CONTRIBUTI DI RICERCA CRENOS

DROP-OUT DECISIONS IN A COHORT OF ITALIAN UNIVERSITY STUDENTS

Gianfranco Atzeni
Luca G. Deidda
Marco Delogu
Dimitri Paolini

WORKING PAPERS

2020/05

CENTRO RICERCHE ECONOMICHE NORD SUD
(CRENoS)
UNIVERSITÀ DI CAGLIARI
UNIVERSITÀ DI SASSARI

CRENOS was set up in 1993 with the purpose of organising the joint research effort of economists from the two Sardinian universities (Cagliari and Sassari) investigating dualism at the international and regional level. CRENoS' primary aim is to improve knowledge on the economic gap between areas and to provide useful information for policy intervention. Particular attention is paid to the role of institutions, technological progress and diffusion of innovation in the process of convergence or divergence between economic areas. To carry out its research, CRENoS collaborates with research centres and universities at both national and international level. The centre is also active in the field of scientific dissemination, organizing conferences and workshops along with other activities such as seminars and summer schools.
CRENoS creates and manages several databases of various socio-economic variables on Italy and Sardinia. At the local level, CRENoS promotes and participates to projects impacting on the most relevant issues in the Sardinian economy, such as tourism, environment, transports and macroeconomic forecasts.

www.crenos.unica.it
crenos@unica.it

CRENOS – CAGLIARI
VIA SAN GIORGIO 12, I-09124 CAGLIARI, ITALIA
TEL. +39-070-6756397; FAX +39-070- 6756402

CRENOS - SASSARI
VIA MURONI 25, I-07100 SASSARI, ITALIA
TEL. +39-079-213511

Title: DROP-OUT DECISIONS IN A COHORT OF ITALIAN UNIVERSITY STUDENTS

Prima Edizione: settembre 2020

ISBN: 978 88 68513 047

Viale Bonaria 98 - 09125 Cagliari
Tel. 070/6848663 - info@arkadiaeditore.it
www.arkadiaeditore.it

Drop-out decisions in a cohort of Italian university students*

Gianfranco Atzeni
University of Sassari & CRENoS
Luca G. Deidda
University of Sassari & CRENoS
Marco Delogu
University of Sassari & CRENoS
Dimitri Paolini
University of Sassari, CRENoS and CORE, Université Catholique de Louvain

Abstract

We study the determinants of student drop-out decisions using data on a cohort of over 230000 students enrolled in the Italian university system. We find that students who leave their homes to enrol at university (off-site students) drop out significantly less than those who study in their home town. We provide significant evidence that off-site students are a self-selected sample of the total population. Accordingly, we use an instrumental variable (IV) approach to identify the causal relationship. The IV estimation finds that studying off-site negatively affects drop-out decisions and more so for students growing up in the south of Italy who typically study off-site in the centre-north of Italy.

Keywords: DropOut, Location Choice, Instrumental Variable, Higher Education.
Jel Classification: A22, C26, I20, I21.

* We wish to thank Bianca Biagi, Claudio Deiana, Claudio Detotto, Alessandra Faggian, Masood Gheasi and Jacques Poot for their precious suggestions. We would like to acknowledge the participants to the workshops "Anvur: III concorso pubblico idee per la ricerca" (2019, Rome), "International and Internal Migration: Challenges and Opportunities in Europe" (2020, L'Aquila). We thank for precious research assistance Paolo Deledda (UNISS). The authors gratefully acknowledge financial support by the Italian Ministero della Ricerca (PRIN), Regione Autonoma della Sardegna (Legge n. 7), Anvur, III Concorso Pubblico Idee di Ricerca, Disea, Dipartimento di Eccellenza 2018-22 and Università degli studi di Sassari (fondo di Ateneo per la Ricerca 2019).

1 Introduction

There is robust evidence that more highly educated individuals earn higher salaries and enjoy a higher employment rates, see OECD (2019). Empirical studies indicate a sizeable effect, with an average increase in annual earnings of around 10% per additional year of education (see Card (2001)). Yet, in "[..] all developed countries the percentage of students dropping out of university or graduating beyond legal terms is very large [..]", see Aina et al. (2018), page 2. In general, delayed completion of studies reduces the average and overall skill levels of the working population. Reducing drop-out rates could, therefore, have a positive impact on the skill composition of the workforce. In turn, this improvement may trigger a positive feedback effect on the economy both in terms of efficiency and inequality. First, a more educated workforce would facilitate technological change and technology adoption, see Acemoglu (2002). Second, it could push down the wage skill premium, thereby reducing inequality, see Katz and Murphy (1992). Along with the US, Italy is one of the OECD countries where the drop-out phenomenon has reached dramatic levels, with more than one student in two dropping out of university before completion, see Aina et al. (2018).

This paper uses the "Anagrafe Nazionale Studenti" (ANS), which is a dataset produced by the Ministry of University and Research, (MUR), to study the determinants of the drop-out rate of undergraduates enrolled in Italian universities. The ANS collects information about all students enrolled in the Italian University system. As fully explained in section 2, we focus on information about undergraduate (i.e. bachelor) students who enrolled in the 2013-14 academic year. In particular, we study the correlation between drop-out rates and student characteristics, courses and universities.

Our empirical analysis reveals that the probability of dropping out of university is negatively correlated with high-school grades and student age. Our benchmark estimation suggests that one additional point in the high-school final grade reduces the probability of dropping out by 4%.[1] Similarly, enrolling one year later at the university increases the probability of dropping out by 9.8%. Consistent with the literature, our results also show that women have a lower probability of dropping out than men, and that individuals who attended a Liceum have a substantially lower probability of dropping out than their peers who attended vocational high school.

Our analysis focuses on the impact of studying off-site on dropping out. We define off-site students as those students who leave

[1]Other studies found an inverse relationship between high school grades and drop-out rates, see Belloc et al. (2010).

their homes to pursue higher education. Although Italian universities are evenly distributed across the national territory, a non-negligible fraction of students enrol in universities located in a region or province different from the one of residence.[2] In our dataset, 22% of individuals enrol in universities located in a region different from that of their homes. Similarly, 53.55% of students study in a province different from that of their homes. Italian inter-regional student mobility is probably eased by the homogeneous distribution of university fees across all public universities, see (Beine et al. (2020)). Indeed, financial barriers to access to education are quite low in Italy as poor students have access to a generous system of government grants (Checchi (2000)). Leaving home and relatives to pursue university education may affect educational outcomes in several ways. On the one hand, studying far from home requires additional effort in terms of organizing daily life, building new relationships, and other factors. On the other hand, studying off-site requires more financial support, often provided by parents, which may be an extra motivation for off-site students, see the insights of Checchi (2000). Also related to self-selection, Faggian et al. (2007) shows that Scottish and Welsh off-site students are more likely to migrate later, and considering US students, Kazakis and Faggian (2017) provides evidence of selectivity, showing that repeated migration is correlated with higher salaries. Regrettably, the ANS dataset does not provide unambiguous information on the off-site status of the student. However, it provides precise information on the place of residence of the student. Linking this information with the geographical location of the university, we construct several indicators that work as proxy variables for the students' off-site status. Using the region of origin to define off-site status, we estimate a reduction of 1.62% in the probability of dropping out associated with off-site status. Other measures for off-site status we use include (i) defining off-site students as students studying in a university outside their home district, and (ii) defining off-site students as the ones studying in a university more than $150km$ or $200km$ from their place of origin.[3] The sign and magnitude of the estimated parameters remain substantially unchanged when adopting these alternative measures of off-site status. The results are also robust to different estimation strategies and when we cluster individuals by macro-area.

We interpret these findings in light of Roy's model of self-selection (see Borjas (1987)). It is widely known that there are sizeable differences between the North and the South of Italy, both in terms of wages and job opportunities. Given those differences, Roy's model predicts self-selection in the flow of migrants. We document that students from the south of Italy are more likely to enrol

[2] 51 out of the 108 Italian provinces host a university. Furthermore, each Italian region hosts at least one university. For all municipalities, the geodesic distance from the nearest university is less than 108km (our computation).

[3] In addition, we provide estimates using an indicator of off-site status as the distance between the university destination and the students' place of origin.

outside their home region or district compared to their peers from the North of the country. Moreover, Southern students tend to move to universities located in the Centre-North of Italy. That is, in line with Roy's model predictions, we show that the flow of students follows mostly a South-Center\North Direction and that very few Northern students move to the South to pursue higher education. We document that off-site students' skills are higher compared to the overall population in terms of high-school grades. Also, students who attended a Liceum are over-represented among off-site students. Evidence of self-selection, as postulated by the Roy model, is reinforced when we run separate estimates by macro-area of origin. For instance, for the North area, we do not obtain a significant negative coefficient for off-site proxies.[4]

Our results are in line with Johnes and McNabb (2004), which is one of the few papers that explicitly addresses the impact of off-site status on drop-out rates. In particular, they find that the probability of dropping out is lower for students attending a university far from the one in the parental home town. Similarly, Modena et al. (2018) report a negative correlation between drop-out rates and studying off-site.[5]

The above discussion leads us to conclude that addressing causality with OLS estimates is problematic for two reasons. First, our significant negative OLS coefficients for the off-site status proxies in our drop-out regression are potentially an artifact of sample-selection bias. Secondly, off-site students go through a significant change in their daily life that, *ceteris paribus*, may affect their studies. We attempt to tackle this issue using an IV procedure; see section 4. Technically, we instrument the distance from the university chosen by the student and her home town with a proxy of the minimum distance from the closest university controlling with fixed-effect characteristics of the districts. Our IV estimates, while being imprecise, still uncover a negative relationship larger in magnitude than the one suggested by the standard OLS procedure. We also implement the IV procedure by splitting our dataset according to the macro-origin of the students. Interestingly, for the sub-sample of Southern students, the off-site status coefficient substantially increases in magnitude while remaining statistically significant and negative. We suggest interpreting this result as evidence that going off-site has a positive effect on the motivation of students coming from more distressed districts. Indeed,

[4] Namely, following David Autor, see https://economics.mit.edu/files/551, the self-selection model predicts a non-obvious relationship in the data. Therefore, our empirical findings suggest another application of Roy's model.

[5] Looking solely at students enrolled at the Università di Sassari, Bussu et al. (2019) find that students who are not from Sassari have a statistically significant lower propensity to drop out. They define students not from Sassari as students whose parental home is located more than 30 km away from Sassari. Zotti (2015) reports a similar relationship focusing on students enrolled at the Università di Salerno.

aside from identification issues, the causal effect of studying off-site is potentially ambiguous. Studying off site is more costly, in terms of organization of daily life and from an economic view point. Extra financial support is therefore necessary, which is often provided by off-site students' parents. The extra costs have two opposing effects. On the one hand, the fact that off-site students face a higher cost of studying compared to their peers who study in their home town undermines the sustainability of the off-site choice, which predicts higher-drop-out rates. On the other hand, the extra costs might provide extra motivation to the off-site students, which would result in a lower drop-out rate. Accordingly, a significant, negative effect is compatible with the idea that the second effect dominates. Nevertheless, as in the landmark study of Dale and Krueger (2002), we are fully aware that uncovering robust causal relationships of the determinants of dropping out requires particular care due to the pervasiveness of self-selection and unobservable variables.

The paper is organized as follows. In section 2, we describe our data and provide some facts on drop-out rates. In section 3, we outline our econometric approach. In section 3.1, we present the empirical OLS estimates along with several robustness checks. In section 4, we describe and implement the IV estimation procedure to tackle the causality issue. Section 5 offers conclusions.

2 Data and variables

In section 2.1, we describe our dataset along with the definition of the variables employed in our empirical analysis. In section 2.2, we provide some descriptive evidence summarized in stylized facts.

2.1 Dataset

Our data from the ANS contains information about all students enrolled in Italian universities. For the cohort in 2013-2014, we follow the students along their academic career until the 21st of March 2018. Abstracting from PhD programs, which we do not deal with in this study, Italian universities offer three types of degrees: (1) "Laurea triennale", which is equivalent to a Bachelor's degree, (2) "Laurea specialistica", which is equivalent to a 2-year Master's degree, and (3) "Laurea a ciclo unico", which combines Bachelor's and Master's degrees.

We focus only on the population of Bachelor's degree students who enrolled in 2013-14 for the first time. We choose to exclude students enrolled in “Laurea specialistica” or “Laurea a ciclo unico” because we lack information about the final grade they got in their previous careers as Bachelor’s degree students. Moreover, we exclude international students as foreign students seem to be selected from a different population compared to national students and constitute a self-selected group. Drop-out mechanisms like differ from those that characterize domestic students. We also exclude students enrolled in online universities.[6] Finally, the above choices lead to a dataset that contains information on $230, 336$ students.

The next step is to provide a precise definition of university drop-out. First of all, notice that due to the peculiar characteristics of the Italian university system, differently from Johnes and McNabb (2004), we can not differentiate between voluntarily and involuntarily drop-out. We proceed as follows. First, we classify students into four main categories: (A) students who successfully completed their degree by the 21st of March 2018), (B) students who were still enrolled by the 21st of March 2018, having not completed their degree yet, (C) students who changed course/university the year after the first year of enrolment, and (D) students who left the Italian university system.

We build a dummy variable, $D_{i,j,c,t}$, that has a value of 1 if student i is enrolled in course c at university j and drops out at time t and is 0 otherwise. Unfortunately, our dataset does not contain direct information on whether the student is actually off-site or not. Hence, to capture off-site status, we combine information on both the place of residence and the student’s origin. We use this information to construct the following three alternative, discrete proxies of off-site status:

1. *OD* (out of district): This variable has a value of 1 if the student enrols in a university located outside her home district. Notice that for a sizable percentage of students, this variable always has a value of 1 given that 52 out of the 110 Italian districts do not have a university.

2. *OR* (out of region): This variable has a value of 1 when student i enrols in a university located outside the home region. Each Italian region hosts at least one university. Therefore, the value of this variable is not prearranged as it is the case for the *OD* variable for a sizable fraction of districts.

3. OFF_{km}: This variable has a value of 1 when student i enrols in a university that is distant more than a threshold distance

[6] Note that, in 2013-14, online universities accounted for only the 4.53% of the total population of students enrolled in Bachelor’s courses. There is no clear definition of off-site status when a student enrols for an online course.

from the student's home. We take advantage of ANS information on students' home residence for all students enrolled in courses at any given university j. Then, after having obtained geographic coordinates for university j, we compute the travel distance between university j and the home of student i.[7] This measure rules out the cases of students whose house is close to the regional border that enrols outside the region without changing residence. In section 3.1, we consider two thresholds, 150 km and 200 km, and create two indicators, OFF_{150} and OFF_{200}.

In addition, to capture the students' off-site status, we also construct two continuous variables. We consider both the travel and geodesic distances between the university j and the home student i.

2.2 Descriptive statistics

Due to missing values on some variables, we end up with a dataset containing information on 226,094 individuals that represents the 98% of the population of students we initially included. Table 1 provides descriptive statistics for the four categories of students defined above. Table 1 reports that 38.40% of the students completed their degree by the 21st of March 2018.

Table 1: Students' group categories

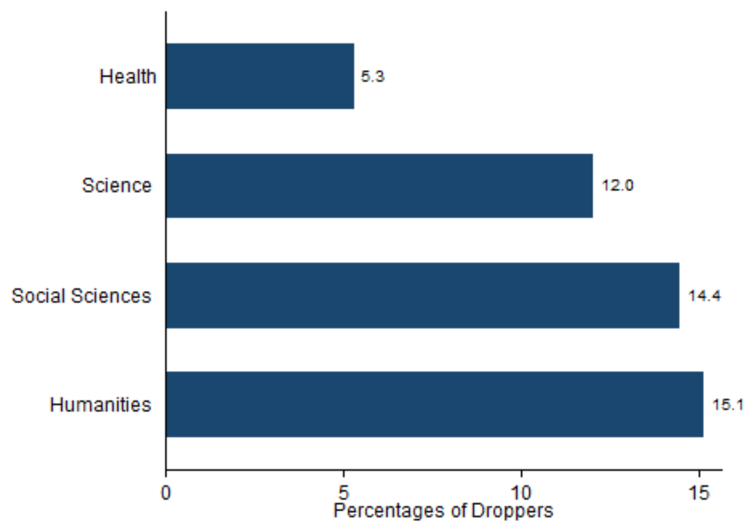
Student Outcome	Number	Percentage
Enrolled and degree not completed ($D_{i,j,c,t} = 0$)	71395	31.00
Changed course\university ($D_{i,j,c,t} = 0$)	41009	17.80
Degree completed ($D_{i,j,c,t} = 0$)	88221	38.30
Left university ($D_{i,j,c,t} = 1$)	29707	12.90

Students enrol in 708 different courses that belong to 46 different classes clustered in the four general subject areas: (1) Health, (2) Science, (3) Social Science and (4) Humanities. Science is the area with the most students, representing 38.4% of the sample. Interestingly, slightly more than the majority of students are enrolled either in Humanities or Social Science. Regarding gender,

[7] We take advantage of the STATA routine developed in Weber and Péclat (2017).

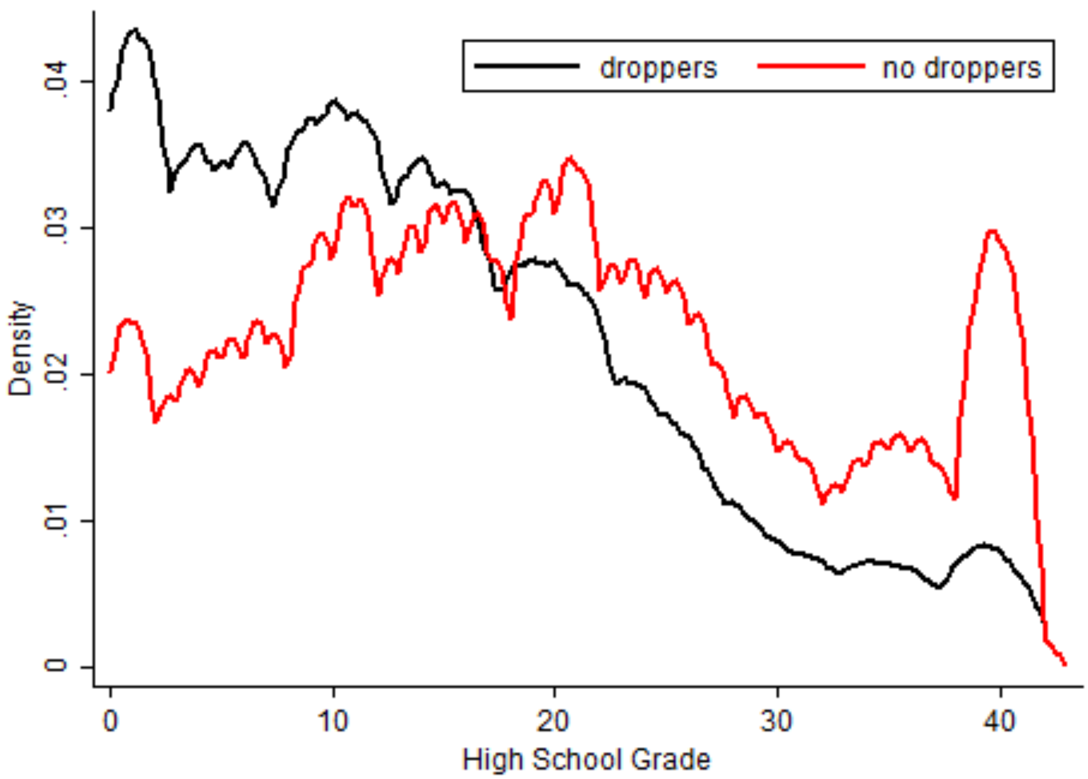
54.2% of students are female, and men are mainly enrolled in Science with only 2.5% enrolled in Health. We find that the percentage of women that leave the university, 14.8%, is lower than that of men, 11.2%. Figure 1 shows that there is a significant difference in the percentage of dropouts across the areas of study. While dropouts are equal only to 5.3% in Health\Medical areas, they reach a sizeable figure of 15.1% in Humanities. To account for these patterns, we include fixed effects for the area of study in our empirical estimations.

Figure 1: Drop-out by Field of Study



Men leave before graduating studies more often than women in any of the four areas of study. For instance, for Science, although women are under-represented, the percentage of men who drop-out is substantially larger than that of women, Another finding is that dropout rates are much larger for students who come from vocational high schools, which holds for all areas. Students coming from a Liceum have a drop-out rate that is 10% lower. Conversely, students coming from vocational schools have a much higher drop-out rate, which reaches 21% for Science students. In Figure 2, we compare the distribution of high-school grades among drop-outs and non-drop-outs, and we find that individuals with low high school grades are over-represented among the drop-outs.

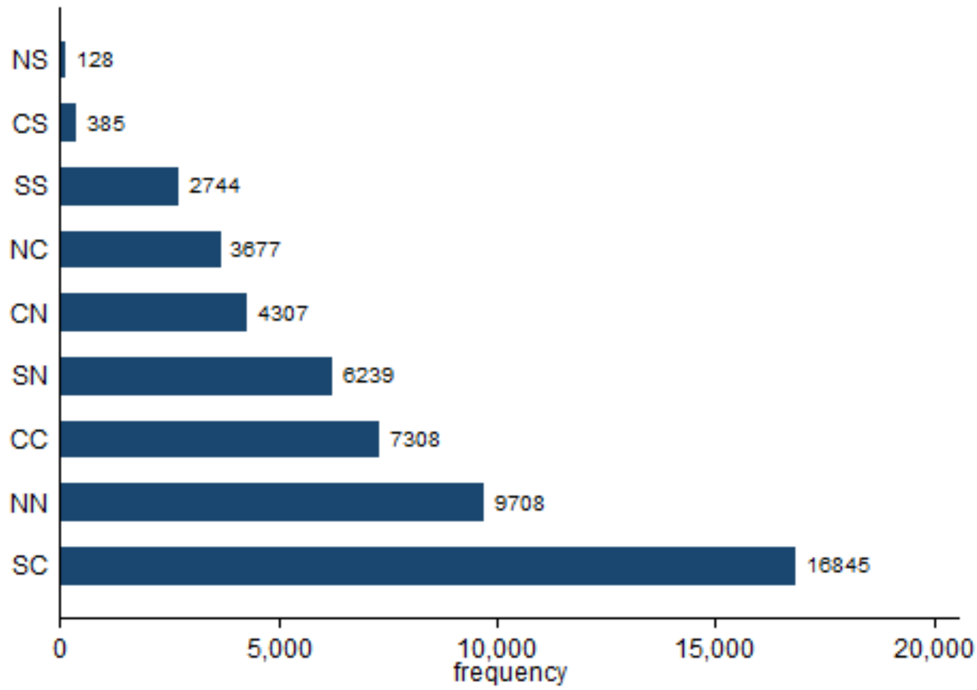
Figure 2: Distribution of High-School Grade by drop-out behaviour



We find that the drop-out rates exhibit significant variation across student home regions. To account for this heterogeneity, fixed effects for the students' district and region of origin are included in our econometric model.

The percentage of students studying off-site is unevenly distributed across Italian districts. Measuring off-site students through the variable OFF_{150}, we find that off-sites students reach 33% among the students who come from the South. Instead, for both those coming from the Centre and the North of Italy, the percentages are much lower at 16% and 17%, respectively. Figure 3 confirms that most of the off-site students move from South Italy to study in the North. Very few individuals (only 128) move from the North to the South. We count 23, 084 students who come the South and enrol in universities located either in the Centre or the North of Italy. Also, we document that the internal mobility of students,[8] is sizeable in the Centre\North of Italy and modest in the South.[9].

Figure 3: Migration Corridors: number of students enrolled out of region by Macro Region - North, Centre and South



NS: North to South; CS: Centre to South; SS: South to South; NC: North to Centre; CN: Centre to North; SN: South to North; CC: Centre to Centre; NN: North to North; SC: South to Centre

The other variables that we use for our estimation are:

- G_i, which is a dummy variable that has a value of 1 if the gender of student i is male and 0 otherwise.

[8]We define intra-mobility as relocation among Italian macro-regions

[9]In figure 3, to capture off-site status, we employ the *OR* variable

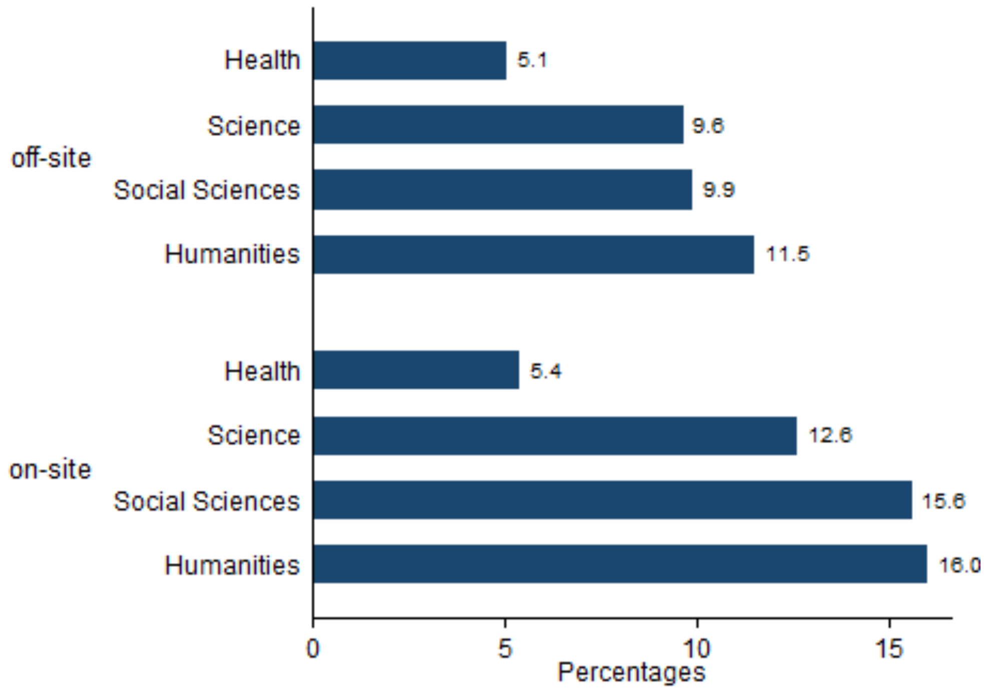
- HT_i is a dummy variable that has a value of 1 if the high school attended by the student is a *Liceum* and 0 otherwise.
- HG_i is high school grades rescaled, see table 16. It is a discrete variable that measures high school grades and has an interval value $[0\ 41]$. A student enrolled in an Italian high school needs to achieve a minimum final grade of $60/100$ to graduate.[10]
- $AGE_i = -1\left(Yearfofbirth - 1995\right)$, which is a variable aimed at capturing late enrolment at the university. Most Italian students end high school at the age of nineteen. However, some students may start university earlier given the possibility to anticipate entrance in primary school.

According to Rosenzweig et al. (2006), two main reasons explain why students move elsewhere to complete higher education.[11] First, individuals move elsewhere due to a lack of higher education institutions in their home region. However, this reason does not really apply to Italy, given that universities are evenly distributed within the country's territory. At the same time, we may expect that the percentage of off-site students is larger at better universities as there is substantial evidence that university quality is a key factor in student mobility (Beine et al. (2020)). Moreover, Italian universities with the best rankings are located in the Centre-North of Italy. The second model explains student migration with individuals intending to move to areas where skilled labour is better paid. This model fits the Italian experience better where many individuals leave the South to join universities located in most the Centre-North area in Italy, which provides better working opportunities after graduation. Figure 4 shows the drop-out rate across the primary area of study disentangled by off-site status, (measured b the dummy OR). Notice that except for the area of Health, the average drop-out rate of off-site students is always considerably lower.

[10] Students may get a mention. In this case, the grade is coded as 101.

[11] Rosenzweig et al. (2006) deals with international students' mobility flows, but similarities with internal student mobility are easily recognized.

Figure 4: Drop-out by Field of Study and off-site status



All our figures seem to suggest that off-site students are a self-selected sub-population. Tables 2 and 3 give additional support to this hypothesis. Table 2 measures off-site students using OR. Differently, table 3 employs the OFF_{150} variable as the indicator of off-site status. Both tables report that the average high-school grade and the percentage of students who earned each high school grade at a *Liceum* are considerably larger among off-site students. For both tables, a *t-test* of the difference in means leads us to reject the null hypothesis of no differences in high-school grades across the two sub-populations of off-site and on-site students. Also, differences in table 3 are larger than in table 2.

Table 2: Means for sub-samples

	HG	Age	HT	$G, M = 1$
$OR = 0$				
mean	18.037	2.13	0.50	0.45
sd	11.51	4.17	0.50	0.50
$OR = 1$				
mean	19.99	2.06	0.57	0.4
sd	11.98	4.07	0.49	0.50
Total				
mean	18.47	2.11	0.52	0.45
sd	11.65	4.15	0.50	0.50

Table 3: Means for sub-samples

	HG	Age	HT	$G, M = 1$
$OFF_{150} = 0$				
mean	18.08	2.12	0.50	0.45
sd	11.52	4.17	0.50	0.50
$OFF_{150} = 1$				
mean	20.60	2.06	0.60	0.46
sd	12.10	4.06	0.49	0.50
Total				
mean	18.47	2.11	0.52	0.45
sd	11.64	4.15	0.50	0.50

3 Empirical analysis

The existing literature provides evidence that the characteristics of universities, the field of study, and the social and economic conditions of the students' home districts are correlated with dropout rates.[12] In this literature, we aim to document the relationship between distance, namely studying off-site, and dropout rates in Italian students. To do so, in this section, we discuss the results of our benchmark estimations complemented with several robustness checks. Then, in section 5, we address causality issues due to self-selection and omitted variables with an instrumental variable approach.

To uncover this relation, we set up the following empirical specification

$$D_{i,u,o,f,c} = \alpha + \mathrm{A}_u + \mathrm{A}_f + \mathrm{A}_o + \beta_1 G_i + \beta_2 AGE_i + \beta_3 HT_i + \beta_4 HG_i + \beta_5 OffSite_{i,t} + \varepsilon_i \tag{1}$$

where ε_i is the error term, and we recall that $D_{i,u,o,f,c,}$ is the dummy variable that captures the student dropout, i, coming from the place of origin, o, enrolled in university, u, the field of study, f, and course c. The variables on the RHS of equation 1 include gender, G_i, age, AGE_I, type of high school, HT_i, and high school grade, HG_i, which section 2.1 already discussed.

- A_u, which is a set of fixed effects that we include to control for differences in university characteristics,
- A_f, which is a set of fixed effects we include to control for the different fields of study, and
- A_o, which is a set of fixed effects controlling for all factors specific to home districts of students. With these fixed effects, we also aim to capture differences in high school education quality among Italian districts.
- $OffSite_i$, which is the measure of the students' off-site status. We code this variable, the focus of our analysis, in four different ways:
    1. OD, which has a value of 1 if the student enrols in a university located outside the home district and zero otherwise;
    2. OR, which has a value of 1 when the student enrols in a university located outside the home region and zero otherwise;
    3. OFF_{km}, which has a value of 1 when the student enrols in a university located more than km away from her home and zero otherwise. We consider two thresholds: 150km and 200km;

[12] See Aina et al. (2018).

4. *TD*, which is the travel distance between the university and the student's place of residence measured in hundreds of kms;
5. *GD*, which is the geodesic distance between the university and the student's place of residence. One unit is equal to 100 km.

Table 16 contains additional information about the above variables, see the Appendix.

Using the above description, 1 controls for university, district of origin, and field of study characteristics through fixed effects and for other individual characteristics in the ANS dataset, including gender, final high school grade, age of the individual, and the type of high school attended.[13] We note that a limitation of the ANS dataset is the lack of information on both family income and parental background.[14] Also, we lack unambiguous information on the amount of tuition charged to each student. [15]

We obtain our baseline estimates for equation 1 through an OLS estimation procedure. Several reasons lead us to stick with the Linear Probability Model (LPM) as a baseline. Among others, Angrist and Pischke (2008) advocates the use of the LPM. [16] Although, non-linear estimation methods may provide an efficiency gain but at the cost of committing to a precise distributional assumption of the error term. Notably, Probit and Logit provide average marginal effect estimates that, quite often, do not differ much from LPM estimates.[17] Also, the interpretation of the regression coefficients is much more straightforward with the LPM. Finally, we evaluate the precision of our estimates through robust standard errors to deal with well-known issues of heteroskedasticity of the LPM.

Section 4 complements our estimation results by considering an IV estimation that attempts to tackle the endogeneity of our

[13] ANS differentiates university courses in 46 distinct fields of studies.

[14] Checchi (2000) highlights the role of both family income and parental background among the determinants of university drop-out rates.

[15] In Italy, tuition fees depend on several factors, such as household income, the field of study, and the year of enrolment. In Italy, private universities are allowed to charge much higher tuition, see Beine et al. (2020). Our fixed effects capture the heterogeneity in fees from different universities' policies. However we do not have unambiguous information on the amount of tuition charged to each student. Modena et al. (2018) considers a similar indicator, and they show that earning an education grant significantly reduces early drop-out

[16] The authors highlight that when *CIA, Conditional Independence Assumption Holds, or Selection on Observables* holds, the LPM estimate is the average treatment effect. Due to the omitted variables and selection issues, we make clear that our empirical model does not fulfil this assumption. However, the same issue would exist if we employ non-linear estimation methods, such as Probit and Logit.

[17] Section 3.2.2 reports AME obtained through a Logit specification. The Logit and AME estimates are very close to the ones from the LPM.

variable for off-site status.

3.1 Results

In this subsection, we present and discuss the empirical estimates of the benchmark model described by equation 1. We consider all different measures of studying off-site status discussed in the previous section.

Table 4 reports the estimation results when we use the *OD* and *OR* dummy variables to measure off-site status. [18] Drop-out rates are negatively correlated with the high-school grades, student age, being a woman, and a diploma from a *Liceum*. Interpreting our coefficient estimates as marginal effects, we find that, *ceteris paribus*, one additional point in high-school grade reduces the probability of dropping out by 0.4%. Graduating from a *Liceum* is correlated with a reduction in drop-outs by 10%. In the correlation between dropping out and being an off-site student, we find a significant negative sign. When we employ *OD*, we find that offsite status is associated with a 1.25% reduction in the probability of dropping out. When we proxy off-site status with the dummy *OR*, the estimated correlation becomes stronger. Neither the sign nor the magnitude of any of the other coefficients change across the two specifications. A comparison of columns (2) and (3) of table 4 shows that our estimates are robust to different measurements of off-site status.

[18] Johnes and McNabb (2004) and Bussu et al. (2019) employ similar indicators.

Table 4: Determinants of Drop-Out Rates. Benchmark (1)

	(1)	(2)	(3)
HG_i	-0.0040***	-0.0040***	-0.0040***
	(0.000)	(0.000)	(0.000)
Age_i	0.0098***	0.0098***	0.0098***
	(0.000)	(0.000)	(0.000)
HT_i	-0.1034***	-0.1030***	-0.1038***
	(0.001)	(0.001)	(0.002)
$G_i, M = 1$	0.0268***	0.0269***	0.0267***
	(0.002)	(0.002)	(0.002)
OD_i	-0.0125***		
	(0.002)		
ORi		-0.0162***	-0.0161***
		(0.002)	(0.002)
University Fixed Effects	yes	yes	yes
Field Fixed effects	yes	yes	yes
Region Fixed effects	yes	yes	no
District Fixed effects	yes	no	yes
R^2	0.0917	0.0917	0.0927
N	226094	226094	226094

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS estimates. Robust standard errors in parentheses.

As pointed out in the introduction, for a large percentage of students, their home district does not host any university, and the only option is to leave the district to pursue university education. Specifically, this finding implies that for students coming from 52 out of the 110 Italian districts, the *OD* dummy variable always has a value of one. In that respect, *OR*, which is based on regions, provides a more conservative definition of off-site status. Still, both *OR* and *OD* might not be meaningful measures of off-site status for various reasons. For instance, using either *OR* or *OD*, we might end up classifying off-site students who enrol in universities that, while located in a different district or region, might be geographically very close to their home location, even close enough to allow daily commuting. Therefore, we also consider alternative measures of off-site status based on travel and geodesic distance between the student's home and student's university. Specifically, in Tables 5 and 6, the dummy variables *OD* and *OR* are replaced with continuous variables *TD* and *GD*, respectively, where *TD* is the travel distance and *GD* is the geodesic distance. Column 1 of Table 5 suggests that a 100 km increase in average travel distance is associated with a 0.3% reduction in the probability of dropping out. Similar results are obtained for geodesic distance (see table 6). In both tables, we also report the results for regression models that include the square of the distance. Including this variable, we test the hypothesis of a non-linear relationship, and we find that the marginal effect of distance diminishes with distance.

Table 5: Determinants of Drop-Out Rates. Benchmark (2)

	(1)	(2)
HG_i	-0.0040***	-0.0040***
	(0.000)	(0.000)
AGE_i	0.0097***	0.0097***
	(0.000)	(0.000)
HT_i	-0.1038***	-0.1038***
	(0.002)	(0.002)
$G_i, M = 1$	0.0268***	0.0267***
	(0.002)	(0.002)
$TD_{u,o}$	-0.0033***	-0.0078***
	(0.000)	(0.000)
$TD^2_{u,o}$		0.000005***
		(0.000)
University Fixed Effects	yes	yes
Field Fixed Effects	yes	yes
District Fixed Effects	yes	yes
R^2	0.0925	0.0926
N	226094	226094

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS estimates. Robust standard errors in parentheses.

Table 6: Determinants of Drop-out Rates. Benchmark (3)

	(1)	(2)
HG_i	-0.0040***	-0.0040***
	(0.000)	(0.000)
Age_i	0.0097***	0.00976587***
	(0.000)	(0.000)
HT_i	-0.1039***	-0.10378***
	(0.002)	(0.002)
$G_i, M = 1$	0.0267***	0.0266***
	(0.002)	(0.002)
$GD_{o,u}$	-0.0035***	-0.0116***
	(0.001)	(0.001)
$GD^2_{o,u}$		0.0011***
		(0.000)
University Fixed Effects	yes	yes
Field Fixed Effects	yes	yes
District Fixed Effects	yes	yes
R^2	0.0925	0.0926
N	226094	226094

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS estimates. Robust standard errors in parentheses.

Finally, we report the results we obtain measuring off-site status with the OFF_{km} dummy variable. We consider two specifications of this indicator: OFF_{150} and OFF_{200}. Notice that OFF_{150} and OFF_{200} have values equal to 1 if the student is enrolled in a university more than 150 and 200 km distant from home, respectively. Table 7 reports the empirical estimates obtained using these two measures of studying off-site.

Table 7: Determinants of Drop-out Rates. Benchmark (4)

	(1)	(2)
HG	-0.0040***	-0.0040***
	(0.000)	(0.000)
AGE	0.0098***	0.0098***
	(0.000)	(0.000)
HT	-0.1037***	-0.1038***
	(0.002)	(0.002)
$G, M = 1$	0.0268***	0.0268***
	(0.002)	(0.002)
OFF_{150}	-0.0195***	
	(0.003)	
OFF_{200}		-0.0165***
		(0.003)
University Fixed Effects	yes	yes
Field Fixed Effects	yes	yes
District Fixed Effects	yes	yes
R^2	0.0926	0.0925
N	226094	226094

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS estimates. Robust standard errors in parentheses.

Two results stand out from table 7. First, the magnitude of the coefficients capturing off-site status is strikingly close to the one delivered by the empirical estimate of OR, see table 4. Also, we notice that the magnitude of the coefficient for OFF_{200} is smaller than the one for OFF_{150}.

In summary, all our measures of studying off-site confirm a strong negative and significant correlation between the drop-out decision and off-site status. The estimates of the other variables of interest are in line with the findings in the literature. Women show a lower propensity to drop out. Also, there is evidence that older individuals tend to leave the university more frequently and that high school grades negatively correlate with dropout rates, with students that earned a better high school grade eventually dropping out less.[19] Finally, students who attend a *Liceum* tend to drop out less than students coming from vocational schools.

The possibility of self-selection and omitted variables suggest that we cannot interpret the correlation between off-site status and dropping out as evidence of a causal relationship. We address causality in section 4 with an IV approach.

3.2 Extension of the benchmark model

In this subsection, we evaluate the robustness of the correlations described in the previous section. We do so considering directional extensions:

1. In section 3.2.1, we estimate equation 1 clustering individuals by macro-area of origin (South-, Centre-, and North- of Italy).

2. In section 3.2.2, we compute the marginal effects by estimating a Logit specification of equation 1.

Our results show that i. the magnitude of our proxy varies substantially once we consider regressions by macro-area, and ii. the marginal effects we obtain with Logit estimation are not different from the ones obtained with the OLS specification.

3.2.1 Regression by Macro-Area

To study whether the correlations reported in section 3.1 remain stable independently of the home macro-area of the off-site students, we run regressions clustering students on their home macro-area. We consider three macro-areas: North, Centre, and South of Italy. Following the previous discussion, we employ OFF_{150} (table 8) and OR as indicators of off-site status, see table 9.

[19] Notice that Belloc et al. (2010) found a positive correlation between high-school grades and drop-out rates.

Table 8: Determinants of Drop-Out Rates: Estimates by Macro-Area (1)

	(South)	(Center)	(North)
HG	-0.0046***	-0.0037***	-0.0036***
	(0.000)	(0.000)	(0.000)
Age	0.0112***	0.0090***	0.0096***
	(0.001)	(0.000)	(0.000)
HT	-0.1071***	-0.1018***	-0.1003***
	(0.003)	(0.003)	(0.002)
$G, M = 1$	0.0285***	0.0270***	0.0248***
	(0.003)	(0.003)	(0.002)
OFF_{150}	-0.0243***	-0.0183***	-0.0068
	(0.006)	(0.005)	(0.004)
R^2	0.1074	0.0888	0.0835
N	77238	67850	80929
University Fixed Effects	yes	yes	yes
Field Fixed Effects	yes	yes	yes
District Fixed Effects	yes	yes	yes

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS estimates. Robust standard errors in parentheses.

Table 9: Determinants of Drop-Out Rates: Estimates by Macro-Area (2)

	(South)	(Center)	(North)
HG	-0.0046***	-0.0037***	-0.0036***
	(0.000)	(0.000)	(0.000)
Age	0.0112***	0.0090***	0.0096***
	(0.001)	(0.000)	(0.000)
HT	-0.1072***	-0.1019***	-0.1003***
	(0.003)	(0.003)	(0.002)
$G, M = 1$	0.0286***	0.0270***	0.0248***
	(0.003)	(0.003)	(0.002)
OR	-0.0311***	-0.0185***	-0.0067
	(0.009)	(0.005)	(0.004)
R^2	0.1073	0.0889	0.0835
N	77238	67850	80929
University Fixed Effects	yes	yes	yes
Field Fixed Effects	yes	yes	
District Fixed Effects	yes	yes	yes

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS estimates. Robust standard errors in parentheses.

Using *OR* or *OFF* yields almost identical results. Interestingly, the students' off-site status is not significantly associated with dropping out when we run the regressions considering only students from the North of Italy. Also, it is interesting to notice that the magnitude of the *HG* coefficient is larger, in absolute value, for the sub-population of students form the South. Remarkably, the coefficient for *HG* is almost identical when we run regressions separately for Centre and North students.

Several reasons may explain the lack of significance of both *OR* and *OFF* coefficients for the sample of North students. First, as shown in section 2.2, the vast majority of off-site students from the North opt to enrol in a university still located in the North, and therefore, a shorter distance from student's home. The distance may be so short that it does not affect students' life in any particular way, and therefore, does not affect their performance.[20]

3.2.2 Logit specifications

The literature often estimates drop-out through non-linear models.[21] We find that the estimated marginal effects do not change significantly when we employ a Logit specification in place of our benchmark LPM Similar to section 3.1, we obtain negative and significant coefficients for our measures of off-site status using *OD* and *OR* Importantly we highlight that using a Logit model does not affect the magnitude of the estimates either.

[20] Section 4 addresses the causal link between off-site status and drop-out decision.

[21] For example, we refer the reader to Belloc et al. (2010) and Zotti (2015).

Table 10: Determinants of Drop-Out. Logit, AME

	(1)	(2)	(3)
HG_i	-0.0042***	-0.0042**	-0.042***
	(0.000)	(0.000)	(0.000)
Age_i	0.0056***	0.0056***	0.0055***
	(0.000)	(0.000)	(0.000)
HT_i	-0.1053***	-0.1049***	-0.1055***
	(0.002)	(0.002)	(0.002)
$G_i, M = 1$	0.0283***	0.0283***	0.0282***
	(0.002)	(0.002)	(0.002)
OD_i	-0.0125***		
	(0.002)		
OR_i		-0.0196***	-0.0200***
		(0.002)	(0.003)
University Fixed Effects	yes	yes	yes
Field Fixed effects	yes	yes	yes
Region Fixed effects	yes	yes	no
District Fixed effects	yes	no	yes
N	226094	226094	226094

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. AME Logit estimates. Robust standard errors in parentheses.

4 Causality: Instrumental Variable approach

The descriptive evidence discussed in section 2.2 suggests the possibility that the sub-population of off-site students is a self-selected group with systematic characteristics different from the overall population. Due to the possibility of self-selection and omitted variables, interpreting the evidence from the regression models presented in section 3.1 is, therefore, problematic.

In other words, the evidence of a strong negative correlation between dropout rates and off-site status does not allow us to conclude anything about the causality or direction of that relationship due to both unobservables variabales and self-selection. Importantly, the decision to study far from home implies sunk costs, both monetary and non-monetary, see Checchi (2000), that affect students' effort. Off-site students leave at home both family and friends, need to acclimate to new city social norms and, last but not least, a substantial monetary investment is required (e.g., renting a room/apartment and transportation costs). These costs are likely to be positively correlated with distance. As Checchi (2000) shows, student effort is sensitive to monetary costs in general, and those studying off-site may exert more effort in their studies because, in the event of dropping out, the sunk cost is higher compared to the ones faced by on-site students Also, Garibaldi et al. (2012) shows that an increase in tuition and fees reduces late graduation, providing evidence that students' effort depends on investments made.[22] Among off-site students, there may be heterogeneity in students' sunk costs. We may have type 1 students, with higher ability and motivation, who choose to study off-site and enrol in a better university, and type 2 students, from high-income households, who may choose to study off-site merely because they can afford it but have average (or below average) motivation and ability. For type 1 students, the decision to study off-site is driven by motivation. For type 2 students, it is driven by family wealth. It is evident that both motivation and wealth are negatively correlated with dropout. As motivation and family income are unobserved in 1, we are not able to say whether the negative correlation between drop-out rates and off-site status is fostered by the link between higher costs and motivation or between higher costs and family wealth, or by both.

The above discussion suggests a need for an appropriate estimation strategy to address bias that self-selection and other omit-

[22] This paper does not find similar evidence for drop-out rates. However, it only considers students enrolled in one of the most expensive Italian private universities.

ted variables generates.[23] Following Card (1993), we exploit information on the distance from the closest university to construct a variable for off-site status. For each student, we determine the distance from her place of residence to the closest university. Taking advantage of this information, we identify two possible variables:

1. The distance from the closest university, which we label *minD*;

2. A dummy variable that we set equal to one if the closest university is more than 20 kilometres from the student's place of residence.[24] We label this instrument *dD*.

We acknowledge that some arguments may be raised about the validity of our variable, similar to the ones mentioned in Card (1993) and Card (2001).[25] The next section collects some evidence on the validity of the exclusion restriction. Section 4.2 presents and discusses our IV estimates.

4.1 Exclusion Restriction and Reduced Form

Our model is just identified, thus preventing us from performing the Sargan-Hansen to check whether the correlations among error terms and the instrument is statistically not different from zero. Despite the impossibility of performing the overid-test, we can check how *minD* correlates with the other drop-out determinants to evaluate the exclusion restriction assumption. Clearly, a good instrument should not be correlated with strong determinants of the dependent variable Table 11 provides evidence that this situation exists for our instruments. *minD* is almost uncorrelated with the determinants of the dropout rate previously discussed.

[23] Focusing on self-selection, one may suggest estimating the model with a Heckman-type correction model. We prefer to stick to an IV procedure. The validity of our estimates do not rely on any assumption concerning the distribution of the error term Angrist and Pischke (2008).

[24] When using this instrument, one may suggest running a Probit model in place of an OLS in the first stage. Angrist and Pischke (2008) and Wooldridge (2010) show that this procedure is incorrect, and we should run a *forbidden regression*. Conversely, another feasible alternative would be a bivariate probit model. However, our rich structure of fixed effects generates collinearity issues. Therefore, we only consider estimations obtained through a two-stage least squares procedure.

[25] Typically, one may argue the validity of the exclusion restriction saying that when deciding where to settle, households internalize children's decision of whether to enrol at the university. However, in Italy, household mobility is very limited, with individuals showing a very low propensity to move once settled.

Table 11: Correlations

	minD		
	correlation	p-value	observations
HT	-.036	1.47e-64	226094
HG	.053	3.5e-143	226094
Age	-.023	5.85e-29	226094

Table 12 reports the reduced form estimates. Column (1) shows *minD*, and column (2) reports similar estimates including *dD* in the set of regressors. Our reduced form estimates are both negative and either not significant or only marginally significant (in the case of *dD*).

Table 12: IV estimates: Reduced Form

	(1)	(2)
HG	-0.0040***	-0.0040***
	(0.000)	(0.000)
Age	0.0097***	0.0097***
	(0.000)	(0.000)
HT	-0.1041***	-0.1041***
	(0.002)	(0.002)
$G, M = 1$	0.0267***	0.0267***
	(0.002)	(0.002)
$minD$	-0.0001	
	(0.000)	
dD		-0.0036*
		(0.002)
University Fixed effects	yes	yes
Field Fixed effects	yes	yes
District Fixed effects	yes	yes
R^2	0.0925	0.0925
N	226094	226094

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. OLS estimates. Robust standard errors in parentheses.

4.2 IV: Results

Table 13 reports our empirical estimates where we instrumented the measures of off-site status. First-stage estimates confirm that our instruments are strong. Column (1) shows the OR dummy variable instrumented with the dD variable. Notice that the sign of the *OR* coefficient is still negative and significant and increases in magnitude compared to the OLS estimation with no instrumental variables.

Importantly, the standard errors increase, which is a typical consequence of the IV procedure. Column (2) reports similar findings. Here, we instrument *OR* with the actual minimum distance, *minD*. Column (3) and column (4) report results when we employ the *TD* variable as a proxy of off-site status. Notice that the magnitude and the sign of all other control variables remain almost unchanged as we vary the variables measuring off-site status. In conclusion, we notice that for all cases, the coefficients for distance lose statistical significance, which may be due to the lower precision implied by IV estimation. To check whether it is sensible to run the IV procedure, we report the Wu-Hausman test. The null hypothesis is that both estimators, OLS and IV, are consistent. We do not obtain strong evidence for non-consistency with the OLS estimates. However, even if we fail to reject the null hypothesis, the test does not allow us to claim that the OLS estimates are consistent. Hence, the WU-Hausman test do not invalidate our IV estimates. Indeed, this situation is typical when the standard error of the IV estimator is as large as it is in table 13.

Table 13: IV estimates (1) - Drop-out decision: instrumented indicator OR

	(1)	(2)	(3)	(4)
OR_i	-0.1068*	-0.0906		
	(0.053)	(0.048)		
HG_i	-0.0040***	-0.0040***	-0.0040***	-0.0040***
	(0.000)	(0.000)	(0.000)	(0.000)
AGE_i	0.0099***	0.0099***	0.0098***	0.0098***
	(0.000)	(0.000)	(0.000)	(0.000)
HT_i	-0.1025***	-0.1027***	-0.1037***	-0.1037***
	(0.002)	(0.002)	(0.002)	(0.002)
$G_i, M = 1$	0.0265***	0.0266***	0.0268***	0.0268***
	(0.002)	(0.002)	(0.002)	(0.002)
TD_i			-0.0093*	-0.0069
			(0.005)	(0.004)
University Fixed effects	yes	yes	yes	yes
Field Fixed effects	yes	yes	yes	yes
District Fixed effects	yes	yes	yes	yes
IV: Dummy (>20km)	yes	no	yes	no
IV: Distance	no	yes	no	yes
F First Stage	502	547	4969	6236
Hausman Test	2.99*	2.38	1.86	1.09
N	226094	226094	226094	226094

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

IV estimates.

Robust standard errors in parentheses.

We complement the results of table 13 with table 14. In columns (1) and (2), we use OFF_{150} as a proxy for off-site status, and columns (3) and (4) employ OFF_{200}. Notice that the results of Table 14 are strikingly similar to the ones reported by columns (1) and (2) of table 13.

Table 14: IV estimates (2) - Drop-out decision: instrumented indicator OFF_{km}

	(1)	(2)	(3)	(4)
OFF_{150}	-0.0941*	-0.0632		
	(0.046)	(0.034)		
HG	-0.0040***	-0.0040***	-0.0040***	-0.0040***
	(0.000)	(0.000)	(0.000)	(0.000)
Age	0.0099***	0.0099***	0.0099***	0.0099***
	(0.000)	(0.000)	(0.000)	(0.000)
HT	-0.1026***	-0.1030***	-0.1024***	-0.1029***
	(0.002)	(0.002)	(0.002)	(0.002)
$G, M = 1$	0.0267***	0.0268***	0.0269***	0.0269***
	(0.002)	(0.002)	(0.002)	(0.002)
OFF_{200}			-0.1294*	-0.0903
			(0.064)	(0.048)
University Fixed effects	yes	yes	yes	yes
Field Fixed effects	yes	yes	yes	yes
District Fixed effects	yes	yes	yes	yes
N	226094	226094	226094	226094
First Stage	717	1024	494	686

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

OLS estimates.

Robust standard errors in parentheses.

Section 3.2.1 suggests that the impact of off-site status on drop-out rates is much stronger among students coming from the

South. In table 15, we report IV estimation clustering individuals by home macro-areas. We employ the OFF_{150} variable as a proxy of off-site status, which we create using $minD$. Column (1) considers only students from the South. It reports a highly significant and negative estimate for the off-site measure, OFF_{150}. This finding suggests that, once we account for the selection effect, going off-site, for a Southern student, has a considerable impact on the decision to not leave the university. Interestingly, this finding is not the same for students originating from the Centre and North of Italy, for whom we do not find any significant impact of off-site status on the decision to drop out. Our results are in line with the model and empirical findings of Checchi (2000). Students moving from the South to the North face larger sunk costs. Large sunk costs appear to have a positive effect on the decision to not drop out. Similar evidence is not obtained once we consider students originating either from the Centre or the North separately. Most of them attend universities located in the same area, therefore, they face lower sunk costs and, as our estimates suggest, the positive effect on drop-out decision does not materialize.

Table 15: IV estimates, by Macro-Area (3)

	(South)	(Center)	(North)
OFF_{150}	-0.2507***	0.1794	-0.0754
	(0.069)	(0.134)	(0.091)
HG	-0.0046***	-0.0037***	-0.0036***
	(0.000)	(0.000)	(0.000)
Age	0.0114***	0.0087***	0.0096***
	(0.001)	(0.000)	(0.000)
HT	-0.1070***	-0.1046***	-0.0992***
	(0.003)	(0.003)	(0.003)
$G, M = 1$	0.0282***	0.0270***	0.0248***
	(0.003)	(0.003)	(0.002)
University Fixed effects	yes	yes	yes
Field Fixed effects	yes	yes	yes
District Fixed effects	yes	yes	yes
N	77238	67850	80929
First Stage	577	179	109

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

IV estimates.

Robust standard errors in parentheses.

5 Conclusions

This paper investigates the determinants of dropping out in the population of students enrolled in the public university system in Italy. We document that off-site students, who left home to pursue university, are a self-selected population with various characteristics that are candidate determinants of the drop-out decision. Then, we show that there is a robust and strong negative correlation between the likelihood of dropping out of university and students' off-site status. To go beyond correlation and assess the causal link between off-site status and the decision to drop out, we employ an instrumental variable approach. The estimates provide strong evidence that off-site status reduces the likelihood of dropping out among Southern students who typically study in universities located in the Centre-North of Italy. The negative effect is still present considering the entire population of students but lower in magnitude and barely significant. Our findings have relevant policy implications.

First, due to the documented self-selection, our estimates suggest that it is not fair to rank university quality through a naive comparison of drop-out rates. We produce abundant evidence that a significant fraction of the best Southern students move to complete higher education at institutions located in the Centre-North of Italy. On the contrary, the flow of students from the Centre-North to the South is negligible. Our empirical results suggest that self-selection among off-site status explains part of the sizable difference in drop-out rates between Northern and Southern institutions.

Second, our results suggest that universities that aim to improve the quality of their student pool set policies to attract off-site students.

Third, we address whether there is a causal relationship between off-site status and drop-out behaviour. We conduct our analysis taking advantage of the instrumental variable approach. We employ variable capturing the proximity from the closest university for our off-site indicators. Our results show that, especially for off-site students originating from the South, there is substantial evidence that going off-site reduces the likelihood of dropping out of university. In line with Checchi (2000), we argue that studying off-site, by requiring substantial investment (not only monetary ones), eventually positively impacts the students' effort. However, we are aware of some shortcomings of our IV approach. Although our sample is large, our IV estimates are relatively imprecise, especially for the sub-sample of Centre and North students. To conclude, we acknowledge the limitations of our IV exercise and call for further research to determine better both the magnitude and significance of the relationship between off-site and drop-out status.

References

Acemoglu, D. (2002). Directed technical change. *The Review of Economic Studies*, 69(4):781–809.

Aina, C., Baici, E., Casalone, G., and Pastore, F. (2018). The economics of university dropouts and delayed graduation: a survey. GLO Discussion Paper Series 189, Global Labor Organization (GLO).

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly harmless econometrics: An empiricist's companion*. Princeton university press.

Beine, M., Delogu, M., and Ragot, L. (2020). The role of fees in foreign education: evidence from Italy. *Journal of Economic Geography*, 20(2):571–600.

Belloc, F., Maruotti, A., and Petrella, L. (2010). University drop-out: an italian experience. *Higher education*, 60(2):127–138.

Borjas, G. (1987). Self-selection and the earnings of immigrants. *American Economic Review*, 77(4):531–53.

Bussu, A., Detotto, C., and Serra, L. (2019). Indicators to prevent university drop-out and delayed graduation: an italian case. *Journal of Applied Research in Higher Education*.

Card, D. (1993). Using geographic variation in college proximity to estimate the return to schooling. NBER Working Papers 4483, National Bureau of Economic Research, Inc.

Card, D. (2001). Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5):1127–1160.

Checchi, D. (2000). University education in italy. *International Journal of Manpower*, 21(3-4):177–205.

Dale, S. B. and Krueger, A. B. (2002). Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *The Quarterly Journal of Economics*, 117(4):1491–1527.

Faggian, A., McCann, P., and Sheppard, S. (2007). Human capital, higher education and graduate migration: an analysis of scottish and welsh students. *Urban Studies*, 44(13):2511–2528.

Garibaldi, P., Giavazzi, F., Ichino, A., and Rettore, E. (2012). College cost and time to complete a degree: Evidence from tuition discontinuities. *The Review of Economics and Statistics*, 94(3):699–711.

Johnes, G. and McNabb, R. (2004). Never give up on the good times: student attrition in the uk. *Oxford Bulletin of Economics and Statistics*, 66(1):23–47.

Katz, L. F. and Murphy, K. M. (1992). Changes in relative wages, 1963–1987: supply and demand factors. *The Quarterly Journal of Economics*, 107(1):35–78.

Kazakis, P. and Faggian, A. (2017). Mobility, education and labor market outcomes for us graduates: Is selectivity important? *The Annals of Regional Science*, 59(3):731–758.

Modena, F., Tanzi, G. M., and Rettore, E. (2018). The effect of grants on university drop-out rates: evidence on the italian case. Temi di discussione (Economic working papers) 1193, Bank of Italy, Economic Research and International Relations Area.

OECD (2019). *Education at a Glance 2015: OECD Indicators*. OECD Publishing, Paris.

Rosenzweig, M. R., Irwin, D. A., and Williamson, J. G. (2006). Global wage differences and international student flows [with comments and discussion]. In *Brookings trade forum*, pages 57–96. JSTOR.

Weber, S. and Péclat, M. (2017). A simple command to calculate travel distance and travel time. *The Stata Journal*, 17(4):962–971.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*. MIT press.

Zotti, R. (2015). Should i stay or should i go? dropping out from university: an empirical analysis of students' performances. In *Youth and the Crisis*, pages 71–88. Routledge.

A Appendix

Table 16 provides a brief description, detailing definition and data source, for all the variables employed in our analysis.

Table 16: Data Sources and Definitions

Variable	Definition	Source	Remarks
Dropout ($D_{\{i,u,c,o\}}$)	Dummy variable that takes one when the student drops out from the course/university and zero otherwise	ANS data. Our computation.	i identifies the individual, u, the university, c the field of study, o the origin of the students.
HG_i	Variable capturing the High school grade of student i	ANS data	The minimum grade to obtain a high school title in Italy is equal to 60 with the maximum equal to 100 (however, students may obtain a mention). We scale subtracting 60 to each vote.
AGE_i	$Addyears_i = -1 * (Yearofbirth - 1995)$	ANS data, our computation	Notice that in Italy students usually finish high school at the age of 19.
HT_i	Dummy variable that captures the type of the high school attended by student i	ANS data	The variable takes value equal to one only if the high school is a *Liceo* of the traditional type, either *Classico* or *Scientifico*. For all the rest of High schools, the variable is set equal to zero.
G_i	Dummy variable that captures the gender of the student i. Takes value 1 for man and 0 otherwise	ANS data, Avvii dataset	

Continued on next page

Table 16 – continued from previous page

Variable	Definition	Source	Remarks
$OD_{i,u,o}$	Dummy variable that takes value 1 when the students enrolls in a university not located in his\her district of residence	ANS, data	
$OR_{i,u,o}$	Dummy variable that takes value 1 when the students enrolls in a university not located in his\her region of residence	ANS data	
$TD_{i,u,o}$	Measures the distance between the student *i* place of residence, *o* and the university of destination *u*	ANS data	Our computation employing the routine developed by Weber and Péclat (2017), one unit is equal to 100km.
$GD_{i,u,o}$	Measures the geodesic distance between the student *i* place of residence, *o*, and the university of destination *u*	ANS data	Our computation, one unit is equal to 100 km
$OFF150_{i,u,o}$	Dummy variable that takes value 1 when the student enrolls in a university distant more than150km, in term of travel distance, from his/place of origin	Our calculation from ANS data	
$OFF200_{i,u,o}$	Dummy variable that takes value 1 when the student enrolls in a university distant more than 150km, in term of travel distance, from his place of origin	Our calculation from ANS data	

Ultimi Contributi di Ricerca CRENoS

I Paper sono disponibili in: http://www.crenos.unica.it

20/04 *Emanuela Marrocu, Raffele Paci, David Rigby, Stefano Usai,* "Smart Specialization Strategy: any relatedness between theory and practice?

20/03 *Giorgio Garau, Stefano Deriu,* "Total Factor Productivity and Relative Prices: the case of Italy"

20/02 *Fabio Cerina, Alessio Moro, Michelle Rendall,* "A Note on Employment and Wage Polarization in the U.S."

20/01 *Elias Carroni, Dimitri Paolini,* "Business models for streaming platforms: content acquisition, advertising and users"

19/16 *Daniela Sonedda*, "Regional variation in apprenticeship and permanent employment rates: which causes?"

19/15 *Daniela Sonedda,* "Regional disparities in the functioning of the labour markets"

19/14 *Bianca Biagi, Barbara Dettori, Raffaele Paci, Stefano Usai,* "Economic development in Sardinia: overcoming the insularity gap"

19/13 *Miguel Casares, Luca Deidda, Jose E. Galdon-Sanchez,* "On financial frictions and firm market power"

19/12 *Massimiliano Bratti, Maurizio Conti, Giovanni Sulis,* "Employment Protection and Firm-provided Training: Quasi-experimental Evidence from a Labour Market Reform"

19/11 *Jessica Goldberg, Mario Macis, Pradeep Chintagunta*, "Incentivized Peer Referrals for Tuberculosis Screening: Evidence from India"

19/10 *Julio J. Elías, Nicola Lacetera, Mario Macis,* "Paying for Kidneys? A Randomized Survey and Choice Experiment"

19/09 *Fabio Cerina, Elisa Dienesch, Alessio Moro, Michelle Rendall,* "Spatial Polarization"

19/08 *Michele Battisti, Massimo Del Gatto, Christopher F. Parmeter,* "Skill Biased Technical Change and Misallocation: a Unified Framework"

19/07 *Fabrizio Fusillo, Francesco Quatraro, Stefano Usai,* "Going Green: Environmental Regulation, eco-innovation and technological alliances"

19/06 *Oliviero A. Carboni, Giuseppe Medda,* "External R&D Acquisition and Product Innovation"

19/05 *José J. Cao-Alvira, Luca G. Deidda,* "Development of Bank Microcredit"

19/04 *Vania Licio,* "When history leaves a mark: a new measure of Roman roads"

19/03 *Simone Franceschini, Gerardo Marletto,* "Reorganization of supply chains as a key for the envision of socio- technical transitions. The case of tourism"

19/02 *Elias Carroni, Dimitri Paolini,* "The business model of a streaming platform"

19/01 *Stefania Capecchi, Marta Meleddu, Manuela Pulina, Giuliana Solinas,* "Mixture models for consumers' preferences in healthcare"

18/13 *Adelaide Baronchelli, Teodora Erika Uberti,* "Exports and FDI: Comparing Networks in the New Millennium"

18/12 *Gabriele Cardullo, Maurizio Conti, Giovanni Sulis,* "Unions, Two-Tier Bargaining and Physical Capital Investment: Theory and Firm-Level Evidence from Italy"

18/11 *Jing Guan, J.D. Tena,* "Estimating the Effect of Physical Exercise on Juveniles' Health Status and Subjective Well-Being in China"

18/10 *Silvia Balia, Rinaldo Brau, Daniela Moro,* "Hospital choice with high long-distance mobility"

18/09 *Luca Deidda, Ettore Panetti,* "Banks' Liquidity Management and Financial Fragility" www.crenos.unica.it

ISBN 9788868513047