Determinants of the transition to upper secondary school: differences between immigrants and Italians

Patrizio Frederic, Michele Lalla

1. Introduction

Education decisions that teenagers in 13-15 age range face are the first important steps in the lifecycle which determine their educational achievements and job trajectory. These choices occur at a particular stage in their lives, when influences inside the home are still strongly felt and knowledge about their interests and abilities or skills is vague and unstable. In this sense, such decisions strongly depend on both individual and family characteristics involving their socio-economic conditions, as well as on the environment or contextual background of the area where they reside.

The objective of this paper is to pinpoint differences with respect to citizenship, a binary variable distinguishing between immigrants and non-immigrants (hereinafter also referred to as Italians), and the *secondary* binary variable, defined as equal to one for individuals who were not enrolled in an upper secondary school and equal to zero otherwise. The Bayesian approach has been applied to investigate the determinants of the secondary variable. The prior distribution was set to be a Laplace distribution with parameter λ. Hence, the Bayesian estimation of the model parameters corresponds to the Lasso estimation procedure. The latter is a popular method that simultaneously allows for the selection of the explanatory variables and their interactions and the estimation of the model coefficients. Starting from an initial model, which includes all the selected quantitative and categorical variables and all the interactions between the categorical variables, the applied method led to a very parsimonious model, but surprisingly it did not include family income.

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 alone, 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 16 to 19, obtaining a sample of 2,702 cases. Then, among the latter, the eligible cases were only those individuals whose highest attained ISCED (International Standard Classification of Education) level was equal to 2 (=lower secondary education). The final target sample was made up of 2,039 individuals.

The relationship between the secondary (binary) dependent variable and the ISCED Level Currently Attended (ILCA) showed that 16.9% of individuals were not enrolled in further education (termed "not-attending"), while 79.7% were currently attending an upper secondary school (Table 1).

Table 1. Absolute frequencies and row percentages of secondary (binary) dependent variable by the ISCED level currently attended (ILCA)

Secondary\ ILCA	Not-attending	Vocational school	Upper secondary	Total
Secondary = 1	344	69		413
	83.3	16.7		100.0
Secondary = 0			1626	1626
			100.0	100.0
Total	344	69	1626	2039
	16.9	3.4	79.7	100.0

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, $CS(2)= 45.177$ ($p<0.000$), where $CS(g)$ stands for Chi-Square with g degrees of freedom: the percentage of immigrants attending upper secondary education was lower than that of Italian citizens (74.3% versus 81.7%), while the percentage of immigrant not in school was higher than that of Italians (24.9% versus 14.0%). There was a significant relationship between the ILCA and self-perceived health, $CS(2)= 8.351$ ($p<0.015$), 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 also related to the index of the total self-perceived health of parents, $CS(6)= 27.356$ ($p<0.000$). The ILCA proved to be linked to the Italian macro-regions $CS(8)= 39.092$ ($p<0.000$), as industrialisation and the possibility of finding employment increased, the percentage of individuals not in school decreased. The ILCA was related to the maximum ISCED level attained by parents, $CS(12)= 179.908$ ($p<0.000$). As the education of parents increased, the percentage of young individuals in school increased. The ILCA yielded significant relationships also with several variables describing the working conditions of the parents, although the strength of such relationships was often weak.

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 four years. Similarly, the mothers of immigrants were younger than the mothers of Italians by about four years and nine months. 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 39.8%.

Table 2. Absolute frequencies, 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

Citizenship\ ILCA	Not-attending	Vocational school	Upper secondary	Total
Italian citizen, *n*	211	65	1229	1505
Means	*8.103*	*9.202*	*10.835*	*10.381*
SD	5.413	9.914	7.867	7.730
Foreign citizen, *n*	133	4	397	534
Means	*5.393*	*2.249*	*6.573*	*6.247*
SD	3.630	1.824	4.338	4.201
Total, *n*	344	69	1626	2039
Means	*7.055*	*8.799*	*9.794*	*9.298*
SD	4.975	9.764	7.396	7.213

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 amounted to about ten thousand euros, i.e., −37.4%. The mothers' disposable personal income presented similar statistically significant differences for both marginal effects, with a gap amounting to about five thousand nine hundred euros, i.e., −39.5%. However, the disposable personal income gender gaps were −51% for Italians and −54% for immigrants.

The size of immigrant families proved to be slightly larger than those of Italians and was statistically significant for both marginal effects, i.e., citizenship and the ILCA.

Citizenship was examined with respect to some other variables, even if it was not a target dependent variable. Its relationship with the maximum ISCED level attained by parents was statistically significant, CS(6)= 97.73 (p<0.000) (Bertolini and Lalla, 2012; Bertolini et al., 2015). Citizenship was significantly related to the degree of urbanisation, CS(2)= 24.225 (p<0.000): immigrants tended to settle in densely populated areas more than Italians (38.4% versus 35.5%) or in moderately populated areas (44.6% versus 39.3%). 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(3)= 29.832 (p<0.000) (Ichou and Wallace, 2019). Citizenship proved to be associated with many variables describing working conditions and revealed a significant relationship with the maximum position of parents on the job, CS(4)= 173.877 (p<0.000).

3. Model by Bayesian Lasso selection of regressors

Let Y be the binary variable coding if the i-th individual is not attending upper secondary education, or he/she is. Let $\mathbf{x}_i$ be a vector of 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

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

A common method that performs estimation and model selection at the same time is the *Lasso* method (Tibshirani, 1996), which 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 named λ, $\lambda \geq 0$. Many penalized methods 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 $p(\boldsymbol{\beta}|\lambda) \propto \exp\left(-\lambda \sum_{j=0}^{K} |\beta_j|\right)$ be the Laplace prior distribution on coefficients $\boldsymbol{\beta}$; then the posterior distribution is

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

The choice of parameter λ plays a crucial role in the estimation procedure. Many different studies have focused on this issue. Besides the classic AIC and BIC criteria, a k-fold Cross Validation (CV) procedure and the One Standard Error Rule (1SE) have been proposed. The applied estimation method consists in 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 consists in 2039 rows by 943 columns, and classical methods can be affected by the *curse of dimensionality*. Instead, the Lasso method is very stable and quick, and shrinks 923 values (out of 943) of $\hat{\boldsymbol{\beta}}_{\lambda_{1SE}}$ to zero; thus only 20 betas have a posterior distribution which is not symmetric to zero.

4. Outcomes of the logistic model

The interpretation of coefficients in a logit model is not easy. The odds ratios (OR) are reported in Table 3, which presents only 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".

A binary variable having an odds ratio 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. The binary variables $(\mathbf{x}_b)$ with an odds ratio greater than 1 were observed for interactions only. For example, the odds ratio of the interaction term the "father is limited in activity because of health problems" $(x_1) \times$ "father with a permanent contract" (x_2), denoted by x_{12}, was equal to 1.826 meaning that the odds of the event *y*=1, when x_{12}=1 (both x_1 and x_2 are equal to 1), are +82.6% greater than the odds of the event *y*=1, when x_{12}=0. Let $\mathbf{x}_c = \boldsymbol{\mu}$ be the mean values of the continuous regressors. Note that: (1) the product of two binary variables is a binary variable again, (2) the percentage of increment of the reference probability, $\pi_{i|\mathbf{x}_b=\mathbf{0} \wedge \mathbf{x}_c=\boldsymbol{\mu}}$, is given by [100*(1–OR)] 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 discontinuing their education) was equal to $\pi_{i|\mathbf{x}_b=\mathbf{0} \wedge \mathbf{x}_c=\boldsymbol{\mu}}$ = **0.160**, calculated at the mean values of the continuous regressors $(\mathbf{x}_c = \boldsymbol{\mu})$ and the binary variables equal to 0 $(\mathbf{x}_b = \mathbf{0})$. Therefore, for x_{12} the result was a probability of $\pi_{x_{12}|\bullet}$ = 1.826×0.160= 0.292, nearly double the probability for x_{12}=0. Similarly, significant high probabilities of discontinuing one's education or dropping out were observed for other interaction terms: "father is limited in activity because of health problems" × "family living in the macro-region Islands" (+149.0%), "father is limited in activity because of health problems" × "family living in a moderately populated area" (+56.3%), "assets reduction for needs" × "young individual with self-perceived bad health" (+234.0%), "assets reduction for needs" × "mother with self-perceived bad health" (+56.3%), "family living in a densely-populated area" × "parents are unemployed or inactive" (+122.7%), "mother only is employed" × "parents skill level on the job is labourer" (+82.3%). In synthesis, real and self-perceived health conditions heavily affect the probability of discontinuing one's education in the transition from lower secondary to upper secondary school and throughout all the secondary school years, although this happens through the interactions with other factors. Note that the "number of helps requests for aid because the family lives in need", which is formally a continuous variable, interacts with "mother suffering from any chronic (long-standing) illness or condition" and yielded an odds ratio greater than 1 at the mean of the first term.

The binary variables having an odds ratio lower than 1 implied that the represented group had a lower probability of having *y*=1 than the complementary group. In Table 3 there are only two binary variables with an odds ratio lower than 1. For example, the binary variable "both parents employed" (BPE) had an odds ratio equal to 0.736 and hence the corresponding complement to one, expressed as a percentage, was equal to [100*(1–0.736)] = 26.4%. Therefore, the probability

of discontinuing one's education amounted to –26.4% (the negative value indicates the reduction quantity) of the probability of the complementary group, which did not have both parents employed, $\pi_{i|\mathbf{x}_b=\mathbf{0}\wedge\mathbf{x}_c=\boldsymbol{\mu}}$. In other words, the group with BPE equal to 1 had a probability $\pi_{\mathrm{BPE}|\bullet}$ = 0.736×0.160= 0.118, implying that the probability of the group with BPE equal to 1 decreased the probability of discontinuing their education by an amount of 100×(1–0.736)= 26.4% with respect to the complementary group, which had a probability given by $\pi_{i|\mathbf{x}_b=\mathbf{0}\wedge\mathbf{x}_c=\boldsymbol{\mu}}$ = 0.160. Similarly, a significant low probability of discontinuing their education was observed for the interaction term the "assets reduction for needs" × "mother with permanent employment contract" (–51.5%). The constant of the model was not statistically significant, even if its magnitude was comparable with other parameters.

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

Variables	OR	SE	*p*	mean
(Intercept)	1.899	1.441	0.1876	
(Age/10)^2	2.766	0.626	0.0000	2.981
(Age father)/10	0.531	0.120	0.0000	4.906
Father: education level in years	0.935	0.328	0.0044	10.598
Mother: education level in years	0.895	0.203	0.0000	10.557
Number of objects owned in home	0.894	0.292	0.0022	4.323
Both parents employed	0.736	0.304	0.0154	0.361
Interactions of first order				
(No. of help requests) × (Mother: chronic illness)	1.146	0.414	0.0056	0.207
(Father: health problems) × (Father: permanent contract)	1.826	0.656	0.0054	0.073
(Father: health problems) × (Macro-region: Islands)	2.490	1.068	0.0198	0.018
(Father: health problems) × (Moderately populated area)	1.563	0.752	0.0376	0.060
(Assets reduction) × (Young individual: bad health)	3.340	1.247	0.0074	0.011
(Assets reduction) × (Mother: bad or poor health)	1.563	0.768	0.0420	0.063
(Assets reduction) × (Mother: permanent contract)	0.485	0.193	0.0118	0.089
(Urban high density) × (Unemployed & inactive)	2.227	0.664	0.0008	0.036
(Only Mother employed) × (labourer)	1.823	0.531	0.0006	0.081
Pseudo-R square	0.171	*n =*	2039	

The continuous variables. The individual age (range 16-19), expressed in decades, showed a parabolic and positive impact on the interruption of education paths before completion of upper secondary school. The high impact may occur for specific reasons: the survey protocols did not interview individuals under the age 16, the vocational school data were not collected well. The other continuous single variables entering the model showed significant effects on the interruption of education. As the ages of fathers and the parents' education levels increased, the probability of discontinuing education decreased. If the number of objects owned in home (dishwasher, refrigerator, telephone, television, and so on) increased, then the risk of interrupting one's education decreased. As indicated above, the increase in the "number of helps requests because the family lives in need" for individuals having a "mother suffering from any a chronic (long-standing) illness or condition" yielded an increase in the risk of dropping out of school. This empirical evidence highlights 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 lack of some income variables in the model because various income components have frequently been found to be significant in the literature (Ochsen, 2011; Krause et al., 2015).

In the applications, the interactions should be supported by social, behavioural, psychological or economic theories. Otherwise, they may be obtained automatically just by using an adaptive procedure like the Lasso method and only as empirical findings. In fact, few models with interactions exist in the literature. Probably, the interactions may 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. (2012). Immigrant Inclusion and Prospects through Schooling in Italy: An Analysis of Emerging Regional Patterns, in *Punishing Immigrants: Policy, Politics and Injustice*, eds. C. E. Kubrin, S. Marjorie, R. M. Jr. Zatz, New York University Press, New York (pp. 178-206).

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.

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

Forster, A.G., van de Werfhorst, H.G. (2020). Navigating Institutions: Parents' Knowledge of the Educational System and Students' Success in Education, *European Sociological Review*, **36**(1), pp. 48-64.

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., and 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.

Krause, A., Rinne, U., Schüller, S. (2015). Kick it like Özil? Decomposing the native-migrant education gap, *International Migration Review*, **49**(3), pp. 757-789.

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.

Ochsen, C. (2011). Recommendation, class repeating, and children's ability: German school tracking experiences, *Applied Economics*, **43**(27), pp. 4127-33.

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.

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