Factors affecting tertiary education decisions of immigrants in Italy

Michele Lalla, Patrizio Frederic

1. Introduction

The decision to enrol in tertiary education is difficult for young people and families if the choice is made without much knowledge about the needs of society. Such decisions may be affected by individual characteristics, the socio-economic conditions of families, and the contextual background of the area. All these aspects may differ among young immigrants and non-immigrants 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 (Paba and Bertozzi, 2017; De Clercq et al., 2017).

The objective of this paper is to point out the differences with respect to citizenship, a binary variable distinguishing between immigrants and non-immigrants (hereinafter also referred to as Italians), and the *tertiary* binary variable, defined as equal to one for individuals who were enrolled in a tertiary education level and equal to zero otherwise. A Bayesian model selection was performed through the Lasso method to investigate the determinants of the tertiary binary variable.

2. Data sources and descriptive statistics

The data were extracted from two surveys, with the reference year being 2009, carried out by the Italian National Institute of Statistics (Istat): one being the European Union Statistics (or Surveys) on Income and Living Conditions (EU-SILC) restricted to Italy, IT-SILC (Istat, 2008; Eurostat, 2009), 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, 2009) that involved families with at least one immigrant component residing in Italy. The IT-SILC sample was added to the IM-SILC sample to obtain a sample with a consistent number of immigrants with respect to non-immigrants. For further details about these two data sets and about the main variables introduced in the model, see Lalla and Frederic (2020). The target sample was obtained by first selecting individuals in the age range of 20 to 25, obtaining a sample of 3,166 cases. Then, among the latter data set, the eligible cases were only those individuals whose highest attained ISCED (International Standard Classification of Education) level was equal to 3 (=upper secondary education) or 4 (=post-secondary non-tertiary education). The final target sample was made up of 2,874 individuals.

The relationship between the tertiary (binary) dependent variable and the ISCED Level Currently Attended (ILCA) showed that 55.3% of individuals, with an ISCED level equal to 3 or 4, were not enrolled in further education (termed “not-attending”), while 44.7% were currently attending a tertiary school (Table 1).

The ILCA was examined with respect to several qualitative variables and revealed many significant relationships. For the sake of brevity, only some of them are cited. The ILCA showed a significant relationship with respect to citizenship, CS(2)= 115.33 ($p<0.000$), where CS(g) stands

Michele Lalla, University of Modena and Reggio Emilia, Italy, michele.lalla.mo@gmail.com, 0000-0002-1639-7300
Patrizio Frederic, University of Modena and Reggio Emilia, Italy, patrizio.frederic@unimore.it, 0000-0001-9073-2878

Referee List (DOI 10.36253/fup_referee_list)
FUP Best Practice in Scholarly Publishing (DOI 10.36253/fup_best_practice)

Michele Lalla, Patrizio Frederic, *Factors affecting tertiary education decisions of immigrants in Italy*, © Author(s), CC BY 4.0, DOI 10.36253/979-12-215-0106-3.32, in Enrico di Bella, Luigi Fabbris, Corrado Lagazio (edited by), *ASA 2022 Data-Driven Decision Making. Book of short papers*, pp. 181-186, 2023, published by Firenze University Press and Genova University Press, ISBN 979-12-215-0106-3, DOI 10.36253/979-12-215-0106-3

for “Chi-Square with g degrees of freedom”, but hereinafter “(g)” is omitted because the corresponding tables do not appear here: the percentage of immigrants attending tertiary education was lower than that of Italian citizens (26.6% versus 50.0%), while the percentage of immigrants not in school was higher than that of Italians (72.4% versus 48.4%). There was a significant relationship between the ILCA and self-perceived health, CS= 10.87 ($p<0.004$), implying that individuals perceiving fair or bad or very bad health tended to discontinue their education with respect to those perceiving good or very good health (Ichou and Wallace, 2019). The ILCA was not related to the index of the total self-perceived health of parents, perhaps its effect operated during the upper secondary education level (Frederic and Lalla, 2021). The ILCA proved to be linked to the Italian macro-regions CS= 24.27 ($p<0.002$), as industrialisation and the possibility of finding employment increased, the percentage of individuals not in school increased. The ILCA was related to the maximum ISCED level attained by parents, CS= 198.80 ($p<0.000$). As the education of parents increased, the percentage of young individuals in school increased. The ILCA was significantly related to several variables describing the working conditions of parents, but the strength of such relationships was generally weak.

Table 1. Absolute and percentage frequencies of tertiary education (EDU) by the ISCED level currently attended (ILCA)

Tertiary\ ILCA	**Not-attending**	**Post-Secondary EDU**	**Tertiary EDU**	**Total**
Tertiary = 1			1285	1285
			100.0	100.0
Tertiary = 0	1546	43	0	1589
	97.3	2.7	0	100.0
Total	1546	43	1285	2874
	53.8	1.5	44.7	100.0

The ILCA was also analysed with respect to the main quantitative variables.

The age of fathers analysed according to the ILCA and citizenship showed that the fathers of immigrants were younger than the fathers of Italians by about twelve years. Similarly, the mothers of immigrants were younger than the mothers of Italians by about twelve years. The Disposable Family Income (DFI) per capita (in thousands of euros) is reported in Table 2 by the ILCA and citizenship. On the average, the DFI per capita for immigrants was significantly lower than that of Italians by about four thousand euros: about 35.7%.

Table 2. Sample size frequencies (n), means, and standard deviations (SD) of the disposable family income per capita (in thousands of euros) by citizenship and by the ISCED level currently attended (ILCA) by their children (E=Education)

Citizenship\ ILCA		**Not-attending**	**Post-Secondary E**	**Tertiary E**	**Total**
Italian citizen.	*n*	1080	36	1114	2230
	Means	*11.389*	*11.508*	*12.543*	*11.967*
	SD	6.999	6.432	9.147	8.153
Foreign citizen.	*n*	466	7	171	644
	Means	*7.777*	*5.868*	*7.563*	*7.699*
	SD	5.315	3.489	5.877	5.452
Total.	***n***	1546	43	1285	2874
	Means	*10.300*	*10.590*	*11.880*	*11.011*
	SD	6.742	6.376	8.942	7.835

The other types of income considered in the models revealed various structures of relationships and levels of significance. For example, the gap between immigrant and Italian fathers' incomes amounted to about eleven thousand euros, i.e., –42.0%. The mothers' incomes also presented significant statistical differences for both marginal effects, with a gap amounting to about five thousand six hundred euros, i.e., –32.5%. However, the disposable personal income gender gaps were –35.9% for Italians and –25.3% for immigrants.

The size of immigrant families proved to be slightly lower than those of Italians, but not statistically significant. The result differed in the population involved in the transition from lower to upper secondary education (Frederic and Lalla, 2021) implying that the size of families who intended to send their children to university was similar to that of the Italians.

Citizenship was examined with respect to some other variables. Its relationship with the maximum ISCED level attained by parents was statistically significant, CS= 217.01 ($p<0.000$) (Bertolini et al., 2015). Citizenship was significantly related to the degree of urbanisation, CS= 19.18 ($p<0.000$): immigrants tended to settle in densely populated areas more than Italians (36.2% versus 35.3%) or in moderately populated areas (46.6% versus 39.6%). Citizenship also showed a significant relationship with the Italian macro-regions and yielded a significant relationship with the index summarising the total self-perceived health of parents, CS= 134.99 ($p<0.000$) (Ichou and Wallace, 2019). Citizenship proved to be associated with many variables describing working conditions; only the relationship with the maximum position of parents on the job, CS= 134.03 ($p<0.000$), is mentioned here.

3. Bayesian Lasso selection of regressors

Let Y be the binary variable coding if the i-th individual is or is not attending tertiary education (i=1, …, n). Let $\mathbf{x}_i$ be a vector of K regressors. Let π_i be the probability that Y=1 given $\mathbf{x}_i$. Let $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_K)$ be the parameters vector of the model. The logit model is

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

The *Lasso* method (Tibshirani, 1996) was applied to carry out the estimation and model selection. In fact, it is a procedure involving an additional penalization term, L_1, summed up to the negative log-likelihood of the model that depends on an additional parameter λ, $\lambda \geq 0$. Many penalized methods can be interpreted as the negative logarithm of a posterior distribution in a purely Bayesian way. Let $p(y_i | \mathbf{x}_i, \boldsymbol{\beta}) = \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}$ be the model in the Bayesian notation and let $p(\boldsymbol{\beta} | \lambda) \propto \exp\left(-\lambda \sum_{j=0}^{K} |\beta_j|\right)$ be the Laplace prior distribution on coefficients $\boldsymbol{\beta}$, where K is the number of regressor coefficients and β_0 is the intercept. Then the posterior distribution is

$$\begin{aligned} p(\boldsymbol{\beta} | \mathbf{x}, \mathbf{y}, \lambda) &\propto p(\mathbf{y} | \mathbf{x}, \boldsymbol{\beta}) \;\; p(\boldsymbol{\beta} | \lambda) \\ &= \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i} \; \exp\left(-\lambda \sum_{j=0}^{K} |\beta_j|\right) \end{aligned} \tag{2}$$

To select λ, the One Standard Error Rule (1SE) procedure was applied. The estimation method consisted of two steps:

1. The model was first estimated using the *glmnet* (Friedman et al., 2010) package in R (R Core Team, 2019). Then the *optimal* lambda (λ_{1SE}) and the mode estimations $\left(\hat{\boldsymbol{\beta}}_{\lambda_{1SE}}\right)$ were evaluated.

2. Using the R package *MCMCpack*, N=10,000 samples were drawn from the posterior distribution $p(\boldsymbol{\beta}|\mathbf{x},\mathbf{y},\lambda_{1SE})$ to perform a full Bayesian analysis, where $p(\boldsymbol{\beta}|\lambda_{1SE})$ was chosen to be Laplace distributed.

Note that the model matrix of the starting model consisted in 2874 rows by 880 columns, and classical methods can be affected by the *curse of dimensionality*. Instead, the Lasso method is very stable and quick, and shrinks 858 values (out of 880) of $\hat{\boldsymbol{\beta}}_{\lambda_{1SE}}$ to zero; thus only 22 betas have a posterior distribution which is not symmetric to zero.

4. Outcomes of the logistic model

The odds ratios (OR) are reported in Table 3, which only presents interaction terms of the first order because the analysis of interactions orders was limited to the first order to simplify interpretation. The interactions are indicated by the symbol ×, which may be read as "by".

Let $\mathbf{x}_b$ be the binary variables. Let $\mathbf{x}_c = \boldsymbol{\mu}$ be the mean values of the continuous regressors, limited to the ages of individuals, which can never be zero in practice. Note that: (1) the product of two binary variables is again a binary variable, (2) the percentage of variation of the reference probability, $\pi_{i|\mathbf{x}_b=\mathbf{0}\wedge\mathbf{x}_c=\boldsymbol{\mu}}$, is given by [100*(OR–1)] and is reported below in parentheses, (3) the corresponding value of OR may be found in Table 3. The probability of having y=1 (i.e., of continuing one's education) was equal to $\pi_{i|\mathbf{x}_b=\mathbf{0}\wedge\mathbf{x}_c=\boldsymbol{\mu}}$ = 0.120, calculated at the mean values of the continuous regressors $(\mathbf{x}_c = \boldsymbol{\mu})$ and the binary variables equal to 0 $(\mathbf{x}_b)$. A binary variable having an OR greater than 1 implied that the group represented by the binary variable equal to 1 had a higher probability of having y=1 than the group identified by the binary variable equal to 0; for example, for women with an OR=1.777, the probability of continuing their education was +77.7% greater than that of men. In other terms, $\pi_{w|\bullet}$= 1.777×0.120= 0.213, which was +77.7% greater than the probability of men. Note that the dot in the index means keeping all other variables fixed, i.e., the binary and the continuous variables other than age equal to zero. The successive binary variable having an OR>1 in Table 3 was "PES (Parents' Employment Status) is inactive" (x_1) × "Family living in a densely populated area" (x_2), denoted by x_{12}, which showed an OR=1.697 meaning that the odds of the event y=1, when x_{12}=1 (both x_1 and x_2 are equal to 1), were +69.7% greater than the odds of the event y=1, when x_{12}=0. Therefore, $\pi_{x_{12}=1|\bullet}$ = 1.697×0.120= 0.204. Similarly, significant high probabilities of continuing one's education were observed for other interaction terms: "Father with permanent contract" × "Only mother employed" (+95.7%), "Father with permanent contract" × "Parents are managers or executives" (+132.1%), "Mother with permanent contract" × "Father is limited by health" (+64.7%), "TSH (Tenure Status of Household): Subtenant" × "Family living in a moderately populated area" (+46.6%), "TSH: Free" × "Assets reduction for needs" (+173.3%), "Father with term contract" × "Mother is limited by health" (+266.5%). This latter appears to be an unbelievable outcome. However, this group (x_{12}=1) only consisted of 30 subjects and whose family income was higher than that of the group consisting of 162 subjects and having "Father with term contract" $(x_1 = 1)$ and "Mother not limited by health" $(x_2 = 0)$. In synthesis, gender, good and stable parents' working conditions, and good actual and self-perceived health deeply affected the probability of continuing one's education in the transition from upper secondary school to tertiary education, although this happened through interactions with other factors, consistent with the literature in any case.

The binary variables having an OR lower than 1 implied that the represented group had a lower probability of having y=1 with respect to the complementary group. In Table 3 there are six (interaction) binary variables with an OR lower than 1. For example, "Father perceives poor health" × "Rent is burdensome" had an OR=0.440 and hence its complement to one, expressed as a percentage, was equal to [100*(0.440–1)] = –56.0%. Therefore, the probability of continuing one's education amounted to –56.0% of the probability of the complementary group, which did not have fathers perceiving poor health and a burdensome rents, $\pi_{i|\mathbf{x}_b=\mathbf{0}\wedge\mathbf{x}_c=\boldsymbol{\mu}}$. In other words, the group with x_{12}=1 had a probability equal to $\pi_{x_{12}=1|\bullet}$ = 0.440×0.120= 0.053, implying that the probability of the group with x_{12}=1 decreased the probability of continuing their education by an amount of –56.4% with respect to the complementary group, which had a probability given by $\pi_{i|\mathbf{x}_b=\mathbf{0}\wedge\mathbf{x}_c=\boldsymbol{\mu}}$ = 0.120. In synthesis, unstable and unfavourable parents' working conditions, poor actual and self-perceived health conditions, and critical and costly tenure status of the household negatively affected the probability of continuing one's education in the transition from upper secondary school to tertiary education, although this happened through the interaction terms.

Table 3. Logistic regression with Lasso method and Bayesian approach: Estimated odds ratio (OR), standard errors (SE), p-values (p), and means

B=Binary/ C=Continuous Variables	OR	SE	p	mean
B - Women	1.777	0.263	0.000	0.530
C - [(Individual's age)/10]^2	0.714	0.044	0.000	5.064
C - (Father's age)/10	1.175	0.073	0.003	4.973
C - (Mother's age)/10	1.548	0.094	0.000	4.727
C - (Education Level of Father: years)^2	1.003	0.001	0.000	1.552
C - FDPI= (Father's DPI)/ 10000	1.452	0.070	0.000	2.372
C - MDPI= (Mother's DPI)/ 10000	1.285	0.062	0.000	1.248
C - FTIPC= (Family's total income per capita)/ 10000	0.314	0.046	0.000	1.101
Interactions of first order				
B - (Father: poor health) × (Burdensome rent)	0.440	0.156	0.011	0.023
B - (PES= Parents' Employment Status: pensioners) × NW*	0.494	0.188	0.031	0.017
B - (PES: inactive) × (Densely populated area)	1.697	0.428	0.048	0.043
B - (PES: part-time) × (North-West= NW*)	0.378	0.237	0.042	0.008
B - (PES: full-time employee) × immigrant	0.531	0.092	0.000	0.121
B - (Father: permanent contract) × (Only mother employed)	1.957	0.544	0.013	0.038
B - (Father: permanent contract) × (Parents: manager/ executive)	2.321	0.737	0.013	0.048
B - (Mother: permanent contract) × (Father: limited by health)	1.647	0.322	0.010	0.065
B - (Father: term contract) × (Mother: limited by health)	3.665	1.776	0.011	0.010
B - (TSH$^+$: Subtenant) × (Moderately populated area)	1.466	0.175	0.001	0.246
B - (TSH$^+$: Free) × (Father: poor health)	0.459	0.186	0.025	0.016
B – (TSH$^+$: Free) × (Assets reduction for needs)	2.733	1.041	0.010	0.016
B – (TSH$^+$ [Tenure Status of Household]: Free) × Savings	0.179	0.220	0.023	0.003
Intercept	0.043	0.029	0.000	
Pseudo-R square	0.180	n =	2874	

The continuous variables. The individual's age (range 20-25), expressed in decades, showed a parabolic and negative impact on education paths, while the ages of both parents revealed a linear positive impact on the probability of continuing one's education. The other continuous single variables (which may be conceptually and concretely equal to 0) entering the model showed significant effects on continuing one's education. As parents' education levels increased, the probability of continuing one's education increased quadratically. The father's (FDPI) and mother's (MDPI) disposable personal income indicated a linear positive effect, while the family's total income per capita (FTIPC) yielded an unexpected negative effect, but perhaps the latter balanced the effect of the former. In fact, FTIPC included both FDPI and MDPI too. However, the algebraic sum of their impacts remained positive implying the importance of welfare programmes to help families experiencing economic (and physical) difficulties, with the specific aim of reducing the number of students interrupting their education.

The main fault of the Lasso method in selecting significant explanatory variables concerns the possibility of selecting a theoretically unjustifiable variable, such as "Father with term contract" × "Mother is limited by health" (+266.5%) or of neglecting some important variables in the model.

The conclusions are similar to those explained in Frederic and Lalla (2021): in the applications, the interactions should be supported by social, behavioural, psychological or economic theories. Otherwise, they may be obtained automatically simply by using an adaptive procedure like the Lasso method and only as empirical findings. In fact, few models with interactions exist in the literature. The interactions may probably be easily found among binary or categorical variables, but this case is relatively interesting because they can be replaced with specific typologies. The same holds true for the interactions of a continuous variable with other explanatory binary variables, but the interaction between two continuous variables is very difficult to grasp immediately. In general, it is useful to find a theoretical justification for the existence of the interactions, instead of blindly searching for interaction terms. However, it is highly plausible that almost all phenomena are outcomes of interactions among many variables, but knowledge about and explanations of these results may become very complicated and challenging.

References

Bertolini, P., Lalla, M., Pagliacci, F. (2015). School enrollment of first- and second-generation immigrant students in Italy: A geographical analysis, *Papers in Regional Science*, **94**(1), pp. 141-160.

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

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

Frederic, P., Lalla, M. (2021). Determinants of the transition to upper secondary school: differences between immigrants and Italians, in *ASA 2021 Statistics and Information Systems for Policy Evaluation – Book of short papers of the on-site conference*, eds. Bruno Bertaccini, Luigi Fabbris, Alessandra Petrucci, Firenze University Press, Firenze, pp. 13-18.

Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, **33**(1), pp. 1-22.

Ichou, M., Wallace, M. (2019). The healthy immigrant effect: The role of educational selectivity in the good health of immigrants, *Demographic Research*, **40**(4), pp. 61-94.

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

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

Lalla, M., Frederic, P. (2020). Tertiary education decisions of immigrants and non-immigrants in Italy: an empirical approach, *DEMB Working Papers Series*, N. 168, pp. 1-39. University of Modena and Reggio Emilia - Department of Economics.

Paba, S., 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.

R Core Team (2019). *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna (AT). URL http://www.R-project.org/. Accessed 18 April 2020.

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