tertiary sector for short degrees holders, with most of the Health sector jobs that can't be accessed without a diploma, while it is more accepted for tertiary dropouts.

The Ph.D. dropouts are also less penalized than the rest of the sample, meaning that these students succeed to value their experience better than short degree holders. This can be explained by the strong competition to get into Ph.D. program and the high rate of dropout in certain fields. The last degree from which it is less penalized to drop out is the Professional Licence, a short degree oriented toward technical jobs. This result is similar to the positive effect of Bac+2 Tertiary, the kind of job these two degrees are leading two are similar.

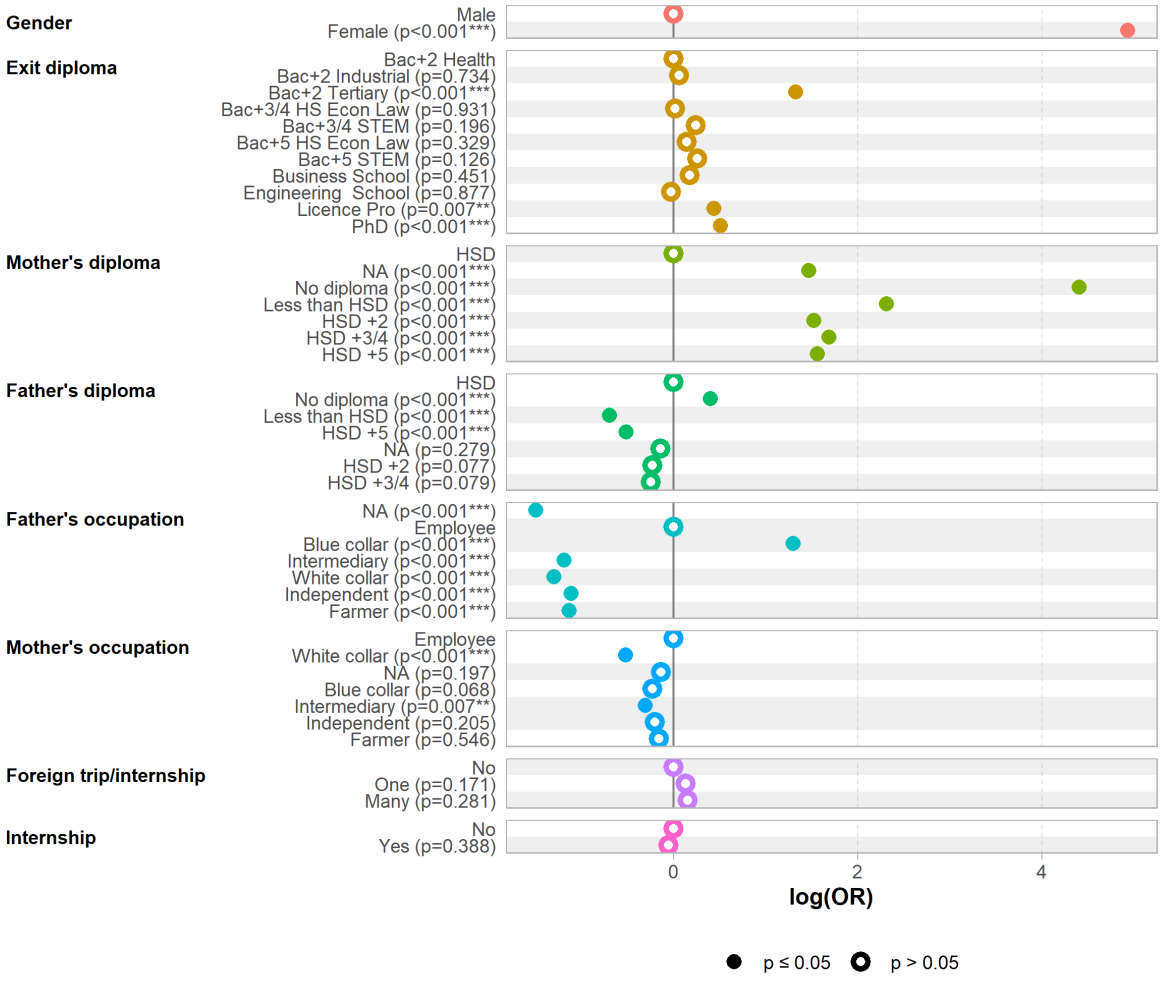


Figure 5: Log(odd ratios) for being in the top 50% of the CATE distribution (Average wage)

The effect of the mother's diploma is as clear as the effect on the time in employment While having a mother with no diploma affects more strongly the likelihood of having a CATE above

the median, the other levels have approximately the same effect, with a strong positive effect for children of mothers without a diploma. A mother having less than an HSD is the second most positive effect compared to the HSD level, while longer degrees have a slight positive effect compared to having a mother with an HSD.

The father's diploma is less important in shaping the CATE distribution for the average wage than for the time in employment. While it was highly decreasing the likelihood of having a CATE above the median to have a father with no diploma, it is actually slightly positive in terms of average wages. Having a father with less than an HSD is still negatively affecting the likelihood of having an above-median CATE. The only significant effect among children of fathers with a higher education diploma is for those with HSD +5, with a negative effect. This effect is similar to the one for the time in employment.

Overall, the mother's diploma is more influencing the CATE distribution than the father's diploma. This result is consistent with the analysis on the effect of individual characteristics on the dropout effect on time in employment.

Regarding the mother's occupation, every level is non significantly different from the employee level, except for the White collars and Intermediaries. White collars and intermediaries children are less likely to end up in the top 50%. These two effects are relatively small compared to those of the mother's diploma for example. While this variable was very significant in shaping the CATE distribution for the time in employment, it is almost not significant for the average wage.

On the contrary, the father occupation is now very significant compared to its role in shaping the dropout effect on time in employment. Compared to employee fathers, only blue collar children have a higher probability of ending in the top 50%. The other levels act negatively in the same range. The mother education has almost no effect on the dropout effect distribution. While being the children of a white-collar or intermediary negatively affects the likelihood of being in the less penalized group, these effects are small.

The other academic characteristics effect doesn't have an effect significantly different from 0, as for the time in employment analysis.

By analyzing the effect of each variable used to build the Instrumental Forest on the likelihood of being less penalized than average by dropout, we got a better understanding of how this effect is shaped across the population. The social origin is mostly driving the heterogeneity of the dropout effect, especially the mother's one. The mother's occupation is more important for shaping the time in employment effect than for the average wage, while the mother's diploma is mostly shaping the CATE distribution for the average wage. The father's diploma is more important for the time in employment than for average wages, while it is the contrary for his professional occupation. Diploma and academic achievements such as travel and internship are, overall, not very affecting compared to the social origin. Concerning the highest tried, it seems that dropping out from a short degree is more penalizing regarding the time in employment,

while less penalizing regarding average wage. Overall, the field seems to be more important than the duration of the highest diploma tried. Gender is driving the heterogeneity for the dropout effect on average wages, while it is not significant for the time in employment.

4.4 Variables importance in determining the causal effect

In order to understand better which type of variables (roughly exogenous and endogenous to the student) is more determinant in shaping the distribution of the CATE. For this aim, we use the variable importance, which is a diagnosis tool usual to Random Forest. The variable importance indicates the number of times a variable x has been split at each depth in the building of the Instrumental Forest. A decaying exponent is introduced, as the variables used for the first split are more important in determining the shape of the CATE distribution than those used for the last split.

We create two groups of variables :

$x_{endogenous}$ = (Highest diploma tried, High school diploma characteristics, secondary academic characteristics, delay at HSD)
$x_{exogenous}$ =(Gender, Professional position and diploma of both parents, number of siblings)

Each variable of $x_{exogenous}$ or $x_{endogenous}$ has its own variable importance expressed as a rate. Since we take into account a different number of variables, with different levels, the sum of each vector's variable importance needs to be divided by the number of levels in the vector. We don't include geographical variables on higher education establishments or high schools because they are used as control. The results are presented in table 8 :

	ROE			AW		
	sum VI	n(levels)	$\frac{VI}{n(levels)}$	sum VI	n(levels)	$\frac{VI}{n(levels)}$
$x_{exogenous}$	0.542	27	0.020	0.517	27	0.019
$x_{endogenous}$	0.431	32	0.013	0.479	32	0.015

Table 8: Variable importance for both Instrumental Forest

As we previously understood in the analysis of the LOGIT estimation, exogenous variables are more important for determining the dropout effect than endogenous ones. This effect is consistent for both indicators, while less pronounced for the average wage estimation. Its means that, while variables on which students don't have any effect are always more important, variables on which they have an effect are more important when determining the effect of dropout on average wages than on the time in employment.

5 Conclusion

In this paper, I showed that the effect of dropping out on labor market outcomes exhibits statistically significant heterogeneity. I also showed that these heterogeneous effects are conditional to certain education or sociodemographic characteristics.

I used the Generalized Random Forest algorithm in an instrumental variable setting to estimate individual Conditional Average Treatment Effects and then grouped these CATE by subgroups around the median to compute the Local Average Treatment Effect on each subsample. The social composition of these subgroups was studied using a logit model to understand which individual characteristics were more likely to minimize the dropout negative effect.

The main finding is that the effect of dropout on the time in employment and the average wage is heterogeneous and that individual characteristics are actively shaping the distribution of the treatment effects. The heterogeneity in these effects has been tested by using the standard property of the Average (Conditional) Local Average Treatment Effect. We observed two subgroups for each indicator, one with a strong negative effect and the other with a slighter negative effect. This result indicates that, while every student are penalized for dropout, some are more penalized than others (going from simple to double the magnitude of the effect).

The diploma from which students drop out and their academic achievements are less important in shaping the heterogeneity of the dropout effect than the social origin. Dropping out from a short degree is more penalizing regarding time in employment, while it is less penalized for average wages. Having other academic experiences such as international study trip or internship doesn't play any role in the heterogeneity of the dropout effect.

Regarding the social origin of both parents, the mother's one is always more important than the father's. The mother's diploma and occupation are actively shaping the heterogeneity of the time in employment, while it is mostly the diploma that is active regarding average wages.

The most interesting finding is in the importance of each type of variable (exogenous or endogenous). The exogenous variable, on which students don't have any effect, is the most overriding in shaping the heterogeneity of the dropout effect. However, this effect is weaker concerning the average wage, indicating that while the time in employment is reflecting potential social inequality, the labor market relies more on academic achievement to set wages.

This paper raises new results on the need for higher education policymakers to consider multi-dimensional effects of dropout and by extension delay in graduation. By showing that some former students are not as penalized as others by dropping out, this paper brings a less common conclusion on the signal brought by the dropout. The importance of exogenous variables in shaping the heterogeneity of the dropout effect stresses the fact that policy-makers should focus on making higher education diplomas and achievements valuable experiences for students, even in the case of dropout.

References

Aina, Carmen, et al. 2018. "The Economics of University Dropouts and Delayed Graduation: A Survey". *SSRN Electronic Journal*. ISSN: 1556-5068, visited on 01/14/2022. doi:10.2139/ssrn.3153385. https://www.ssrn.com/abstract=3153385.

Angrist, J. D., and A. B. Krueger. 1991. "Does Compulsory School Attendance Affect Schooling and Earnings?" *The Quarterly Journal of Economics* 106, no. 4 (): 979–1014. ISSN: 0033-5533, 1531-4650, visited on 06/01/2021. doi:10.2307/2937954. https://academic.oup.com/qje/article-lookup/doi/10.2307/2937954.

Arcidiacono, Peter, Patrick Bayer, and Aurel Hizmo. 2010. "Beyond Signaling and Human Capital: Education and the Revelation of Ability". *American Economic Journal: Applied Economics* 2, no. 4 (): 76–104. ISSN: 1945-7782, 1945-7790, visited on 11/29/2020. doi:10.1257/app.2.4.76. https://pubs.aeaweb.org/doi/10.1257/app.2.4.76.

Aronow, Peter M., and Allison Carnegie. 2013. "Beyond LATE: Estimation of the Average Treatment Effect with an Instrumental Variable". *Political Analysis* 21 (4): 492–506. https://EconPapers.repec.org/RePEc:cup:polals:v:21:y:2013:i:04:p:492-506_01.

Athey, Susan, and Guido Imbens. 2016. "Recursive partitioning for heterogeneous causal effects". *Proceedings of the National Academy of Sciences* 113, no. 27 (): 7353–7360. ISSN: 0027-8424, 1091-6490, visited on 10/06/2020. doi:10.1073/pnas.1510489113. http://www.pnas.org/lookup/doi/10.1073/pnas.1510489113.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2019. "Generalized random forests". *The Annals of Statistics* 47 (2): 1148–1178.

Athey, Susan, and Stefan Wager. 2019. "Estimating Treatment Effects with Causal Forests: An Application". *arXiv:1902.07409 [stat]* (). Visited on 04/28/2021. arXiv: 1902.07409. http://arxiv.org/abs/1902.07409.

— . 2020. "Policy Learning with Observational Data". *arXiv:1702.02896 [cs, econ, math, stat]* (). Visited on 04/25/2021. arXiv: 1702.02896. http://arxiv.org/abs/1702.02896.

Becker, Gary S. 1993. *Human capital: a theoretical and empirical analysis, with special reference to education*. 3rd ed. Chicago: The University of Chicago Press. ISBN: 978-0-226-04119-3 978-0-226-04120-9.

Bjerk, David. 2012. "Re-examining the impact of dropping out on criminal and labor outcomes in early adulthood". *Economics of Education Review* 31, no. 1 (): 110–122. ISSN: 02727757, visited on 01/20/2022. doi:10.1016/j.econedurev.2011.09.003. https://linkinghub.elsevier.com/retrieve/pii/S0272775711001506.

Breiman, L., et al. 1983. “Classification and Regression Trees”.

Breiman, Leo. 2001. “Random Forests”. *Machine Learning* 45 (1): 5–32. ISSN: 08856125, visited on 04/21/2021. doi:10.1023/A:1010933404324. http://link.springer.com/10.1023/A:1010933404324.

Brodaty, Thomas, Robert Gary-Bobo, and Ana Prieto. 2008. “Does Speed Signal Ability? The Impact of Grade Repetitions on Employment and Wages”. *C.E.P.R. Discussion Papers, CEPR Discussion Papers* ().

Card, David. 1993. *Using Geographic Variation in College Proximity to Estimate the Return to Schooling*. NBER Working Papers 4483. National Bureau of Economic Research, Inc. https://EconPapers.repec.org/RePEc:nbr:nberwo:4483.

Chernozhukov, Victor, et al. 2016. “Locally robust semiparametric estimation”. *arXiv preprint arXiv:1608.00033*.

Fang, Hanming. 2006. “DISENTANGLING THE COLLEGE WAGE PREMIUM: ESTIMATING A MODEL WITH ENDOGENOUS EDUCATION CHOICES”. *International Economic Review* 47, no. 4 (): 1151–1185. ISSN: 0020-6598, 1468-2354, visited on 11/25/2020. doi:10.1111/j.1468-2354.2006.00409.x. http://doi.wiley.com/10.1111/j.1468-2354.2006.00409.x.

Flores-Lagunes, Alfonso, and Audrey Light. 2007. “Interpreting sheepskin effects in the returns to education” ().

Fox, John, and Georges Monette. 1992. “Generalized Collinearity Diagnostics”. *Journal of the American Statistical Association* 87, no. 417 (): 178–183. ISSN: 0162-1459, 1537-274X, visited on 01/19/2022. doi:10.1080/01621459.1992.10475190. http://www.tandfonline.com/doi/abs/10.1080/01621459.1992.10475190.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York. ISBN: 978-0-387-84857-0 978-0-387-84858-7, visited on 11/28/2020. doi:10.1007/978-0-387-84858-7. http://link.springer.com/10.1007/978-0-387-84858-7.

Imbens, Guido W., and Joshua D. Angrist. 1994. “Identification and Estimation of Local Average Treatment Effects”. *Econometrica* 62, no. 2 (): 467. ISSN: 00129682, visited on 11/26/2020. doi:10.2307/2951620. https://www.jstor.org/stable/2951620?origin=crossref.

Mahjoub, Mohamed-Badrane. 2017. “The treatment effect of grade repetitions”. *Education Economics* 25, no. 4 (): 418–432. ISSN: 0964-5292, 1469-5782, visited on 05/31/2021. doi:10.1080/09645292.2017.1283006. https://www.tandfonline.com/doi/full/10.1080/09645292.2017.1283006.

Matkovic, T., and I. Kogan. 2012. "All or Nothing? The Consequences of Tertiary Education Non-Completion in Croatia and Serbia". *European Sociological Review* 28, no. 6 (): 755–770. ISSN: 0266-7215, 1468-2672, visited on 01/20/2022. doi:10.1093/esr/jcr111. https://academic.oup.com/esr/article-lookup/doi/10.1093/esr/jcr111.

Navarro, Salvador, Jane Fruehwirth, and Yuya Takahashi. 2016. "How The Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects". *Journal of Labor Economics* 34. doi:10.1086/686262.

Psacharopoulos, George, and Harry Anthony Patrinos. 2018. "Returns to investment in education: a decennial review of the global literature". *Education Economics* 26, no. 5 (): 445–458. ISSN: 0964-5292, 1469-5782, visited on 06/01/2021. doi:10.1080/09645292.2018.1484426. https://www.tandfonline.com/doi/full/10.1080/09645292.2018.1484426.

Reisel, Liza. 2013. "Is more always better? Early career returns to education in the United States and Norway". *Research in Social Stratification and Mobility* 31 (): 49–68. ISSN: 02765624, visited on 01/31/2022. doi:10.1016/j.rssm.2012.10.002. https://linkinghub.elsevier.com/retrieve/pii/S0276562412000558.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66 (5): 688–701. ISSN: 0022-0663, visited on 11/28/2020. doi:10.1037/h0037350. http://content.apa.org/journals/edu/66/5/688.

Schnepf, Sylke V. 2014. *Do Tertiary Dropout Students Really Not Succeed in European Labour Markets?* IZA Discussion Papers 8015. Bonn: Institute for the Study of Labor (IZA). http://hdl.handle.net/10419/96694.

Scholten, Mirte, and Nicole Tieben. 2017. "Vocational qualification as safety-net? Education-to-work transitions of higher education dropouts in Germany". *Empirical Research in Vocational Education and Training* 9 (). doi:10.1186/s40461-017-0050-7.

Spence, Michael. 1973. "Job Market Signaling". *The Quarterly Journal of Economics* 87, no. 3 (): 355. ISSN: 00335533, visited on 11/29/2020. doi:10.2307/1882010. https://academic.oup.com/qje/article-lookup/doi/10.2307/1882010.

Vignoles, Anna F., and Nattavudh Powdthavee. 2009. "The Socioeconomic Gap in University Dropouts". *The B.E. Journal of Economic Analysis & Policy* 9 (1). doi:doi:10.2202/1935-1682.2051. https://doi.org/10.2202/1935-1682.2051.

Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". *Journal of the American Statistical Association* 113, no. 523 (): 1228–1242. ISSN: 0162-1459, 1537-274X, visited on 04/23/2021. doi:10.1080/01621459.2017.1319839. https://www.tandfonline.com/doi/full/10.1080/01621459.2017.1319839.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data*. 2nd ed. Cambridge, Mass: MIT Press. ISBN: 978-0-262-23258-6.

6 Appendix

6.1 GVIF : Assessing potential multicolinearity

	GVIF	Df	GVIF^(1/(2*Df))	GVIF^(1/(2*Df))^2
Gender	1.47	1.00	1.21	1.47
Highest diploma tried	2.44	10.00	1.05	1.09
Diploma (mother)	3.91	6.00	1.12	1.25
Diploma (father)	4.86	6.00	1.14	1.30
Occupation (father)	4.09	6.00	1.12	1.26
Occupation (mother)	2.92	6.00	1.09	1.20
Region of High School diploma	68479.13	22.00	1.29	1.66
Region of highest diploma	71799.94	22.00	1.29	1.66
Foreign trip during study	1.32	2.00	1.07	1.15
Internship	1.12	1.00	1.06	1.12

Table 9: Various measure of VIF

As suggested in Fox and Monette 1992, using GVIF^(1/(2*Df) allows to compare the value of GVIF across different number of parameters. I elevate this measure to the square to use the standard rule of thumb of GVIF. Here, no GVIF goes above 2, so I can safely include and interpret all the parameters in the Logit model.

6.2 Logit table : top 50% for CATE on time in employment

	Dependent variable:
	median_2
Female	0.047
	(0.051)
Highest diploma tried	
Bac+2 Industrial	0.077
	(0.146)
Bac+2 Tertiary	-0.759^{***}
	(0.100)
Licence Pro	-0.123
	(0.127)
Bac+3/4 HS Econ Law	-0.409^{***}
	(0.112)
Bac+3/4 STEM	-0.137
	(0.138)
Bac+5 HS Econ Law	-0.217^{*}
	(0.111)
Bac+5 STEM	-0.039
	(0.127)
Business School	-0.229
	(0.179)
Ingeniering School	-0.396^{***}
	(0.131)
PhD	-0.204^{*}
	(0.112)
Other academic characteristics	
Foreign trip : One	0.056
	(0.073)
Foreign trip : Many	0.189^{*}
	(0.102)
Internship : Yes	-0.083
	(0.052)
Mother's diploma	
NA	1.236^{***}
	(0.105)
No diploma	2.962^{***}
	(0.088)
Less than HSD	1.815^{***}
	(0.085)
HSD +2	1.411^{***}
	(0.090)
HSD +3/4	1.483^{***}
	(0.092)
HSD +5	1.387^{***}
	(0.110)
Father diploma	
NA	0.062
	(0.107)
No diploma	-1.867^{***}
	(0.094)
Less than HSD	-0.877^{***}
	(0.090)
HSD +2	-0.212^{**}
	(0.105)
HSD +3/4	-0.154
	(0.111)
HSD +5	-0.225^{**}
	(0.102)
Mother's occupation	
NA	1.153^{***}
	(0.083)
Blue collar	1.477^{***}
	(0.095)
Intermediary	1.522^{***}
	(0.092)
White collar	1.360^{***}
	(0.072)
Independent	1.556^{***}
	(0.130)
Farmer	1.856^{***}
	(0.230)
Father's occupation	
NA	-0.383^{***}
	(0.109)
Blue collar	0.486^{***}
	(0.085)
Intermediary	-0.161^{*}
	(0.095)
White collar	-0.290^{***}
	(0.079)
Independent	-0.160^{*}
	(0.096)
Farmer	-0.446^{***}
	(0.167)
Observations	17,094
Log Likelihood	-6,174.392
Akaike Inf. Crit.	12,606.780
Note:	*p<0.1; **p<0.05; ***p<0.01

Reference levels : Bac +2 Health (Highest diploma tried), High School Diploma/Baccalaureate (father/mother diploma), Employee (father/mother occupation).

6.3 Logit table : top 50% for CATE on average wage

	Dependent variable:
	$\hat{\tau} > Median(\hat{\tau})$
Female	4.930***
	(0.107)
Highest diploma tried	
Bac+2 Industrial	0.060
	(0.177)
Bac+2 Tertiary	1.324***
	(0.130)
Licence Pro	0.438***
	(0.163)
Bac+3/4 HS Econ Law	0.012
	(0.141)
Bac+3/4 STEM	0.236
	(0.183)
Bac+5 HS Econ Law	0.137
	(0.140)
Bac+5 STEM	0.252
	(0.165)
Business School	0.171
	(0.228)
Ingeniering School	−0.029
	(0.189)
PhD	0.506***
	(0.145)
Mother's diploma	
NA	1.467***
	(0.130)
No diploma	4.402***
	(0.125)
Less than HSD	2.312***
	(0.105)
HSD +2	1.520***
	(0.109)
HSD +3/4	1.687***
	(0.115)
HSD +5	1.560***
	(0.144)
Father diploma	
NA	−0.144
	(0.133)
No diploma	0.397***
	(0.112)
Less than HSD	−0.698***
	(0.111)
HSD +2	−0.232*
	(0.131)
HSD +3/4	−0.248*
	(0.141)
HSD +5	−0.516***
	(0.131)
NA	−1.500***
	(0.139)
Mother's occupation	
NA	−0.142
	(0.110)
Blue collar	−0.232*
	(0.127)
Intermediary	−0.309***
	(0.115)
White collar	−0.524***
	(0.088)
Independent	−0.207
	(0.163)
Farmer	−0.163
	(0.270)
Father's occupation	
NA	−1.500***
	(0.139)
Blue collar	1.297***
	(0.115)
Intermediary	−1.193***
	(0.118)
White collar	−1.304***
	(0.099)
Independent	−1.113***
	(0.120)
Farmer	−1.137***
	(0.194)
Observations	17,094
Log Likelihood	-3,782.291
Akaike Inf. Crit.	7,822.582
Note:	*p<0.1; **p<0.05; ***p<0.01

Reference levels : Bac +2 Health (Highest diploma tried), High School Diploma/Baccalaureate (father/mother diploma), Employee (father/mother occupation).

6.4 Time in employment distribution conditional on dropout and individual characteristics

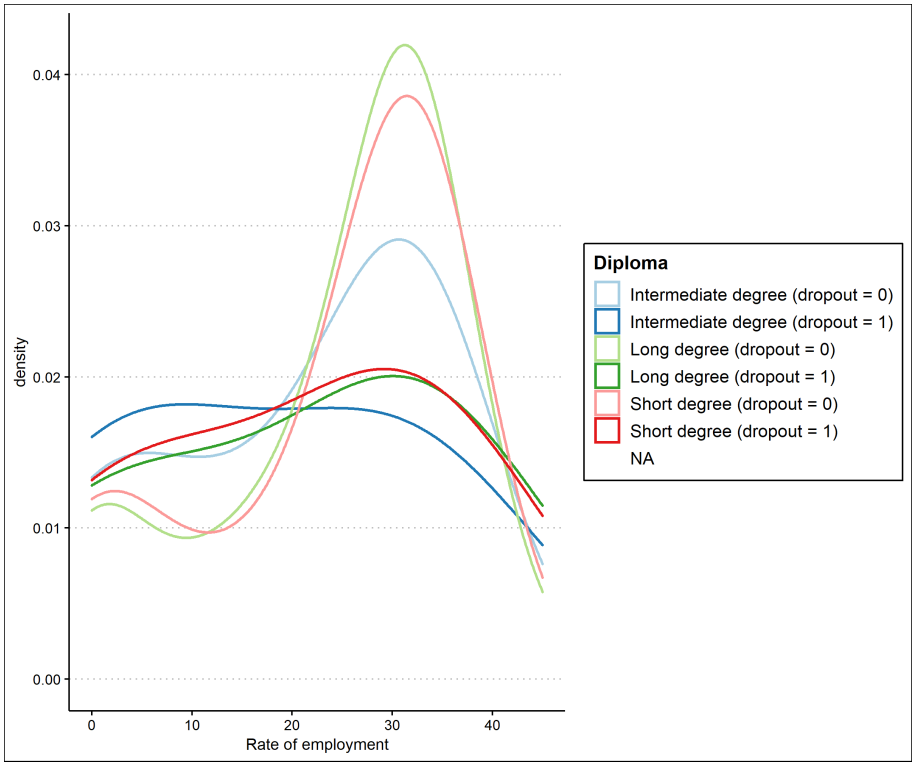


Figure 6: Time in employment distribution conditional on highest diploma tried

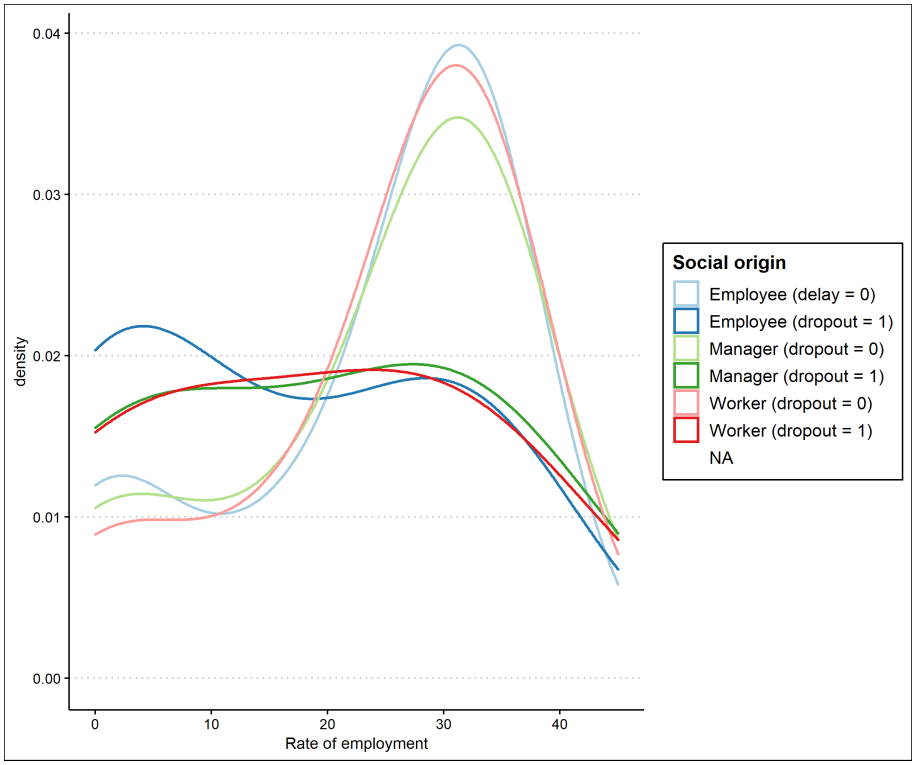


Figure 7: Time in employment distribution conditional on the mother's occupation

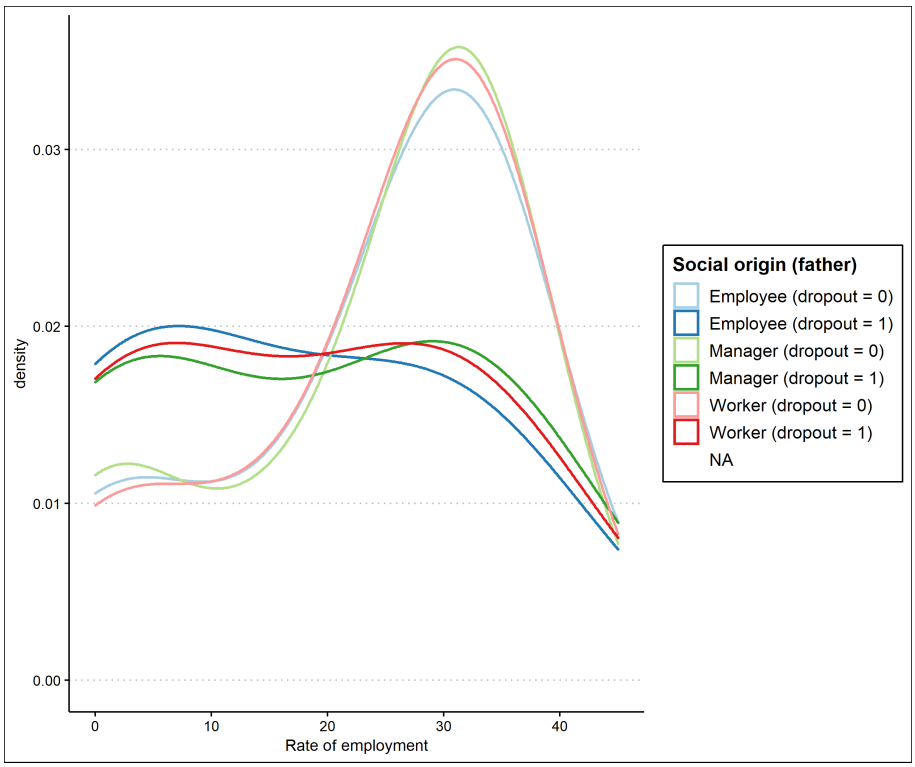


Figure 8: Time in employment distribution conditional on the father's occupation

6.5 Time in employment distribution conditional on dropout and individual characteristics

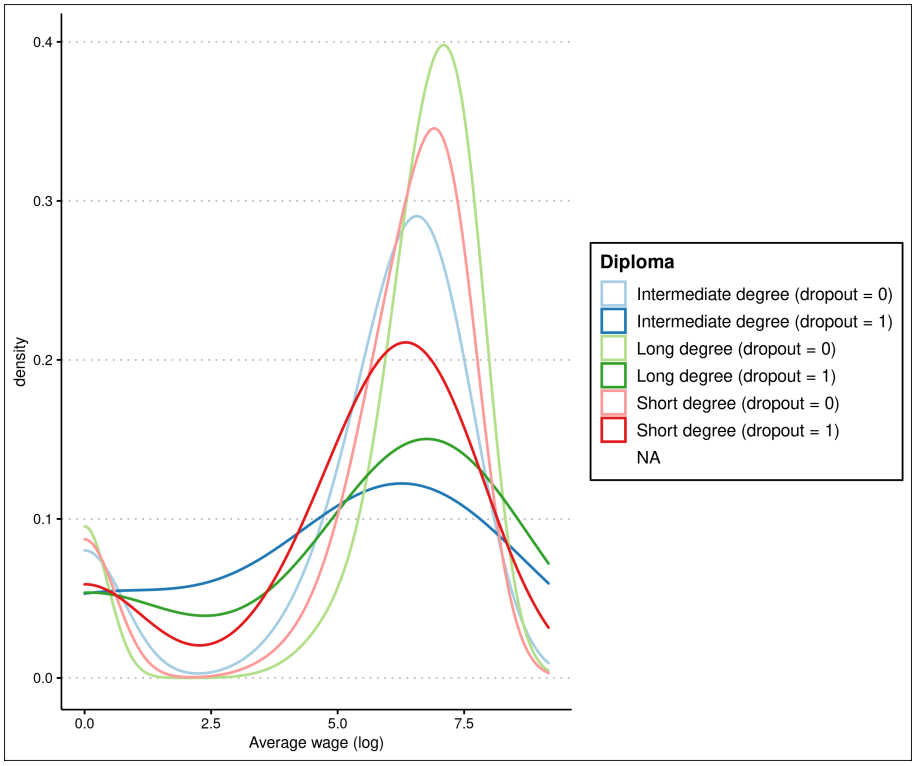


Figure 9: Average wage distribution conditional on highest diploma tried

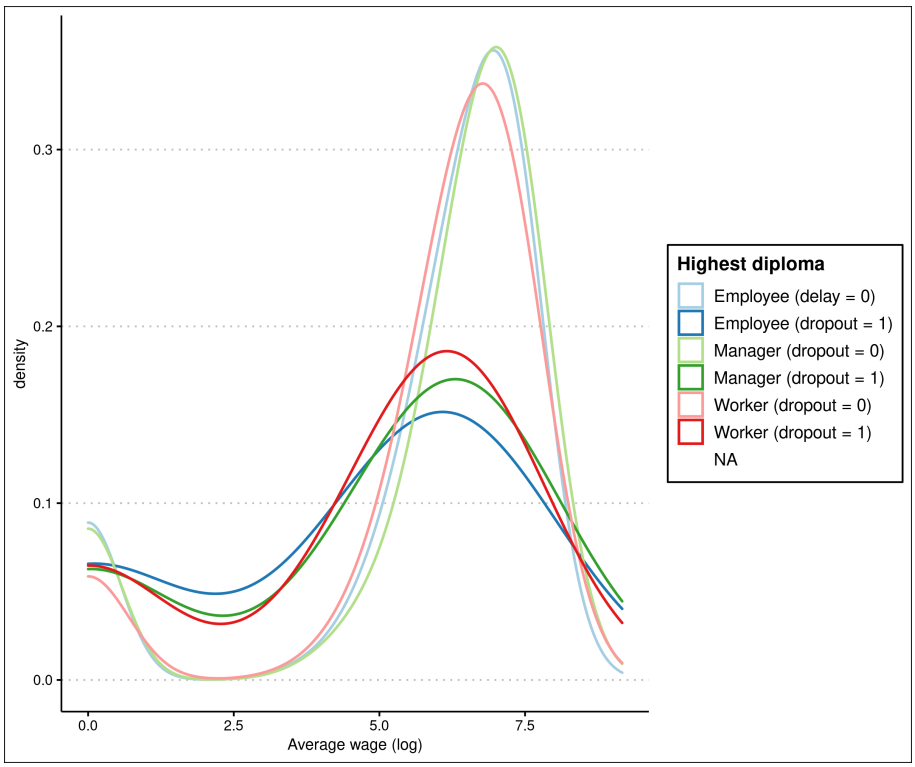


Figure 10: Average wage distribution conditional on the mother's occupation

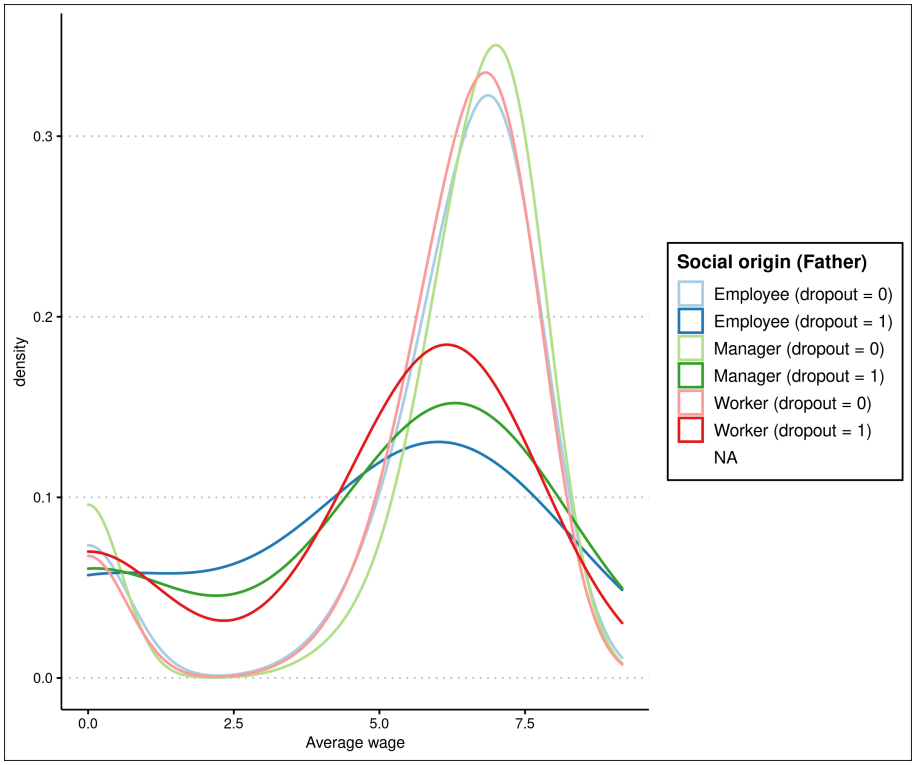


Figure 11: Average wage distribution conditional on the father's occupation