

Analysis of the Effectiveness of Degree Programmes by means of Principal Stratification

*Analisi di Efficacia dei Corsi di Laurea per mezzo della
Stratificazione Principale*

Leonardo Grilli, Fabrizia Mealli
Dipartimento di Statistica “G. Parenti”
Viale Morgagni, 59 Firenze
grilli@ds.unifi.it, mealli@ds.unifi.it

Riassunto: Il contributo indica come studiare l’efficacia esterna di due corsi di laurea per mezzo della metodologia degli strati principali, focalizzando l’attenzione sulla costruzione, tramite metodi non parametrici, di intervalli di valori ammissibili per gli effetti causali oggetto dell’inferenza.

Keywords: causal effects, effectiveness, potential outcomes, principal strata.

1. Introduction

The present work jointly analyses the careers and the job opportunities of university students. The main aim is to investigate, for a cohort of freshmen of the University of Florence enrolled in two distinct degree programmes, the factors determining the success in the academic context (achieving graduation or not) and the success on the job market (getting employed or not), with special emphasis on the analysis of direct causal effects of the degree programmes on the employment status.

Traditional analyses of the effect of degree programmes on the employment status (external effectiveness) are performed on the sole basis of graduated students, neglecting the fact that the set of students who are able to graduate in a certain degree programme is in principle different from the set of students who are able to graduate in another degree programme. In other words, two degree programmes might select different kinds of students with different attitudes and propensity to be employed. An analysis of employment status based only on graduated students mixes the direct effect of the degree programme on employment status with the indirect effect through the graduation status.

To study the direct effect of a degree programme on employment status it is convenient to put the problem in the framework of principal strata (Frankgakis & Rubin 2002, Barnard *et al.*, 2003), which is a development of Rubin’s approach to casual analysis. In the present application the intermediate variable (i.e. a post-treatment variable) defining the principal strata is the graduation status, which assumes only two values (graduated vs. not graduated). The key point is that if a student does not graduate the outcome variable is not defined for the purpose of assessing the effectiveness of the degree programmes with respect to the job market.

The present analysis is limited to the comparison of only two degree programmes, Economics (Economia e Commercio) and Political Science (Scienze Politiche), but the extension to many degree programmes is conceptually simple.

2. The data and the principal strata framework

The joint analysis of the careers and job opportunities of the university students requires merging two data sources: (a) the administrative database of the 1992's cohort of freshmen enrolled in the degree programmes in Economics and Political Science of the University of Florence; and (b) three census surveys on the occupational status of the graduates of the University of Florence of the years 1998, 1999 and 2000, respectively. Overall, 1941 freshmen belong to the examined 1992's cohort: 1068 in Economics and 873 in Political Science. The academic status of the students by the end of the year 2000 and the employment status at the interview for the subset of graduated students are summarized in Table 1. For the students still enrolled there are no survey data to assess their employment status if they eventually graduate. Moreover, graduation after nine years can be considered somewhat anomalous. Therefore for the purpose of the present analysis graduation is defined as "graduation within nine years from enrollment".

Table 1: Cohort of the year 1992: academic status by the end of the year 2000 and employment status at the interview

<i>Degree Prog.</i>	<i>Dropped</i>	<i>Graduated</i>	<i>Still enrolled</i>	<i>Interviewed</i>	<i>Permanent job</i>
<i>Economics</i>	545 51.0%*	270 25.3%*	253 23.7%*	186 68.9%**	96 51.6%***
<i>Political Science</i>	532 60.9%*	176 20.2%*	165 18.9%*	99 56.2%**	36 36.4%***

* % on Enrolled

** % on Graduated

*** % on Interviewed

All the interviewed graduates responded to the question on the employment status. Apart from 21 students who graduated before 1998, who are out of the target of the surveys, almost all missing interviews are due to missing contact.

The *treatment variable* Z_i is defined as: $Z_i = 1$ if student i was enrolled in Economics; $Z_i = 0$ if student i was enrolled in Political Science. Under the standard Stable Unit Treatment Value Assumption (SUTVA: Frankgakis and Rubin, 2002), the post-treatment variables are defined as follows. The *intermediate variable* $S_i(z_i)$ is 1 or 0 if student i graduated or did not graduate by the end of 2000 (i.e. within 9 years) when enrolled in degree programme z_i . The *response variable* $Y_i(z_i)$ is 1 or 0 if student i had or did not have a permanent job (provided he/she graduated) at the time of the interview, i.e. from one to two years after the degree, when enrolled in degree programme z_i and graduated.

Since for each individual the treatment variable assumes a single value, for every post-treatment variable only one of the two potential versions can be observed: therefore it is useful to introduce the notation $S_i^{obs} = S_i(Z_i)$, $Y_i^{obs} = Y_i(Z_i)$. Since both the treatment variable and the intermediate variable are dichotomous, four *principal strata* can be defined in the following way through the latent variable L_i : $L_i = 'GG'$ (Graduated, Graduated) if $S_i(1)=1$ and $S_i(0)=1$; $L_i = 'GN'$ (Graduated, Not graduated) if $S_i(1)=1$ and $S_i(0)=0$; $L_i = 'NG'$ (Not graduated, Graduated) if $S_i(1)=0$ and $S_i(0)=1$; $L_i = 'NN'$ (Not graduated, Not graduated) if $S_i(1)=0$ and $S_i(0)=0$. The principal strata are latent classes and the data only allow to estimate the probability that a given student belongs to a certain latent class. If the purpose is to evaluate the effectiveness of the degree programme with respect to the job market, the response variable Y is defined only for

the graduates, so the causal effect $Y_i(1)-Y_i(0)$ is properly defined only for the GG stratum, i.e. the students who are able to graduate in both degree programmes.

The administrative database includes some variables for the whole cohort, used to build the covariate vector $\mathbf{x}_i$. All the covariates are dichotomous, with the first category being 1 and the second category being 0: Female vs. Male; Residence in Florence vs. others; Gymnasium vs. others; High grade (final mark 50-60) vs. Low grade; Late enrollment (i.e. did not enroll in the first possible academic year) vs. Regular enrollment.

Since Z is not randomised there may be some confounders that influence both Z and S or Y : in such a case the effect of Z on Y cannot be interpreted as a causal effect. The available covariates may help alleviate this problem and this is what underlies the assumption of *conditional unconfoundedness of treatment assignment*: $Z_i \perp S_i(0), S_i(1), Y_i(0), Y_i(1) \mid \mathbf{x}_i$. Moreover it is assumed that the response mechanism is ignorable, so the analysis can be safely based on the available responses.

Under the assumptions of SUTVA and conditional unconfoundedness of the treatment and of the response mechanism, the whole data generating process can be defined in terms of the following two sets of probabilities. (a) Probabilities of the principal strata:

$\pi_{GG;i}, \pi_{GN;i}, \pi_{NG;i}, \pi_{NN;i}$, where for example $\pi_{GG;i} = \Pr(L_i = 'GG' \mid \mathbf{x}_i)$; (b) Probabilities of the response variable, conditional on the principal stratum: $\gamma_{1,GG;i}, \gamma_{0,GG;i}, \gamma_{1,GN;i}, \gamma_{0,GN;i}$, where for example $\gamma_{1,GG;i} = \Pr(Y_i(1) = 1 \mid Z_i = 1, L_i = 'GG', \mathbf{x}_i)$. The probabilities π s and γ s are not directly estimable from the data without further assumptions. Note that the γ s corresponding to other combinations of degree programme and principal stratum are not defined in the present application.

3. Large sample bounds for the causal effects

As a first step of analysis, before trying to fit a model, it is useful to determine the range of admissible probabilities for the principal strata and the corresponding bounds on the causal effect of interest. Let us first assume that the population is homogenous, i.e. the covariates are ignored. Following a procedure similar to that of Zhang and Rubin (2004), we first fix one of the π s, in this case $\pi_{GG;i}$ (which can vary between 0 and 0.202), and then using the sample proportions we estimate the bounds for $\gamma_{1,GG;i} - \gamma_{0,GG;i}$, which could be plotted as functions of $\pi_{GG;i}$. The bounds widen as $\pi_{GG;i}$ becomes smaller: for high values of $\pi_{GG;i}$ (between 0.196 and the maximum 0.202) the bounds are both positive, so the sign of the causal effect is determined; then the bounds widen until they reach the interval $[-1,1]$, becoming useless. In large samples these bounds estimate the true bounds nearly without error and there is no need to consider the confidence limits. The bounds convey the uncertainty involved in the estimation of the average causal effect for the GG stratum irrespective of the sample size.

The bounds can be sharpened by making some suitable assumptions. A sensible restriction on the probabilities of the principal strata is that the students who are able to graduate in both degree programmes are a majority in the group of students who are able to graduate in at least one degree programme. This leads to the assumption of *Relative majority of the GG stratum*: $\pi_{GG;i} \geq \pi_{NG;i} + \pi_{GN;i}$. Since the bounds widen as $\pi_{GG;i}$ diminishes, under this assumption the widest bounds correspond to

$\pi_{GG:i} = \pi_{NG:i} + \pi_{GN:i}$, which is 0.152 in the present application, leading to the bounds $[-0.290, 0.708]$, which are obviously more informative than the interval $[-1, 1]$. As for the probabilities of the response variable it is sensible to assume that the students who are able to graduate in both degree programmes (GG stratum) have more chances to get a permanent job than the students who are able to graduate in a degree programme but not in the other one (NG and GN strata). This consideration leads to the *Stochastic dominance* assumption: $\gamma_{1,GG:i} \geq \gamma_{1,GN:i}$ and $\gamma_{0,GG:i} \geq \gamma_{0,NG:i}$. Under stochastic dominance the bounds are narrower than in the general case: in particular, when $\pi_{GG:i} = \pi_{NG:i} + \pi_{GN:i}$ the estimated bounds become $[0.030, 0.494]$, i.e. the average causal effect is necessarily positive. This shows that two weak assumptions may be sufficient to determine the sign of the effect without relying on parametric models. Since the unconfoundedness assumption is much more sensible when expressed conditional on the covariates, it is advisable to derive the bounds for each cell defined by a combination of the covariates. Since the cells may have low frequencies, we replace the cell sample proportions with the estimates obtained with logit regression. To study the dependence of the bounds on the covariates it is useful to restrict the attention to a single scenario, for example the one under stochastic dominance and $\pi_{GG} = \pi_{NG} + \pi_{GN}$ for each cell (see Table 2, where *Baseline* refers to the individual with all covariates equal to zero and each subsequent line refers to the individual who differs from the baseline one only for a single covariate). Note that for the baseline individual the sign of the causal effect is quite uncertain, but a single change in the covariate *Late enrollment* is sufficient to constrain the effect to be positive. The unconditional bounds can be reconstructed from the conditional ones through an average weighted by the cell frequencies, obtaining an interval $[-0.014, 0.438]$, which is shifted downward with respect to the homogenous population bounds $[0.030, 0.494]$. To get more powerful results, we plan to continue the analysis with a suitable parametric model to be fitted with likelihood or Bayesian methods.

Table 2: Some lower and upper bounds for the causal effect for the GG stratum

Type	Frequency	Prob. of principal strata				Bounds	
		GG	NG	GN	NN	Lower	Upper
Baseline	305	0.07	0.02	0.05	0.85	-0.21	0.26
Gymnasium	194	0.18	0.06	0.12	0.65	-0.10	0.31
Female	140	0.08	0.02	0.05	0.84	-0.07	0.43
High grade	118	0.19	0.06	0.13	0.63	-0.05	0.43
Residence in Florence	85	0.08	0.05	0.03	0.84	-0.29	0.23
Late enrollment	64	0.02	0.00	0.02	0.96	0.21	0.60

References

- Barnard J., Frangakis C.E., Hill J.L. & Rubin D.B. (2003) Principal Stratification Approach to Broken Randomized Experiments: A Case Study of School Choice Vouchers in New York City, *Journal American Statistical Association*, 98, 299-323.
- Frankgakis, C. E. & Rubin, D. B. (2002) Principal Stratification in Causal Inference, *Biometrics*, 58, 21-29.
- Zhang, J. & Rubin, D. B. (2004) Censoring due to death via principal stratification. To appear in the *Journal of Educational and Behavioral Statistics*.