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

Gaspard Tissandier - Université Paris 1 Panthéon Sorbonne *

20/05/2022

Abstract:

For higher education system with different tracks in higher education, the question of how to target dropout policy is fundamental. In France, one of the two main tracks (the university) concentrate most of the focus and resources compared to the other (the Technical track). This paper questions this setting by estimating heterogeneous effect of dropout on labor market outcomes conditional on the followed tracks and other socio-demographic characteristics of former students. These effects are measured on the time in employment and average wages at an individual's entry into the labor market, and are obtained by using a highly dimensional identification methods to account for the heterogeneity of individuals' characteristics in each track. I use the Generalized Random Forest algorithm with the distance to the closest higher education institution at 6th grade as an instrument to estimate Local Average Treatment Effect of dropping out. I find that using 2SLS lead to underestimate the overall effect of dropping by 9 percentage point for the rate of employment, and by 10 percentage point for the average wage. I also find that the Technical track dropouts are actually more penalized than university dropout on the average wage, but not on the time in employment. Finally, using the higher education track and socio-economic status when considering dropout effects can lead to better targeted dropout policies with widely available information for the higher education institutions.

JEL code : J01, J24, I2, I24

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

*I am thankful to Pierre Kopp, Marc Arthur Diaye, Robert Gary-Bobo and Carmen Aina for their helpful reading and remarks. I also thank the participants of the JMA 2022, IWAEE 2022 and Ifo Dresden WLE 2022 for their comments and useful remarks

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), especially when the dropout happens at the beginning of the higher education period. 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 through 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].

The French higher education system proposes two main tracks after the high school diploma: the technical path, from which students obtain a BTS or a DUT degree, is labeled STS/IUT, and the general path, from which students obtain a Licence degree, mostly done at the University[2]. The first one is dedicated to training students for technical jobs, and offers mostly two years degrees, while the University path proposes three years degrees, mostly general, and aims mainly at bringing students to the Master's level. As the 2007 national policy against dropouts shows, most of the efforts and funds are concentrated on the University dropouts, as a large part of the Licence students exit this degree before graduating. (Morlaix & Perret, 2013). Moreover, the STS/IUT degrees are historically known to have less difficulty to integrates the labor market in case of dropout, mainly because of the technical nature of the jobs the students are trained for. However, recent studies show that since 2007, this statistical fact does not seem to hold anymore, and so policies regarding dropout must be re-calibrated to account for the actual structure of the effect of dropping out conditional on the degree of the students and other socio-demographic characteristics[3].

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. It has been pointed out recurrently that dropping out is a complex and multifaceted phenomenon, which depends on dense interactions between individuals' characteristics (Ménard, 2018). Knowing that the social composition of the students in STS or University or even the geographical situation of the establishment varies greatly between these two main paths, simply assessing the effect of dropping out on labor market outcomes by only considering single-dimensional effects, as the actual policy does, will lead to misestimating its

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

[2] BTS stands for Brevet de technicien supérieur, DUT for Diplôme universitaire de technologie and the Licence is the equivalent of a Bachelor

[3] Merlin Fanette, Le « décrochage » en STS : l'autre échec dans l'enseignement supérieur, Céreq Bref, n° 366, 2018, 4 p.

real consequences. Acknowledging that the effect of dropping out is not the same depending on the students is crucial to understand the educational choices of students, and designing better policies against student dropout, especially concerning the first years in the higher education system.

By studying the heterogeneous effect of dropping out on labor market outcomes, this paper asks if the actual design of the french policy to fight against higher education dropout is right in targeting mostly University students. I study the effect of dropping from the main two first degrees in french higher education and then explore the heterogeneity of these effects conditional on the diploma and socio-economic status of the parents to propose a better targeting of dropout policies.

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

As highlighted by Schnepf (2014), the literature about labor market performances of dropouts is scarce but indicates an overall negative effect. 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 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) is highly conditional on the number of years of schooling. In France, the study from Brodaty et al., 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".

Except for the overall earnings, dropping out can have a negative effect on many labor market outcomes as the opportunities or the rate of employment (Flores-Lagunes and Light, 2007, Matkovic and Kogan, 2012 ,Reisel, 2013, Schnepf, 2014). The negative effect of dropping out can be heterogeneous conditional on the motivation of the dropout (Bjerk, 2012) or even on the diploma or the age of the individual (Brodaty et al., 2008, Navarro et al., 2016, Scholten and Tieben, 2017). It has been shown that the social origin and individual characteristics of students are also 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.

The main issue in estimating the effect of dropping out or having a delay in graduation is the endogeneity of the event with the underlying ability of the student: following Spence (1973), dropping out sends a negative signal to the labor market about the ability of the individual, where ability is defined as an underlying variable reflecting the capacity of a worker to perform well in a task, or a set of tasks. Propensity score matching can be 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 et al. (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) method, developed by Athey et al., 2019, on a French database of 12000 young workers who finished their education in 2010. Their work records are surveyed from 2010 to 2013, which helps us to construct two indicators of the average wages and the 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). While being indicative of the potential treatment effect, the CATE is not an unbiased estimator of the actual treatment effect. The analysis relies on the Average (Conditional) Local Average Treatment Effect, estimated on multiple sub-samples to understand the heterogeneity of dropping out on labor market outcomes. The A(C)LATE is computed per quartiles of CATE effects, per diplomas, per SES and parents' diploma, and the interaction of diploma and social origin.

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 in 6th grade as an instrument. Paired with a vector of controls to estimate the predicted probabilities of dropout and using the three steps methods proposed in (Adams et al., 2009), 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 et al., 2008), and used as a polynomial

function.

I succeed to find heterogeneity of the effect of dropping out on the rate of employment and the average wage on the overall distribution. After splitting the overall sample into quartiles according to individual Condition Average Treatment Effect, I found that the lower quartile has an effect of -32% in the time in employment, while the higher quartile has an effect of -22%. The difference is wider for the average wage, as the most penalized quartile has an effect of -70% while the less penalized quartile doesn't exhibit any significant effect.

I find that, while STS/IUT dropouts are less penalized in terms of time in employment, they are far more penalized regarding their average wage. STS/IUT dropouts have a penalty of around -24% in time in employment, while it is -27% for university dropouts. However, technical track dropouts have an effect of -35% on the average wage, while University dropouts don't have a significant effect after around three years on the labor market.

Finally, considering the socio-economic of the parents within each degree reveals another layer of heterogeneity, especially for University dropouts. Students from low SES status who drop out from STS/IUT track have a negative effect of -38% on their average wage over three years, while these same students don't exhibit any significant effect when they drop out from the university track. Students from high SES backgrounds are more penalized regarding the time in employment when they drop out from the University than from STS/IUT degree, while the effect for STS/IUT dropouts is almost homogeneous conditional on the SES status.

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. These results will help to understand better the path of higher education dropouts and to design policies 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) [4]. 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

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

education to Ph.D. graduates. This represents a data set of 21829 individuals.

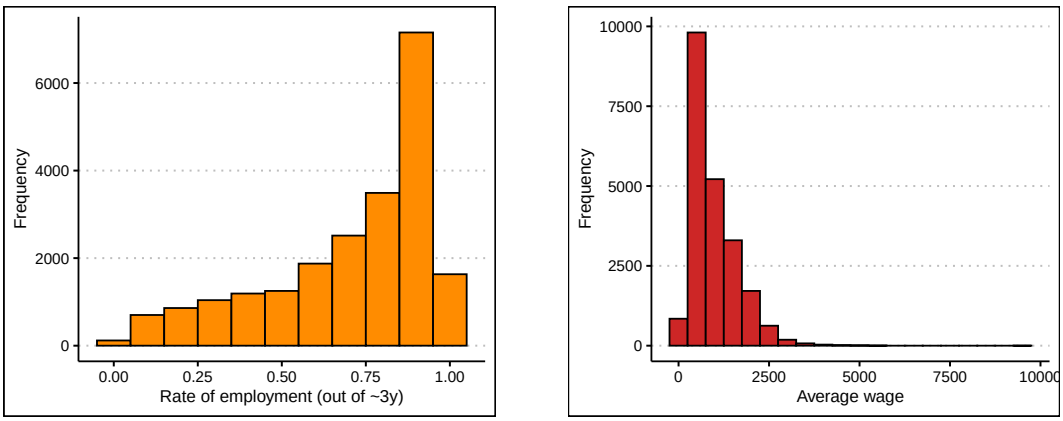
The chosen methodology relies, among others, on propensity score for estimating the treatment effect of dropping out. However, due the propensity score distribution with the chosen variable, I have to discard 42 % of the sample, with the final dataset counting around 12600 observations (see section 4.1). Since the first stage is implemented on the initial dataset (before the discard step), the descriptive statistics will be presented on the 21829 individuals included in the database

I create indicators variables for dropout, the number of months worked as a rate, 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 4923 individuals who dropped out, and 16906 who didn't (23% 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, as the duration in months. This setting allows us to create two variables in order to test our hypothesis.

The first outcome variable, *Rate of employment*, consists in the number of months worked over the period spent on the labor market. I use this definition since not every degree finish at the same time of the year, or student who drop out spend more time on the labor market. The second outcome variable, *Average Wages*, is an average of the wages on the whole labor market period observer. Then, if an individual works 12 months with a salary of 1200€ while spending 36 months on the labor market, *Average Wages* will be equal to 400€.

In order to work on percentage differential between individuals, and not percentage point or monetary difference, I use the logarithm of the rate of employment and average wage as dependent variables. The distribution of *roe* and *aw* are presented in figures 1.



(a) Time in Employment (b) Average wages

Figure 1: Distribution of dependent variables

The following variables are kept : highest diploma tried on 6 levels, if the individuals has done a foreign study travel or an internship during her higher education, the 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 and diplomas of both parents, and information about past education such as the discretized grade of the high school diploma, the type of high school diploma (general, technical or professional). The descriptive statistics are presented in table 10. For commodity reasons, the social origin is presented only for the highest among the both parents.

	Dropout rate	Frequency	Percentage
Gender			
Male	27.3 %	10092	46.2%
Female	18.5 %	11737	53.8%
Highest diploma tried			
Bac +2 (STS/IUT)	32.3%	10095	46.2%
Bac +3 (university)	21.4%	3262	14.9%
Bac +4 (university)	51.3%	943	4.3%
Bac +5 (university/Grande Ecole)	4.7%	5051	23.1%
PhD	9.9%	2478	11.4%
Parents' highest social category			
Disadvantaged	26.1%	3132	14.3%
Intermediate	25.3%	5718	26.2%
Advantaged	23.2%	4277	19.6%
Highly Advantaged	19.1%	8702	39.9%
Parents' highest diploma			
No diploma	28.2 %	3819	17.5%
Bac or below	25.4%	8119	37.2%
Short degree	19.8%	5915	27.1%
Long degree	15.3%	3976	18.2%
Other			
Foreign trip (= yes)	8.8%	4513	20.7%
Internship (= yes)	15.0%	14717	67.4%

Table 1: Summary statistics by dropout status

While females represent around 54% (and thus the majority) of the sample, they also drop

out less than male, with a dropout rate of 18.5%. It is a well known fact that women tend to undertake more often higher education, and we can observe that female also tend to drop out less in secondary education.

Bac (or Baccalauréat) corresponds to the High School Diploma, and is the reference for the time spent in higher education. The time needed to acquire the diploma is counted as "+y" : Bac +2 corresponds to two years of study after the HSD, and correspond here to vocational degree (labeled STS/IUT), which lead to a precise field, and are considered as "professional degrees". The Bac +3, obtained at the university, are general diploma which lead to a broad array of jobs, and are organized around field (such as STEM, law, economics, management). STS/IUT students show a dropout rate of 32% while 21% of Bac +3 students are concerned by dropout. The high dropout level for Bac +4 can be explained by the fact that most student which start a Master's usually undertake the full program, in two years, and not only the first one. The dropout rate in Master 2 is very low, as for most higher education path, it is the last year of studying. Finally, the PhD students present a dropout rate of 11.4%, which is quite high for the longest degree possible.

Concerning parents' occupation, the levels are defined using the type of occupation. Disadvantaged social category corresponds to factory worker and unemployed individuals. The intermediate category gathers employee and farmer, the advantaged category gathers intermediary profession, craftsman and independent while the highly advantaged gathers CEO, managers and executives. The parents' diploma are self explanatory, except for long degree which gathers individuals with 5 years or more of higher education. I use the maximum of these variables among the both parents in order to account for the global family environment, and not only the father's or the mother's background. The dropout rate is decreasing with the increase of the parents' highest social category or highest diploma. These results are fully in line with the literature documenting the heterogeneity of the dropout rate among different social origin.

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[5]. 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.

[5]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/

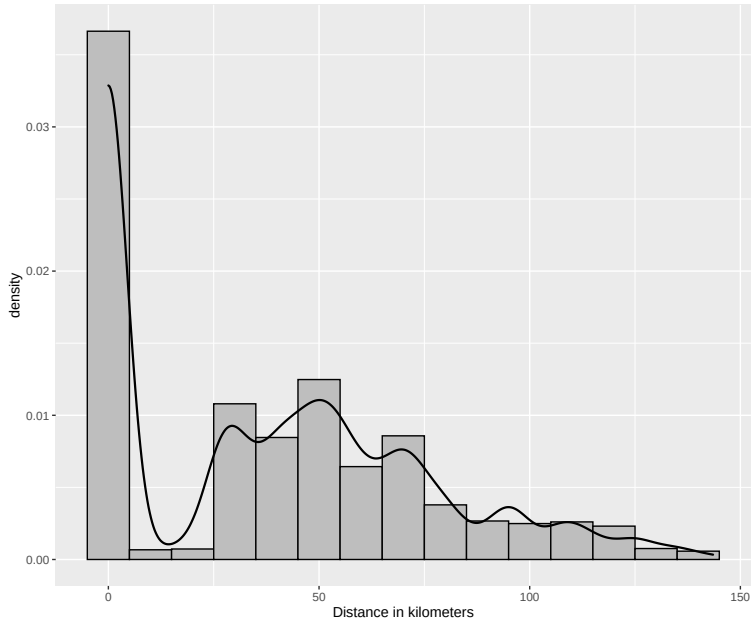


Figure 2: Distribution of the distance from 6th grade home to the closest university

While widely used as an instrument for educational attainment, the distance to the closest university can also be used as an instrument for dropout or delay in graduation. The distance to the closest university affects the dropout probability in two main ways. The considered distance captures either the cost of education (or the effort produced to acquire education), but also a part of the sunk costs in case of dropout. For students who have a university in their surrounding ($distance = 0$), the cost are lower that for those who have to commute to the university, reducing the sunk cost in case of dropout, and then increasing the probability of dropping out. For students who either have to commute to the university, or to live in another city, the potential effects of distance are plurals. The distance is increasing the cost of acquiring degrees, thus increasing the probability of dropping out. However, for students having to live in another city, the sunk cost of housing and transportation will acts negatively on the dropout probability. Thus, I will include a polynomial characterization of the distance to the closest university in order to predict the probability of dropping 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 et al., 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 et al., 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 et al. (2009), Athey and Imbens (2016), Wager and Athey (2018) or Athey et al. (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 et al. (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 6 levels, if the student made internship or international travel, the higher education institution 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 highest professional occupation and diploma of both parents, and the gender.

3.2 Orthogonalization

The Generalized Random Forest rely on an orthogonalization step which regress out the effect of the covariates X on Z, W and Y. The objective of this step is to obtain accurate treatment effect estimation, and to increase the efficiency of the learning phase of the forest. By regress-

ing out the effect of X, the forest is trained on an dependent variable vector $\tilde{Y}$ which doesn't depend on X, thus concentrating the learning on the heterogeneity actually depending on W, and not on X. The same procedure is applied to Z by regressing it on X.

To do so, the conditional marginal expectations of Y, W and Z are computed and used to obtain the conditionally centered outcomes :

$$\tilde{Y}_i = Y_i - \hat{y}^{(-1)} \quad \text{with} \quad \hat{y}^{(-1)} = \mathbb{E}[Y_i|X = x] \tag{3}$$
$$\tilde{W}_i = W_i - \hat{w}^{(-1)} \quad \text{with} \quad \hat{w}^{(-1)} = \mathbb{E}[W_i|X = x] \tag{4}$$
$$\tilde{Z}_i = Z_i - \hat{z}^{(-1)} \quad \text{with} \quad \hat{z}^{(-1)} = \mathbb{E}[Z_i|X = x] \tag{5}$$
$$\tag{6}$$

The forest is training using the set of transformed outcomes $(\tilde{Y}_i, \tilde{W}_i)$ (also called the centered outcomes). This step also helps to reduce the training time dedicated to estimating the propensity of treatment conditional on X, since it is already sorted out with this step.

The structure of transformed outcomes conditional on certain covariates are presented in section **??**.

3.3 The instrumental variable setting

As explained before, we cannot consider students who drop out as randomly selected, even conditionally on coviates. Thus, we need to rely an instrumental variable setting to identify the effect of dropping out on labor market outcomes. As precised in the section 2, the used instrumental variable is the distance to the closest university at 6th grade. I don't use the distance between the high school and the closest university, since some students move from their initial high school to a better one, usually based on performance or merit, thus inducing endogeneity between the distance and the labor market outcomes (Brodaty et al., 2008).

In order to obtain stable and consistent estimate of the effects of dropout on labor market outcomes, I adapt the methodology proposed by Adams et al., 2009 : a four step instrumental variable process. The steps are :

1. Estimate $Pr(w = 1|x_c, z) = \phi(\gamma_0 + \gamma z + \theta x_c)$, with ϕ a cumulative distribution function (here the logistic cumulative distribution function)
2. Compute the fitted probability $\hat{w_1}$ using the precedent step estimation
3. Estimate $w_i = \phi(\eta + \eta\hat{w_1} + \theta x) + \epsilon_i$, this model is estimated with a linear model
4. Compute the fitted probabilities $\hat{w_2}$ using the precedent step estimation, and use these fitted probabilities as the instrumental variable

With z the vector of instrumental variable, x_c the first step vector of control, x the vector of covariates used to build the forest. The instrumental variables include a dichotomous variable indicating if an individual has an university in her area (distance to closest university = 0), and a polynomial expression of the distance : $z = (dist_0, dist, dist^2)$. The vector of control x_c consists in fixed effects for all the french regions.

In the presence of multiple instrumental variable and control, this methodology has advantages compared to the pseudo-IV methods : it can smoothly include many instrumental variable, and doesn't need the first step to be correctly specified. The only requisite for the first step is for the instrumental variables to be correlated with the dropout indicator, and for $\hat{w_1}$ to keep a strong correlation with this indicator in step 3. The results of step 1 and 3 are presented in section 4.2.

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 there exists instruments z satisfying all the IV assumptions, the dropout effect can be estimated as :

$$\tau(x) = \frac{Cov[Y_i, Z_i \mid X_i = x]}{Cov[W_i, Z_i \mid X_i = x]} \tag{7}$$

In this setting, the IV can be implemented with a binary or continuous instrumental variable. However, since the A(C)LATE is used to estimate treatment effect on group of former students, a dichotomous instrumental variable is needed. The identification strategy rely on the estimation of doubly robust scores, as proposed by Athey and Wager (2020), and average them over subsamples to get the unbiased A(C)LATE. As precised in Athey and Wager (2020), we need a binary instrument to compute the doubly robust scores. I proceed to a dichotomization of the fitted probability $\hat{w_2}$:

$$\begin{cases} \tilde{w_2} = 1 \quad \text{if} \quad \hat{w_2} > p(\alpha) \\ \tilde{w_2} = 0 \quad \text{if} \quad \hat{w_2} \leq p(\alpha) \end{cases}$$

With $p(\alpha)$ the value corresponding to the α^{th} percentile and $\hat{w_2}$ the fitted probability of dropping out computed at step 3. 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 $(\alpha) = 0.80$. After this step, the dichotomized instrument is equal to 1 for 4246 observations, and equal to 0 for 8324 observations. The distribution of the instrument

is not exactly 20% positive because a part of the sample is dropped due to too low propensity score (see section 4.1).

It is possible that certain students with high effect of dropping out are below the $(\alpha) = 0.80$ threshold, thus potentially overturning the results of this estimation. Thus, I perform the same analysis with $(\alpha) = 0.60$ as a robustness check. The results can be found in section 6.5.

3.4 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., 2022 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 :

$$\Gamma = \tau(X) + g(X, Z)\left(Y - \mathbb{E}[Y|X] - (W - \mathbb{P}[W = 1|X])\tau(X)\right) \tag{8}$$

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

$$g(X, Z) = \frac{1}{\Delta(X)} \frac{Z - \mathbb{P}[Z = 1|X]}{\mathbb{P}[Z = 1|X](1 - \mathbb{P}[Z = 1|X])} \tag{9}$$

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 **beyond_late_aronow_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 Preliminary steps : overlapping and orthogonalization of the outcomes

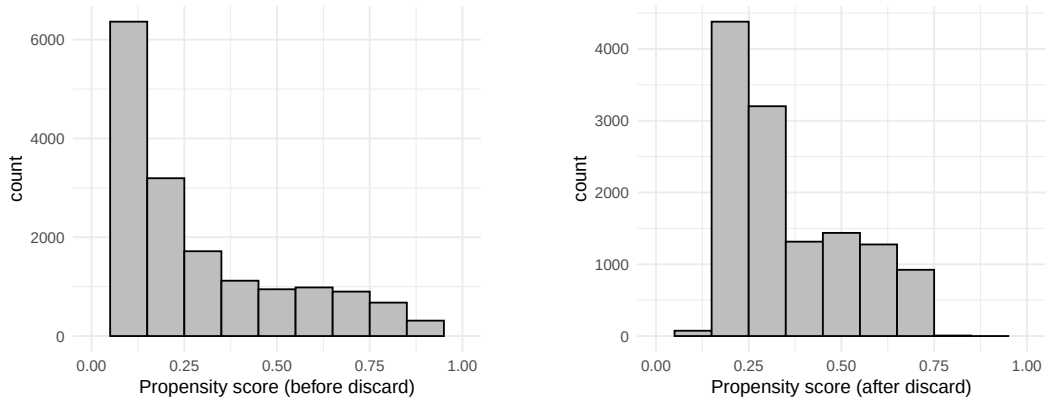
Before estimating the Generalized Random Forest, we need to address the question of overlapping and show the preliminary step of the GRF defined as orthogonalization (or residualisation) defined in section 3.2.

Overlapping

The Conditional Average Treatment Effect proposed by Athey et al., 2019 relies (among others) on propensity score to estimate the treatment effect. The propensity score, defined as $p_s = \mathbb{E}[W_i = 1|X_i = x]$, is estimated using a regression forest in order to account for the highly dimensional predictable power of X on W. However, in the presence of propensity scores close to 0 or 1, this lead to unstable estimator of the CATE, and then to the A(C)LATE.

Following Crump et al., 2009, we discard observations with a propensity score such that $p_s \notin [0.1; 0.9]$. As seen in the data part, certain covariates' levels exhibit very low dropout rate, leading to very low propensity score, or the inverse. Fortunately, these levels (mostly Bac +4 and Bac +5) do not represent the core target of this work.

This discard phase leads to drop 9215 observations, and let 12614 observations in the dataset. The result of this step is presented in figure 3.



(a) Before discard (b) After discard

Figure 3: Propensity score $\mathbb{E}[W_i = 1|X_i = x]$

This step leads to drop 9215 observations, including 3030 Bac +2 and 201 Bac +3. While dropping more Bac +2 than Bac +3, the final database include a reasonable number of observations for each levels, and should not biased the estimation of the LATE for each type of degrees. The final database counts 7065 Bac +2 and 3061 BAc +3.

Orthogonalization

As described in section 3.2, the outcomes Z, W and Y are centered (or orthogonalized) by subtracting the expected outcomes conditional on X : $\mathbb{E}[Z_i|X], \mathbb{E}[W_i|X = x], \mathbb{E}[Y_i|X = x]$. This preliminary step, except from ensuring the efficiency of the LATE estimators, also focus the splitting part on the treatment effect conditional on X, and not on splits related to the direct effect of X on Y, W or Z. The orthogonalized outcomes are obtains using :

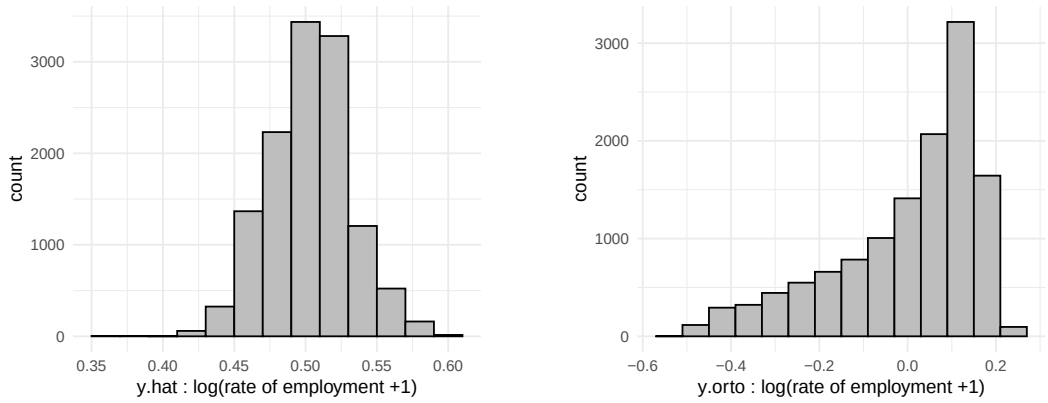
$$\tilde{Y}_i = Y_i - \hat{y}^{(-1)} \quad \text{with} \quad \hat{y}^{(-1)} = \mathbb{E}[Y_i|X = x] \tag{10}$$

$$\tilde{W}_i = W_i - \hat{w}^{(-1)} \quad \text{with} \quad \hat{w}^{(-1)} = \mathbb{E}[W_i|X = x] \tag{11}$$

$$\tilde{Z}_i = Z_i - \hat{z}^{(-1)} \quad \text{with} \quad \hat{z}^{(-1)} = \mathbb{E}[Z_i|X = x] \tag{12}$$

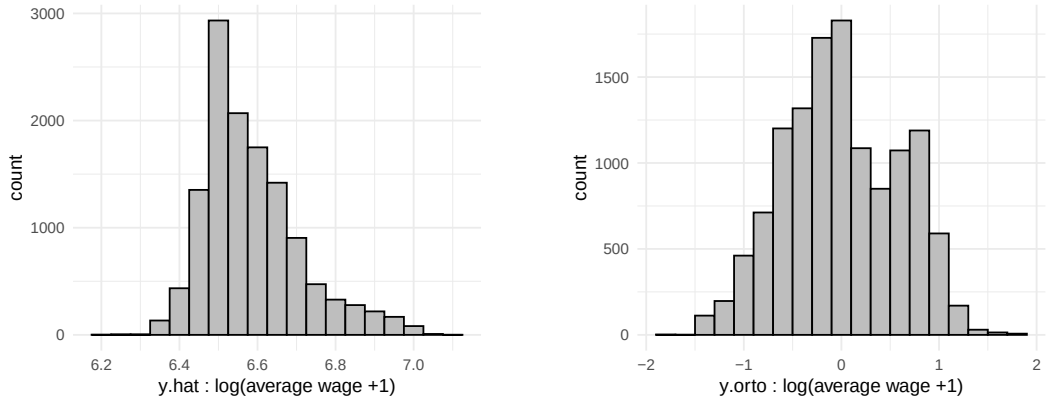
$$\tag{13}$$

In this section are presented the predicted and orthogonalized outcome Y for the rate of employment and the average wage, and the orthogonalized outcomes for the average wage conditional on two levels of the maximum social category of the parents.



(a) $\hat{Y}$ (b) Y_{ortho}

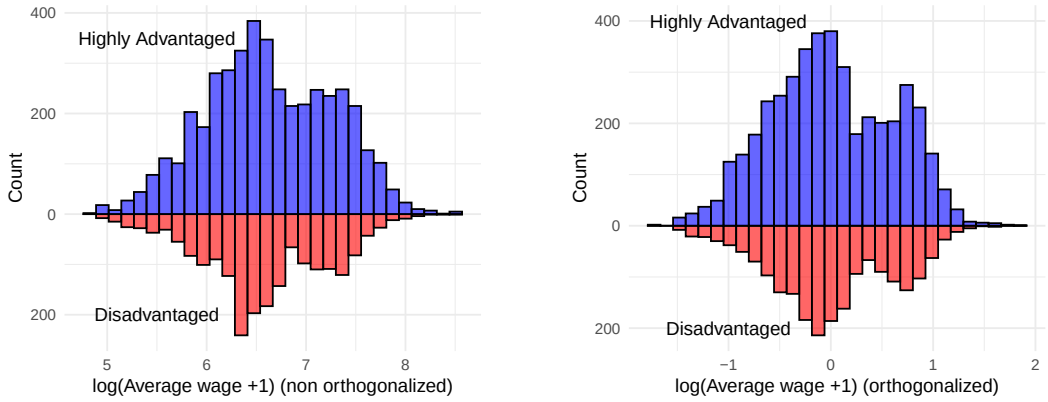
Figure 4: Predicted and orthogonalized outcome Y : rate of employment



(a) $\hat{Y}$ (b) Y_{ortho}

Figure 5: Predicted and orthogonalized outcome Y : average wage

We can observe that the orthogonalization center the Y distribution on 0, but keep the overall shape of the distribution.



(a) Non orthogonalized (b) Orthogonalized

Figure 6: Orthogonalized outcome Y conditional on the social category of the parents : average wage

	Min	Q1	Median	Mean	Q3	Max
$[Y\|X = \text{Disadvantaged}]$	4.95	6.19	6.53	6.57	7.06	8.46
$[Y\|X = \text{Highly advantaged}]$	4.86	6.18	6.58	6.63	7.15	8.54
$[Y_{ortho}\|X = \text{Disadvantaged}]$	-1.49	-0.40	-0.05	-0.00	0.48	1.62
$[Y_{ortho}\|X = \text{Highly advantaged}]$	-1.78	-0.43	-0.03	0.01	0.49	1.80

Table 2: Distribution of the initial and orthogonalized outcome conditional on the social origin : average wage

In figure 6 and table 2, we can observe that the orthogonalization step helps to smooth the distribution and to close the median and mean difference between disadvantaged and highly advantaged background student. The step generate the same effect for every variables included in X, and the interaction generated by the regression forest used to estimate $\mathbb{E}[Y_i|X = x]$. Thus, the built forest will be around the treatment effect of dropping out conditional on the vector of covariables X.

4.2 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. In this section, I present the results from the step 1 and 3 described in section 3.3.

The results table 3 indicate a strong correlation of the three part of the instrumental variable with the dropout indicators. When the regional fixed effects are included, living in an area hosting a university is positively correlated with dropping out, while the distance to the closest

	Dependent variable:	
	Dropout	
	(1)	(2)
Intercept	-1.447^{***}	-1.496^{***}
	(0.108)	(0.114)
$dist_0$	0.179	0.308^{**}
	(0.112)	(0.125)
$dist$	1.049^{***}	1.281^{***}
	(0.339)	(0.380)
$dist^2$	-0.009^{***}	-0.009^{***}
	(0.002)	(0.003)
Regional FE	No	Yes
Observations	21,829	21,829
Log Likelihood	-11,638.390	-11,520.570
Akaike Inf. Crit.	23,284.790	23,091.150
Note:	*p<0.1; **p<0.05; ***p<0.01	

For the readability of the parameters, the distance and distance squared have been divided by 100. The regional fixed effects include 22 indicators

Table 3: First stage LOGIT regression

university follow a concave parabolic curve. This indicate a threshold of the distance effect : under a certain distance, living far from a university increase the likelihood of dropping out, while after this threshold, the distance tends to decrease the probability of dropping out. Since regional fixed effect are included, and every region have a university, this definition of distance measure the effect within each region, cleaning out the potential heterogeneity proper to each region.

As described in the Data section, the minimum distance to the closest higher education establishment is negatively correlated with the probability of dropping out. The fitted probabilities of the stage 2 and 4 are presented in figure 7.

The stage 3 results are presented in table 4. When all the controls are included, the first stage fitted probabilities are still very significant and the t-stat : $t_{stat} = 7.66 \Leftrightarrow F_{stat} = 59$ is high enough to rule out the weak instrument bias. The controls included are those used to build the forest : the gender, the highest diploma tried (6 levels), the maximum socio-economic status of the parents, the maximum parents' diploma, the type of HSD and the distinctions at the final exam (proxy of the grade), the region of the university the individual went at, if she

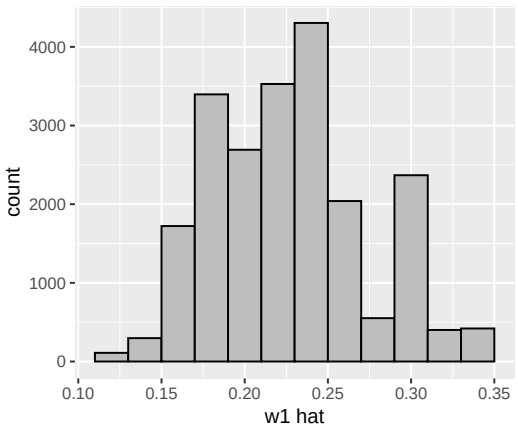
	Dependent variable:	
	dropout	dropout
	(1)	(2)
Constant	−0.0002	0.312***
	(0.014)	(0.022)
$\hat{w_1}$	1.001***	0.595***
	(0.061)	(0.076)
Controls	No	Yes
Observations	21,829	21,829
R^2	0.012	0.183
Adjusted R^2	0.012	0.181
Residual Std. Error	0.415 (df = 21827)	0.378 (df = 21781)
F Statistic	269.573*** (df = 1; 21827)	103.482*** (df = 47; 21781)
Note:		*p<0.1; **p<0.05; ***p<0.01

The controls include : Gender, highest diploma tried (6 levels), Min/Max SES, Min/Max parents' diploma, type of HSD, HSD grade, region of the university, went to study abroad, did an internship

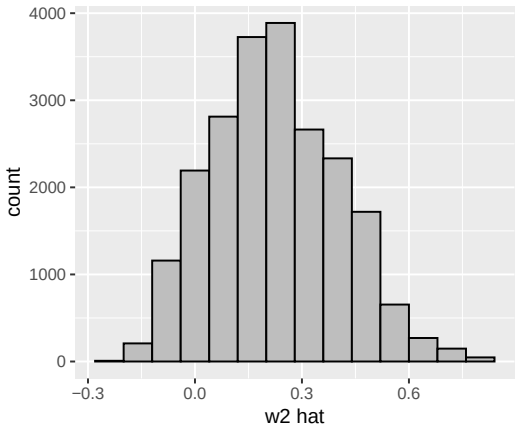
Table 4: First stage linear regression

went to study abroad and if she did an internship.

The final step is to dichotomize the obtain instrument $\hat{w_2}$: individuals in the top 20% of $\hat{w_2}$ are assigned a $Z_i = 1$, the others 0. I also perform the same analysis with $Z_i = 1$ for the top 40% of the distribution; the results can be found in section 6.5.



(a) Stage 2 : fitted probability of dropping out (w1 hat)



(b) Stage 4 : fitted probability of dropping out (w2 hat)

Figure 7: Distribution of fitted probabilities

4.3 Instrumental variable regression

In order to understand how the highly dimensional setting used by the GRF is useful to obtain an accurate estimation of the effect of dropping out, I present the estimation of the local average treatment effect using a two stage least square approach. The TSLS estimation is preceded by the first stage treatment fitted probabilities estimation described in section 3.3 and 4.2. This methodology is also described in Adams et al. (2009).

An emphasis should be put on why the LATE obtained with TSLS is different from the LATE obtained with GRF. While TSLS estimate the effect of dropping out everything else equal, the GRF algorithm average all the individuals CATE for $X = x$ obtained by estimating the treatment effect in each terminal nodes of the trees, thus using the highly dimensional setting proposed by the tree structure to estimate a conditional treatment effect. Then, the $LATE_{GRF}$ can differs from $LATE_{TSLS}$ mainly because of the tree structure of the CATE estimation. The results of both TSLS estimation are presented in table 5, the full tables can be found in section 6.4. The included controls are : the highest diploma tried (and eventually obtained) on 6 levels, the maximum SES of the parents, the maximum diploma of the parents, the discretized grade obtained at the high school diploma, if the students did an internship, if she did an foreign study trip, and a indicators of the higher education institution region on 22 levels.

	Dependent variable:	
	log(Rate of employment + 1)	log(Average wage + 1)
	(1)	(2)
Constant	0.566***	6.816***
	(0.011)	(0.039)
dropout	−0.183***	−0.357***
	(0.025)	(0.086)
Weak instruments	229.17***	239.899***
Wu-Hausman	22.73***	4.612**
Observations	12,614	12,614
R^2	-0.001	0.100
Adjusted R^2	-0.005	0.097
Residual Std. Error (df = 12572)	0.171	0.594
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 5: TSLS estimation of dropout LATE

The $LATE_{TSLS}$ of dropping out on the rate of employment is significantly different from 0 and show an effect of $e^{-0.183} - 1 = -16.7\%$. The weak instrument and Wu-Hausman test are both significantly different from 0, assuring that we can safely rule out the weak instrument

bias eventuality.

The $LATE_{TSLS}$ of dropping out on the average wage exhibits an effect of $e^{-0.357}-1 = -30\%$ of dropping out on the average wage, and is significantly different from 0 at the 1% level. The weak instrument and Wu-Hausman test are again significantly different from 0. Those estimated effects are in line with the aforementioned literature, as their magnitude.

4.4 Assessing the dropout effect 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 dropping out on labor market outcomes. 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 the doubly unbiased estimation of the A(C)LATE and its statistical properties to test for the presence of actual heterogeneity in the distribution.

In section 4.4.1, the LATE obtained with GRF for the whole sample is compared to the LATE obtain with TSLS in section 4.3 in order to test if the highly dimensional structure used by GRF helps to catch previously uncovered effect. Then, the sample is split around the CATE median and quartiles, and the A(C)LATE are computed on each subsamples. I compute the t-statistic for the difference between each subsamples to assess if the effect heterogeneity between each part is generated by the data structure or by noise.

In section 4.4.2, the A(C)LATE per diploma, socio-economic status and the interaction of the both are computed to assess which subsamples is the most penalized when dropping out. The heterogeneous effect per diploma answer the question of the legitimacy of targeting mostly the university in dropout policy, while the interaction of both covariates helps us to assess which population is an optimal target for these policies.

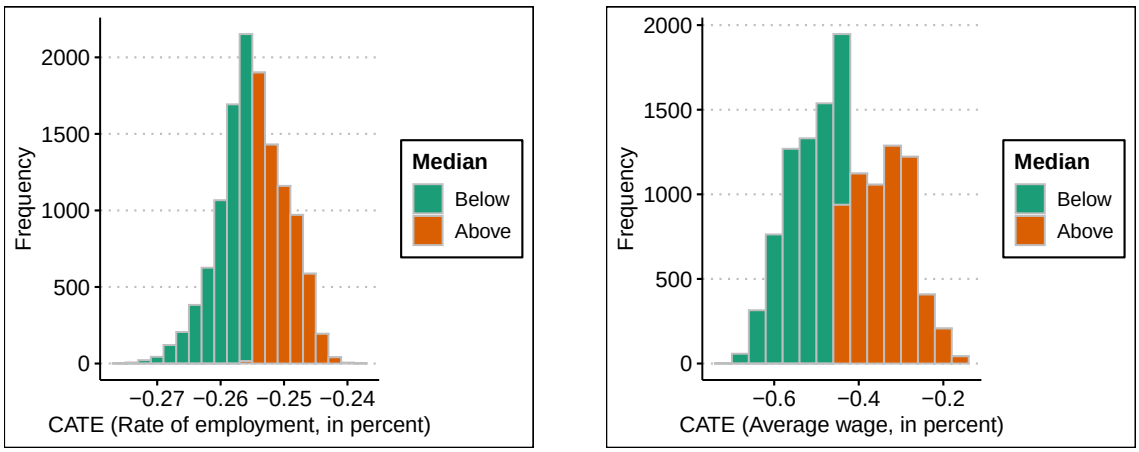
4.4.1 LATE estimation with GRF

The estimated CATE follow the distributions showed in figure 8 and table 6. Since the variation measured by the CATE or the LATE are not small, the log transformation doesn't measure correctly the variation in percentage, thus I apply the $e^{\text{CATE or LATE}} - 1$ transformation for every results presented in this section.

	Min	Q1	Median	Mean	Q3	Max
Rate of Employment	-28%	-26%	-26%	-25%	-25%	-24%
Average Wages	-72%	-52%	-44%	-43%	-34%	-15%

Table 6: Distribution of CATE of dropout on the Rate of employment and the Average wage

The Conditional Average Treatment Effect of dropping out on the rate of employment ranges from -28% to -24% of the time spent on the labor market at the end of the education



(a) Time in Employment (b) log(Average wages)

Figure 8: Conditional Average Treatment Effect of dropout

period. The individual estimated effect of dropping out on the time in employment is concentrated with only a 4 percentage points difference between the strongest and the smallest estimated CATE. The average (and the median) of the CATE is equal to -26%, which for a 36 months period represent a reduction of 9 months of the time in employment for dropouts.

The CATE on the average wage range from -72% to -15%, with a median at -44% and mean at -42 % on monthly wage. The CATE distribution exhibits a strong heterogeneity in individual conditional average treatment effect, far greater than the distribution of CATE for the rate of employment. With an average wage of 877€ per month, a average decrease of 42% is equivalent to a reduction of 385€ per month.

As explained in section 3.4, even if the CATE distribution can be indicative of the dropout effect, we need to rely on the doubly robust score average (the A(C)LATE) to estimate the unbiased effect of dropping out on the indicators. First, I compare the estimated LATE obtained with TSLS and GRF and test for their difference. Then, as proposed by Athey and Wager, 2019, the individuals are split following their individual CATE around the median. I define $\tau_1 \leq Median(\tau_i)$ the group of individuals showing a CATE below the median of estimated CATE and $\tau_2 > Median(\tau_i)$ the group of individuals showing a CATE equal or above the median of estimated CATE. The τ_1 corresponds to a group showing a strong negative effect of dropout, while the $\hat{\tau}_2$ correspond to a group showing a less negative effect. I also present the A(C)LATE for the four quartile (Q_1, Q_2, Q_3, Q_4). A student test for the difference between the A(C)LATE of both groups is performed and the T-Statistic is given in the table. For convenience, the A(C)LATE will be designated as LATE in this section.

The A(C)LATE on each subgroups are presented in table 7 for A(C)LATE on time in employment and in table 8 for the A(C)LATE on average wages.

The first interesting result is that the GRF estimate a LATE of dropping out of -27% of the rate of employment on the whole sample, and that this effect is significantly different from

	$TSLS$	GRF_1	GRF_2	GRF_3	t-stat
Overall sample	-0.175***	-0.266***			-15.670
	(0.025)	(0.020)			
$\tau_1 \leq Median(\tau_i)$			-0.303***		-1.919
			(0.029)		
$\tau_2 > Median(\tau_i)$			-0.227***		
			(0.027)		
Q_1				-0.321***	-1.713
				(0.048)	
Q_2				-0.284***	-1.330
				(0.033)	
Q_3				-0.234***	-0.275
				(0.040)	
Q_4				-0.220***	-
				(0.033)	
Note:					*p<0.1; **p<0.05; ***p<0.01

The standard error is precised in parenthesis. The two first columns indicate the LATE obtain with TSLS and GRF, while the last column shows the t-statistics testing the difference between the both estimators. The third column shows the A(C)LATE for the 50% lower than the median and for the top 50%. The fourth column indicates the A(C)LATE per quartiles. For the effect per quartiles, the t-test is computed with respect to Q_4, with the smaller effect. $LATE \times 100$ gives the estimated effect in percentage.

Table 7: Estimation of dropout LATE (rate of employment)

0. By comparing the LATE obtained with TSLS and GRF, it appears that using a linear instrumental variable method to estimate the effect of dropping out on the rate of employment lead us to underestimate the effect by 9 percentage point. The difference between the TSLS and GRF estimators is significant at the 0.1% level, comforting the use of a highly dimensional method to estimate correctly the dropout effect.

The LATE on half sample is of -30% for the most penalized group, and of -23% for the less penalized group. Both LATE are significantly different from 0 but their difference is only significant at the 10% level, indicating a very small heterogeneity of the dropout effect on the rate of employment, which is in line with the CATE distribution presented in figure 8.

The LATE estimation per quartiles ranges from -32% for the lowest quartile to -22% for the highest quartile. While all these effects are significantly different from 0, only the difference between Q_1 and Q_4 is significant at the 10% level, reinforcing the fact that dropout doesn't show a strong significant effect on the rate of employment.

GRF estimate an overall LATE of -40% of dropping out on the average wage. This effect is significantly different from 0 and from the $LATE_{TSLS}$ at the 0.1% level. Relying on TSLS for this step results in an underestimation of the effect of dropping out on the average wage of

	$TSLS$	GRF_1	GRF_2	GRF_3	t-stat
Overall sample	-0.297***	-0.402***			-6.813
	(0.085)	(0.058)			
$\tau_1 \leq Median(\tau_i)$			-0.584***		-3.861
			(0.091)		
$\tau_2 > Median(\tau_i)$			-0.141**		
			(0.071)		
Q_1				-0.707***	-4.609
				(0.140)	
Q_2				-0.411***	-3.269
				(0.114)	
Q_3				-0.319***	-2.845
				(0.101)	
Q_4				-0.084	-
				(0.099)	
Note:			*p<0.1; **p<0.05; ***p<0.01		

The standard error is precised in parenthesis. The two first columns indicate the LATE obtain with TSLS and GRF, while the last column shows the t-statistics testing the difference between the both estimators. The third column shows the A(C)LATE for the 50% lower than the median and for the top 50%. The fourth column indicates the A(C)LATE per quartiles. For the effect per quartiles, the t-test is computed with respect to Q_4, with the smaller effect. $LATE \times 100$ gives the estimated effect in percentage.

Table 8: Estimation of dropout LATE (average wage)

10 percentage point.

The lowest 50% of the CATE distribution exhibits an effect of -58% while the top 50% shows an effect -14%. These two effect are strongly different with a t-statistic of -3.9, indicating a strong heterogeneity of the dropout effect on the average wage.

The LATE of each quartile go from -71% for Q_1 to a non significant effect for Q_4. Quartile Q_1 to Q_3 have a statistically different effect from Q_4 at the 1% level. This indicates that, while 25% of the sample has a decrease of their monthly wage of more than 50%, the top 25% of the sample doesn't show any significant effect, and thus are not penalized when dropping out. We can observe that Q_2 and Q_3 have an effect closer than from Q_1 and Q_4, indicating that while some part of the distribution show extreme effect, most of the distribution has an effect centered around the overall LATE estimated with GRF_1.

4.4.2 Dropout heterogeneity per diploma, SES and Dip x SES

The objective of the paper is to assess the heterogeneity of the effect dropping out, and then to explore which population is the most penalized when dropping out of higher education. We have seen that most of the effort is targeted toward the university dropouts, while the Technical section dropouts could be as penalized as the university's ones. Finally, we emitted

the hypothesis that targeting dropout policies with respect to a multidimensional understanding of student can be beneficial.

In this section, I present the LATE estimation of dropping out with respect to the highest diploma tried (or obtained) by the student, her socio-economic status and finally the interaction of the both. I also present the LATE for different characteristics such as the gender or academic achievement as if the student made a foreign study trip or an internship.

The diploma indicator is divided in 5 levels : Bac +2 (STS/IUT), Bac +3 (university), Bac +4, Bac +5 and PhD. The first two levels are those of interest for the analysis, with the Bac +3 level concentrating dropout policy resources. I present the results for the other levels to get a broad understanding of the dropout effect in the french higher education. The higher professional occupation among both parents (noted SES) is defined as follow : disadvantaged corresponds to factory worker and unemployed parents. The intermediate category gathers employee and farmer, the advantaged category gathers intermediary profession, craftsman and independent while the highly advantaged gathers CEO, managers and executives. For the interaction $Diploma \times SES$, the SES are grouped as low and high SES, with disadvantaged and intermediate in the low level, and the rest in the high level.

The main results are highlighted in table 15. We can observe that both subgroups (STS/IUT and university) have a negative and significant effect of dropping out on the rate of employment. However, the effect is slighter for the STS/IUT, sustaining the historical hypothesis of STS/IUT students being less penalized when they drop out. The STS dropout have a negative effect of -24% of time in employment, while the university dropouts have an effet of -27%. Knowing that these effects are computed on the total period observed, which is on average 3 years, these effects are strong and can influence considerably the path of young workers in the labor market.

Regarding the dropout effect on the average wage, STS/IUT dropouts exhibit a negative effect of -35% on the monthly average wage, while the university dropouts show a non significant effect. This results indicate that Bac +2 dropout are highly penalized by dropout, while Bac +3 dropout are only penalized on their time in employment, but not in terms of wage.

These results indicate that while university student may undergo more time unemployed in their entry on the labor market, they fill the gap with STS/IUT students, are those are actually highly penalized on their average wage, but university dropouts doesn't exhibits any significant effect of dropping out after three years. This setting can indicate that dropout policies actually implemented in France are potentially mistargeted, and STS/IUT students should be the focus of more intense dropout policies.

The other objective of the paper is to understand if it could be efficient for policy makers to rely on other variables than the higher education to design dropout policies. I present results regarding the social origin (SES) of the dropout and the interaction of the higher education degree and the SES in table 15

		Rate of employment	Average wage
Diploma	**Bac +2 (STS/IUT)**	**-0.237*****	**-0.349*****
		(0.022)	**(0.073)**
	Bac +3 (university)	**-0.272*****	**-0.172**
		(0.051)	(0.148)
	Bac +4	-0.249***	-0.397
		(0.073)	(0.292)
	Bac +5	-0.281***	-0.755***
		(0.051)	(0.198)
	PhD	-0.278	-0.853***
		(0.043)	(0.152)
SES	Disadvantaged	-0.287***	-0.452***
		(0.067)	(0.167)
	Intermediate	-0.219***	-0.360***
		(0.028)	(0.098)
	Advantaged	-0.255***	-0.318***
		(0.036)	(0.120)
	Highly advantaged	-0.259***	-0.481***
		(0.030)	(0.115)
DIP × SES	Bac +2 × Low SES	-0.236***	-0.378***
		(0.031)	(0.102)
	Bac +2 × High SES	-0.238***	-0.319***
		(0.030)	(0.105)
	Bac +3 × Low SES	-0.227**	0.162
		(0.097)	(0.206)
	Bac +3 × High SES	-0.302***	-0.344*
		(0.053)	(0.206)

Table 9: LATE for Diploma, SES and Diploma × SES (Rate of employment and Average wage)

The LATE of dropping out on the rate of employment of SES levels indicate that the most penalized dropouts are those from either disadvantaged or highly advantaged background. This result, while quite surprising, can be explained by the difference in constraint undergo by these students : individuals coming from highly advantaged SES have on the average more parental resources to choose their path after dropping out, thus allowing less time in employment. However, the strongest negative effect is for dropouts of disadvantaged background with an effect -29%.

Regarding the effect of dropping out on average wage conditional on the SES of the dropout, the structure of the LATE follows the same pattern as those from the rate of employment : the most penalized categories are disadvantaged and highly advantaged. The disadvantaged student can face a penalty of -45% on their average wage while those from highly advantaged background can undergo a penalty of -48%.

While it is easy to target dropout policy to certain degrees, it is more complicated to target student only based on their SES. However, it is clearly easier to target student from a certain SES inside a given degree, as the SES is most of the time known by the institution. Thus, I estimate the LATE of dropping out for Bac +2 and Bac +3, and for low and high SES.

The LATE for every levels are between -23% to -30%, the most penalized category being the Bac +3 $\times$ high SES. We can observe that the negative average effect for the whole category Bac +3 is mostly driven by this level, as the LATE for Bac +3 is -27%. The effect of dropping out for Bac +2 is sensibly similar between both SES levels. Regarding the effect on the average wage, the most penalized level is Bac +2 $\times$ low SES with an effect of -38%, followed by Bac +3 $\times$ high SES with an effect of -34% on the average wage. The difference between low and high SES for Bac +2 dropouts is of 6 percentage point, indicating that it is beneficial to consider the dropout as a multidimensional process, and especially targeting from lower SES in Bac +2 would generate higher social benefit.

5 Conclusion

In this paper, the objective was to answer if targeting most of the resources dedicated to reduce dropout on one of the main two tracks in the higher education system in France was legitimated by poorer performances on the labor market for those dropouts. The second objective was to state if using a multidimensional measure based on the diploma and the social origin can helps to create targeting categories to increase the efficiency of dropout policy. The effect of dropout was estimated on the rate of employment and the average wage, for a period of around three years following the exit of the education system.

To answer those questions, I applied the Generalized Random Forest algorithm in an instrumental variable setting to estimate individual Conditional Average Treatment Effects. Then, observations were grouped conditional either on the CATE or other characteristics such as the highest diploma tried or the SES of the parents, and the average conditional local average treatment effect was computed on each subgroups.

I answer this question in three steps. First, I considered if the dropout effect on labor market outcomes exhibited actual heterogeneity, and if this heterogeneity was statistically significant for different sub-samples (subs-sampling around the median and in quartiles). The distributions of the dropout effect is effectively heterogeneous for both considered labor market outcomes, thus justifying to test for the presence of heterogeneity conditional on more precise categories. Second, I tested if Bac +2 (STS/IUT) dropouts were actually less penalized in the long run, compared to Bac +3 university dropouts, which could justify the focus of dropout policy on the laters. I found that it was not the case, as Bac +2 dropouts were more penalized either on the rate of employment or the average wage, over a period of around three years on the labor market. Finally, I estimated the effect of dropping out conditional on the highest diploma tried and the SES of the parents, and found that including this second layer in the targeting of dropout policies can helps to generate more useful categories to target, instead of only basing the dropout policies on the degree.

Certain results, especially regarding the fact that individuals from advantaged background tend to be more penalized when dropping out from the university than students from lower SES origins, lead to consider the question of dropout as a chosen or undergone events, and potential strategic behavior of students when investing time and energy in their education. A great extension of this work can be an analysis of the effect of dropping out conditional on the chosen/undergone paradigm, for different type of students, and could helps to split the "dropout" category into plural definitions.

This paper opens the discussion about the necessity of considering the potential heterogeneity of educational event such as the dropout, the accumulation of delay or even re-orientation on educational and labor market outcomes of students, to design adapted and efficient edu-

cational policies. While the considered layers of analysis in this paper are reduced, it already shows that taking into account already available information when targeting dropout policy can helps to better understand students paths, behaviors and potential issues in the higher education system.

References

Adams, R., Almeida, H., & Ferreira, D. (2009). Understanding the relationship between founder–CEOs and firm performance. *Journal of Empirical Finance*, *16*(1), 136–150. https://doi.org/10.1016/j.jempfin.2008.05.002

Aina, C., Baici, E., Casalone, G., & Pastore, F. (2018). The economics of university dropouts and delayed graduation: A survey. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3153385

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. https://doi.org/10.2307/2937954

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2). https://doi.org/10.1214/18-AOS1709

Athey, S., & Wager, S. (2019). Estimating treatment effects with causal forests: An application. *arXiv:1902.07409 [stat]*. Retrieved April 28, 2021, from http://arxiv.org/abs/1902.07409

Athey, S., & Wager, S. (2020). Policy learning with observational data. *arXiv:1702.02896 [cs, econ, math, stat]*. Retrieved April 25, 2021, from http://arxiv.org/abs/1702.02896

Becker, G. S. (1993). *Human capital: A theoretical and empirical analysis, with special reference to education* (3rd ed). The University of Chicago Press.

Bjerk, D. (2012). Re-examining the impact of dropping out on criminal and labor outcomes in early adulthood. *Economics of Education Review*, *31*(1), 110–122. https://doi.org/10.1016/j.econedurev.2011.09.003

Breiman, L., Friedman, J., Olshen, R., & Stone, C. J. (1983). Classification and regression trees.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brodaty, T., Gary-Bobo, R., & Prieto, A. (2008). Does speed signal ability? the impact of grade repetitions on employment and wages. *C.E.P.R. Discussion Papers, CEPR Discussion Papers*.

Card, D. (1993). *Using geographic variation in college proximity to estimate the return to schooling* (NBER Working Papers No. 4483). National Bureau of Economic Research, Inc. https://EconPapers.repec.org/RePEc:nbr:nberwo:4483

Chernozhukov, V., Escanciano, J. C., Ichimura, H., Newey, W. K., & Robins, J. M. (2022). Locally robust semiparametric estimation. *Econometrica*, *90*(4), 1501–1535. https://doi.org/10.3982/ECTA16294

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika*, *96*(1), 187–199. https://doi.org/10.1093/biomet/asn055

Fang, H. (2006). DISENTANGLING THE COLLEGE WAGE PREMIUM: ESTIMATING a MODEL WITH ENDOGENOUS EDUCATION CHOICES. *International Economic Review*, *47*(4), 1151–1185. https://doi.org/10.1111/j.1468-2354.2006.00409.x

Flores-Lagunes, A., & Light, A. (2007). Interpreting sheepskin effects in the returns to education.

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association*, *87*(417), 178–183. https://doi.org/10.1080/01621459.1992.10475190

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7

Mahjoub, M.-B. (2017). The treatment effect of grade repetitions. *Education Economics*, *25*(4), 418–432. https://doi.org/10.1080/09645292.2017.1283006

Matkovic, T., & Kogan, I. (2012). All or nothing? the consequences of tertiary education non-completion in croatia and serbia. *European Sociological Review*, *28*(6), 755–770. https://doi.org/10.1093/esr/jcr111

Ménard, B. (2018). Le décrochage dans l'enseignement supérieur à l'aune de l'approche par les capabilités. *Formation emploi*, (142), 119–141. https://doi.org/10.4000/formationemploi.5684

Morlaix, S., & Perret, C. (2013). L'évaluation du plan réussite en licence : Quelles actions pour quels effets ? analyse sur les résultats des étudiants en première année universitaire. *Recherches en éducation*, (15). https://doi.org/10.4000/ree.7390

Navarro, S., Fruehwirth, J., & Takahashi, Y. (2016). How the timing of grade retention affects outcomes: Identification and estimation of time-varying treatment effects. *Journal of Labor Economics*, *34*. https://doi.org/10.1086/686262

Psacharopoulos, G., & Patrinos, H. A. (2018). Returns to investment in education: A decennial review of the global literature. *Education Economics*, *26*(5), 445–458. https://doi.org/10.1080/09645292.2018.1484426

Reisel, L. (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. https://doi.org/10.1016/j.rssm.2012.10.002

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. https://doi.org/10.1037/h0037350

Schnepf, S. V. (2014). *Do tertiary dropout students really not succeed in european labour markets?* (IZA Discussion Papers No. 8015). Institute for the Study of Labor (IZA). Bonn. http://hdl.handle.net/10419/96694

Scholten, M., & Tieben, N. (2017). Vocational qualification as safety-net? education-to-work transitions of higher education dropouts in germany. *Empirical Research in Vocational Education and Training*, *9*. https://doi.org/10.1186/s40461-017-0050-7

Spence, M. (1973). Job market signaling. *The Quarterly Journal of Economics*, *87*(3), 355. https://doi.org/10.2307/1882010

Vignoles, A. F., & Powdthavee, N. (2009). The socioeconomic gap in university dropouts. *The B.E. Journal of Economic Analysis & Policy*, *9*(1). https://doi.org/doi:10.2202/1935-1682.2051

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

6 Appendix

6.1 Descriptive statistics database after discard

	Dropout rate	Frequency	Percentage
Gender			
Male	39.1%	6472	51.5%
Female	32.1%	6098	48.5%
Highest diploma tried			
Bac +2 (STS)	43.0%	7065	56.2
Bac +3 (university)	22.6%	3061	24.4%
Bac +4 (university)	51.3%	943	7.5%
Bac +5 (university/Grande Ecole)	16.5%	588	4.7%
PhD	19.0%	913.00	7.3%
Parents' highest social category			
Disadvantaged	35.8%	2045	16.3%
Intermediate	38.2%	3527	28.1%
Advantaged	34.6%	2661	21.2%
Highly Advantaged	34.2%	4337	34.5%
Parents' highest diploma			
No diploma	36.8%	2606	20.7%
Bac or below	37.4%	5197	41.3%
Short degree	33.9%	3108	24.7%
Long degree	31.8%	1659	13.2%
Other			
Foreign trip (= yes)	42.2%	6166	49.1%
Internship (= yes)	29.3%	6404	51.0%

Table 10: Summary statistics by dropout status

6.2 GVIF : Assessing potential multicolinearity

	GVIF	Df	GVIF^(1/(2*Df))	(1/(2*Df))^2
Gender	1.46	1.00	1.21	1.46
Highest diploma tried (6 levels)	4.70	4.00	1.21	1.47
SES max parents	1.47	3.00	1.07	1.14
Diploma max parents	1.50	3.00	1.07	1.14
Type of HSD	2.40	2.00	1.24	1.55
HSD grade discretize	1.70	3.00	1.09	1.19
Region higher education institution	1.82	22.00	1.01	1.03
Foreign travel	1.31	1.00	1.14	1.31
Internship	2.54	1.00	1.60	2.54

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 TSLS model.

6.3 Compliance score : distribution

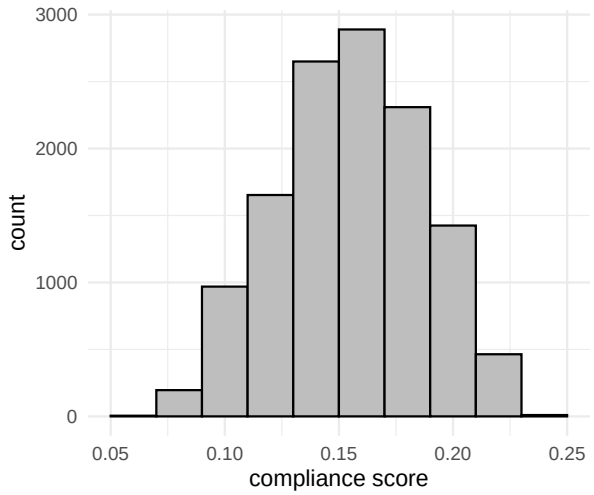


Figure 9: Distribution of the compliance score

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
compliance score	0.06	0.13	0.16	0.16	0.18	0.24

Table 11: Distribution of the compliance score $\mathbb{E}[W_i|Z_i = 1, X_i = x]$

6.4 Full tables TSLS

	Dependent variable:	
	log(Rate of employment + 1)	log(Average wage + 1)
	(1)	(2)
Constant	0.566^{***}	6.816^{***}
	(0.011)	(0.039)
dropout	-0.183^{***}	-0.357^{***}
	(0.025)	(0.086)
gender2	-0.017^{***}	-0.076^{***}
	(0.004)	(0.013)
nisor_6Bac +4	0.024^{**}	0.134^{***}
	(0.010)	(0.033)
nisor_6Bac +5	0.035^{***}	0.254^{***}
	(0.008)	(0.028)
nisor_6Bac+2	0.019^{***}	0.024
	(0.006)	(0.019)
nisor_6PhD	0.044^{***}	0.445^{***}
	(0.007)	(0.023)
cat_max_tDisadvantaged	0.003	-0.036^{**}
	(0.005)	(0.018)
cat_max_tHighly Advantaged	0.017^{***}	-0.016
	(0.005)	(0.016)
cat_max_tIntermediate	0.002	-0.009
	(0.004)	(0.015)
dip_max_tLong degree	-0.024^{***}	0.060^{***}
	(0.005)	(0.019)
dip_max_tNo diploma	-0.010^{**}	0.042^{***}
	(0.004)	(0.015)
dip_max_tShort degree	-0.011^{***}	-0.003
	(0.004)	(0.014)
typeBAC2	0.005	-0.069^{***}
	(0.004)	(0.013)
typeBAC3	0.038^{***}	-0.129^{***}
	(0.007)	(0.024)
mentionBAC2	0.003	0.040^{***}
	(0.004)	(0.014)
mentionBAC3	0.020^{***}	0.113^{***}
	(0.006)	(0.022)
mentionBAC4	0.035^{***}	0.154^{***}
	(0.013)	(0.045)
foreign_travel1	-0.025^{***}	-0.043^{**}
	(0.006)	(0.020)
internship1	-0.015^{***}	-0.022
	(0.005)	(0.017)
Observations	12,614	12,614
R^2	-0.001	0.100
Adjusted R^2	-0.005	0.097
Residual Std. Error (df = 12572)	0.171	0.594
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 12: TSLS estimation of dropout LATE

6.5 Robustness check Δ IV threshold discretization

After lowering the threshold $\alpha = 0.60$, the number of observations with $Z_i = 1$ is 8193, while the instrument is equal to 0 for 4519 observations. This unbalance is due to the discard phase needed to avoid too low propensity score.

Rate of employment

	GRF_1	GRF_2	t-stat
Overall sample	-0.174***		
	(0.021)		
$\tau_1 \leq Median(\tau_i)$		-0.139***	-2.07
		(0.027)	
$\tau_2 > Median(\tau_i)$		-0.181***	
		(0.031)	
Note:	*p<0.1; **p<0.05; ***p<0.01		

Hello

Table 13: Estimation of dropout LATE (rate of employment) with $\alpha = 0.6$

Average wage

	GRF_1	GRF_2	t-stat
Overall sample	-0.579***		
	(0.100)		
$\tau_1 \leq Median(\tau_i)$		-0.737***	-2.07
		(0.148)	
$\tau_2 > Median(\tau_i)$		-0.326**	
		(0.132)	
Note:	*p<0.1; **p<0.05; ***p<0.01		

Hello

Table 14: Estimation of dropout LATE (average wage) with $\alpha = 0.6$

By diploma

		Rate of employment	Average wage
Diploma	**Bac +2 (STS)**	**-0.186*****	**-0.5718*****
		(0.025)	**(0.12052)**
	Bac +3 (university)	**-0.016**	**-0.427***
		(0.050)	(0.236)
	Bac +4	-0.206***	0.372
		(0.064)	(0.337)
	Bac +5	-0.192**	-0.710*
		(0.088)	(0.429)
	PhD	-0.312***	-0.948**
		(0.084)	(0.400)

Table 15: LATE for Diploma, SES and Diploma × SES (Rate of employment and Average wage)