Convegno della Società Italiana di Statistica
Bari, 9-11 giugno 2004

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

Leonardo Grilli
grilli@ds.unifi.it

Fabrizia Mealli
mealli@ds.unifi.it

Università degli Studi di Firenze
Dipartimento di Statistica "Giuseppe Parenti"

Outline

- The structure of our application
- Causal effects with intermediate variables and principal strata
- Non parametric bounds
- Likelihood-based analysis

Our application: effectiveness of degree programmes

Joint analysis of the *careers* and the *job opportunities* of university students

- 1992's cohort of freshmen of the University of Florence
- two distinct degree programmes, Economics and Political Sciences

Our application: effectiveness of degree programmes

Why do we need a joint analysis?

- employment status is observed only for graduated students, while the effect of interest concerns all enrolled students
- it is possible that the two d.p. "select" the indivudals in a different way, so a comparison based only on graduated students is not fair

Data

A. Administrative database of the 1992's cohort of freshmen enrolled in the degree programmes in Economics (Economia e Commercio) and Political Sciences (Scienze Politiche) of the University of Florence

B1-B3. Three census surveys on the occupational status of the graduates of the University of Florence of years 1998, 1999 and 2000, respectively

Datasets A and B1-B3 are merged

Data

1941 freshmen belong to the examined 1992's cohort: 1068 in *Economics* and 873 in *Political Sciences*. By the end of the year 2000 the status of the students is the following:

Degree Programme	Dropped	Graduated	Still enrolled	Total
Economics	545 51.03%	270 25.28%	253 23.69%	1068
Political Sciences	532 60.94%	176 20.16%	165 18.90%	873

Data

After the merge with the survey data the situation is:

Degree Programme	Graduated	Interviewed	Permanent job
Economics	270	186 68.89%*	96 51.61%**
Political Sciences	176	99 56.25%*	36 36.36%**

* Interviewed/Graduated **Permanent job/Interviewed

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

Data

Covariate	*Economics (n=1068)*	*Political Science (n=873)*
Female	0.41	0.54
Residence in Florence	0.23	0.31
Gymnasium	0.34	0.45
Late enrollment	0.06	0.22
High grade	0.37	0.25

Covariates are important since the treatment is not randomized!

Effectiveness of degree programmes

Treatment variable Z:

$$Z = \begin{cases} 1 & \text{if enrolled in Economics} \\ 0 & \text{if enrolled in Political Sciences} \end{cases}$$

- No active vs. placebo → values of Z on an equal footing
- No randomisation → possible confounders

Effectiveness of degree programmes

The main aim is to investigate the factors determining

- the success in the academic context (*achieving graduation or not* - intermediate variable S) and
- the success on the job market (*getting employed or not* - response variable Y)

with special emphasis on the analysis of *causal effects* of the degree programmes on the job status

Effectiveness of degree programmes

Intermediate variable S:

$$S = S(Z) = \begin{cases} 1 & \text{if graduated when } Z \\ 0 & \text{if not graduated when } Z \end{cases}$$

S is the observed version of the potential outcomes $S(0)$, $S(1)$

Response variable Y:

$$Y = Y(Z) = \begin{cases} 1 & \text{if job (after graduation) when } Z \\ 0 & \text{if not job (after graduation) when } Z \end{cases}$$

Y is the observed version of the potential outcomes $Y(0)$, $Y(1)$

Effectiveness of degree programmes

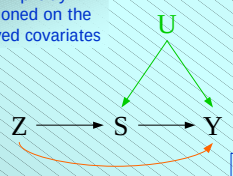
For our purposes Y is defined only when $S=1$

Causal effect of Z on Y for an individual:
$Y_i(1)-Y_i(0)$

Is this causal effect defined for all individuals?

Intermediate variables

Graph implicitly conditioned on the observed covariates



Z = treatment

Y = response

S = intermediate

U = unobs. variables

arrow Z → Y: direct effect

Assumption: conditionally on the observed covariates there are no unobserved confounders, i.e. no arrow U → Z

Causal effects with an intermediate variable

The arrow Z → Y represents the direct effect of Z on Y, i.e. the effect non mediated by S

When the arrow Z → Y is present, it may be that the effect is defined *only* for a subset of individuals, e.g. those individuals who have a certain value of S irrespective of the value of Z

This idea is based on the concept of potential outcomes (counterfactual reasoning)

Principal strata

In our case both Z and S are dichotomous → 4 possible strata

Z	GG	GN	NG	NN
1	G	G	N	N
0	G	N	G	N

G=Graduated

N=Not graduated

Principal strata are defined by the values of the two potential versions of the intermediate variable S (counterfactual)

Principal strata are not influenced by Z (nor S)

The membership indicator of the principal stata is a partially observed covariate (in general data cannot reveal which principal stratum an individual belongs to)

Causal inference with principal strata

Principal causal effect of Z on Y:

a comparison of p(Y(1)) vs. p(Y(0))

for the individuals of a given principal stratum

Causal effects across principal strata are nonsense

It may be that the causal effect is defined only for some principal strata: *in our case only for the GG stratum*

Relationships between observed and latent groups

Observed group $O(Z, S^{obs})$	Z_i	S_i^{obs}	R_i^{obs}	Y_i^{obs}	Latent group L_i (principal stratum)
$O(1,1)$	1	1	$\in \{0,1\}$	$\in \{0,1\}$	GG or GN
$O(1,0)$	1	0	not defined	not defined	NG or NN
$O(0,1)$	0	1	$\in \{0,1\}$	$\in \{0,1\}$	GG or NG
$O(0,0)$	0	0	not defined	not defined	GN or NN

$P_{S,11}$ = **0.253** sample proportion of graduates among students in Economics (Z_i=1)

$P_{S,01}$ = **0.202** sample proportion graduates among students in Political Science (Z_i =0)

$P_{Y,11}$ = **0.516** sample proportion of individuals with a permanent job among students in Economics (Z_i=1) who gradutated (S_i^{obs} = 1) and responded to the interview (Y_i^{obs} = 1)

$P_{Y,01}$ = **0.364** sample proportion of individuals with a permanent job among students in Political Science (Z_i =0) who gradutated (S_i^{obs} = 1) and responded to the interview (Y_i^{obs} = 1)

Calculation of the bounds

Probabilities of the principal strata: π_{GG}, π_{GN}, π_{NG}, π_{NN}

Probabilities of having job: $\gamma_{1,GG}$, $\gamma_{0,GG}$, $\gamma_{1,GN}$, $\gamma_{0,NG}$

Two sensible assumptions:

Relative majority of the GG stratum: $\pi_{GG} \geq \pi_{NG} + \pi_{GN}$

Stochastic dominance: $\gamma_{1,GG} \geq \gamma_{1,GN}$ $\quad \gamma_{0,GG} \geq \gamma_{0,NG}$

Calculation of the bounds

Large sample non parametric bounds

(under the assumption that the treatment is assigned at random and the population is homogenous)

$P_{S,11}$ estimates $\pi_{GG} + \pi_{GN}$ $\quad$ $P_{S,01}$ estimates $\pi_{GG} + \pi_{NG}$

$P_{Y,11}$ estimates $\gamma_{1,GG}\frac{\pi_{GG}}{\pi_{GG}+\pi_{GN}} + \gamma_{1,GN}\frac{\pi_{GN}}{\pi_{GG}+\pi_{GN}}$

$P_{Y,01}$ estimates $\gamma_{0,GG}\frac{\pi_{GG}}{\pi_{GG}+\pi_{NG}} + \gamma_{0,NG}\frac{\pi_{NG}}{\pi_{GG}+\pi_{NG}}$

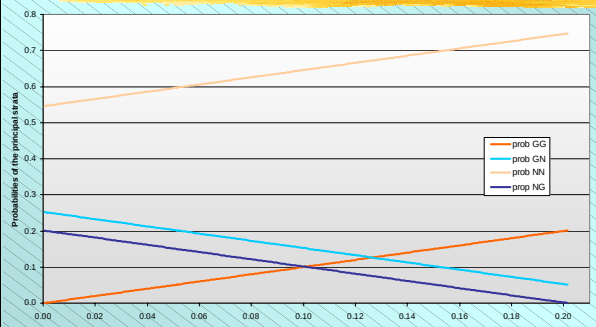
Calculation of the bounds

Large sample non parametric bounds

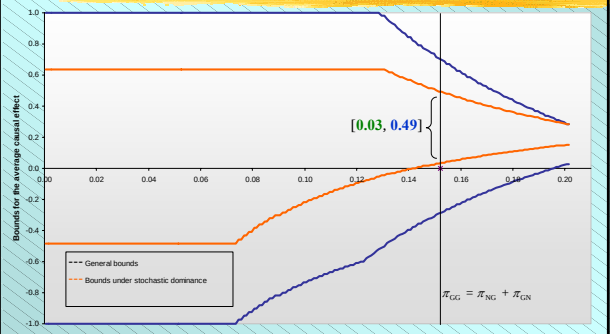
Fixing π_{GG} it is possible to estimate the probabilities of the principal strata (π_{GG}, π_{GN}, π_{NG}, π_{NN}) and calculate the bounds of the *average causal effect in the GG stratum*

$$\gamma_{1,GG} - \gamma_{0,GG}$$

Admissible values of the probabilities of the principal strata – homogenous population



Bounds for the causal effect in the GG stratum – homogenous population



Bounds for the causal effect in the GG stratum – conditional bounds

Bounds under Stochastic Dominance and $\pi_{GG} = \pi_{NG} + \pi_{GN}$

		Prob. of principal strata				Bounds		Indexes	
Type	Frequency	GG	NG	GN	NN	Lower	Upper	Det.	Width
Baseline	305	0.07	0.02	0.05	0.85	-0.21	0.26	0.11	0.24
Gymnasium	194	0.18	0.06	0.12	0.65	-0.10	0.31	0.52	0.20
Female	140	0.08	0.02	0.05	0.84	-0.07	0.43	0.71	0.25
High grade	118	0.19	0.06	0.13	0.63	-0.05	0.43	0.78	0.24
Residence in Florence	85	0.08	0.05	0.03	0.84	-0.29	0.23	-0.12	0.26
Late enrollment	64	0.02	0.00	0.02	0.96	0.21	0.60	1.00	0.20

Bounds for the whole population reconstructed through an average weighted by the cell frequencies: [**-0.01**, **0.44**]

Likelihood-based inference

$$L\left(\boldsymbol{\theta} \mid \mathbf{Z}, \mathbf{S}^{obs}, \mathbf{R}^{obs}, \mathbf{Y}^{obs}, \mathbf{X}\right) =$$

$$\prod_{i\in O(1,1)} \left\{ \pi_{GGi} \left[\left(\gamma_{1,GGi}\right)^{Y_i^{obs}} \left(1-\gamma_{1,GGi}\right)^{1-Y_i^{obs}} \right]^{R_i^{obs}} + \pi_{GNi} \left[\left(\gamma_{1,GNi}\right)^{Y_i^{obs}} \left(1-\gamma_{1,GNi}\right)^{1-Y_i^{obs}} \right]^{R_i^{obs}} \right\}$$

$$\times \prod_{i\in O(1,0)} \left\{ \pi_{NGi} + \pi_{NNi} \right\}$$

$$\times \prod_{i\in O(0,1)} \left\{ \pi_{GGi} \left[\left(\gamma_{0,GGi}\right)^{Y_i^{obs}} \left(1-\gamma_{0,GGi}\right)^{1-Y_i^{obs}} \right]^{R_i^{obs}} + \pi_{NGi} \left[\left(\gamma_{0,NGi}\right)^{Y_i^{obs}} \left(1-\gamma_{0,NGi}\right)^{1-Y_i^{obs}} \right]^{R_i^{obs}} \right\}$$

$$\times \prod_{i\in O(0,0)} \left\{ \pi_{GNi} + \pi_{NNi} \right\}$$

Likelihood-based inference

Principal strata submodel (π's)

$$\pi_{GG:i} = \frac{\exp(\eta^{\pi}_{GG:i})}{1+\exp(\eta^{\pi}_{GG:i})+\exp(\eta^{\pi}_{GN:i})+\exp(\eta^{\pi}_{NG:i})}$$

$$\pi_{GN:i} = \frac{\exp(\eta^{\pi}_{GN:i})}{1+\exp(\eta^{\pi}_{GG:i})+\exp(\eta^{\pi}_{GN:i})+\exp(\eta^{\pi}_{NG:i})}$$

$$\pi_{NG:i} = \frac{\exp(\eta^{\pi}_{NG:i})}{1+\exp(\eta^{\pi}_{GG:i})+\exp(\eta^{\pi}_{GN:i})+\exp(\eta^{\pi}_{NG:i})}$$

$$\pi_{NN:i} = \frac{1}{1+\exp(\eta^{\pi}_{GG:i})+\exp(\eta^{\pi}_{GN:i})+\exp(\eta^{\pi}_{NG:i})}$$

$$\eta^{\pi}_{GG:i} = \alpha^{\pi}_{GG} + \boldsymbol{\beta}^{\pi}_{GG}\,'\mathbf{x}_i$$

$$\eta^{\pi}_{GN:i} = \alpha^{\pi}_{GN} + \boldsymbol{\beta}^{\pi}_{GN}\,'\mathbf{x}_i$$

$$\eta^{\pi}_{NG:i} = \alpha^{\pi}_{NG} + \boldsymbol{\beta}^{\pi}_{NG}\,'\mathbf{x}_i$$

Likelihood-based inference

Outcome submodel (γ's)

$$\gamma_{1,GG:i} = \frac{1}{1+\exp(-\eta^{\gamma}_{1,GG:i})}$$

$$\gamma_{0,GG:i} = \frac{1}{1+\exp(-\eta^{\gamma}_{0,GG:i})}$$

$$\gamma_{1,GN:i} = \frac{1}{1+\exp(-\eta^{\gamma}_{1,GN:i})}$$

$$\gamma_{0,NG:i} = \frac{1}{1+\exp(-\eta^{\gamma}_{0,NG:i})}\,.$$

$$\eta^{\gamma}_{1,GG:i} = \alpha^{\gamma}_{1,GG} + \boldsymbol{\beta}^{\gamma}\,'\mathbf{x}_i$$

$$\eta^{\gamma}_{0,GG:i} = \alpha^{\gamma}_{0,GG} + \boldsymbol{\beta}^{\gamma}\,'\mathbf{x}_i$$

$$\eta^{\gamma}_{1,GN:i} = \alpha^{\gamma}_{1,GN} + \boldsymbol{\beta}^{\gamma}\,'\mathbf{x}_i$$

$$\eta^{\gamma}_{0,NG:i} = \alpha^{\gamma}_{0,NG} + \boldsymbol{\beta}^{\gamma}\,'\mathbf{x}_i\,.$$

Likelihood-based inference

- Model has 27 parameters
- The treatment and the five covariates lead to 128 theoretical sample proportions
- The available sample proportions are 99

- Maximization algorithm: quasi-Newton with a BFGS update of the Cholesky factor of the approximate Hessian.
- Software: SAS proc nlmixed

Likelihood-based inference

- Some parameters of the Principal strata submodel (π's) have
  - highly negative estimates and
  - huge standard errors

for certain values of the covariates some principal strata are empty

some constraints are needed

Principal strata submodel results

	Initial model		*Final model*	
Number of parameters	27		21	
Deviance (-2*logL*)	2231.8		2231.8	
Principal strata submodel (π's)				
α^{π}_{GG}	-4.403	(0.449)	-4.402	(0.448)
α^{π}_{GN}	-2.644	(0.749)	-2.647	(0.752)
α^{π}_{NG}	-3.206	(0.836)	-3.207	(0.835)
$\beta^{\pi}_{GG,gymnasium}$	1.275	(0.157)	1.275	(0.157)
$\beta^{\pi}_{GN,gymnasium}$	-5.757	(n.a.)	$-\infty$	
$\beta^{\pi}_{NG,gymnasium}$	-15.041	(n.a.)	$-\infty$	
$\beta^{\pi}_{GG,high\_grade}$	1.204	(0.146)	1.205	(0.146)
$\beta^{\pi}_{GN,high\_grade}$	1.113	(0.653)	1.113	(0.652)
$\beta^{\pi}_{NG,high\_grade}$	-8.092	(114.022)	$-\infty$	
$\beta^{\pi}_{GG,regular\_enrolment}$	2.024	(0.425)	2.023	(0.425)
$\beta^{\pi}_{GN,regular\_enrolment}$	-0.012	(0.788)	-0.009	(0.792)
$\beta^{\pi}_{NG,regular\_enrolment}$	-8.140	(64.473)	$-\infty$	
$\beta^{\pi}_{GG,female}$	0.117	(0.137)	0.117	(0.137)
$\beta^{\pi}_{GN,female}$	-0.617	(0.753)	-0.622	(0.755)
$\beta^{\pi}_{NG,female}$	0.988	(1.112)	0.991	(1.111)
$\beta^{\pi}_{GG,Florence}$	0.280	(0.144)	0.280	(0.144)
$\beta^{\pi}_{GN,Florence}$	-13.499	(559.599)	$-\infty$	
$\beta^{\pi}_{NG,Florence}$	-10.353	(533.855)	$-\infty$	

Outcome submodel results

	Initial model		*Final model*	
Number of parameters	27		21	
Deviance (-2*logL*)	2231.8		2231.8	
Outcome submodel (γ's)				
$\alpha^{\gamma}_{1,GG}$	1.257	(1.240)	1.262	(1.241)
$\alpha^{\gamma}_{0,GG}$	-1.357	(1.561)	-1.365	(1.568)
$\alpha^{\gamma}_{1,GN}$	0.593	(1.185)	0.596	(1.185)
$\alpha^{\gamma}_{0,NG}$	0.498	(1.057)	0.484	(1.058)
$\beta^{\gamma}_{gymnasium}$	-0.405	(0.374)	-0.410	(0.374)
$\beta^{\gamma}_{high\_grade}$	-0.035	(0.262)	-0.036	(0.263)
$\beta^{\gamma}_{regular\_enrolment}$	-0.933	(0.979)	-0.932	(0.979)
β^{γ}_{female}	0.072	(0.272)	0.070	(0.272)
$\beta^{\gamma}_{Florence}$	0.106	(0.333)	0.104	(0.333)
Causal effect $\alpha^{\gamma}_{1,GG} - \alpha^{\gamma}_{0,GG}$	0.664	(0.301)	0.666	(0.301)

Estimated probabilities (per cent) for some covariates' patterns

Probability	00000	00100	00110	00101	01100	10100	11100	11111
π_{GG3}	1.1	8.0	9.1	10.9	20.3	24.9	52.5	62.2
π_{GN3}	6.3	6.0	3.3	0.0	14.0	0.0	0.0	0.0
π_{NG3}	3.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
π_{NN3}	89.0	86.0	87.6	89.1	65.7	75.1	47.5	37.8
$\gamma_{1,GG3}$	77.9	58.2	59.9	60.7	57.3	48.0	47.1	51.5
$\gamma_{0,GG3}$	64.5	41.7	43.4	44.2	40.8	32.2	31.4	35.3
$\gamma_{1,GN3}$	61.9	39.0	40.7	41.5	38.1	29.8	29.0	32.8
$\gamma_{0,NG3}$	20.3	9.1	9.7	10.0	8.9	6.3	6.1	7.1
Causal effect $\gamma_{1,GG3}-\gamma_{0,GG3}$	13.5	16.5	16.5	16.4	16.5	15.8	15.7	16.2

Note: the pattern (x_1,x_2,x_3,x_4,x_5) stands for $Gymnasium=x_1$, $High\ grade=x_2$, $Regular\ enrolment=x_3$, $Female=x_4$, $Florence=x_5$.

Principal strata submodel results

- The estimated proportion of students belonging to the *GG* group varies a lot with the covariates, from a minumum of 1.1% to a maximum of 62.2%
- the proportion of students belonging to the *GN* and *NG* groups (i.e. the students able to graduate in only one degree programme) tends to diminish as the *GG* stratum grows even if the *NN* stratum goes down

Principal strata submodel results

- the two degree programmes have a *differential effect on the probability of graduation* only for students having a weak background. Orientation policies should then be designed especially for this kind of students.
- the assumption of relative majority of the *GG* stratum used in the construction of the conditional bounds generally holds, though with the exception of the individuals who enrolled late.

Outcome submodel results

- the causal effect on the GG group (on the logit scale) is estimated as 0.666 (s.e. 0.301), so it is significantly different from zero at the 5% level
- the reliability and also the substantive importance of the causal effect depends on the size of the *GG* stratum: for example, the causal effect for the *GG* group for the baseline individual has little meaning

Outcome submodel results

- the assumption of stochastic dominance holds
- The level of the probability of being employed varies a lot with the covariates:
  - 47.1% to 77.9% for Economics
  - 31.4% to 64.5% for Political Science

Further developments

- Sensitivity analysis
- Bayesian analysis
- Model for the missing outcomes
- Implications for policy