UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Dipartimento di Economia Marco Biagi

DEMB Working Paper Series

N. 168

Tertiary education decisions of immigrants and non-immigrants in Italy: an empirical approach

Michele Lalla[1], Patrizio Frederic[2]

March 2020

[1] University of Modena and Reggio Emilia and CAPP, Center for the Analysis of Public Policies
Address: Viale Berengario 51, 41121, Modena, Italy
E-mail: michele.lalla@unimore.it
[2] University of Modena and Reggio Emilia and RECent, Center for Economic Research
Address: Viale Berengario 51, 41121, Modena, Italy
E-mail: patrizio.frederic@unimore.it

ISSN: 2281-440X online

Tertiary education decisions of immigrants and non-immigrants in Italy: an empirical approach

Michele Lalla, Patrizio Frederic

Abstract

Decisions regarding tertiary schooling are important for young people as it affects future opportunities for employment and social mobility. Tertiary schooling also plays a role in the social integration of immigrants. To determine differences in the choices of young Italian natives and immigrants concerning education, two datasets for 2009 were used: European Union Statistics on Income and Living Conditions (EU-SILC) and the Italian Survey on Income and Living Conditions of Families with Immigrants in Italy (IT-SILCFI). Analysing a sub-sample of young Italians and immigrants, between 18 and 29 years of age, the association of both individual and family explanatory variables in the choice of secondary schooling (yes/no) was assessed using logistic models. The results show that young immigrants tend to interrupt their schooling earlier than their Italian peers. However, differences disappear when family background and parental characteristics are taken into account.

Keywords: educational inequality, peer effects, educational territorial pattern

JEL codes: I21, I24, I25, J15

Acknowledgements: This research was developed within the project "*Determinant and economic effects of international migration*" of the Department of Economics "Marco Biagi" and the Marco Biagi Foundation, with financial support from the University of Modena and Reggio Emilia UNIMORE FAR 2017 Grant, which is gratefully acknowledged. Moreover, the illustrated results are based on data from Istat and the European Union - Statistics on Income and Living Conditions (EU-SILC). Responsibility for all conclusions drawn from the data lies entirely with the authors.

Index

1. Introduction

Tertiary schooling is not compulsory in almost all educational systems and enrolment decisions represent a difficult step for students because they are making decisions for their future without knowing much about themselves and/or the evolution and needs of society. Such decisions may be affected by differences in individual behaviour or the socio-economic conditions of families. Additionally, such decisions may impact opportunities for future employment and upward mobility, and may also lead to dramatic decreases in their grades (Grove et al., 2006; Wintre et al., 2011; Armstrong and Biktimirov, 2013). All these aspects may differ among immigrant and non-immigrant youths and, in the case of the former, tertiary schooling plays an important role not only in terms of investing in human capital, the cultural formation process, and social integration, but also as an instrument of social mobility and transformation, development through attuned interactions and collective healing through cooperation (Entwisle and Alexander, 1993; Paba and Bertozzi, 2017; De Clercq et al., 2017).

The objective of this paper is to determine the differences between the two groups, immigrants and non-immigrants (hereinafter sometimes referred to as Italians for the sake of simplicity), with respect to their decision to continue on with tertiary education or to interrupt their studies after finishing their upper secondary schooling, taking into account individual, social and demographic characteristics and family background. The data were extracted from two surveys, with reference year 2009, carried out by the Italian National Institute of Statistics (Istat): one being European Union Statistics on Income and Living Conditions (EU-SILC) restricted to Italy only (IT-SILC) – an annual survey since 2004 under the coordination of Eurostat (Istat, 2008; Eurostat, 2009; Atkinson and Marlier, 2010) – and the other being the Italian Survey on Income and Living Conditions of families with Immigrants (IM-SILC), which is a single cross-sectional survey (Istat, 2009a)[1] that involved families with at least one immigrant component resident in Italy.

The binary nature of the dependent variable implies that it is equal to 1 when an individual attends or has achieved tertiary education or a post-tertiary education level, and is equal to zero otherwise, *i.e.*, when he/she has achieved an upper secondary level of education. It directly involves some specific techniques, such as ordinary logistic regression in the classical approach or a Bayesian approach, both of which have been applied here. In the latter case, the independent variables set may be and have been defined using the Lasso method, which simultaneously allows for the selection of the explanatory variables and the estimation of the model coefficients.

The paper is organised as follows. Section 2 concisely describes the theoretical background, and Section 3 illustrates the sample, the data and some descriptive results concerning the main variables used in the subsequent analyses. Section 4 describes the ordinary logistic model and includes comments on the results. Section 5 reports the model obtained through the peculiar Lasso techniques for selection of the independent variable and a Bayesian approach for the estimation of parameters. Finally, Section 6 briefly concludes with some comments and remarks.

[1] Note that the letter S in the acronym EU-SILC is often assumed to mean "Survey", rather than "Statistics". The same has been done here to provide correspondence with the acronym for the Italian Survey on Income and Living Conditions of families with immigrants, where the term "immigrants" refers to individuals without Italian citizenship. The name of the latter appears here as "Immigrant Survey on Income and Living Conditions" (IM-SILC) to obtain a similar structure of the acronyms given to the two surveys.

2. Background

Education decisions that young people face are made at a particular stage in their lives, when the influences of inside and outside the home are strongly felt. In this sense, such decisions strongly depend on both individual and family characteristics, as well as on the social and contextual background of the area where they reside. Therefore, they are examined from multiple points of view, including social and economic conditions, personal and family background, psychological and school-related situations.

The purposes of the present analysis are mainly aimed at obtaining an empirical description of facts and the relationships existing in the target population apart from theories underlying each association and interaction. In fact, each perspectival approach may involve many theories and strategies concerning data collection methods and data analysis. Thus, three main subgroups of explanatory variables have been considered to explain current enrolment in or completion of a tertiary or post-tertiary education.

Firstly, for each young individual, gender, age and health conditions were considered because they have proved to be associated with the choice to continue one's education and training (Lalla and Pirani, 2014; Contini, 2013). In this context, immigrant status and the length of stay in the country also clearly play a role.

Secondly, educational choices reflect and originate from the family context of young people, including both immigrants and non-immigrants. The effect of family background on assimilation and expectations has been thoroughly analysed for both immigrants and Italians, and different factors have been identified as relevant in these processes: household size and family composition, educational level of parents, socioeconomic status, language and expectations of parents, parent support and involvement, cultural background and income. The influence of these factors in the education decisions of young people has also been investigated (Luciano et al., 2009) with an extensive comparison of a group of individuals at various steps of their careers and many explanations have been given for employment and income inequalities (Algan et al., 2010; Zwsyen and Longhi, 2018).

Lastly, the social context of the community and the area of residence may be also relevant. Schooling has been analysed as a source of inequality between immigrants and Italians and/or among different groups of immigrants as well, and includes attending kindergarten, previous experiences of success and failures, advice of teachers and peers, and availability of schools in the area. The school environment can provide strong stimuli for integration in the community as a source of potential comparison with others and induce motivation for all to improve their knowledge and education. The context of the community of residence may refer to social characteristics of the neighbourhood (Pong and Hao, 2007) and to economic characteristics. The former have often been represented considering crime levels, characteristics of peers, companionship and so on, while the economic factors may refer to the employment/ unemployment rate in the area of residence, the local gross domestic product, the value added by sector (Bertolini et al., 2015). The local area may provide an important indicator summarising many effects such as segregation (Sleutjes et al., 2018) and good or bad economic conditions, thus affecting decisions on continuing education. The degree of urbanisation is another useful indicator of the ethnic concentrations, sometimes as a result of people's settlement preferences, as some regions and towns attract more immigrants than others.

3. Data sources and descriptive statistics

In order to obtain a consistent sample and comparable information, two datasets were used, as mentioned above: the IT-SILC sample was considered together with the IM-SILC sample.

3.1. European Union Survey on Income and Living Conditions

The European Community Household Panel (ECHP) was designed to satisfy the need to study social exclusion within the European Union (EU) using a multidimensional approach, to obtain a high level of harmonisation for statistics on income and living conditions in the EU, and to produce a multidimensional dataset centred primarily on income, but also on housing, labour, health, education, demography, and deprivation. The ECHP was launched in 1994 in 14 of the 15 Member States (Sweden was the exception) and ended in 2001 for many reasons, such as technical problems that emerged over time and the need to adapt the survey to the enlargement of the EU which occurred in the 1990s.

Under framework regulation adopted by the European Council (2003) and the European Parliament, a new project termed EU-SILC was planned in 2003 on the basis of a "gentlemen's agreement" (Eurostat, 2005; Wolff et al., 2010) and implemented from 2004 onwards (Eurostat, 2009, p. 15). This annual survey was aimed at gathering information based on nationally representative random samples of private households in each European country concerning individual socio-demographic characteristics, micro-level data on income, poverty, social exclusion and living conditions using a unique sampling design and identical definitions of the concepts used in the ECHP survey.

The EU-SILC covers people living in private households only, excluding persons living in collective households and in institutions, which involves possible under-representation of some vulnerable groups living in private households because they are not easy to reach (Wolff et al., 2010). Therefore, the target population refers to all private households and all persons aged 16 years old and over. The data set is composed of nationally representative probability samples of the population residing in private households within the country, independently of nationality, language, or legal residence status. Representativeness[2] must concern both households and individual persons, so that each individual and each household in the target population should have a known and positive probability of being selected (Wolff et al., 2010). EU-SILC has a rotated sample and the structure recommended by Eurostat has an overlapping proportion of ¾, implying that the initial survey sample (first year/ round) is subdivided into four subsamples (a, b, c, d); in the second year/ round, the subsample (a) is dropped from the sample and replaced with a new fresh sample (e) of equal size drawn from the target population and added to the remaining subsamples, thus obtaining the sample (b, c, d, e). The process continues over time in a similar manner: third year (c, d, e, f), fourth year (d, e, f, g). In the fifth year/ round, the sample is totally renewed, as all four subsamples (e, f, g, h) differ from the initial subsamples.

The total sample in the EU, as a whole, should have a minimum effective size of about 137,000 households, including Iceland and Norway. Allocations among countries

[2] Note that representativeness has no statistical meaning, unless it is intended as a synonym for a probabilistic sample, which has a specific meaning and properties (Cicchitelli et al., 1997; Särndal et al., 1992).

are based on a compromise between the significance of the results at the level of individual countries and the significance of the results at the level of the EU. The longitudinal component has less compulsory requirements and an effective sample size of about 103,000 was planned at the level of the EU. The minimum effective sample size refers to a design effect related to the "risk of poverty rate" variable equal to 1. However, the actual sample sizes should be larger than the minimum to compensate for all kinds of non-response (Eurostat, 2009; Verma and Betti, 2006).

The weighting scheme is described in the EU-SILC guidelines and articles (Eurostat, 2009; Verma et al., 2007). A description of errors has been illustrated by Verma and Betti (2010).

The EU-SILC provides two types of *annual data*: (1) cross-sectional data containing information on the statistical units for each year, (2) longitudinal data concerning individual-level changes over time involving a maximum interval of four years (Eurostat, 2009) as determined by the rotation scheme (see above).

The target variables are distributed in four different groups or datasets, each one grouping different variables: (D) Household Register, (H) Household Data, (R) Personal Register, and (P) Personal Data.

The household register file (D) contains every selected household, implying that it also includes the households that could not be contacted (non-contacts) or that refused to be interviewed (refusals). Moreover, it comprises the substituted and the split-off (longitudinal only) households as well. In the other files, the records associated with a household refer to a contacted household *having* a completed household interview in the household data file (H) *and* at least one member having complete data in the personal data file (P). In the selected reference year 2009, the number of records in the D file was 20,492 and each record corresponded to a family.

The Household Data, (H), contains the core of the information on households, such as the total household gross income, total disposable household income, imputed rent, housing conditions, and so on. Obviously, in the selected reference year 2009, there were 20,492 records again and the data were gathered under the specified conditions and/or restrictions provided by the sampling design.

The personal register file (R) contains a record for every person living in the household or temporarily absent during the interview. In the longitudinal component there should also be a record of every person who has moved out or died since the preceding occasion "and for every person who lived in the household at least three months during the income reference period and was not recorded otherwise in the register of this household" (Eurostat, 2009). It does not contain many data/ variables about individuals. In the selected reference year 2009, the number of records was equal to 51,196 in the R file and each record corresponded to an individual. Of the 51,196 individuals, 43,566 individuals were at least 16 years old.

The personal data file, (P), contains the core of the information about individuals, which should be matched with the other three files to obtain complete information about each individual. In 2009, the number of useful personal records in the P file was equal to 43,636 (individuals at least 16 years old of age), which was naturally lower than the number of records in the personal register file (R).

The selected IT-SILC reference year, 2009, was a necessary choice because the IM-SILC (see below) was carried out by Istat only in that year. The four files were matched to obtain a complete file with information at a different level. Table 1 gives a synthesis of the sample size and the distributions of individuals by the age classes and education levels, grouped according to a restricted form of the International Standard Classification of Education (ISCED) levels (UNESCO, 2012).

In the resulting matched file, the total number of cases was equal to the number in the personal register file (R): 51,196. However, the number of useful and manageable records remained the same as the number of personal records, each one corresponding to an interviewed individual, for a total of 43,636. The 0-14 year age class was empty with respect to the highest ISCED level attained because the individuals under the age of 16 were not interviewed, as established by the survey design. In fact, the 15-19 year age class had many missing data, almost all referring to 15-year-olds because they were not interviewed. Without data weighting, the distribution of age class in the sample was approximately similar to that obtained in the Census carried out in 2011 and reported in Table 1. The sample distribution of the highest ISCED level attained (again without data weighting) was still approximately similar to that obtained in the 2011 Census, even if the differences observed for lower secondary education and for post-secondary non-tertiary education were greater than those observed for the other education levels. Using weights, these differences diminished, but did not disappear, suggesting the possibility of processing the data without using weighting procedures.

Table 1. Absolute frequencies and row percentages of the highest ISCED level attained (ILA) by age classes in the IT-SILC data of 2009

ILA\ Age	**0-14**	**15-19**	**20-29**	**30-39**	**40-49**	**50-64**	**>=65**	**Total**	**C%**	**IT-C%**
PE		39	124	242	464	2322	6813	10004	23.1	23.2
		0.4	1.2	2.2	4.8	25.0	66.5	100		
LSE		1576	972	1822	2827	3123	1821	12141	28.1	31.4
		13.0	8.0	15.0	23.3	25.7	15.0	100		
USE		457	3235	3015	3339	3261	1625	14932	34.5	33.1
		3.1	21.7	20.2	22.4	21.8	10.9	100		
Post-SNTE		6	251	429	364	283	77	1410	3.3	1.3
		0.4	17.8	30.4	25.8	20.1	5.5	100		
TE		1	769	1346	1092	1087	487	4782	11.1	11.0
		0.0	16.1	28.2	22.8	22.7	10.2	100		
Missing	7043	520	12	26	20	35	271	7927		
	88.8	6.6	0.2	0.3	0.3	0.4	3.4	100		
Total	**7043**	**2599**	**5363**	**6880**	**8106**	**10111**	**11094**	**51196**		
	13.8	5.1	10.5	13.4	15.8	19.7	21.7	100	100	100
Census ’11	14.0	4.8	10.6	14.0	16.1	19.5	20.8	100		

Legend: ISCED=International Standard Classification of Education. C%= Column percentage. IT= Italy. PE= Primary Education. LSE= Lower Secondary Education. USE= Upper Secondary Education. Post-SNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education.

3.2. The Italian Survey on Income and Living Conditions of Families with Immigrants

The Italian Survey on Income and Living Conditions of Families with Immigrants (IM-SILC) was funded by the Ministry of Labour and Social Policies and conducted by Istat in 2009. The design of the IM-SILC was similar to the IT-SILC project described above, in terms of contents and methodological aspects. Therefore, it covered immigrant people living in private households and the entire target population referred to private households and persons who were 16 years of age and over. The data set was composed of a one-shot national probability sample of the population residing in private households within the country. Representativeness had to concern both households and individual persons. The specificity of the reference population in the IM-SILC, compared to the IT-SILC, involved numerous expedients to improve the representativeness of the sample: (A) The sample design at the origin of the IM-SILC was based on the extraction of municipalities as primary sampling units, taking into account the distribution of the main groups of foreign nationals in Italy, reducing the risk of excluding some groups of foreign nationals that may be particularly concentrated in some areas. (B) Non-respondent families were replaced with other families of the same citizenship, minimizing self-selection of the most collaborative citizenships and the consequent bias. (C) The questionnaires were translated into the ten most common languages among foreigners residing in Italy, to support the interviewers and facilitate the interviewees' understanding of the questions and encourage their collaboration. (D) The sample was post-stratified at a geographical distribution level, taking into account, in addition to the usual constraints on the known total population, the number of families with immigrants and the foreign population classified into the 13 main nationalities residing in Italy, for better calibration with respect to the reference population (Istat, 2009b). These expedients and a sample of families with immigrants that was more numerous than that surveyed with the IT-SILC made it possible to obtain results that could also be analysed by nationality and area of residence of families with immigrants (Donadio et al., 2014). The IT-SILC representative of the entire population of Italy and therefore including families composed of Italians only, constituted the term of comparison for the results of the living conditions of families with immigrants.

The target variables were distributed in four different groups as in the IT-SILC: (D) Household Register, (H) Household Data, (R) Personal Register, and (P) Personal Data.

The Household Register file (D) contains every selected household, implying that it also includes the households that could not be contacted (non-contacts) or that refused to be interviewed (refusals). Moreover, it comprises the substituted households too, as in the IT-SILC. In the other files, the records associated with a household refer to a contacted household *having* a completed household interview in the Household Data file (H) *and* with at least one member having complete data in the Personal Data file (P). The reference year was 2009 only and the number of records was equal to 6,014 (families).

The Household Data file, (H) contains the core of the information on households, such as total household gross income, total disposable household income, imputed rent, hosing conditions, and so on. Obviously, once again the number of records was equal to 6,014.

The Personal Register file (R) contains a record for every person living in the household or temporarily absent during the interview. The reference year was 2009 only and the number of records was equal to 15,036 (individuals and among them there were 11,535 individuals who were at least 16 years old).

The Personal Data file (P) contains the core of the information about individuals, which was matched with the other three files to obtain complete information about each individual. The number of useful personal records was equal to 11,611 (individuals at least 16 years old), which was lower than the number of the records in the personal register file (R), as above in the case of the IT-SILC.

The four files were matched to obtain a complete file with information at a different level. Table 2 provides a synthesis of the sample size and the distributions of individuals according to the age classes and education levels, grouped in a restricted form, but according to the International Standard Classification of Education (ISCED) levels.

Table 2. Absolute frequencies and row percentages of the highest ISCED levels attained (ILA) by age classes in the 2009 IM-SILC data

ILA\ Age	0-14	15-19	20-29	30-39	40-49	50-64	>=65	Total	C%	IT-C%
PE		98	345	546	508	385	214	2096	18.3	12.8
		4.7	16.5	26.0	24.2	18.4	10.2	100		
LSE		463	929	1182	728	428	85	3815	33.3	37.5
		12.1	24.4	31.0	19.1	11.2	2.2	100		
USE		54	1003	1377	1080	558	75	4147	36.2	40.2
		1.3	24.2	33.2	26.0	13.5	1.8	100		
Post-SNTE		1	16	29	25	8	1	80	0.8	(*)
		1.3	20.0	36.3	31.3	10.0	1.3	100		
TE		3	177	420	358	283	65	1306	11.4	9.5
		0.2	13.6	32.2	27.4	21.7	5.0	100		
Missing	3258	168	12	37	42	44	31	3592		
	90.7	4.7	0.3	1.0	1.2	1.2	0.9	100		
Total	**3258**	**787**	**2482**	**3591**	**2741**	**1706**	**471**	**15036**		
	21.7	5.2	16.5	23.9	18.2	11.3	3.1	100	100	100
Census '11	18.8	4.9	18.6	25.8	18.3	11.2	2.4	100		

Legend: ISCED=International Standard Classification of Education. C%= Column percentage. IT= Italy. PE= Primary Education. LSE= Lower Secondary Education. USE= Upper Secondary Education. Post-SNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education.

(*) Note that this information was not available for public use of the 2011 Census data.

In the resulting matched file, the total number of cases was equal to the number of Personal Register file (R): 15,036. However, the number of useful and manageable records remained the same as the number of personal records, which was 11,611 and each one corresponding to an interviewed individual. Again, the 0-14 year age class was empty with respect to the highest ISCED level attained because the individuals under the age of 16 years old were not interviewed. Consequently, the 15-19 year age class presented many missing data; almost all missing data referred to the 15-year-old. Without data weighting, the sample distribution by age classes was approximately similar to that of the 2011

Census reported in Table 2. The sample distribution of the highest ISCED level attained (again without data weighting) was only roughly similar to that obtained in the 2011 Census. In fact, the observable differences in the levels beyond tertiary education may also appear negligible, but the data refer to a two-year period after the survey date, making them less relevant and only indicative. Using weights, the differences decreased slightly but did not disappear. The differences in the distributions of highest ISCED level attained in the weighted and unweighted data remained negligible, but dichotomisation of education level used in the analysis carried out below eliminated the possible effects of these inequalities. The weighted and unweighted age distributions were adequately similar and the differences were insignificant in the age classes under analysis. Therefore, handling data without the weighting procedures is a possibility.

3.3. The sample selected for tertiary education

The rules for the selection of individuals were based on the identification of factors determining the decision to attend/ achieve tertiary education or to interrupt schooling in order to work or the decision to attend/ achieve a post-secondary non-tertiary education degree. Therefore, a first limit was imposed on the age range of 20 to 29 years because within this range it was possible to distinguish between individuals interrupting their education path and individuals currently attending tertiary education degree programmes or who already had a bachelor's, graduate or doctoral degree. Following this selection, the remaining samples described by age are shown in Table 3, where it possible to note that the IM-SILC has a proportion of 31.6% (given by 100 × 2,482/ 7,845), but the proportion of immigrants was greater than the latter because the IT-SILC contains a representative proportion of immigrants only. However, among the IT-SILC there were low values of frequencies for immigrants in each year of age and some ages were strongly under- or over-represented as was the case for 27 and 28 years of age.

Table 3. Absolute frequencies and row percentages of the type of survey (TOS) by age

TOS\ Age	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**Total**
Non-immigrants	487	548	484	505	551	512	509	515	516	570	5197
	9.4	10.5	9.3	9.7	10.6	9.9	9.8	9.9	9.9	11.0	100
Immigrants	17	13	13	14	14	15	18	8	29	25	166
	10.2	7.8	7.8	8.4	8.4	9.0	10.8	4.8	17.5	15.1	100
IT-SILC Total	504	561	497	519	565	527	527	523	545	595	5363
	9.4	10.5	9.3	9.7	10.5	9.8	9.8	9.8	10.2	11.1	100
IM-SILC	172	183	197	214	235	288	247	320	301	325	2482
	6.9	7.4	7.9	8.6	9.5	11.6	10.0	12.9	12.1	13.1	100
Total	**676**	**744**	**694**	**733**	**800**	**815**	**774**	**843**	**846**	**920**	**7845**
	8.6	9.5	8.9	9.3	10.2	10.4	9.9	10.8	10.8	11.7	100

A description of current education and the achieved education level is provided in Table 4. The individuals who were still studying amounted to 25.8% (given by 100 × 2,024/ 7,845). The data in Table 4 suggest that the ISCED levels lower than 3 (=upper secondary education) should be dropped because our interest is centred on various behaviours concerning tertiary education as opposed to the alternative of avoiding tertiary education involving individuals who had achieved an ISCED=3. The column of missing values should be dropped too.

After dropping the non-eligible cases, which was carried out by applying the previously listed conditions, the remaining sample was that described in Table 5, where there are 11 individuals who had an upper secondary education level and were attending another type of upper secondary school. This peculiar condition led to their elimination. Moreover, the classification distinguishing bachelor's and graduate degrees from other levels was not carried out exemplarily and these degrees were generally aggregated in the category "Tertiary Education" level. Therefore, the *final target sample* was made up of 5,440 individuals. Table 6 illustrates reporting frequencies for the latter and percentages of the highest ISCED levels attained by age. Therefore, summarising, the sample analysed below contains individuals in the 20-29 year age range having at least an upper secondary education level; hereinafter it is referred to simply as the target sample.

Table 4. Absolute frequencies and row percentages for continuing education (CE) by the highest ISCED level attained (ILA)

CE\ ILA	**PPE**	**PE**	**LSE**	**USE**	**PSNTE**	**TE**	**Missing**	**Total**
In Education	3	22	147	1508	59	285	0	2024
	0.2	1.1	7.3	74.5	2.9	14.1	0.0	100
Not in Education	119	325	1754	2730	208	661	24	5821
	2.0	5.6	30.1	46.9	3.6	11.4	0.4	100
Total	**122**	**347**	**1901**	**4238**	**267**	**946**	**24**	**7845**
	1.6	4.4	24.2	54.0	3.4	12.1	0.3	100

Legend: ISCED=International Standard Classification of Education. PPE= Pre-Primary Education. PE= Primary Education. LSE= Lower Secondary Education. USE= Upper Secondary Education. PSNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education.

Table 5. Absolute frequencies and column percentages in target sample for continuing education (CE) by the highest ISCED level attained (ILA)

CE\ ILA	**USE**	**PSNTE**	**BD**	**GD**	**PGMD**	**Ph.D**	**MISS**	**Total**
USE	11	47	101	1349			2730	4238
	100	97.9	94.4	94.8			75.8	77.7
PSNTE			1	58			208	267
			0.9	4.1			5.8	4.9
TE		1	5	16	235	28	661	946
		2.1	4.7	1.1	100.0	100.0	18.4	17.4
Total	**11**	**48**	**107**	**1423**	**235**	**28**	**3599**	**5451**
	100	100	100	100	100	100	100	100

Legend: ISCED=International Standard Classification of Education. USE= Upper Secondary Education. PSNTE= Post-Secondary Non-Tertiary Education. BD= Bachelor's Degree. GD= Graduate Degree. PGMD=Post-Graduate Master's Degree. PhD= Doctor of Philosophy degree. MISS= Missing data. TE= Tertiary Education.

Table 6. Absolute frequencies and column percentages for the highest ISCED level attained (ILA) by age in the *final target sample*

ILA\ Age	20	21	22	23	24	25	26	27	28	29	Total
USE	455	509	469	450	447	408	360	378	350	401	4227
	96.6	95.0	92.7	84.8	80.3	72.2	66.2	68.0	61.7	66.1	77.7
PSNTE	13	21	20	28	27	27	23	30	43	35	267
	2.8	3.9	4.0	5.3	4.9	4.8	4.2	5.4	7.6	5.8	4.9
TE	3	6	17	53	83	130	161	148	174	171	946
	0.6	1.1	3.4	10.0	14.9	23.0	29.6	26.6	30.7	28.2	17.4
Total	**471**	**536**	**506**	**531**	**557**	**565**	**544**	**556**	**567**	**607**	**5440**
	100	100	100	100	100	100	100	100	100	100	100

Legend: ISCED=International Standard Classification of Education. USE= Upper Secondary Education. PSNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education.

3.4. Univariate and bivariate description of the selected sample

The objective dependent variable concerned the highest ISCED level attained, which proved to be a binary variable in the sample and termed *tertiary*, distinguishing between individuals with an upper secondary education (USE) level for whom it assumed a value equal to 0, and individuals who had already achieved a tertiary education (TE) level or who were attending tertiary education courses and for whom it assumed a value equal to 1. Table 7 reports the frequencies for the highest ISCED level attained by the ISCED level currently attended. The main figures emerging concerned individuals who had achieved an upper secondary education level who were not enrolled in continuing education (64.6%), termed "*not-attending*", while only 34.3% of them were currently attending a tertiary education programme. Individuals who had achieved a tertiary education degree and not currently attending continuing education amounted to 69.9%.

Table 7. Absolute frequencies and row percentages of the highest ISCED level attained (ILA) by the ISCED level currently attended (ILCA)

ILA\ ILCA	Not-attending	PSNTE	TE	Master	PhD	Total
USE	2730	47	1450			4227
	64.6	1.1	34.3			100
PSNTE	208		59			267
	77.9		22.1			100
TE	661	1	21	235	28	946
	69.9	0.1	2.2	24.8	3.0	100
Total	**3599**	**48**	**1530**	**235**	**28**	**5440**
	66.2	0.9	28.1	4.3	0.5	100

Legend: ISCED=International Standard Classification of Education. USE= Upper Secondary Education. PSNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education. PhD= Doctor of Philosophy degree.

Table 7 reports the frequencies of the dependent variable, *tertiary education*, whereas its distribution is reported in Table 8.

Table 8. Absolute frequencies and row percentages of tertiary education by the ISCED level currently attended (ILCA)

Tertiary\ ILCA	**Not-attending**	**PSNTE**	**TE**	**Master**	**PhD**	**Total**
Non-tertiary education	2938	47				2985
	98.4	1.6				100
Tertiary education	661	1	1530	235	28	2455
	26.9	0.0	62.3	9.6	1.1	100
Total	**3599**	**48**	**1530**	**235**	**28**	**5440**
	66.2	0.9	28.1	4.3	0.5	100

Legend: ISCED=International Standard Classification of Education. PSNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education. PhD= Philosophy Doctor degree.

The relationship between gender and the ISCED level currently being attended was statically significant (χ^2_4 = 28.185, p<0.000). Women tended to be attending more than men (36.5 versus 30.6), with the exception of Doctor of Philosophy (PhD) degree category (Table 9), in which the percentage of men (60.7%) was greater than that of women (39.3%), highlighting a well-known gender discrimination in favour of men versus women and involving higher positions and even different life choices.

Table 9. Absolute frequencies and row percentages of gender by ISCED level currently attended (ILCA)

Gender\ ILCA	**Not-attending**	**PSNTE**	**TE**	**Master**	**PhD**	**Total**
Men	1713	22	634	84	17	2470
	69.4	0.9	25.7	3.4	0.7	100
Women	1886	26	896	151	11	2970
	63.5	0.9	30.2	5.1	0.4	100
Total	**3599**	**48**	**1530**	**235**	**28**	**5440**
	66.2	0.9	28.1	4.3	0.5	100

Legend: ISCED=International Standard Classification of Education. PSNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education. PhD= Doctor of Philosophy.

The relationship between the non-immigrant/ immigrant condition and the ISCED level currently attended was statically significant (χ^2_3 = 248.300, p<0.000; γ = –0.517, Kendall's tau-b= –0.208). Non-immigrants tended to continue their education more than immigrants (39.7% versus 17.2%), as expected (Table 10), supporting well-known empirical difficulties in studies involving immigrants in the integration process, who are conditioned by scarce economic resources to be used for education.

The relationship between the total indicator of self-perceived health (SPH; see below) and the ISCED level currently attended (Table 11) was not statistically significant (χ^2_9 = 14.683, p<0.100). Individuals presenting some health-related problems or difficulties amounted to 10.5% of the sample. Therefore, the impact of the SPH variable on binary *tertiary* dependent variable was negligible, as expected, *i.e.*, SPH did not affect the higher education decisions to enrol at a university to continue one's education after the upper secondary education level or to avoid tertiary education.

Table 10. Absolute frequencies and row percentages of immigrants by the ISCED level currently attended (ILCA)

Immigrant\ ILCA	Not-attending	PSNTE	TE	PTE	Total
Non-immigrant	2423	40	1308	249	4020
	60.3	1.0	32.5	6.2	100
Immigrant	1176	8	222	14	1420
	82.8	0.6	15.6	1.0	100
Total	3599	48	1530	263	5440
	66.2	0.9	28.1	4.8	100

Legend: ISCED=International Standard Classification of Education. PSNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education. PTE= Post-Tertiary Education.

Table 11. Absolute frequencies and row percentages of the totals for self-perceived health (SPH) by the ISCED level currently attended (ILCA)

SPH\ ILCA	Not-attending	PSNTE	TE	PTE	Total
No difficulty	3186	43	1394	244	4867
	65.5	0.9	28.6	5.0	100
One problem	261	3	80	15	359
	72.7	0.8	22.3	4.2	100
Two problems	88	2	33	2	125
	70.4	1.6	26.4	1.6	100
Three problems	64	0	23	2	89
	71.9	0.0	25.8	2.3	100
Total	**3599**	**48**	**1530**	**263**	**5440**
	66.2	0.9	28.1	4.8	100

Legend: ISCED=International Standard Classification of Education. PSNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education. PTE= Post-Tertiary Education.

The age of fathers according to the non-immigrant/ immigrant condition (hereinafter referred to simply as *immigrants*, a binary variable, which is equal to 1 when the individual is an immigrant, and equal to 0 otherwise) and the ISCED level currently attended (ILCA) is reported in Table 12. The fathers of immigrants were younger than those of non-immigrants by about ten years. The differences were statistically highly significant for both marginal effects (immigrants with $F_{1;\,5432}$ = 58.94 and p<0.000, ILCA with $F_{3;\,5432}$ = 57.20 and p<0.000). Their interaction was also significant ($F_{3;\,5432}$ = 8.32 and p<0.000), implying that there was nonlinearity with respect to the ordinal variable ILCA. For the sake of brevity, the ages of mothers are not reported, but they did reveal the same structure in terms of relationships and significance. The mean ages of the mothers of immigrants were lower than those of non-immigrants by about 3 years.

Table 12. Absolute frequencies, means, and standard deviations (SD) of the ages of fathers of immigrants/ non-immigrants by the ISCED level currently attended (ILCA)

Immigrant\ ILCA	Not-attending	PSNTE	TE	PTE	Total
Non-immigrant, *n*	2423	40	1308	249	4020
Means	*50.2*	*54.4*	*53.3*	*54.9*	*51.6*
SD	11.5	7.2	7.7	9.4	10.4
Immigrant, *n*	1176	8	222	14	1420
Means	*37.0*	*48.4*	*44.1*	*44.9*	*38.3*
SD	10.7	7.8	11.4	15.4	11.2
Total, *n*	**3599**	**48**	**1530**	**263**	**5440**
Means	*45.9*	*53.4*	*52.0*	*54.4*	*48.1*
SD	12.8	7.6	9.0	10.0	12.1

Legend: ISCED=International Standard Classification of Education. PSNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education. PTE= Post-Tertiary Education.

The disposable family income (DFI) per capita (in thousands of euros, DFI divided by the number of components of the family) by immigrants/ non-immigrants and by the ISCED level currently attended (ILCA) is reported in Table 13. The DFI per capita for immigrants on the average was significantly lower than that of non-immigrants by about four thousand euros. In fact, the differences were statistically highly significant for the marginal effects concerning immigrants with $F_{1;\,5432}$ = 17.14 and p<0.000, while the variations of DFI per capita with respect to ILCA showed a borderline p-value ($F_{3;\,5432}$ = 2.06, p<0.103) and their interaction was not significant ($F_{3;\,5432}$ = 0.50, p<0.682).

Table 13. Absolute frequencies, means, and standard deviations (SD) of the disposable family income per capita (in thousands of euros) by immigrant/ non-immigrants and by ISCED level currently attended (ILCA)

Immigrant\ ILCA	Non-attending	PSNTE	TE	PTE	Total
Non-immigrant, *n*	2423	40	1308	249	4020
Means	*12.776*	*11.497*	*12.506*	*13.418*	*12.715*
SD	7.635	6.218	8.808	7.559	8.020
Immigrant, *n*	1176	8	222	14	1420
Means	*8.701*	*6.447*	*8.023*	*11.422*	*8.609*
SD	7.450	3.622	6.105	7.405	7.246
Total, *n*	**3599**	**48**	**1530**	**263**	**5440**
Means	*11.444*	*10.655*	*11.855*	*13.312*	*11.643*
SD	7.812	6.136	8.614	7.550	8.030

Legend: ISCED=International Standard Classification of Education. PSNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education. PTE= Post-Tertiary Education.

The other aspects of income concerning the source of income were not considered explicitly. These would include income from rental property or land, interest, dividends, profit from capital investments, and so on. The other types of income considered in the models were disposable personal income (DPI), the fathers' disposable personal income (FDPI), and the mothers' disposable personal income (MDPI). For the sake of brevity, they are not reported in a table, but they did reveal various structures of relationships and levels of significance. For example, FDPI revealed statistically high significance for both marginal effects (immigrants with $F_{1;5432} =$ 22.54 and p<0.000, ILCA with $F_{3;5432} =$ 8.52 and p<0.000) and for their interaction as well ($F_{3;5432} =$ 4.28 and p<0.000), implying that there was nonlinearity with respect to the ordinal variable ILCA. However, the gap between immigrant and non-immigrant fathers varied by about ten thousand euros. MDPI presented similar statistically significant differences for both marginal effects (immigrants with $F_{1;5432} =$ 12.64 and p<0.000, ILCA with $F_{3;5432} =$ 8.45 and p<0.000), but their interaction showed a borderline p-value ($F_{3;5432} =$ 2.23 and p<0.083). The gap between immigrant and non-immigrant mothers varied by about five thousand euros.

The size of immigrant families proved to be lower than those of non-immigrants and was statistically significant (Table 14) for both marginal effects (immigrants with $F_{1;5432} =$ 7.39 and p<0.007 and ILCA with $F_{3;5432} =$ 14.57 and p<0.000), but their interaction was not significant ($F_{3;5432} =$ 1.13 and p<0.336). Given that the total fertility rate of immigrant women is higher than that of non-immigrants, one might expect that the size of immigrant families would be greater than that of non-immigrants. However, many immigrants stay in Italy without their families and presumably this results in a decrease in the size of immigrant families.

Table 14. Absolute frequencies, means, and standard deviations (SD) of the number of family components by immigrants/ non-immigrants and by ISCED level currently attended (ILCA)

Immigrant\ ILCA	**Not-attending**	**PSNTE**	**TE**	**PTE**	**Total**
Non-immigrant, *n*	2423	40	1308	249	4020
Means	*3.51*	*3.83*	*3.75*	*3.62*	*3.60*
SD	1.23	1.17	1.02	1.06	1.16
Immigrant, *n*	1176	8	222	14	1420
Means	*2.99*	*3.75*	*3.39*	*2.93*	*3.06*
SD	1.49	1.16	1.63	1.38	1.51
Total, *n*	3599	48	1530	263	5440
Means	*3.34*	*3.81*	*3.70*	*3.58*	*3.45*
SD	1.34	1.16	1.13	1.09	1.28

Legend: ISCED=International Standard Classification of Education. PSNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education. PTE= Post-Tertiary Education.

The relationship between the immigrant/ non-immigrant condition and the maximum ISCED level attained by parents was statistically significant ($\chi^2_6 = 468.218$, p<0.000; $\gamma = 0.345$, Kendall's tau-b= 0.171). The two-sample Kolmogorov-Smirnov test

(K-S) for equality of distribution functions showed that they were statistically different (combined K-S= 0.280, p<0.000). Immigrants attained upper secondary education levels more frequently than non-immigrants (70.3% versus 43.1%, Table 15) as expected because other empirical findings have revealed this tendency (Bertolini and Lalla, 2012; Bertolini et al., 2015). In fact, this was also the case with professional qualifications achieved through post-secondary non-tertiary education (2.4% versus 0.7%). This behaviour is reflected in post-tertiary education too, as immigrants tended to avoid this type of education (0.9% versus 3.1%), seeking employment immediately after a degree because of scarce economic resources compared to non-immigrants.

Table 15. Absolute frequencies and row percentages for immigrants (IMM.)/non-immigrants by the maximum ISCED level attained by parents (MPILA)

IMM.\ MPILA	PE	LSE	V/PE	USE	PSNTE	TE	PTE	Total
Non-immigrant	289	971	314	1731	26	563	126	4020
	7.2	24.2	7.8	43.1	0.7	14.0	3.1	100
Immigrant	40	79	40	998	34	217	12.0	1420
	2.8	5.6	2.8	70.3	2.4	15.3	0.9	100
Total	**329**	**1050**	**354**	**2729**	**60**	**780**	**138**	**5440**
	6.1	19.3	6.5	50.2	1.1	14.3	2.5	100

Legend: ISCED=International Standard Classification of Education. PE= Primary Education or lower. LSE= Lower Secondary Education. V/PE= Vocational/ Professional Education. USE= Upper Secondary Education. PSNTE= Post-Secondary Non-Tertiary Education. TE= Tertiary Education. PTE= Post-Tertiary Education.

The relationship between the immigrants and the degree of urbanisation (DOU), without the two missing values (see Table 16), was statistically significant ($\chi_2^2 = 34.511$, p<0.000; $\gamma = -0.107$). The two-sample Kolmogorov-Smirnov test (K-S) for equality of distribution functions showed that the distributions of immigrants and non-immigrants were statistically different (combined K-S= 0.076, p<0.000). Immigrants tend to settle in densely populated areas more than non-immigrants (38.4% versus 35.7%) or in intermediate areas (44.5% versus 39.6%). As expected, the reverse was observed in the thinly populated areas (17.1% versus 24.7%), as it may be seen in Table 16. In fact, integration for immigrants is facilitated in highly populated cities.

Table 16. Absolute frequencies and row percentages of immigrants/ non-immigrants by the degree of urbanisation (DOU)

Immigrant\ DOU	High	Medium	Low	Missing	Total
Non-immigrant	1434	1593	991	2	4020
	35.7	39.6	24.7	0.1	100
Immigrant	545	632	243		1420
	38.4	44.5	17.1		100
Total	**1979**	**2225**	**1234**	**2**	**5440**
	36.4	40.9	22.7	0.0	100

The ISCED level currently attended was equally strongly related with the degree of urbanisation, without the two missing values ($\chi_6^2 = 17.821$, p<0.007). The strength of the relationship was weak, but significant: as the density of the area increases, the ISCED level currently attended increases. These results are not reported in a table.

The relationship between immigrants and Italian macro-regions (IMR), a geographical subdivision of Italy into five zones (the North-West, North-East, Centre, South, and the Islands) was statically significant ($\chi_4^2 = 128.362$, p<0.000; $\gamma = -0.118$). The immigrants tended to establish themselves in the North-East (25.2%), in the Centre (23.9%) where Rome attracts many immigrants, and in the North-West (22.0%). The data are reported in Table 17.

Table 17. Absolute frequencies and row percentages of immigrants/ non-immigrants by the ordinary geographical subdivision of Italy (IMR)

Immigrant\ IMR	**NW**	**NE**	**Centre**	**South**	**Islands**	**Total**
Non-immigrant	700	834	910	1166	410	4020
	17.4	20.8	22.6	29.0	10.2	100
Immigrant	313	358	340	204	205	1420
	22.0	25.2	23.9	14.4	14.4	100
Total	**1013**	**1192**	**1250**	**1370**	**615**	**5440**
	18.6	21.9	23.0	25.2	11.3	100

The ISCED level currently attended was equally strongly related with the Italian macro-regions ($\chi_{12}^2 = 63.732$, p<0.007). The strength of the relationship was weak, but significant: as industrialisation and the possibility to find employment increases, the percentage of individuals continuing their education decreases. In fact, over 30% of individuals were attending a tertiary or post-tertiary education level in the South, with respect to an expected 25.2%. These results are not reported in a table.

The relationship between immigrants and the index summarising the total self-perceived health of parents (SPHP), expressed as the number of health-related problems, was statistically significant ($\chi_3^2 = 248.787$, p<0.000; $\gamma = -0.561$). The two-sample Kolmogorov-Smirnov test (K-S) for equality of distribution functions showed that they were statistically different (combined K-S= 0.213, p<0.000), implying that when SPHP increased (*i.e.*, the number of problems), the percentage of non-immigrants decreased and was always higher than that of immigrants, although in slight nonlinear way. For example, immigrants with parents without problems showed a percentage greater than that of non-immigrants: 89.5% versus 68.2% (Table 18).

The ISCED level currently attended (ILCA)[3] was also related to the total self-perceived health of parents (SPHP): $\chi_9^2 = 38.356$, p<0.000, $\gamma = 0.142$. The strength of the relationship between the two variables, ILCA and SPHP, was weak, but significant: as the number of problems of the total SPHP increases, the number of individuals in education increases. These results are not reported in a table.

[3] Note that some acronyms are defined at various points to facilitate the reader.

Table 18. Absolute frequencies and row percentages of the immigrants/non-immigrants by the total self-perceived health of parents (SPHP) expressed as the number of health-related problems (PR)

Immigrant\ SPHP	**Zero PR**	**One PR**	**Two PRs**	**Three PRs**	**Total**
Non-immigrant	2741	606	469	204	4020
	68.2	15.1	11.7	5.1	100
Immigrant	1271	68	47	34	1420
	89.5	4.8	3.3	2.4	100
Total	**4012**	**674**	**516**	**238**	**5440**
	73.8	12.4	9.5	4.4	100

The relationship between immigrants and the self-declared current "main activity status" of parents (MASP) was weak, but statistically significant ($\chi_4^2 = 232.887$, p<0.000; $\gamma = 0.025$), as may be seen in Table 19. Note that MASP expressed a synthesis of the positions of the father and the mother intended to capture the individual's own perception concerning the main activity of parents. Therefore, its definition differed from the (International Labour Organization) ILO definition (Eurostat, 2009). Immigrants presented percentages lower than those of non-immigrants for the category "both parents employed" and the category "at least one parent is retired": 23.1% and 1.2% versus 30.9% and 14.8%, respectively. Immigrants presented percentages greater than those of non-immigrants for "employment of father only" and for "employment of mother only": 20.8% and 1.2% versus 14.8% and 12.9%, respectively.

Table 19. Absolute frequencies and row percentages of immigrants/ non-immigrants for the main activity status of parents (MASP)

Immigrant\ MASP	**Both**	**Father**	**Mother**	**Retired**	**Others**	**Total**
Non-immigrant	1242	1267	596	517	398	4020
	30.9	31.5	14.8	12.9	9.9	100
Immigrant	328	597	295	17	183	1420
	23.1	42.0	20.8	1.2	12.9	100
Total	1570	1864	891	534	581	5440
	28.9	34.3	16.4	9.8	10.7	100

Legend: Both= Both parents are employed. Father= Only the father is employed. Mother= Only mother is employed. Retired= At least one parent is retired.

The relationship between immigrants and the maximum position of parents (MPOP), expressed as the maximum position of the father and mother, was statistically significant ($\chi_5^2 = 726.857$, p<0.000; $\gamma = 0.784$), without missing values. The two-sample Kolmogorov-Smirnov test (K-S) for equality of distribution functions yielded a strong statistically significant difference (combined K-S= 0.492, p<0.000), implying that when MPOP increases (*i.e.*, when one of the parents has a high position), the percentage of non-immigrants increases, although in a slight nonlinear way. For example, there was a lower percentage of immigrants in managerial positions with respect to non-immigrants: 0.6% versus 3.1% (Table 20). The difference for the position of executive director was 1.1%

versus 6.5%. Note that employment position had a reverse order with respect to its importance involving a positive γ coefficient, otherwise it would have been negative.

The ISCED level currently attended (ILCA) was related with the maximum position of parents (MPOP): $\chi^2_{15} = 317.780$, $p<0.000$, $\gamma = -0.468$. The strength of the relationship between the two variables, ILCA and MPOP, was strong and significant: as the employment position increases, the ISCED level currently attended increases, implying that the higher the education level of parents, the more frequently their children attended or achieved high education levels. As above, the negative sign of the γ coefficient depended on the reverse order of the MPOP variable (results not reported in a table).

Table 20. Absolute frequencies and row percentages of immigrants/non-immigrants by the maximum position of parents (POP)

Immigrant\ POP	MAN	EXEC	EMPL	LAB	APPR	HW	MISS	Total
Non-immigrant	123	261	1168	872	16	2	1.578	4.02
	3.1	6.5	29.1	21.7	0.4	0.1	39.3	100
Immigrant	8	15	131	891	22	5	348	1420
	0.6	1.1	9.2	62.8	1.6	0.4	24.5	100
Total	131	276	1299	1763	38	7	1926	5440
	2.4	5.1	23.9	32.4	0.7	0.1	35.4	100

Legend: MAN= Manager. EXEC= Executive director. EMPL= Employee. LAB= Labourer. APPR= Apprentice. HW= Home Worker. MISS= Missing data.

The last variables examined combine various concepts and may be considered as the working conditions of parents (WCP). The two variables, one referring to the father and the other referring to the mother, were rearranged into a single variable with few modalities, as illustrated in Table 21. The relationship between immigrants and the WCP, expressed as a combination of the working conditions of both parents, was statistically significant ($\chi^2_5 = 376.668$, $p<0.000$; $\gamma = -0.237$). The negative sign of the γ coefficient depended on the WCP variable, implying that moving rightwards on the columns lowers the percentage of immigrants. The reliability of WCP was low because it proved to be incoherent with other similar variables for some categories such as the type of employment contract or the work hours. In fact, the other few variables related to the labour conditions of parents often represented subjective perceptions and statements, which may differ in other similar or overlapped items, as is well known.

The ISCED level currently attended (ILCA) was related to WCP: $\chi^2_{15} = 56.476$, $p<0.000$, $\gamma = 0.105$. The strength of the relationship between the two variables, ILCA and WCP, was weak but significant. These results are not reported in a table here.

In summary, non-immigrants tended to be attending a tertiary education level more than immigrants (39.7% versus 17.2%), as expected (Table 10), supporting the evidence of empirical difficulties of immigrants in the integration process and scarce economic resources available to invest in human capital. The parents of immigrants were younger than those of the non-immigrants: fathers on the average were about ten years younger and mothers were about three years younger (Table 12) than those of non-immigrants, respectively. on the average, disposable family income (DFI) per capita for immigrants was significantly lower than that of non-immigrants by about four thousand euros (Table 13). Compared to the parents of non-immigrants, the fathers of immigrants had a lower income by about ten thousand euros and the income of mothers was five thousand euros less, respectively. The immigrants attained an ISCED level that was generally lower than those of non-immigrants (Table 15). The immigrants tended to settle in densely populated areas more than non-immigrants (Table 16) to be facilitated in integration paths and socially with other immigrants. They established themselves prevailingly in the Centre-North of Italy. With respect to the self-perceived health of parents, immigrants had parents without problems at a percentage greater than that of non-immigrants (Table 18). Immigrants revealed lower percentages than those of non-immigrants in the case of both parents being employed and at least one parent who had retired, while the percentages for the employment of only the father and the employment of only the mother were higher than those of non-immigrants (Table 19).

Table 21. Absolute frequencies and row percentages of immigrants/ non-immigrants by working conditions of parents (WCP)

Immigrant\ WCP	**FTD**	**FTSE**	**PTDSE**	**F/PTS**	**PENS**	**U-OLF**	**Total**
Non-immigrant	1640	542	126	756	590	366	4020
	40.8	13.5	3.1	18.8	14.7	9.1	100
Immigrant	803	140	136	140	25	176	1420
	56.6	9.9	9.6	9.9	1.8	12.4	100
Total	2443	682	262	896	615	542	5440
	44.9	12.5	4.8	16.5	11.3	10.0	100

Legend: FTD= Full-Time Dependent worker: one or both parents. FTSE= Full-Time Self-Employed worker: one or both parents. PTDSE= Part-Time Dependent or Self-Employed worker: one or both parents. F/PTS= Full-/ Part-Time dependent or Self-employed worker including the remaining possible combinations. PENS= Pensioner: one or both parents. U-OLF= Unemployed or Out of Labour Force.

3.5. The selected variables for choice models

The data set contained many variables describing different aspects of each individual: personal, family (separately for father and mother), and household information. Given the mix of variables, a factor analysis might have aggregated them into a reduced set. However, in general, there are difficulties in understanding and interpreting these factors. Therefore, only the original variables illustrated below were included in the models, sometimes with modifications and/or adaptations.

Gender was dichotomised in 0 (men) and 1 (women) and termed *women*.

Age was introduced into the model through a second-degree polynomial form $(a\,x^2 + b\,x + c)$ to capture some nonlinearities in the behaviours of individuals of different ages and income values. However, the impact of age on the choice of tertiary education was expected to be ineffective, while the ages of mothers or fathers were expected to be significant in some way. In the models, to have age values comparable with other regressors, which were prevailingly binary variables, the original age values were divided by 10, although it is not necessary for the age of individuals in the 20-29 year age range, while it was useful for the ages of parents.

Nationality was distinguished as non-immigrant (0) or immigrant (1) and the binary variable was termed *immigrants*. IT-SILC made it possible to estimate the Italian and immigrant population, while IM-SILC estimated the total immigrant population only. Therefore, the weights referring only to immigrants were divided by 2, given that the target sample was obtained by appending the IT-SILC sample and the IM-SILC immigrant sample.

Income concerned many variables and components and was introduced into the model through a second-degree polynomial form $(a\,x^2 + b\,x + c)$ again, to capture some nonlinearities in the behaviours of individuals of different ages and income values. However, the expected impact of income on the choice of tertiary education may represent an intriguing issue because disposable personal income (DPI) should have a negative impact on the choice of tertiary education, as a student may be more interested in going to work rather than attending a degree programme. The income of parents or of his/her family may have a positive impact because if disposable family income increases, then the probability of choosing to attend tertiary education school or to achieve tertiary education level should increase. The variables considered were the father's disposable personal income (FDPI), the mother's disposable personal income (MDPI), disposable family income (DFI), and family income per capita (FIPP). The income variables were mutually correlated, and the correlation coefficients differed significantly from zero, but the values were surprising low, except for the coefficient between the total income of the family and the father's income (r=0.776, p<0.000). In the models, the original income values were divided by 20,000 to obtain income values comparable with other regressors, which were binary variables. However, DPI should be used in the model with caution because its value was zero in the case of 1611 individuals (29.6%) and 1037 of the latter (64.4%) had achieved or were currently attending a tertiary level of education.

Individual health data may presumably be neglected because young people are generally in good health. In fact, the self-perceived health (SPH) was measured through a Likert scale (1=very good, 2=good, 3=fair, 4=bad, 5=very bad) and the individuals having a bad or very bad perception constituted only 1.2% of the sample. The intermediate category 'fair', which was the neutral term (neither good, nor bad), as suggested by Eurostat guidelines (Eurostat, 2009), was selected by 5.3% of the individuals. To explore the effect of SPH, the Likert scale was dichotomised into SPH1 assuming the value of 1 when SPH was problematic (6.4%), *i.e.*, when the answer was fair or bad or vary bad, and the value of 0 otherwise. Another binary variable was SPH2, which was equal to 1 when the individual suffered from any chronic (long-standing) illness or condition (5.2%), and equal to 0 otherwise. Another indicator, SPH3, was equal to 1 when the individual was subject to limitations in activities because of health problems (5.1%) and equal to 0 otherwise. The unmet need for medical treatment or examination (5.0%) and the unmet need for dental examination or treatment (8.1%) were not included in the classical logistic model, in order to reduce the set of the explanatory variables, but also because this information should be captured by the income of the family. The three binary variables were summed to obtain a unique indicator of the SPH of each individual *i*, $\mathrm{SPH}_i = \mathrm{SPH}_{1i} + \mathrm{SPH}_{2i} + \mathrm{SPH}_{3i}$, ranging from 0 to 3 (Table 22). Presumably, SPH may only provide a generic indication. Therefore, in the first step of the model estimation SPH1, SPH2, and SPH3 were included first and directly in the model.

The local and geographical variables were limited to two variables.

The first was the set-up of the macro-region (MR) subdivision of Italy. The North-West of Italy (NW) was a binary variable equal to 1 when it included Valle d'Aosta, Piedmont, Liguria, and Lombardy. The North-East of Italy (NE) was a binary variable equal to 1 when it included Trentino Alto Adige, Veneto, Friuli Venice Giulia, and Emilia-Romagna. The Centre of Italy (C) was a binary variable equal to 1 when it included Tuscany, Umbria, Marche, and Latium. The South of Italy (S) was a binary variable equal to 1 when it included Abruzzo, Molise, Campania, Basilicata, Puglia, and Calabria. The Islands of Italy (I) constituted a binary variable equal to 1 when they included Sicily and Sardinia (Table 22).

The second variable concerned the degree of urbanisation (DOU), providing three modalities, transformed into three dummies: (1) DOU_HD being equal to 1 when the location had a high density of population, (2) DOU_MD being equal to 1 when the location had a medium density, and (3) DOU_LD being equal to 1 when the location had a low density. The three binary variable was equal to 0, otherwise (Table 22). MR and DOU variables may prefigure a sort of embryonal segregation (Andersson et al., 2018).

The health of parents may strongly affect the decision to continue one's education. However, given that the SPH options of bad or very bad for mothers (4.8%) and fathers (4.9%), together with the SPH of parents, SPH-P1, was only 8.0%, a global indicator of the health of the family was generated and introduced into the models. Here, the item "3=fair" or neither good nor bad, was not included because the frequencies were high for both fathers (22.4%) and mothers (23.0%). The SPH-P2 denoted parents suffering from a chronic (long-standing) illness or condition: 15.9% (fathers) and 14.1% (mothers). SPH-P3 referred to parents subjected to limitations in activities because of health problems: 17.7% (fathers) and 18.6% (mothers). The three binary variables were summed to obtain a unique indicator of the SPH of the parents, SPH-P, of each individual *i*, $\mathrm{SPH\text{-}P}_i = \mathrm{SPH\text{-}P}_{1i} + \mathrm{SPH\text{-}P}_{2i} + \mathrm{SPH\text{-}P}_{3i}$, ranging from 0 to 3, as above (Table 22).

Table 22. Mean (M) and standard deviation (SD), minimum (Min) and maximum (Max) of the individual variables examined in the models by tertiary education (TE) level with the indication of the reference group [RG]

Variables\ TE	**Tertiary=0**		**Tertiary=1**			
	M	**SD**	**M**	**SD**	**Min**	**Max**
Individual characteristics						
Women (men [**RG**])	0.490	0.500	0.614	0.487	0	1
Age	2.497	0.284	2.428	0.284	2	2.9
Immigrant (non-immigrant [**RG**])	0.464	0.443	0.266	0.445	–1.32	6.51
DPI= (Disposable personal income)/ 20,000	0.341	0.474	0.164	0.370	0	1
Self-perceived health (SPH): binary	0.079	0.270	0.046	0.209	0	1
Suffer from chronic illness/ condition	0.050	0.217	0.054	0.226	0	1
Limitations in activities because of health pr.	0.055	0.228	0.046	0.210	0	1
Total self-perceived health	0.179	0.555	0.139	0.494	0	3
Number of components of the family (**NCF**)	3.389	1.356	3.534	1.184	1	10
Area of residence						
Macro-region: North-West	0.195	0.396	0.176	0.381	0	1
Macro-region: North-East	0.228	0.420	0.208	0.406	0	1
Macro-region: Centre [**RG**]	0.229	0.421	0.230	0.421	0	1
Macro-region: South	0.228	0.419	0.281	0.450	0	1
Macro-region: Islands	0.120	0.325	0.105	0.307	0	1
Degree of urbanisation: high density	0.333	0.471	0.401	0.490	0	1
Degree of urbanisation: average density	0.420	0.494	0.396	0.489	0	1
Degree of urbanisation: low density [**RG**]	0.247	0.431	0.202	0.402	0	1
Parental characteristics						
Father’s age	4.543	1.277	5.132	1.036	2	8.1
Mother’s age	4.291	1.286	4.880	1.073	1.9	8.1
Maximum education level of parents: years	11.5	3.30	13.58	4.35	0	22
FDPI= (Father’s DPI)/ 20,000	0.964	0.723	1.347	1.106	–1.90	17.53
MDPI= (Mother’s DPI)/ 20,000	0.449	0.596	0.697	0.788	–0.41	14.37
FTI= (Family’s total income)/ 20,000	1.758	1.244	2.177	1.602	–0.50	21.19
FIPP= Family income per capita= FTI/ NCF	0.541	0.363	0.633	0.439	–0.31	6.69
Maximum total self-perceived health	0.411	0.806	0.486	0.870	0	3
Sample size	*n* =	**2985**	*n* =	**2455**	**Total**	**5440**

The educational attainment of parents was defined as a unique variable representing the highest level of an education programme that the person (the mother or father) had successfully completed. The value of the variable “parents’ education level” was the highest ISCED level (UNESCO, 2012) of either the mother or the father, but it was transformed in years officially required to achieve the attained level.

The characteristics concerning the labour market situation of the parents involved several categorical variables, which were transformed into binary variables as indicated in Table 23. The parents’ activity status (PAS) as defined by Eurostat guidelines (2009) provided four possible options, as reported in Table 19: employed, unemployed, retired, and other. The combination of the father’s and mother’s conditions generated five disjoint binary variables: (1) PAS-FM equal to 1 when the father and mother were both employed and 0 otherwise, (2) PAS-F equal to 1 when only the father was employed and 0 otherwise, (3) PAS-M equal to 1 when only the mother was employed and 0 otherwise, (4) PAS-R equal to 1 when at least one of the parents was retired and 0 otherwise, (5) PAS-O equal 1 when both parents were classifiable under other conditions and 0 otherwise (Table 23).

The employment position of parents (POP) was set-up retaining the maximum between the positions of the father and the mother. The resulting categorical variable (see Table 20) was transformed into the following binary variables: (1) POP_ME if at least one of the parents was a manager or executive director and the other in a lower position, (2) POP_E if at least one of the parents was an employee and the other in a lower position, (3) POP_L if at least one of the parents was a labourer and the other was unemployed, (4) POP_A if at least one of the parents was an apprentice or a home worker, and (5) POP_O was the residual category containing any other situations not included in the previous binary variables (Table 23).

The working conditions of parents (WCP) was not entirely reliable, but it was constructed combining the conditions of the father and those of the mother. The resulting categorical variable (see Table 21) was transformed into the following binary variables: (1) WCP_FTD if only one or both parent were full-time dependent workers, (2) WCP_FTSE if only one or both parents were full-time self-employed workers, (3) WCP_PT if only one or both parents were part-time dependent or self-employed workers, (4) WCP_MIX if only one or both parents were full-/part-time dependent or self-employed workers but different from the previous WCP binary variables, (5) WCP_PENS if at least one of the parents was a pensioner and the other was employed part-time, unemployed or out of labour force, and (6) WCP_O if at least one of the parents was unemployed or out of the labour force (Table 23).

The type of job contract of parents (TOC-P) was a binary variable equal to 1 when at least one parent had a work contract of unlimited duration. The job skill level and the domain specialization of parents was reduced to two binary variables: one was equal to 1 when at least one of the parents was a highly skilled employee (HSE-P) or had a white collar job, and 0 otherwise; another was equal to 1 when at least one of the parents was medium-skilled or low-skilled or unskilled (LSE-P), and 0 otherwise (Table 23).

In summary, the main independent variables described above and included in the models for the data analysis were subdivided into three categories, as follows: (1) The *socio-demographic characteristics of the young people* were gender, age, immigrant status, disposable personal income, general indicator of self-perceived health, and the number of family components (Table 22). (2) The geographic *area of residence* was simply defined by the macro-region of residence and the degree of urbanisation due to a scarcity of detailed information on this topic (Table 22). (3) The *parental and family characteristics* consisted of the general information reported in Table 22: father’s age, mother’s age, maximum education level of parents, family’s total income, family income per capita, and a general indicator of self-perceived health. The labour market information reported in Table 23 mainly concerned the parent’s activity status, their employment positions, and job conditions.

Table 23. Mean (M) and standard deviation (SD), minimum (Min) and maximum (Max) of the labour market category variables for parents, transformed into binary variables and examined in the models, by tertiary education level with the indication of the reference group [RG]

Variables\ TE	**Tertiary=0**		**Tertiary=1**			
	M	**SD**	**M**	**SD**	**Min**	**Max**
Parents' activity status (PAS)						
PAS: Both employed	0.263	0.440	0.320	0.466	0	1
PAS: Only father employed	0.383	0.486	0.293	0.455	0	1
PAS: Only mother employed	0.154	0.361	0.176	0.381	0	1
PAS: at least one pensioner (*****)	0.113	0.317	0.136	0.3412	0	1
PAS: Other conditions [**RG**]	0.105	0.306	0.109	0.312	0	1
Employment position of parents (POP)						
POP: Manager or executive director	0.034	0.181	0.125	0.330	0	1
POP: Employee	0.194	0.396	0.293	0.455	0	1
POP: Labourer	0.416	0.493	0.212	0.409	0	1
POP: Apprentice [**RG**]	0.008	0.087	0.006	0.078	0	1
POP: Home worker [**RG**]	0.002	0.045	0.000	0.020	0	1
Job conditions of parents (WCP)						
WCP: Full-Time Dependent – one or both	0.480	0.500	0.412	0.492	0	1
WCP: Full-Time Self-employed – one or both	0.130	0.336	0.120	0.325	0	1
WCP: Part-Time – one or both	0.052	0.223	0.043	0.203	0	1
WCP: Full-/ Part-Time – residuals	0.138	0.345	0.197	0.398	0	1
WCP: Pensioner – one or both (*****)	0.113	0.317	0.136	0.3412	0	1
WCP: Unemployed/ Non-Labour Force [**RG**]	0.094	0.292	0.106	0.308	0	1
Other job information (OWI)						
OWI: Highly-skilled employed	0.228	0.420	0.418	0.493	0	1
OWI: Low-skilled employed	0.097	0.296	0.047	0.212	0	1
OWI: Type of contract	0.445	0.497	0.525	0.499	0	1
Sample sizes	*n* =	**2985**	*n* =	**2455**	**Total**	**5440**

Note (*) The two binary variables were suitably set up by pooling the original data to be coincident.

4. Outcomes of the logistic model

The decision to achieve or attend tertiary level of education was analysed for young immigrants and non-immigrants. A binary variable, Y, denoting the dichotomised choice with respect to tertiary education, “attained or attending a tertiary level of education ” (y=1) versus “an upper secondary level of education ” (y=0) was considered for each individual i in the sample $(i=1,\ldots,n)$ with respect to a vector of covariates $\mathbf{x}_i$. Let π_i be the probability that Y=1 depending on the vector of covariate values $\mathbf{x}_i$. The logit model is

$$\pi_i = \frac{\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x}_i'\boldsymbol{\beta}\right)} \tag{1}$$

The vector of coefficients $\boldsymbol{\beta}=\left(\beta_0,\ldots,\beta_K\right)$ describes the effect of the covariates $\mathbf{x}_i=\left(1,x_{i1},\ldots,x_{iK}\right)$ on $\pi_i=\pi_i(\mathbf{x}_i,\boldsymbol{\beta})$, for $i=1,\ldots,n.$

A multinomial logit model might be an alternative to the logistic model (Nguyen and Taylor, 2003) assuming as the dependent variable a combination of the ISCED level attained (UNESCO, 2012) and the status of currently attending an education programme, involving different disjoint modalities: (a) non-attending with an upper secondary education indicated by ISCED level 3 and including the post-secondary non-tertiary education as well (attending and non-attending, with ISCED level 4) because they proved to be few in number, (b) attending a short-cycle tertiary education programme corresponding to ISCED level 5, (c) attending a bachelor’s or equivalent degree programme termed ISCED level 6, (d) attending post-tertiary education programme defined as ISCED level 7 and including individuals attending doctoral degree programmes denoted by ISCED level 8, (e) non-attending having attained a tertiary education programme with ISCED equal to 5, and (f) non-attending having achieved a tertiary education level with ISCED equal to 7 and 8. However, the unreliability of the classification of students attending an ISCED level equal to 5 and 6 and the low sizes of ISCED levels higher than 4 led to a reduction of these six groups down to three groups: (*i*) individuals who had achieved an upper secondary education only or ISCED equal to 3 and 4, (*ii*) attending a tertiary education programme designated by an ISCED higher than 4, and (*iii*) having attained a tertiary level of education an ISCED level higher than 4. Moreover, with respect to the decision to continue one’s education following achievement of an ISCED level equal to 3 or 4, the distinction of the last two groups seemed to be not important for the aims of this study. Therefore, the target variable was binary, distinguishing the group (*i*), for which it a value equal to 0 was assumed, from the other two groups, (*ii*) and (*iii*), for which a value equal to 1 was assumed, and leading to the simple logistic model.

The regressors were selected based on the literature, the available data described above and specifically reported in Table 22 and Table 23, and depending on their statistical significance. In fact, the actual set of regressors did not contain the binary variables assumed as the reference group (RG) because they were derived from polytomous categorical variables: men, non-immigrants, macro-region Centre, low degree of urbanisation, parent activity status (PAS) defined as others, employment position of parenst (POP) equal to apprentice or home worker, job conditions of parents

(WCP) equal to unemployed or out of the labour force. In addition, WCP equal to "one or both pensioners" (WCP_P) was not included because it is equivalent to the binary variable PAS_P denoting a PAS equal to pensioner. However, the latter was suitably combined with the former, *i.e*., PAS_P was assumed equal to 1, when it was equal to 0 and the WCP_P was equal to 1. Another critical point concerned missing values for the father's and/or mother's disposable personal income. The simple solution adopted in order to avoid losing cases consisted in replacing them with the family's total income. This set of variables proved to be satisfactory immediately. However, the binary variable "parents in the high-skilled employment condition" was automatically eliminated by the STATA (2005) estimation procedure because it generated collinearity. Moreover, the variables with a p-value greater than 0.5 were eliminated to simplify reporting the results in a table, as the number of variables in the set was high. The variables involved in this step were: the number of components of the family, the three parent indicators of self-perceived health, parents in the low-skilled job condition, type of contract, the squared term of the father's age, PAS equal to both father and mother being employed, POP equal to manager or executive director positions, and POP equal to employee.

Considering the set of regressors in the model, the final reference group consisted of young man, non-immigrant, living in the Centre of Italy, in a location with a low degree of urbanisation, with a very good or good self-perceived health, not suffering from any chronic illness or condition, without limitations in activities due to health problems, having parents in the following conditions: PAS equal to father and mother being employed or "other conditions", POP equal to manager and executive director positions or employee, apprentice or home worker, WCP equal to unemployed or out of labour force. With respect to continuous variables, the reference individuals had average values. For the sake of brevity, they are not listed here, but they can be derived from Table 24, which reports the estimation of the odds ratios (OR) of the model with standard errors and p-values.

The binary variables having an odd ratio greater than 1 implied that the represented group had a higher probability than that of the reference group. Therefore, women with an odd ratio equal to 2.009 had a probability almost double that of the men of attending or achieving a tertiary or a post-tertiary level of education. Similarly, significant high probabilities of attending or achieving a tertiary or post-tertiary education level were observed for those suffering from a chronic illness or condition (1.406), living in the macro-regions of the North-East (1.299) or South (1.185) with a borderline probability (0.079), living in a place with a high (1.185) or average (1.154) degree of urbanisation, both with a borderline probability: (0.053) and (0.094), respectively.

The binary variables having an odd ratio lower than 1 implied that the represented group had a lower probability than the complementary group. Therefore, immigrants with an odd ratio equal to 0.649 had a probability equal to 35.1% (which is the complement to one of the estimated OR expressed as a percentage) less than that of non-immigrants of attending or achieving a tertiary or a post-tertiary education level. Similarly, significant low probabilities of attending or achieving a tertiary or post-tertiary education level was observed for young people with bad self-perceived health (0.583), and some labour market variables of parents: PAS at least one pensioner (0.577), POP labourer (0.776), WCP Full-Time Dependent for one or both parents (0.711), WCP Full-Time Self-employed for one or both parents (0.630).

Table 24. Estimated odds ratios (OR) with standard errors (SE) and corresponding p-values

Variables\ TE	**OR**	**SE**	**p-value**
Women (men [**RG**])	2.009	0.142	0.000
Age/10	9.752	21.426	0.300
$(\mathrm{Age}/10)^2$	0.636	0.284	0.311
Immigrant	0.649	0.064	0.000
DPI= (Disposable personal income)/ 20,000	0.119	0.019	0.000
DPI^{2}	1.671	0.120	0.000
Self-perceived health (SPH)	0.583	0.093	0.001
SPH: Suffering from chronic illness or condition	1.406	0.237	0.043
SPH: limitations in activities due to health problems	0.802	0.142	0.212
Macro-region: North-West	1.092	0.114	0.400
Macro-region: North-East	1.229	0.122	0.039
Macro-region: South	1.185	0.115	0.079
Macro-region: Islands	0.842	0.101	0.152
Degree of urbanisation: high density	1.185	0.104	0.053
Degree of urbanisation: average density	1.154	0.098	0.094
Parental characteristics			
(Father's age)/10	1.220	0.056	0.000
(Mother's age)/10	3.572	0.673	0.000
$[(\mathrm{Mother's\ age})/10]^2$	0.907	0.018	0.000
MELP= Maximum education level of parents in years	0.804	0.035	0.000
MELP^2	1.017	0.002	0.000
FDPI= (Father's DPI)/ 20,000	1.467	0.153	0.000
FDPI^{2}	0.945	0.015	0.001
MDPI= (Mother's DPI)/ 20,000	1.509	0.158	0.000
MDPI^{2}	0.920	0.019	0.000
FTI= (Family's total income)/ 20,000	0.612	0.056	0.000
FTI^{2}	1.044	0.012	0.000
FIPP= Family income per capita= FTI/ NCF	2.601	0.850	0.003
FIPP^{2}	0.849	0.082	0.089
Labour market variables of Parents			
PAS: Only father employed	0.870	0.079	0.128
PAS: Only mother employed	0.862	0.092	0.164
PAS: At least one Pensioner	0.577	0.085	0.000
POP: Labourer	0.776	0.070	0.005
WCP: Full-Time Dependent (one or both)	0.711	0.092	0.008
WCP: Full-Time Self-employed (one or both)	0.630	0.093	0.002
WCP: Part-Time (one or both)	0.852	0.162	0.401
WCP: Full-/ Part-Time (residuals)	0.828	0.119	0.192
Constant	0.001	0.002	0.007

Several continuous variables yielded significant ORs and showed different behaviours in the observable range values of the regressors. The maximum education level of parents (MELP) expressed in years had a highly significant and nonlinear impact on the probability of attending or having achieved a tertiary and post-tertiary education, π_i : the squared term had a positive coefficient, involving a negative and decreasing effect on attending or achieving a tertiary or post-tertiary education level of up to 6.8125 years of education, which is the abscissa vertex of the parabola given by $-b/2a$, *i.e.*, $x_{\mathrm{MELP;T}} = -(-0.218)/[2 \times 0.016] \approx 6.8125$, but its effect on tertiary level education became positive only after about 14 years of education. The effect on π_i strongly increased thereafter. The coefficients of the model are used to describe the parabolic forms, instead of the ORs, but they are not reported here for the sake of brevity.

At the individual level, disposable personal income (DPI) proved to be highly significant. The squared term had a positive coefficient, involving a negative and decreasing effect on attending or achieving a tertiary or post-tertiary education level up to €41,520.47, given by $x_{\mathrm{DPI};T} = -(-2.130)/[2 \times 0.513] \approx 2.0760$ multiplied by 20,000. The effect increased thereafter and became positive after about €78,000.

The father's age had an effect only through the linear term of the parabola, involving an average increase in π_i of about 0.05 with each decade of age. The mother's age (MA) yielded significant coefficients for both the linear and quadratic (negative) terms of the parabola, which had a vertex in the $x_{\mathrm{MA;T}} = -(1.273)/[2 \times (-0.098)] \approx 6.49$ decades, involving only the increasing branch of the parabola in the observable age range. It induced an average increase in π_i of about 0.11 with each decade of age.

The father's disposable personal income (FDPI) showed significant coefficients for the squared (negative) term and linear term, involving a positive and increasing effect on attending or achieving a tertiary or post-tertiary education level, up to €68,392.86, given by $x_{\mathrm{FDPI;T}} = -(0.383)/[2 \times (-0.056)] \approx 3.4196$ multiplied by 20,000; thereafter the effect decreased, as the original values were divided by 20,000.

The mother's disposable personal income (MDPI) showed significant coefficients for the squared (negative) term and linear term, involving a positive and increasing effect up to €49,638.55, given by $x_{\mathrm{MDPI;T}} = -(0.412)/[2 \times (-0.083)] \approx 2.4819$ multiplied by 20,000. The effect decreased thereafter.

The family's total income (FTI) proved to be equally highly significant. The squared term had a positive coefficient, involving a negative and decreasing effect up to €114,186.05, given by $x_{\mathrm{FTI;T}} = -(-0.491)/[2 \times 0.043] \approx 5.7093$ multiplied by 20,000 T he effect increased thereafter. In practice, for the great majority of individuals, the family's total income revealed the same behaviour as DPI, but differed from the other income variables, presumably because there were many variables measuring the same concept involving an overestimation, balanced by an opposite behaviour of the FTI.

The family's income per capita (FIPP) proved to be highly significant, as were the other components, yielding a negative coefficient for the squared term, which involved an increasing and decreasing effect on attending or achieving a tertiary or post-tertiary education level up to €58,292.68, given by $x_{\mathrm{FIPP;T}} = -(0.956)/[2 \times (-0.164)] \approx 2.9146$ multiplied by 20,000. The effect increased thereafter.

5. Model by Lasso selection of regressors

Another way of modelling data through equation (1) considers a *Machine Learning* approach to the selection of independent variables and the estimation of parameters. The covariates were selected based on the *Lasso* method (Tibshirani, 1996; Zou and Hastie, 2005) which is a procedure that enables simultaneous estimation and model selection. Roughly speaking, the Lasso method consists of adding a penalization term to the negative log-likelihood of the model that depends on an additional parameter named λ, $\lambda \geq 0$. More precisely, let $\Phi(\cdot)$ be the objective function of the logit model, hence

$$\Phi(\boldsymbol{\beta}, \lambda; \mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \left[y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \right] + \lambda \sum_{j=0}^{K} \left| \beta_j \right| \tag{2}$$

where $\mathbf{x} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$, $\mathbf{y} = (y_1, \dots, y_n)$, and $\pi_i = \pi_i(\mathbf{x}_i, \boldsymbol{\beta})$. Finally, $\Phi(\cdot)$ is minimized for different values of parameter λ. It can be noted that when $\lambda = 0$, then $\Phi(\cdot)$ is the negative log-likelihood of the logit model. On other hands, larger values of λ yield many β's exactly equal to zero.

In many penalized methods, Φ can be interpreted as the negative logarithm of a posterior distribution in a purely Bayesian fashion. Let $p(y_i | \mathbf{x}_i, \boldsymbol{\beta}) = {\pi_i}^{y_i} (1 - \pi_i)^{1 - y_i}$ be the usual logit model in the usual Bayesian notation, and let $p(\boldsymbol{\beta}|\lambda) \propto \exp\{-\lambda \sum_{j=0}^{K} |\beta_j|\}$ be the Laplace prior distribution on coefficients $\boldsymbol{\beta}$; then the posterior distribution is

$$\begin{aligned} p(\boldsymbol{\beta} | \mathbf{x}, \mathbf{y}, \lambda) &\propto p(\mathbf{y} | \mathbf{x}, \boldsymbol{\beta})\, p(\boldsymbol{\beta} | \lambda) \\ &\propto \prod_{i=1}^{n} p(y_i | \mathbf{x}_i, \boldsymbol{\beta})\, p(\boldsymbol{\beta} | \lambda) \\ &= \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i} \exp\left(-\lambda \sum_{j=0}^{K} \left| \beta_j \right| \right) \end{aligned} \tag{3}$$

Note that $\Phi(\boldsymbol{\beta}, \lambda; \mathbf{x}, \mathbf{y}) = -\log p(\boldsymbol{\beta}|\mathbf{x}, \mathbf{y}, \lambda)$. Hence the Lasso method can be interpreted as a maximum posterior Bayesian estimation method where the prior distribution on β's is Laplace, and λ plays the role of the hyper-parameter. Let $\widehat{\boldsymbol{\beta}}_\lambda$ be the minimizing of $\Phi(\cdot)$, then $\widehat{\boldsymbol{\beta}}_\lambda$ is the maximum posterior estimation of $\boldsymbol{\beta}$ conditional to the data and λ.

$$\hat{\boldsymbol{\beta}}_\lambda = \underset{\boldsymbol{\beta}}{\arg\min}\, \Phi(\boldsymbol{\beta}, \lambda; \mathbf{x}, \mathbf{y}) = \underset{\boldsymbol{\beta}}{\arg\max}\, p(\boldsymbol{\beta} | \mathbf{x}, \mathbf{y}, \lambda) \tag{4}$$

The choice of parameter λ plays a crucial role in the estimation procedure. Many different research studies have focused on this issue, see Zou, Hastie, and Tibshirani (2007) for an extensive review. Beside the classic *AIC* and *BIC* criteria, a *k-fold Cross Validation* (*CV*) procedure and a *One Standard Error Rule* (*1SE*) have been proposed (Arlot et al., 2010). The *CV* procedure consists in randomly partitioning the original sample into *k* equal-sized subsamples (usually *k* = 5 or, *k* = 10). Of the *k* subsamples, a single subsample is retained as validation data for testing the model and the remaining $k-1$ subsamples are used as training data. The process is repeated k times, and each of the k subsamples is used exactly once as validation data. The CV for a given λ is the average of binomial deviance in each step. The optimal value of λ is

$$\lambda_{CV} = \arg\min_{\lambda} CV(\lambda) \tag{5}$$

In order to achieve more regularization, the *1SE* rule consists in choosing $\lambda_{1SE} > \lambda_{CV}$ such that $CV(\lambda_{1SE}) = CV(\lambda_{CV}) + SE(CV(\lambda_{CV}))$, where $SE(CV(\lambda_{CV}))$ is the standard error estimated in the k steps.

It is known (Hastie et al., 2015) that *CV* estimates prediction error at any fixed value of the tuning parameter, and thus by using it, it is implicitly assumed that achieving the minimal prediction error is the goal, which is not the case here. The *1SE* rule is the best candidate for achieving the goal of recovering the *true* model. Actually, *1SE* adds more regularization than *CV*. We used the *1SE* rule for selecting the variables.

Given the large number of missing values, categorical covariates that presented missing values where recoded, and the level termed *not available* (NA) were added to the taxonomy. Moreover, for such a variable, the NA level was absorbed by the intercept.

We chose a starting model that considers the larger set of covariates compatible with the first computational step of the algorithm. The covariates are described in Table 25. The starting model considered 55 variables, many of them factors with more than 2 levels, and the corresponding design matrix had 229 columns, and thus the number of components of $\boldsymbol{\beta}$ was equal to 229.

The model was estimated using the *glmnet* (Friedman et al., 2010) package in **R** (R Core Team, 2019). Figure 1 shows the *CV* values and the corresponding estimated *SE* for different values of $\log\lambda$. The lower *x* axis shows $\log\lambda$ and the upper *x* axis enumerates the number of βs different from zero for a given λ. Binomial deviance is evaluated on the *y* axis. The red points represent the evaluated *CV* given λ, and the grey segments represent the corresponding standard errors.

The *glmnet* package, like many other penalized likelihood packages, provides point estimation for coefficients $\boldsymbol{\beta}$ and many statistics for evaluating the *CV*, but it does not provide confidence intervals for the parameters nor standard errors. However, it is possible to draw samples from the posterior distribution $p(\boldsymbol{\beta}|\mathbf{x}, \mathbf{y}, \lambda_{1SE})$, and then to perform a full Bayesian analysis.

A simple *Gibb Sampler* (Gilks et al., 1995) was built. This is a special kind of *MCMC* algorithm with an *acceptance rate* equal to one. The main idea of the algorithm is simple. It consists in drawing samples from the posterior distribution, from one dimension at a time, with all other dimensions being fixed. A good starting point for the algorithm is the estimation of $\boldsymbol{\beta}$ provided by the *glmnet* package: Let $\widehat{\boldsymbol{\beta}}_{\lambda_{1SE}}$ be such a vector. Let $\boldsymbol{\beta}^{(sim)}$ be the sample from $p(\boldsymbol{\beta}|\mathbf{x}, \mathbf{y}, \lambda_{1SE})$, at step *sim* of the algorithm. The *Gibbs Sampler* is indicated in Table 26.

Table 25. Variables included in the starting model, their type, and the intercept level

	Acron.	Type	No. lvl	Intercept level
Gender		Factor	2	Males
Age		Numeric		
Age^2 (squared)		Numeric		
No. of components of the family	ncomp	Numeric		
Immigrants		Factor	2	NA
Parental relationships		Factor	18	HELP Reference person: 01
Private work		Factor	4	Employed
Marital status		Factor	7	Not married
Year of education level attained	PE030	Factor		
General health (Liker scale)	PH010	Factor	6	NA
Suffering from chronic illness/ cond.	PH020	Factor	3	NA
Limitations due to health problems	PH030	Factor	4	NA
Unmet need for medical examination	PH040	Factor	2	Yes, on at least one occasion
Reason for unmet medical examination	PH050	Factor	9	NA
Unmet need for dental examination	PH060	Factor		Yes, on at least one occasion
Reason for unmet dental examination	PH070	Factor	9	NA
Professional training (Regional)	Formaz	Factor	2	Yes
Net monthly salary + overtime	Dipnt_e	Numeric		
Type of contract	Tipcon	Factor	5	NA
Type of term contract in main job	Tipte_p_c	Factor	9	NA
Type of term contract (≠ set of answers)	Tipte_c	Factor	8	NA
Employment position	Posdip	Factor	8	NA
Job position in main job	Posdip_p	Factor	7	NA
Net monthly salary	Dnet_e	Numeric		
Income from dependent work		Numeric		
Income from independent work		Numeric		
Properties		Factor	5	NA
Contract location	Contratto	Factor	7	NA
Aid requested for food, etc.	Difcib	Factor	4	Yes, often
Helpers: Parents	Cgen	Factor	12	NA
Helpers: brothers	Cfrat	Factor	12	NA
Helpers: others	Caltri	Factor	12	NA
Family's total income (FTI)	Fytot	Numeric		
Macro-region	Ripgeo	Factor	5	North-West
Father's age	Age1fa	Numeric		
Father's squared age	Age2fa	Numeric		
Father's marital status	PB190fa	Factor	6	NA
Father's education level	Istr_cfa	Factor	11	NA
Father's chronic illness/ conditions	Ph_chronfa	Factor	3	NA
Father's unmet medical examination	Ph_unfa	Factor	3	NA
Father's limitations due to health probl.	Ph_limitfa	Factor	3	NA
Father's job position in main job	Posdipfa	Factor	7	NA
Father's job conditions	Condizfa	Factor	13	NA
Father's disposable personal income	Yind1_fa	Numeric		
(Mother's age)	Age1mo	Numeric		
(Mother's age)^2	Age2mo	Numeric		
(Mother's age)^2	PB190mo	Factor	6	NA
Mother's education level	Istr_cmo	Factor	11	NA
Mother's chronic illness/ conditions	Ph_chronmo	Factor	3	NA
Mother's unmet medical examination	Ph_unmo	Factor	3	NA
Mother's limitations due to health probl.	Ph_limitmo	Factor	3	NA
Mother's job conditions	Condizmo	Factor	13	NA

Table 26. Statements of the Gibbs Sampler

```
$\boldsymbol{\beta}^{(sim=0)} = \widehat{\boldsymbol{\beta}}_{\lambda_{1SE}}$ ,
for $sim$ in $1,2,\dots,N_{sim}$ {
    for $m$ in $0,1,\dots,k$ {
        compute numerically $p(\beta_m|\boldsymbol{\beta}_{-m})$ the marginal distribution of $\beta_m$ given $\boldsymbol{\beta}_{-m} = \left(\beta_0^{(sim)},\dots,\beta_{m-1}^{(sim)},\beta_{m+1}^{(sim-1)},\dots,\beta_k^{(sim-1)}\right)$, and draw $\beta_m^{(sim)}$ from $p(\beta_m|\boldsymbol{\beta}_{-m})$
    }
}
```

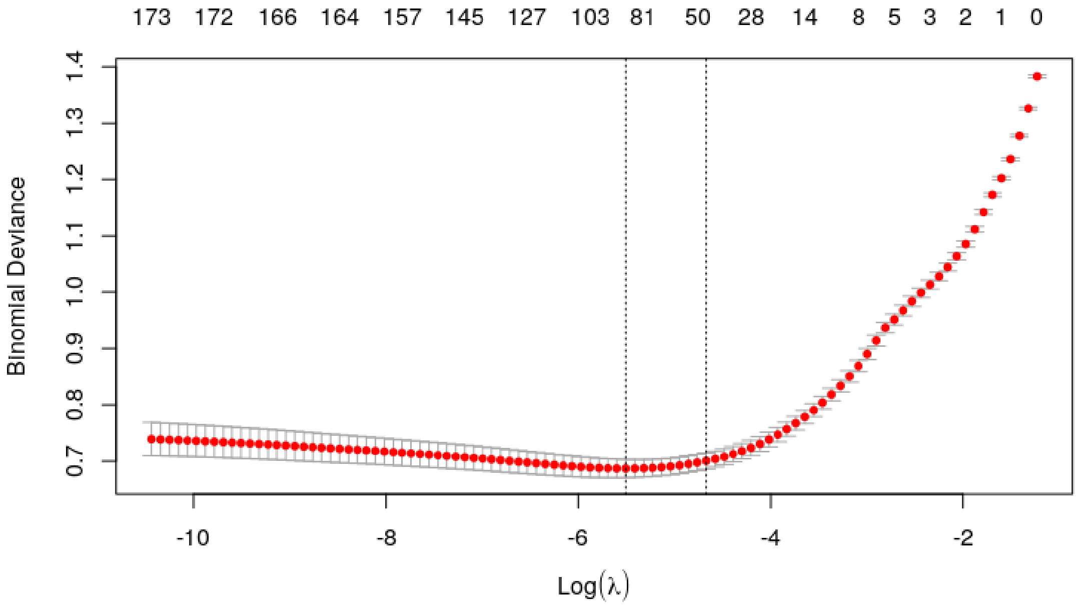

Figure 1. CV values for different choices of λ: The minimizing values λ_{CV} and λ_{1SE} are marked by dashed segments

The size of the sample was $N_{sim} = 2000$ points from $p(\boldsymbol{\beta}|\mathbf{x},\mathbf{y},\lambda_{1SE})$ and the 95% credibility intervals were computed. The latter are intervals delimitated by the 0.025 and 0.975 quantiles of the marginal posterior distribution.

Table 27 shows the nonzero estimated parameters via *glmnet*, the estimated standard errors, the 0.025 quantile of the distribution $q_{0.025}$, the median, and the 0.975 quantile of the distribution $q_{0.975}$. Note that the procedure shrinks to zero 179 values of $\widehat{\boldsymbol{\beta}}_{\lambda_{1SE}}$ over 229; thus only 50 values were different from zero. Note also that not all credibility intervals were symmetric around the estimate. This is because some marginal posterior distributions may present some cusps and asymmetry, due to the Laplace prior distribution.

Table 27. Estimated $\hat{\boldsymbol{\beta}}_{\lambda_{1\mathrm{SE}}}$, standard errors (SE), the 0.025 quantile of the distribution $q_{0.025}$, the median, and the 0.975 quantile of the distribution $q_{0.975}$

	$\hat{\boldsymbol{\beta}}_{\lambda_1 \mathrm{SE}}$	SE	$q_{0.975}$	*Median*	$q_{0.025}$
Intercept	–839.8381	0.0462	–839.9314	–839.8408	–839.7502
Women	0.3900	0.0629	0.2622	0.3853	0.5090
Age	3.5323	0.0189	3.4937	3.5308	3.5679
Consort of the head of the family 01	–0.9176	0.2042	–1.3489	–0.9402	–0.5478
Son of the head of the family 01	0.2580	0.0504	0.1590	0.2578	0.3569
Nephew of 01-03	–1.4197	0.7135	–2.9546	–1.4848	–0.1415
Consort of brother/ sister of 01-03	6.1247	10.7297	1.2591	17.8681	37.0055
Unemployed	–0.8299	0.1396	–1.1172	–0.8380	–0.5696
Other	2.3080	0.0659	2.1776	2.3056	2.4360
Attainment of education level in years	0.4133	0.0000	0.4132	0.4133	0.4133
SPH: Fair = ph010	–0.4835	0.2132	–0.9093	–0.4884	–0.0733
SPH: Very bad = ph010	–2.2696	1.4181	–5.9866	–2.5796	–0.3573
Other reasons: ph050	–2.4105	1.0136	–4.5138	–2.4890	–0.5307
Professional training: NO = formaz	0.3064	0.0498	0.2035	0.3011	0.3990
Fixed–term contract: tipcon	–0.2608	0.1296	–0.5290	–0.2724	–0.0203
Permanent contract: tipcon	–0.4506	0.1042	–0.6616	–0.4547	–0.2527
Agreement for temporary work: tipcon	1.4675	0.7293	–0.1399	1.3665	2.7313
Apprentice: tipte_c	–0.5768	0.3093	–1.2156	–0.5910	–0.0017
Executive: posdip	0.9534	0.4126	0.1459	0.9362	1.7658
Labourer: posdip	–0.9022	0.1406	–1.2008	–0.9188	–0.6494
Apprentice: posdip	–0.7320	0.2804	–1.3162	–0.7503	–0.2155
Net monthly salary: dnet_e	–0.0002	0.0001	–0.0003	–0.0002	–0.0001
Private : property	–0.3366	0.1376	–0.6103	–0.3384	–0.0708
Agreed or conventional: contratto_b	–0.8013	0.2835	–1.4008	–0.8349	–0.2883
Father’s squared age: age2fa	0.0045	0.0015	0.0013	0.0044	0.0074
Father: Primary education | istr_cfa	–0.3006	0.1375	–0.5736	–0.3022	–0.0343
Father: Lower secondary education | istr_cfa	–0.2527	0.0819	–0.4163	–0.2555	–0.0951
Father: Upper secondary education | istr_cfa	0.1907	0.0759	0.0338	0.1823	0.3315
Father: Tertiary education | istr_cfa	0.8084	0.1755	0.4592	0.7970	1.1475
Father: Post–tertiary education | istr_cfa	0.8492	0.3964	0.0696	0.8198	1.6254
Father: executive | posdipfa=2	0.6226	0.2398	0.1490	0.6064	1.0898
Father: unemployed | condizfa=5	–0.5707	0.2766	–1.1300	–0.5821	–0.0448
Father: housekeeper | condizfa=8	1.9272	1.3679	0.0479	2.1399	5.4775
Father’s disposable personal income	0.0808	0.0161	0.0493	0.0805	0.1125
Mother’s age | age1mo	0.2143	0.0094	0.1951	0.2135	0.2319
Mother: Illiterate and no schooling | istr_cmo	–12.1978	93.2999	–361.2774	–77.5962	–12.7579
Mother: Literate and no schooling | istr_cmo	–1.2284	0.6654	–2.6897	–1.3012	–0.0742
Mother: Lower secondary educ. | istr_cmo	–0.1636	0.0821	–0.3273	–0.1661	–0.0054
Mother: Upper secondary educ. | istr_cmo	0.1553	0.0739	0.0092	0.1539	0.2991
Mother: Post-secondary non-tert. | istr_cmo	1.4300	0.6168	0.2341	1.4280	2.6577
Mother: Tertiary education | istr_cmo	1.4011	0.1759	1.0629	1.4015	1.7526
Mother: Post-tertiary education | istr_cmo	1.1626	0.6375	0.0283	1.1762	2.5313
Mother: Full-time dependent | condizmo=1	0.3712	0.0863	0.2022	0.3710	0.5409
Mother: Student | condizmo=7	–0.2370	0.0770	–0.3965	–0.2454	–0.0943
Mother: Other conditions | condizmo=12	–0.6203	0.2081	–1.0386	–0.6279	–0.2225

Finally, the classification of error rates was computed for $\widehat{\boldsymbol{\beta}}_{\lambda_{1SE}}$ by assigning $\hat{y}_i = 0$ if $\hat{\pi}_i = \exp\{\mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{\lambda_{1SE}}\}/\left(1 + \exp\{\mathbf{x}_i'\widehat{\boldsymbol{\beta}}_{\lambda_{1SE}}\}\right) < 0.5$, and $\hat{y}_i = 1$, if $\hat{\pi}_i \geq 0.5$. The misclassified number of $\hat{y}_i$ equal to zero was 203 out of 1868, which is a false negative error rate equal to 11.7%, and the misclassified number of $\hat{y}_i$ equal to one was 267 out of 1799, which is a false positive error rate equal to 13.8%, and there was an overall misclassification error rate equal to 12.8% (Table 28). The performance of the Lasso model seemed to be better than that of the logistic model: the false negative rate was 30.6%, while the false positive rate was 19.5%. In the logistic model, the overall misclassification error rate was equal to 24.5%. Table 28 reports the false plus or minus rates of individuals classified plus or minus. For these probabilities, the performances of the Lasso model were also better than those of the logistic model. However, the two models were not comparable, given that the number of cases differed. Moreover, the starting set of variables was different too.

Table 28. Performance classification of the logistic and Lasso models

	Logistic model			**Lasso model**		
Classified\ Tertiary	**T=1**	**T=0**	**Total**	**T=1**	**T=0**	**Total**
Positive +	1704	583	2287	1532	267	1799
Negative –	751	2402	3153	203	1665	1868
Total	**2455**	**2985**	**5440**	**1735**	**1932**	**3667**
False ± rate for true T=0/ 1	30.6	19.5		11.7	13.8	
	$P(- \mid T=1)$	$P(+ \mid T=0)$		$P(- \mid T=1)$	$P(+ \mid T=0)$	
False ± rate for classified ±	23.8	25.5		10.9	14.8	
	$P(- \mid \hat{y} = -)$	$P(+ \mid \hat{y} = +)$		$P(- \mid \hat{y} = -)$	$P(+ \mid \hat{y} = +)$	

Note: Classified + if predicted Pr(T) ≥ 0.5

In spite of the difference in the number of individuals included in the Lasso model for missing values in the continuous variables, given that for qualitative variables the missing values constituted their corresponding reference group, it revealed approximately the same tendencies as the logistic model, except for some different behaviours of some explanatory variables. In the logistic model, the variables such as the relationships of the individual with the head of the family and their conditions on the job were excluded *a priori*: the former because they could have highlighted spurious effects and the latter because they also indicated *a posteriori* effects of the working individual having achieved a tertiary education level confounding their impact on the choice to continue on with a tertiary education or to go to work. In the Lasso model, the two groups of variables were included at least in an explorative way and revealed significant impacts, as supposed *a priori*. Women and self-perceptions of health showed the same effects in both models. The characteristics of the family again played a relevant role in explaining differences in continuing education in both models, although with some changes in the Lasso model, such as the squared term of the father's age, the linear term of the mother's age, their education levels showing an increasing nonlinear impact of the coefficients referring to the various levels, the father's income, and the conditions of parents in the labour market. All them proved to have a highly significant effect on attending or having achieved a tertiary or post-tertiary level of education.

6. Conclusions

An empirical analysis was performed to investigate differences in tertiary education enrolment or achieved level of education among immigrant and non-immigrant young people. The main empirical evidence may be summarised as follows. In general, women tended to continue their education longer than men. Women attended tertiary levels of education (30.2%) or post-tertiary levels of education (5.5%) more than men (25.7% and 4.1%, respectively). The percentage of women not in school was lower than that of men: 63.5% versus 69.4%. Young immigrants attended education programmes less than young non-immigrants: 17.2% versus 39.7%. Only 15.6% of immigrants attended tertiary education in 2009, with respect to 32.5% of non-immigrants, and only 1.0% attended post-tertiary education with respect to 6.2% of non-immigrants. Consequently, the percentage of immigrants not enrolled in schools was higher than that of non-immigrant young people (82.8% versus 60.3%).

In the model, among the young people’s general self-perception of health and to a large extent, their suffering from a chronic (long-standing) illness or condition was associated with their enrolment in school. On the one hand, those with uncertain or bad health tended to interrupt their education with a probability that was 41.7% less than those without health problems. On the other hand, the presence of chronic illnesses or conditions surprisingly revealed a tendency to increase the odds ratio by 40.6%, *i.e.*, the probability of attending or achieving a tertiary and post-tertiary level of education.

Differences between immigrants and non-immigrants were also found for parental background. The age of the fathers and mothers of immigrants was significantly lower than that of the fathers and mothers of non-immigrants, showing on average a difference equal to about ten years and three years, respectively. Immigrant parents seemed to be affected by chronic illness less than non-immigrant parents. The level of education of parents had a significant impact on the probability of young people continuing their education. The employment status of immigrant fathers/ mothers was significantly lower than that of non-immigrant parents. The same was true for the various disposable personal income items and for total income of families, all well represented by a parabolic form. Moreover, for young people attending or who had attained a tertiary or post-tertiary level of education, this income was approximately 24% higher than that of young people with an upper secondary education level and who were not enrolled in a tertiary education programme. These descriptive tendencies were confirmed in the models: the young immigrants revealed a significant lower probability of continuing their education than young non-immigrants, mainly because the education level of their parents was lower than that of the parents of non-immigrants, the various components of income were lower than those of non-immigrants, and the labour conditions of immigrant parents were worse than those of non-immigrants.

The present empirical results are coherent with those reported in the literature and suggest that an “immigration” gradient is present in educational decisions also in Italy. Differences in education enrolment/ attainment at the tertiary level among immigrants and non-immigrants were explained by the socio-economic status of parents, *i.e.*, their level of education, employment status, and occupational position. These results highlight the need for integrated policies in educational programs, directed both at sustaining youth and helping their families, in order to enhance and improve enrolment of young immigrants in education programmes and to foster a complete integration process.

References

Algan Y., Dustmann C., Glitz A., and Manning A. (2010). The Economic Situation of first and Second-Generation Immigrants in France, Germany and the United Kingdom. *The Economic Journal*, 15(2), 353–385.

Andersson E. K., Lyngstad T. H., and Sleutjes B. (2018). Comparing Patterns of Segregation in North-Western Europe: A Multiscalar Approach. *European Journal of Population*, 34(2), 151–168.

Arlot S., and Celisse A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4, 40–79.

Armstrong M. J. and Biktimirov E. N. (2013). To Repeat or Not to Repeat a Course. *Journal of Education for Business*, 88(6), 339–344.

Atkinson A. B. and Marlier E. (eds.) (2010). *Income and living condition in Europe*. Luxembourg: cat. No KS-31-10-555-EN-C, Publication Office of the European Union.

Bertolini P. and Lalla M. (2012). Immigrant Inclusion and Prospects through Schooling in Italy: An Analysis of Emerging Regional Patterns, in Charis E. Kubrin, Marjorie S. Zatz, Ramiro Martinez Jr., *Punishing Immigrants: Policy, Politics and Injustice*, (pp. 178–206). New York: New York University Press.

Bertolini P., Lalla M., and Pagliacci F. (2013). School enrollment of first- and second-generation immigrant students in Italy: A geographical analysis, *Papers in Regional Science*, 94(1), 141–160.

Cicchitelli G., Herzel A., and Montanari G. E. (1997). *Il campionamento statistico*, II edizione. Bologna: il Mulino.

Contini D. (2013). Immigrant background peer effects in Italian schools. *Social Science Research*, 42(4), 1122–1142.

De Clercq M., Galand B., and Frenay M. (2017). Transition from high school to university: a person-centered approach to academic achievement. *European Journal of Psychology of Education*, 32(1), 39–59.

Donadio P., Gabrielli G., and Massari M. (a cura di) (2014). *Uno come te. Europei e nuovi europei nei percorsi di integrazione*. Milano: Franco Angeli.

Entwisle D. R. and Alexander K. L. (1993). ENTRY INTO SCHOOL: The Beginning School Transition and Educational Stratification in the United States. *Annual Review of Sociology*, 19, 401–423.

European Council (2003). Regulation (EC) No 1177/2003 of the European Parliament and of the Council of 16 June 2003 concerning Community statistics on income and living conditions (EU-SILC) (Text with EEA relevance), OJ L 165, 3.7.2003, p. 1–9, https://eur–lex.europa.eu/legal–content/EN/ALL/?uri=CELEX%3A32003R1177. Accessed 3 January 2020.

Eurostat (2005). *The continuity of indicators during the transition between ECHP and EU-SILC*. Working Papers and Studies. Cat. No.: KS-CC-05-006-EN-N. ISBN: 92-894-9931-1. Luxembourg: Eurostat.

Eurostat (2009). *Description of Target Variables: Cross–section and Longitudinal*, EU–SILC 065 (2009 operation). Directorate F, Unit F–3. Luxembourg: Eurostat.

Friedman J., Hastie T., and Tibshirani R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1), 1–22.

Gilks W. R., Richardson S., and Spiegelhalter D. J. (1995). *Markov Chain Monte Carlo in Practice*. London: Chapman & Hall/CRC.

Grove W. A., Wasserman T., and Grodner A. (2006). Choosing a proxy for academic aptitude. *Journal of Economic Education*, 37(2), 131–147.

Hastie T., Tibshirani R., and Wainwright M. (2015). *Statistical learning with sparsity: The Lasso and generalizations*. London: Chapman & Hall/CRC.

Istat (2008). Ceccarelli C., Di Marco M., and Rinaldelli C. (Eds.). *L'indagine europea sui redditi e le condizioni di vita delle famiglie (Eu-Silc)*. Metodi e Norme n. 37. Rome: Istat.

Istat (2009a). Reddito e condizioni di vita delle famiglie con stranieri [electronic resource]. Rome: Istat. Weblink, https://www.istat.it/it/archivio/52405. Accessed 3 January 2020.

Istat (2009b). Nota informativa sull’utilizzo dell’UDB (User Data Base) CVS 2009. Released by Istat together with data sets (pp. 1–3). Rome: Istat.

Lalla M. and Pirani E. (2014). The secondary education choices of immigrants and non-immigrants in Italy. Rivista Italiana di Economia Demografia e Statistica, LXVIII (3-4), lug.-dic. 2014, pp. 39–46.

Luciano A., Demartini M., and Ricucci R. (2009). L’istruzione dopo la scuola dell’obbligo. Quali percorsi per gli alunni stranieri? In Zincone G. (ed.), *Immigrazione: segnali di integrazione. Sanità, scuola e casa*, (pp. 113–156). Bologna: il Mulino.

Nguyen A. N. and Taylor J. (2003) Post-high school choices: New evidence from a multinomial logit model. *Journal of Population economics*, 16(2), 287–306.

Paba S. and Bertozzi R. (2017). What happens to students with a migrant background in the transition to higher education? Evidence from Italy. *Rassegna Italiana di Sociologia*, 58(2), 315–351.

Pong S. and Hao L. (2007). Neighborhood and School Factors in the School Performance of Immigrants’ Children. *International Migration Review*, 41(1), pp. 206–241.

R Core Team (2019). *R: A language and environment for statistical computing*. Vienna (Austria): R Foundation for Statistical Computing. URL http://www.R-project.org/.

Särndal C. E., Swensson B., and Wretman J. (1992). *Model Assisted Survey Sampling*. Berlin: Springer-Verlag.

Sleutjes B., de Valk H. A. G., Ooijevaar J. (2018). The measurement of Ethnic Segregation in the Netherlands: Differences Between Administrative and Individualized Neighbourhoods. *European Journal of Population*, 34(2), 195–224.

STATA (2005). *Stata Statistical Software: Release 9*, volumes 1-4. College Station (TX): StataCorp LP.

Tibshirani R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B, 58(1), 267–288.

Wintre M. G., Dilouya B., Pancer S. M., Pratt M. W., Birnie–Lefcovitch S., Polivy J., and Adams G. (2011). Academic achievement in first–year university: Who maintains their high school average? *Higher Education*, 62(4), 467–481.

UNESCO Institute for Statistics (2012). *International Standard Classification of Education ISCED 2011*, Montreal (Quebec – Canada): Ref. UIS/2012/INS/10/REV, UNESCO–UIS.

Verma V. and Betti G. (2006). EU Statistics on Income and Living Conditions (EUSILC): Choosing the survey structure and sample design, *Statistics in Transition*, 7(5), pp. 935–970.

Verma V., Betti G., and Ghellini G. (2007). Cross–sectional and longitudinal weighting in a rotational household panel: applications to Eu–Silc, *Statistics in Transition*, 8(1), pp. 5–50.

Verma V. and Betti G. (2010). Data accuracy in EU–SILC. In Atkinson A. B., Marlier E. (eds.) (2010). *Income and living condition in Europe*, (pp. 57–77). Luxembourg: cat. No KS–31–10–555–EN–C, Publication Office of the European Union

Wolff P., Montaigne F., and González G. R. (2010). Investing in statistics: EU–SILC. In Atkinson A. B., Marlier E. (eds.) (2010). *Income and living condition in Europe*, (pp. 37–55). Luxembourg: cat. No KS–31–10–555–EN–C, Publication Office of the European Union.

Zou H., and Hastie T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*, Series B, 67(2), 301–320.

Zou H., Hastie T., and Tibshirani R. (2007). On the degrees of freedom of the lasso. *The Annals of Statistics*, 35(5), 2173–2192.

Zwysen W. and Longhi S. (2018). Employment and earning differences in the early career of ethnic minority British graduates: the importance of university career, parental background and area characteristics. *Journal of Ethnic and Migration Studies*, 44(1), pp. 154–172.