Social selection in higher education. Enrolment, dropout and timely degree attainment in Italy

Supplementary material (Online Appendix)

A. Using or not using sample selection models

In this section we explain why we choose to adopt a descriptive approach and estimate single logit models for enrolment, dropout and timely completion, rather than use models that aim at controlling unobservable traits, like sample selection models. (For example, Montmarquette 2001, Di Pietro 2003 and Di Pietro, Cutillo 2008 model the joint likelihood of enrolment and dropout with bivariate probit, to account for the fact that enrolment and dropout decisions may be influenced by the same individual-specific unobserved factors. Similarly, Cingano, Cipollone 2007 and Ghignoni 2016 use sample selection methods, acknowledging that the dropout probability is estimated on the subgroup of university entrants, so children of lower backgrounds are likely to be positively selected in terms of unobservable characteristics like innate ability or motivation).

Answering different research questions

Consider the typical sample selection model (for simplicity, consider a continuous *y*):

$$z = \gamma + \delta w + u \tag{1a}$$

$$y = \alpha + \beta x + \varepsilon \tag{1b}$$

Let y be observed only if $z > 0$. (u, ε) are possibly correlated errors, so we may write $u = \theta + v_1$, $\varepsilon = \theta + v_2$, where v_1 and v_2 are mutually independent random disturbances, and independent of explanatory variables x and *w*. The parameter of interest is β, the causal effect of x on y, on the entire population. Various estimation methods are proposed (e.g. Heckman's two stage estimator) to account for the fact that the estimation of (1b) on the observed sample produces biased estimates of β.

In our case study, z is university enrolment (propensity) and y is dropout (propensity). x is a vector of explanatory variables including social background (x_1) and prior schooling characteristics (x_2). These variables are likely to affect the enrolment choice as well, so these explanatory variables also enter model (1a). θ can be conceived as innate ability or motivation. For simplicity, imagine a binary x_1 taking value 1 for high social origin and 0 for low social origin.

The parameter of interest in sample selection models is:

$$\beta = E(y|x_1 = 1, x_2, \theta) - E(y|x_1 = 0, x_2, \theta) \tag{2}$$

Thus, β represents the difference in the expected value of the dropout propensity between high ($x_1 = 1$) and low ($x_1 = 0$) social background, *given* observed prior schooling characteristics x_2 and unobserved individual traits θ such as innate ability and motivation.

Instead, the estimation of (1b) on the observed sample of university entrants provides information on:

$$E(y|x_1 = 1, x_2, z > 0) - E(y|x_1 = 0, x_2, z > 0) = \beta + \big(E(\theta|x_1 = 1, x_2, z > 0) - E(\theta|x_1 = 0, x_2\ z > 0)\big) \tag{3}$$

The left hand side term in (3) represents the difference in the dropout propensity between and low social background given prior schooling characteristics, among the group of enrolled students ($z > 0$). This difference is given by β, plus a term accounting for the difference in average motivation/innate ability between enrolled students ($z > 0$) of different social origin with the same prior schooling characteristics. Due to sample selection, the last term is generally negative: the reason is that in order to overcome their disadvantage, individuals of low social origin need to be better in terms of unobserved traits than their peers of high social origin. Hence, in general (3) will be smaller than (2).

Expression (3) provides an answer to the question:

How do university entrants from different family backgrounds with the same prior schooling history behave?

In our perspective, this is the quantity of main interest.
Consider two ideal-typical young individuals just enrolled in university, one of low and one of high social background, having the same schooling history in terms of school-types, curricula and grades. To catch up the disadvantage, the first has needed more effort than the second one to obtain these results, so she will usually have better unobserved personal traits. Being aware of this difference, we are interested in comparing *their* dropout probabilities.

Instead, sample selection methods aim at comparing individuals sharing the same prior schooling history *and* unobserved characteristics, and answer the question:

How would individuals from different family backgrounds with the same prior schooling history, innate ability and motivation behave, if they enrolled in university?

This question would be relevant for prospective students, (i.e. high school diplomats) wishing to evaluate their probability of success in university, given all the relevant characteristics describing them, including ability (provided they have correct knowledge on it). However, in our view it is not particularly salient to evaluate actual inequalities in education. We now attempt to explain why.

The general idea of causal reasoning is to produce counterfactual evidence, i.e. to compare identical individuals differing only in that some are exposed to a treatment and others are not. However, when we make a counterfactual argument on social origin (conceiving parental education or social class as a "treatment"), we should acknowledge that exposition *begins at birth*.

Consider two new born of different family backgrounds, identical in terms of innate abilities. Because of the "treatment", as they grow older they start differentiating (the fact that social origin has a large effect on children's schooling outcomes is demonstrated by an enormous literature). Consequently, students of different family backgrounds may have identical schooling histories if low class individuals compensate their disadvantage with more effort, motivation or higher innate ability. Put it differently, this implies that individuals of different family backgrounds with identical unobserved personal traits may experience identical educational careers only by pure luck (i.e., due to the effect of the idiosyncratic random term component). Consequently, it is unusual to find comparable individuals. In this sense, the contrast between students of

different family backgrounds with identical prior schooling history and identical unobserved personal traits seems to be rather odd.[1]

Statistical reasons

There are two additional reasons for preferring a descriptive approach in this context.

The first is that applying the sample selection model on high school graduates we are ignoring a previous stage of selection, because only a subgroup of the initial birth cohort eventually attains the high school diploma (approximately 80% in Italy).

The second is that, as emphasized by Brandt and Schneider (2007) and Kennedy (2003; pg. 291), sample selection estimators perform very poorly if model assumptions are violated – the neglect of the first stage of selection is an example – or when the degree of collinearity between the explanatory variables in the regression and the selection equations is high.[2] This concern applies to our case study, since it is difficult to think of determinants of enrolment not influencing also the dropout and completion probability.

Additional References

Brandt P.T., C. J. Schneider (2007) So the reviewer told you to use a selection model? Selection models and the study of international relations, http://pages.ucsd.edu/~cjschneider/working_papers/pdf/Selection-W041.pdf

Kennedy P. (2003) *A Guide to Econometrics*, MIT Press.

[1] It is worth noticing that this argument may not apply to family income, as income is more variable over time and may be exogenously increased by scholarships or loans.

[2] We have also developed a simulation study confirming these findings. Results are available upon request.

B. Comparison with official rates (cf. note 9 main manuscript)

Dropout rates obtained by all waves of the Survey on High-School Graduates are lower than the corresponding aggregate rates reported by the Ministry of Education – for example, the official 1-year dropout rate was around 16% for 3-year programs and 9% for 5-year programs in 2008-9 (ANVUR, 2013). In analysing earlier waves of the survey, Cingano and Cipollone (2007) offer some potential explanations of the discrepancy, suggesting that there might be some misreporting in the survey data. In particular, some students might omit declaring university enrolment if it is a transient state not followed by actual participation, or they might omit declaring dropout if they hope resuming their studies in the near future. We find no empirical support for the former explanation, as enrolment rates in the survey are very similar to official ones. Instead, we agree with Cingano, Cipollone on the potential relevance of the latter, but we cannot check it empirically. We suggest two additional explanations: (i) Official dropout rates refer to the entire student-body, while our figures refer to the subsample of students enrolled within few years after high school graduation. There is wide evidence that older students and working-students, accounting for a significant share of the Italian university student population, have substantially higher dropout rates. (ii) The survey sample does not seem fully representative in terms of student ability and motivation, as the final high school examination grade distribution is more favourable among survey participants than according to official statistics. Overall, since disadvantaged students are less likely to enter the survey, we will consider our results as conservative estimates of socioeconomic inequalities.