PSET 3 STATS II RIPS

Vaibhavi Sharma Pathak, Marcelo Piemonte Ribeiro, Fredrik Wallin

05/12/2022

Question 1

i)

Estimate the effect of job training grant (*grant*) on the hours of job training per employee (*hrsemp*). This simple relation can be summarized by the equation $hrsemp = \beta_0 + \beta_1 grant + \beta 2 employ + u$. The OLS results are reported below.

Table 1: Cross section

	Dependent variable:
	hrsemp
grant	
employ	-0.042^{*} (0.022)
Constant	11.272^{***} (1.956)
Observations	129
R^2	0.029
Adjusted R^2	0.021
Residual Std. Error	17.221 (df = 127)
F Statistic	3.727^{*} (df = 1; 127)
Note:	$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

The initial results present no effect of the *grant*. This happens because because no firms received grants in 1987 - Wooldridge, J. M. (2019). Introductory econometrics: A modern approach. Cengage learning, Ch. 13, p.445.

ii)

In a panel context, the previous relation can be summarized by the following equation: $hrsemp_{it} = \beta_0 + \beta_1 grant_{it} + \beta_2 employ_{it} + a_i + u_{it}$, t=1987, 1988 (t=1,2), where a_i is term for unobserved heterogeneity.

iii)

The simple regression performed in *i)* likely suffers from omitted variable issues. This happens especially if the latter does not contain all possible control variables, which is very often the case. The use of panel data allow us to overcome such issue without the need of additional variables. This is possible because the unobserved non-time varying effects present in the error term and affecting the dependent variable can be accounted for in a panel setting. Furthermore, running OLS in this case violates the GM assumption of independent observations. This is due the non-independence of firms (i) across the time periods.

iv)

The first difference equation is characterized by $\Delta hrsemp_i = \beta_0 + \beta_1 \Delta grant_i + + \beta_2 \Delta employ_i + \Delta u_i$

Table 2: First differences

	Dependent variable:
	hrsemp
grant	28.632^{***}
	(3.141)
employ	-0.139
	(0.084)
Constant	0.980
	(1.574)
Observations	125
R^2	0.405
Adjusted R^2	0.396
F Statistic	41.594^{***} (df = 2; 122)
Note:	*p<0.1; **p<0.05; ***p<0.01

The estimated equation $\Delta \widehat{hrsemp} = 0.98 + 28.63 \Delta grant - 0.14 \Delta employ$ indicates no effects of *employ*, but it does for *grant*. Having a grant signficantly increases the hours of job trainning per employee. An unit increase in the grant reflected around 28 more hours of trainning per employee *ceteris paribus*.

v)

The time demeaned equation performs the difference between the units of observations and the mean of them across time, as follows: $hrsemp_{it} - \overline{hrsemp_i} = \beta_1(grant_{it} - \overline{grant_i}) + \beta_2(employ_{it} - \overline{employ_i}) + (u_{it} - \overline{u_i})$. The fixed effect equation could be summarized as folows: $\ddot{hrsemp}_{i,t} = \beta_1 \ddot{grant}_{it} + \beta_2 \ddot{employ}_{it} + \ddot{u_{it}}$, t=1987, 1988. We should expect the same results between FE and FD estimations because in this case T=2, in other words only two years are being considered. While the FD estimation performs the difference between grant in the year of 1988 and in 1987, the FE estimation performs the difference between each observation and the mean. Because T=2 the FD and FE estimates should be equivalents

However, the estimates are slightly different without including the dummy regarding the year of 1988. This could be a signal for the violation of the strict exogeneity. Time dummies account for time related effects which are common to all individuals/firms and that are not capture already in the model.

Table 3: First-differences and Fixed-effects with and without dummy regarding 1988

	Dependent variable:		
	hrsemp		
	First-differences	Fixed-effects with 1988	Fixed-effects
	(1)	(2)	(3)
Constant	0.9795 (0.6358)		
grant	28.6317 (3.3006)***	28.6317 (4.7236)***	29.5274 (4.6607)***
employ	−0.1390 (0.0635)	−0.1390 (0.0910)	−0.1295 (0.0925)
d88		0.9795 (0.9132)	
Observations	125	256	256
R^2	0.4054	0.4826	0.4809
Adjusted R^2	0.3957	−0.0815	−0.0761
F Statistic	41.5937*** (df = 2; 122)	37.9271*** (df = 3; 122)	56.9806*** (df = 2; 123)
Note:	*p<0.1; **p<0.05; ***p<0.01 datasets::freeny lm() function vcovHC(type = 'HC1')-Robust SE		

vi)

The table below presents the FE and FD estimates. Although both estimates are significant, FE presented a higher magnitude. Also, dummy for the year 1989 showed to be significant.

Table 4: First-differences and Fixed-effects full sample with and without time fixed-effects

	Dependent variable:			
	hrsemp			
	FD	FE	FD dummies	FE dummies
	(1)	(2)	(3)	(4)
Constant	2.8665 (0.7953)**		5.1116 (2.2688)	
grant	31.1179 (2.8385)***	34.9330 (3.5604)***	32.3103 (2.7245)***	34.3910 (3.4105)***
employ	−0.0820 (0.0389)	−0.0418 (0.0354)	−0.0917 (0.0412)	−0.0812 (0.0363)
d88			−5.2680 (2.3563)	−0.7887 (1.0756)
d89			−4.6040 (3.4976)	4.9776 (1.8376)***
Observations	255	390	255	390
R^2	0.4682	0.4723	0.4801	0.4958
Adjusted R^2	0.4640	0.1886	0.4718	0.2186
F Statistic	110.9479*** (df = 2; 252)	113.2062*** (df = 2; 253)	57.7155*** (df = 4; 250)	61.7010*** (df = 4; 251)
Note:	*p<0.1; **p<0.05; ***p<0.01 datasets::freeny lm() function vcovHC(type = 'HC1')-Robust SE			

vii)

Strict exogeneity implies that u_{it} should not correlate with the independent variables of all time periods. A serial correlation test allows to identify if such assumption holds. The test below has the alternative hypothesis of serial correlation and the null of non serial correlation (strict exogeneity). The results for the FD model present a p-value<0.05 and we reject the null of non serial correlation, indicating the inverse.

While for the FE model the conclusion is the opposite. This results corroborates the different results from FD and FE with T=2 found in *v).*

```
Wooldridge's first-difference test for serial correlation in panels
```

data: fd_full F = 15.584, df1 = 1, df2 = 122, p-value = 0.0001325 alternative hypothesis: serial correlation in differenced errors

```
Wooldridge's test for serial correlation in FE panels
```

data: fe_within_full F = 0.87834, df1 = 1, df2 = 253, p-value = 0.3495 alternative hypothesis: serial correlation

viii)

The model to estimate either using FE or FD then becomes $hrsemp_{it} = \beta_0 + \beta_1 grant_{it} + \beta_2 employ_{it} + \beta_3 union_{it} + a_i + u_{it}$, t=1987, 1988, 1989.

Table 5: First-differences and Fixed-effects full sample and union and with and without time fixed-effects

	Dependent variable:			
	hrsemp			
	FD	FE	FD dummies	FE dummies
	(1)	(2)	(3)	(4)
Constant	2.8665 (0.7953)**		5.1116 (2.2688)	
grant	31.1179 (2.8385)***	34.9330 (3.5604)***	32.3103 (2.7245)***	34.3910 (3.4105)***
employ	−0.0820 (0.0389)	−0.0418 (0.0354)	−0.0917 (0.0412)	−0.0812 (0.0363)
d88			−5.2680 (2.3563)	−0.7887 (1.0756)
d89			−4.6040 (3.4976)	4.9776 (1.8376)***
Observations	255	390	255	390
R^2	0.4682	0.4723	0.4801	0.4958
Adjusted R^2	0.4640	0.1886	0.4718	0.2186
F Statistic	110.9479*** (df = 2; 252)	113.2062*** (df = 2; 253)	57.7155*** (df = 4; 250)	61.7010*** (df = 4; 251)
Note:	*p<0.1; **p<0.05; ***p<0.01 datasets::freeny lm() function vcovHC(type = 'HC1')-Robust SE			

The results are the same as the previous estimation because they rely on within variation, but *union* does not vary across individuals/firms. In other words, there is no within variation in *union*, individuals associated to an union were linked to the latter in all three years.

Question 2

i)

The IV used in the paper "is an indicator variable for whether or not the country is a former colony of the EU Council presidency in the second 6 months of the year t-2, when the budget is determined", Carneige and Marinov (2017), p.677. The authors employed such strategy because the original relation they aimed to estimate contained an endogenous variable, logged net EU official development assistance (ODA). This happened because the previous variable is not randomly assigned as “aid disbursements are made in ways that are systematically related to the recipient countries’ human rights” p.677, where recipient countries’ human rights is the dependent variable. This strategy was necessary to respond such estimation, otherwise the inital results could mask reverse causality.

ii)

First, they assume that $Colony_{i(t-2)2}$ only affects DV_{it} through the explanatory variable ($log(ODA)i(t-1)$), which is the exclusion restriction (p. 677) that they test in their supplemental code. They therefore need two statistical assumptions to be met: Random assignment of colony status; and a significant effect of the $Colonyi(t-2)2$ on $log(ODA)_{i(t-1)}$.

They show that the first assumption is met through a quasi-random assignment of the presidency of the EU Council. As it rotates due to a mechanism decided upon long before, the randomness seems given and valid, which is further corroborated through the use of an array of control variables in the 2SLS, which do not lead to a big shift of the instrumental effect. Thereby, they manage to show exogeneity quite convincingly. However, the use of a J-test could have further underlined this.

The second assumption holds in the paper due to a significant effect of the instrument in the first stage regression (which we later also observe) as well as an F-value 10< (10.85). To be more specific, the effect of the instrumental on the explanatory variable is such that an increase of the independent variable (aid) by 18% for every unit increase of the IV (p. 678). Moreover, they use the supporting information to underline that the colony status is only statistically significantly linked to the aid allocation if the EU Council presidency is in the second term, where budgets are discussed. Consequently, they show quite neatly that the instrument is significant (after having shown that it is also needed theoretically to avoid reverse causality).

Additionally, and somewhat underlying to every statistical study is the risk of measuring the wrong or inconsistent values through their identification strategy. This includes, e.g., (i) the changing composition of the Council, (ii) recipient countries that are not former colonies of any of the eligible Council members, and (iii) amendments to the rotation rules (p. 677). However, they address these by restricting the samples or statistically correcting the analysis via the inclusion of year fixed effects, which is plausible. Finally, they also mention that the linearity and constant effect assumptions are not necessary for the estimation of the causal effect.

In terms of theoretical assumptions, the paper claims that the Council presidency matters for the budget allocation (in favour of the presidency’s former colonies), the Commission then ties aid to advancements in human rights and democracy issues, and that the recipient of the aid also implements changes accordingly. They proceed to show these results with the IV as well as Figure 1, where they show the effect of aid over the years (t + 5).

iii)

Equation for the first stage:

$$log(ODA)_{i(t-i)} = \theta_0 + \theta_1 Colony_{i(t-2)2} + \sum_{k\in K} \theta_k I(i=k) + \sum_{j\in J} \theta_j I(t=j) + e_{it}$$

Equation for the second stage:

$$DV_{it'} = \beta_0 + \beta_1 log(ODA)_{i(t-i)} + \sum_{k\in K} \theta_k I(i=k) + \sum_{j\in J} \theta_j I(t=j) + u_{it}$$

Note that in the data set:

$$log(ODA)_{i(t-i)} = EV$$

$$Colony_{i(t-2)2} = l2CPcol2$$

and

$$DV\_it' = new\_empinxavg$$

iv)

When running the first stage regression, we find an estimated relationship of $Colony_{i(t-2)2}$ on $log(\widehat{ODA})_{i(t-i)}$ is 0.160 (SE = 0.073, p < 0.05), with an F-statistic of 40.03 (far above 10). Consequently, one cannot reject the assumption that the instrument is strong. The results are summarized in Table 6.

Table 6: 2SLS manual estimates

	Dependent variable:		
	EV	new_empinxavg	
	1st Stage	2nd Stage (no controls)	2nd (controls)
	(1)	(2)	(3)
l2CPcol2	0.1604 (0.0726)**		
fitted(tsls_first)		1.8852 (0.7972)**	1.6414 (0.7788)**
Observations	1,792	1,792	1,792
R^2	0.7640	0.8961	0.9036
Adjusted R^2	0.7449	0.8877	0.8948
Residual Std. Error	0.7170 (df = 1657)	1.2617 (df = 1657)	1.2211 (df = 1642)
F Statistic	40.0333*** (df = 134; 1657)	106.6454*** (df = 134; 1657)	103.2415*** (df = 149; 1642)
Note:	*p<0.1; **p<0.05; ***p<0.01		

In the second stage regression, we manage to replicate the first result displayed by Carnegie and Marinov in table 1 (p. 679), whereby the coefficient for the regression without controls is 1.885 (SE = .80, p < 0.05) and with controls is 1.641 (SE = 0.78, p < 0.05). Moreover, we cannot report the standard errors from this analysis, as they are not robust.

v)

Using the `ivreg`-command, we were able to replicate the results in the paper. Both for the IV regression without control variables (see column 1 below), as with control variables (columns 2 & 3) and then also through our manual estimation (column 4). These results are summarized in Table 7.

% Table created by stargazer v.5.2.3 by Marek Hlavac, Social Policy Institute. E-mail: marek.hlavac at gmail.com % Date and time: Thu, May 12, 2022 - 08:58:21

vi)

Here, it shows that there is a difference between the iv regression and OLS, with or without inclusion of controls (columns 2 & 3 in table below are OLS). While the OLS regressions yielded statistically significant results (coefficients 0.190 with controls or 0.216 without), the instrumental variable did not. This indicates an added value of the IV approach. The results are summarized in Table 8.

vii)

When running an endogeneity tests with the residuals of the first stage regression (`tsls_first` in the table below), we find that we can reject the null hypothesis (H_0) at $p < 0.1$ (with controls) or $p < 0.05$ (without controls) that there is no statistically significant difference of the residuals from 0. Based on this, it appears the explanatory variable is indeed not exogenous. The results are summarized in Table 9.

Table 7:

	Dependent variable:			
	new_empinxavg			
	instrumental variable		*panel linear*	*OLS*
	No controls	With controls	with controls	with controls
	(1)	(2)	(3)	(4)
EV	1.885 (1.096)	1.705 (1.077)	1.705 (1.077)	
fitted(tsls_first)				1.641* (0.779)
Observations	1,792	1,792	1,792	1,792
R^2	0.804	0.829	0.050	0.904
Adjusted R^2	0.788	0.814	−0.036	0.895
Residual Std. Error	1.735 (df = 1657)	1.625 (df = 1642)		1.221 (df = 1642)
F Statistic			138.032***	103.241*** (df = 149; 1642)
Note:				*p<0.05; **p<0.01; ***p<0.001

Table 8: IV and OLS

	Dependent variable:		
	new_empinxavg		
	instrumental variable	*OLS*	
	with controls	with controls	without controls
	(1)	(2)	(3)
EV	1.7054 (1.0765)	0.1903 (0.0421)***	0.2157 (0.0429)***
Observations	1,792	1,792	1,792
R^2	0.8293	0.9045	0.8973
Adjusted R^2	0.8138	0.8958	0.8890
Residual Std. Error	1.6246 (df = 1642)	1.2153 (df = 1642)	1.2543 (df = 1657)
F Statistic		104.3486*** (df = 149; 1642)	108.0538*** (df = 134; 1657)
Note:	*p<0.1; **p<0.05; ***p<0.01		

Table 9: Endogeneity test with residuals from the 1st-stage

	Dependent variable:	
	new_empinxavg	
	without controls	with controls
	(1)	(2)
EV	1.8852 (0.7917)**	1.6484 (0.7745)**
resid(tsls_first)	$-$1.6744 (0.7928)**	$-$1.4623 (0.7755)*
Observations	1,792	1,792
R^2	0.8976	0.9047
Adjusted R^2	0.8892	0.8960
Residual Std. Error	1.2530 (df = 1656)	1.2143 (df = 1641)
F Statistic	107.5104*** (df = 135; 1656)	103.8379*** (df = 150; 1641)
Note:	*p<0.1; **p<0.05; ***p<0.01	