Econometrics part 2, PS 6

Mangiante Giacomo

May 3, 2019

Discussed with: Aleksandra Malova

1 6.1: Matching vs Regressions

1.1 a)

Yes, gender can be considered a treatment. Indeed, it allows to perfectly distinguishing between who are treated (male in our case) and who are not (female). However, it is not a good treatment to evaluate the effects on admission since it is probably to be correlated with other characteristics that are likely to influence the likelihood of being admitted.

1.2 b)

If we assume full independence between treatment and potential outcome then the average treatment effect can be computed by the naive estimator:

$$ATE = \Delta^{naive} = \bar{Y}_1 - \bar{Y}_0 = 44\% - 30\% = 15\%$$

1.3 c)

No, other characteristics, apart from gender, could influence tha admission.

1.4 d)

Since treatment and potential outcome are independent conditional on the field of study we can compute the overall ATE as the weighted average among the ATEs computed at field level. The results is -4%.

On the other hand, to obtained the ATT we weight using the number of observation in the treatment group (being a male). The results is -7%.

1.5 e)

The results are different since we use different assumptions about the independence between treatment and potential outcome. Moreover, in the computation of the average treatment effect we use different weights in point b) and point d).

1.6 f)

Table 1: Paramater estimates from OLS

	(1) OLS
male	-0.0184***
	(-17.49)
dum2	-0.0103***
	(-6.44)
dum3	-0.303***
	(-199.43)
dum4	-0.311***
	(-205.08)
dum5	-0.403***
	(-235.68)
dum6	-0.586***
	(-376.02)
_cons	0.660***
	(484.40)
N	4526
R^2	0.978

t statistics in parentheses

$^{*}p < 0.10,^{**}\ p < 0.05,^{***}\ p < 0.010$

1.7 g)

The results from matching and from OLS are not the same. This is due to the fact that these two techniques weight differently the covariate-specific effects in order to compute the average causal effect.

2 6.2: Instrumental variable

2.1 a)

In order to consistently estimate the coefficient we need to assume linearity, that the independent variables (both $Educ$ and X) are uncorrelated with the error term μ (exogeneity), no multicollinearity and the spherical error terms.

However, in our setting the exogeneity assumption is likely to fail since there could be other individual characteristics that influence both education and health.

2.2 b)

Yes, if a valid instrument is found we can solve the problem .

2.3 c)

The assumptions we need to make for the instrumental variable strategy to yield consistent estimate are:

- SUTVA
- exogeneity: Cov(M, ν) = 0
- esclusion restriction: Cov(Z, μ) = 0
- relevance: Cov(M, Z) $\neq$ 0
- monotonicity

2.4 d)

Compliers are the students who attended one additional year of school due to the compulsory law. There are no never-takers since all the students stayed in school at least until they were 15 and the always takers are the students who stayed longer.

The monotonicity assumption eliminates defliers and guarantees existence of compliers.

2.5 e)

BJB95 and BJ96 argue that the birth date is actually a weak instrument since it has been proven in the literature that when a person is born has some correlation with personality, mental health and parental income which affect future earnings.

2.6 f)

Compulsory schooling law can be used as an instrument only if it is strongly correlated with the endogenous variable (education in our case) and does not affect the dependent variable (future earnings) in other way but through the endogenous variable.

However, it could be the case that the choice of introducing or expanding the compulsory schooling law is endogenously determined by economic condition: richer countries (and therefore healthier) could afford to leave their young generation in school longer with respect to poorer countries.

2.7 g)

Several are the channels through which education may effect health. First of all, higher educated people are expected to be on average wealthier and therefore they can take care more of their health.

Moreover, people who decide to stay in school longer may be more forward looking when it comes to future earnings than people who decide to drop out for school. The same long-term approach could be applied to health decision and so more educated people decide to practise sport more regularly, to eat healthier etc. expeting to be paid off in the long run.

Finally, it could also be the other way around: people more educated accept more demanding and more stressful job positions resulting in a lower health level.

2.8 h)

Table 2: Paramater estimates from OLS

	(1) OLS 1	(2) OLS 2	(3) OLS 3
educrec	-0.0270***	−0.0256***	−0.0252***
	(-118.57)	(−112.02)	(−109.90)
smsa		0.00126	0.000690
		(0.96)	(0.52)
married		−0.0923***	−0.0930***
		(−77.93)	(−78.61)
reg_dum1		−0.00900***	−0.00937***
		(−4.14)	(−4.31)
reg_dum2		−0.0189***	−0.0195***
		(−11.57)	(−11.95)
reg_dum3		−0.0144***	−0.0145***
		(−9.24)	(−9.25)
reg_dum4		−0.0101***	−0.0101***
		(−5.23)	(−5.26)
reg_dum5		0.00445***	0.00454***
		(2.72)	(2.78)
reg_dum6		0.0201***	0.0207***
		(9.69)	(9.97)
reg_dum7		0.00135	0.00138
		(0.75)	(0.77)
reg_dum8		0.00251	0.00262
		(1.14)	(1.19)
N	446241	446241	446241

t statistics in parentheses

$^{*}p < 0.10,^{**}\ p < 0.05,^{***}\ p < 0.010$

2.9 i)

$$Health_i = \beta_1 + \beta_2 Educ_i + \beta_3 X_i + \epsilon_i \tag{1}$$

$$Educ_i = \gamma_1 + \gamma_2 Z_i + \gamma_3 X_i + u_i \tag{2}$$

In equation (1) we separate the exogenours variable (X_i) from the endogenous one ($Educ_i$). Since the OLS estimator in case of endoneity would be inconsistent we perform a two stage least square: in the first stage we estimate equation (2) and then we replace $Educ_i$ in equation (1) with the predicted values $\widehat{Educ_i} = \hat{\gamma} Z_i$ from the first regression.

I would not include income as a covariate since I would expect it to cause additional biased due to endoneity.

2.10 j)

Table 3: Paramater estimates from OLS and IV

	(1) OLS 1	(2) OLS 2	(3) OLS 3	(4) IV 1	(5) IV 2	(6) IV 3	(7) IV 4
educrec	-0.0270***	−0.0256***	−0.0252***	−0.0725***	−0.0703***	−0.0675***	−0.0493***
	(-118.57)	(−112.02)	(−109.90)	(−6.49)	(−6.23)	(−6.71)	(−7.23)
smsa		0.00126	0.000690		0.0116***	0.0110***	0.00595***
		(0.96)	(0.52)		(3.80)	(3.92)	(3.09)
married		−0.0923***	−0.0930***		−0.0769***	−0.0779***	−0.0843***
		(−77.93)	(−78.61)		(−18.20)	(−20.44)	(−30.51)
reg_dum1		−0.00900***	−0.00937***		−0.0225***	−0.0217***	−0.0197***
		(−4.14)	(−4.31)		(−5.64)	(−5.87)	(−4.86)
reg_dum2		−0.0189***	−0.0195***		−0.0344***	−0.0335***	−0.0363***
		(−11.57)	(−11.95)		(−8.39)	(−8.96)	(−7.79)
reg_dum3		−0.0144***	−0.0145***		−0.0403***	−0.0387***	−0.0333***
		(−9.24)	(−9.25)		(−6.05)	(−6.46)	(−6.82)
reg_dum4		−0.0101***	−0.0101***		−0.0336***	−0.0322***	−0.0224***
		(−5.23)	(−5.26)		(−5.41)	(−5.73)	(−4.95)
reg_dum5		0.00445***	0.00454***		−0.0256***	−0.0237***	−0.00238
		(2.72)	(2.78)		(−3.31)	(−3.42)	(−0.97)
reg_dum6		0.0201***	0.0207***		−0.0331**	−0.0298**	0.00378
		(9.69)	(9.97)		(−2.43)	(−2.44)	(0.89)
reg_dum7		0.00135	0.00138		−0.0260***	−0.0244***	−0.00283
		(0.75)	(0.77)		(−3.66)	(−3.81)	(−1.07)
reg_dum8		0.00251	0.00262		−0.00689**	−0.00630**	−0.000778
		(1.14)	(1.19)		(−2.08)	(−2.02)	(−0.29)
N	446241	446241	446241	446241	446241	446241	446241

t statistics in parentheses

$^{*}p < 0.10,^{**}\ p < 0.05,^{***}\ p < 0.010$

2.11 k)

BJB95 argue that if the chosen instrument is weakly correlated with the endogenous variable than the IV estimator will be inconistent. How strong an instrument is can be evaluated by looking at the magnitude of the F test relative to the first stage of the 2SLS procedure. As a rule of thumb, the existing literature suggests to consider an instrument weak if the relative F statistics is smaller than 10.

In our case the F-tests are respectively equal to 67, 703, 316 and 154 suggesting that the risk that our chosen instrument is weak is quite low.

Finally, in our dateset we have more than 440 thousands observation so the finite sample bias is unlikely to be an issue.

2.12 l)

If the instrument is valid I would espect that the magnitude of the coefficient would be smaller: indeed if one of the control variable is suspected to be endogenous then once I control for this through the 2SLS procedure the effect on the dependent variable should be smaller.

2.13 m)

In case of Heterogeneous treatment effects the IV method estimates the average treatment effect for the subpopulation of the compliers also called LATE.

Therefore, it could actually be the case that the average treatment effect for the compliers is actually stronger than for the rest of the treated observations. This would result in a bigger coefficient in absolute value in the 2SLS with respect to the OLS.

2.14 n)

The results of our analysis seem to suggest that there is a positive relationship between education and health. With the 2SLS we were able to correct for the endogenity bias however to be fully confident abour our results other things could be tried.

For instance, the dependent variable is a dummy so using the linear model we used is not the correct one. It would have been more appropriated to use a probit or logit model.

Moreover, the definition of healthness we used is quite specific: in our framework a person is considered healthy is has not disability. In could be more interesting to evaluate the effect of education on a broader definition of wellbeing.

Finally, even if our point estimates were correct we did not considered any alternative approach for the standard errors we computed to tackle heteroskedasticity of cluster correlation for instance.