EXCERSISES IN APPLIED PANEL DATA ANALYSIS #8

CHRISTOPHER F. PARMETER

1. Introduction

This `R` example will introduce you to estimation of the unobserved effects model in the presence of endogeneity. We will discuss both classical two-stage least squares estimation as well as Hausman-Taylor estimation. Both of these estimators are easily handled within the `plm` library.

2. Endogeneity in the Unobserved Effects Model

2.1. **An Economic Model of Crime.** Cornwell & Trumbull (1994) estimated a classic model of crime, based on rational behavior. Their data comes from 90 counties in North Carolina over the 1987-1987 period, hence a balanced panel. The main model of Cornwell & Trumbull (1994) is

$$\begin{aligned}\ln(crmrte)_{it} =& \beta_0 + \beta_1 \ln(prbarr)_{it} + \beta_2 \ln(prbconv)_{it} + \beta_3 \ln(prbpris)_{it} + \beta_4 \ln(avgsen)_{it} \\ &+ \beta_5 \ln(polpc)_{it} \cdots + c_i + \varepsilon_{it} \end{aligned} \tag{1}$$

where *crmrte* is the rate of crime in the county, *prbarr* is the probability that someone is arrested (# of arrests to reported offenses), *prbconv* is the probability of conviction (# of convictions to # of arrests), *prbpris* is the probability of receiving a prison sentence (measured as # of convictions that result in prison sentence compared to number of convictions), *avgsen* is the average length of a prison sentence (measured in days) and *polpc* is the number of police per capita. The $\cdots$ captures other variables measuring the economic state of the county, but the key variables for a model of rational crime are those appearing in the above equation.

Each of these variables to some degree measure the 'cost of crime' and the rational model of crime implies that criminals respond rationally to increased costs of crime by decreasing the amount of crime they commit. If this model is true then this generates predictable effects from the model, namely that β_1, β_2, β_3, β_4 and β_5 are all negative. Cornwell & Trumbull (1994) argued that both *polpc* and *prbarr* are endogenous. Their instruments for these two variables were the logarithm of *taxpc* the tax revenue per capita of a county, and the logarithm of *mix*, the proportion of crimes that involved face-to-face contact.

The argument behind the validty of these instruments is that counties with high taxes per capita have a 'preference' for law enforcement and so higher taxes reflect this fact (more police). The

---

University of Miami

Date: September 12, 2013.

Christopher F. Parmeter, Department of Economics, University of Miami; e-mail: cparmeter@bus.miami.edu.

[*] These exercises have been prepared for the "Applied Panel Data Econometrics" course in Dakar, Senegal sponsored by IFPRI and AGRODEP..

proportion of crimes involving face-to-face contact is a good instrument for *prbarr* because the probability of being identified as having committed a crime is higher if the crime is face-to-face. To begin our analysis we will ignore endogeneity and estimate the unobserved effects model of crime, using both the fixed and random effects estimator. We will also estimate the model using the between and pooled OLS models for consistency.

```
> library(plm)
> data("Crime")
> Crime.data <- Crime
> ## Add in logs of main variables
> Crime.data$lcrmrte  <- log(Crime.data$crmrte)
> Crime.data$lprbarr  <- log(Crime.data$prbarr)
> Crime.data$lprbconv <- log(Crime.data$prbconv)
> Crime.data$lprbpris <- log(Crime.data$prbpris)
> Crime.data$lavgsen  <- log(Crime.data$avgsen)
> Crime.data$lpolpc   <- log(Crime.data$polpc)
> Crime.data$ldensity <- log(Crime.data$density)
> Crime.data$lwcon    <- log(Crime.data$wcon)
> Crime.data$lwtuc    <- log(Crime.data$wtuc)
> Crime.data$lwtrd    <- log(Crime.data$wtrd)
> Crime.data$lwfir    <- log(Crime.data$wfir)
> Crime.data$lwser    <- log(Crime.data$wser)
> Crime.data$lwmfg    <- log(Crime.data$wmfg)
> Crime.data$lwfed    <- log(Crime.data$wfed)
> Crime.data$lwsta    <- log(Crime.data$wsta)
> Crime.data$lwloc    <- log(Crime.data$wloc)
> Crime.data$lpctymle <- log(Crime.data$pctymle)
> Crime.data$lpctmin  <- log(Crime.data$pctmin)
> Crime.data$ltaxpc   <- log(Crime.data$taxpc)
> Crime.data$lmix     <- log(Crime.data$mix)
> Crime.data$urban    <- ifelse(Crime.data$smsa=="yes",1,0)
> Crime.data$west     <- ifelse(Crime.data$region=="west",1,0)
> Crime.data$central  <- ifelse(Crime.data$region=="central",1,0)
> ## Pooled Model
> pool.crime <- plm(lcrmrte˜lprbarr+lprbconv+lprbpris+
+                            lavgsen+lpolpc+
+                                  ldensity+lwcon+lwtuc+lwtrd+lwfir+
+                                  lwser+lwmfg+lwfed+lwsta+lwloc+
+                                  lpctymle+lpctmin+west+central+
+                                  urban,
+                                  model="pooling",
+                                  data=Crime.data)
> ## Generic fixed effects estimation
>
> fe.crime <- plm(lcrmrte~lprbarr+lprbconv+lprbpris+
+                            lavgsen+lpolpc+
+                                  ldensity+lwcon+lwtuc+lwtrd+lwfir+
+                                  lwser+lwmfg+lwfed+lwsta+lwloc+
+                                  lpctymle+lpctmin+west+central+
+                                  urban+factor(year),
+                                  model="within",
+                                  effect="individual",
+                                  data=Crime.data)
> b.crime <- plm(lcrmrte~lprbarr+lprbconv+lprbpris+
+                                  lavgsen+lpolpc+
+                                  ldensity+lwcon+lwtuc+lwtrd+lwfir+
+                                  lwser+lwmfg+lwfed+lwsta+lwloc+
+                                  lpctymle+lpctmin+west+central+
+                                  urban,
+                                  model="between",
+                                  effect="individual",
+                                  data=Crime.data)
> ## Random Effects, using SWAR
> rd.crime <- plm(lcrmrte~lprbarr+lprbconv+lprbpris+
+                            lavgsen+lpolpc+
+                                  ldensity+lwcon+lwtuc+lwtrd+lwfir+
+                                  lwser+lwmfg+lwfed+lwsta+lwloc+
+                                  lpctymle+lpctmin+west+central+
+                                  urban,
+                                  model="random",
+                                  effect="individual",
+        random.method="swar",
+                                  data=Crime.data)
```

Table 1 reports the results from these four models. As is apparent, the cost of crime variables across the four models have negative signs and in most cases are statistically significant. The one exception is the average sentence. One possible explanation for this is that average sentence is a low cost variable to a criminal given that they would need to be arrested, convicted and then sent to prision before the cost of the sentence affected them. A Hausman test for the fixed versus random effects frameworks yields a p-value of 0, which clearly rejects the random effects framework suggesting the presence of county specific heterogeneity that is correlated with (some) of the regressors.

Table 1.

	Dependent variable:			
	lcrmrte			
	Pool	FE	BW	Random
lprbarr	-0.545^{***}	-0.355^{***}	-0.648^{***}	-0.415^{***}
	(0.030)	(0.032)	(0.088)	(0.030)
lprbconv	-0.439^{***}	-0.282^{***}	-0.528^{***}	-0.325^{***}
	(0.021)	(0.021)	(0.067)	(0.020)
lprbpris	-0.129^{***}	-0.173^{***}	0.297	-0.196^{***}
	(0.048)	(0.032)	(0.231)	(0.033)
lavgsen	-0.060	-0.002	-0.236	0.019
	(0.038)	(0.026)	(0.174)	(0.026)
lpolpc	0.362^{***}	0.413^{***}	0.364^{***}	0.415^{***}
	(0.022)	(0.027)	(0.060)	(0.025)
Observations	630	630	90	630
R^2	0.809	0.463	0.880	0.566
Adjusted R^2	0.782	0.381	0.675	0.547
Note:			*p<0.1; **p<0.05; ***p<0.01	

To estimate the model of crime accounting for endogeneity we make use of the `|` notation in our formula inside of `plm`. All variables before `|` are the variables in the model, while the variables after `|` are the instruments. Any variable which instruments for itself appears both before and after the `|`.

```
> ## 2SLS fixed effects estimation
>
> fe2SLS.crime <- plm(lcrmrte~lprbarr+lprbconv+lprbpris+
+                                         lavgsen+lpolpc+
+                                         ldensity+lwcon+lwtuc+lwtrd+lwfir+
+                                         lwser+lwmfg+lwfed+lwsta+lwloc+
+                                         lpctymle+lpctmin+west+central+
+                                         urban+factor(year)|ltaxpc+lmix+
+                                         lprbconv+lprbpris+lavgsen+
+                                         ldensity+lwcon+lwtuc+lwtrd+lwfir+
+                                         lwser+lwmfg+lwfed+lwsta+lwloc+
+                                         lpctymle+lpctmin+west+central+
+                                         urban+factor(year),
+                                         model="within",
+                                         effect="individual",
+                                         data=Crime.data)
> ## 2SLS between estimation
>
> b2SLS.crime <- plm(lcrmrte~lprbarr+lprbconv+lprbpris+
+                                         lavgsen+lpolpc+
+                                         ldensity+lwcon+lwtuc+lwtrd+lwfir+
+                                         lwser+lwmfg+lwfed+lwsta+lwloc+
+                                         lpctymle+lpctmin+west+central+
+                                         urban|ltaxpc+lmix+
+                                         lprbconv+lprbpris+lavgsen+
+                                         ldensity+lwcon+lwtuc+lwtrd+lwfir+
+                                         lwser+lwmfg+lwfed+lwsta+lwloc+
+                                         lpctymle+lpctmin+west+central+
+                                         urban,
+                                         model="between",
+                                         effect="individual",
+                                         data=Crime.data)
> ## EC2SLS fixed effects estimation
>
> EC2SLS.crime <- plm(lcrmrte~lprbarr+lprbconv+lprbpris+
+                                         lavgsen+lpolpc+
+                                         ldensity+lwcon+lwtuc+lwtrd+lwfir+
+                                         lwser+lwmfg+lwfed+lwsta+lwloc+
+                                         lpctymle+lpctmin+west+central+
+                                         urban+factor(year)|ltaxpc+lmix+
+                                         lprbconv+lprbpris+lavgsen+
+                                         ldensity+lwcon+lwtuc+lwtrd+lwfir+
+                                         lwser+lwmfg+lwfed+lwsta+lwloc+
+                                         lpctymle+lpctmin+west+central+
+                                         urban+factor(year),
+                                         model="random",
+                                         effect="individual",
+                                         inst.method="baltagi",
+                                         data=Crime.data)
```

Table 2 reports the results from the three different invocations for 2SLS estimation of the economic model of crime, including the original fixed effects estimates for comparison. As is before, the cost of crime variables across the four models have negative signs and in most cases are statistically significant. The one exception is the average sentence, which is negative, but statistically insignificant in all of the models (except for FE2SLS). The FE2SLS estimates suggest that none of the cost of crime variables belong in the model. This would suggest the rational behavior model is incorrect. However, the EC2SLS estimates are in line with economic theory, aside from the positive sign on police per capita.

A Hausman test for the fixed versus random effects frameworks yields a p-value of 0.6136, which fails to reject the random effects framework suggesting that perhaps the endogeneity we detected earlier was because *polpc* and *prbarr* were endogenous in general and not some correlation between the coariates and unobserved county specific heterogeneity. This is of course dependent upon validity of our instruments.

Table 2.

	Dependent variable:			
	lcrmrte			
	FE	FE2SLS	BW2SLS	EC2SLS
lprbarr	-0.355^{***}	-0.575	-0.503^{**}	-0.413^{***}
	(0.032)	(0.802)	(0.241)	(0.097)
lprbconv	-0.282^{***}	-0.423	-0.525^{***}	-0.323^{***}
	(0.021)	(0.502)	(0.100)	(0.054)
lprbpris	-0.173^{***}	-0.250	0.187	-0.186^{***}
	(0.032)	(0.279)	(0.318)	(0.042)
lavgsen	-0.002	0.009	-0.227	-0.010
	(0.026)	(0.049)	(0.179)	(0.027)
lpolpc	0.413^{***}	0.657	0.408^{**}	0.435^{***}
	(0.027)	(0.847)	(0.193)	(0.090)
Observations	630	630	90	630
R^2	0.463	0.444	0.874	0.598
Adjusted R^2	0.381	0.365	0.670	0.573
Note:		$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$		

2.2. **Estimating an Earnings Equation.** Cornwell & Ruppert (1988) analyzed returns to schooling for 595 individuals taken from the Panel Study of Income Dynamics (PSID), observed over 1976-1982. Their model was

$$\begin{aligned}\ln(wage)_{it} =& \beta_0 + \beta_1 weeks_{it} + \beta_2 D_{South,it} + \beta_3 D_{SMSA,it} + \beta_4 D_{MARR,it} + \beta_5 Exper_{it} \\ &+ \beta_6 Exper^2_{it} + \beta_7 D_{OCC,it} + \beta_8 D_{IND,it} + \beta_9 D_{UNION,it} + \beta_{10} D_{FEM,i} \\ &+ \beta_{11} D_{BLK,i} + \beta_{12} D_{EDUC,it} + c_i + \varepsilon_{it}. \end{aligned} \tag{2}$$

For the Hausman-Taylor approach $South$, OCC, IND, $SMSA$ were treated as exogenous time varying covariates and FEM and BLK were treated as exogenous time constant variables.

```
> ## Hausman-Taylor Estimation of a wage equation
> Wages <- read.csv(file="Wages.csv",h=T)
> Wages$exp2 <- (Wages$exp)^2
> pdata.Wages <- pdata.frame(Wages,index=c("ID","YEAR"))
> fe.wage <- plm(lwage~wks+south+smsa+married+
+                  exp+exp2+bluecol+ind+
+                  union+sex+black+ed,
+                                          model="within",
+                                          effect="individual",
+                                          data=pdata.Wages)
> rd.wage <- plm(lwage~wks+south+smsa+married+
+                  exp+exp2+bluecol+ind+
+                  union+sex+black+ed,
+                                    model="random",
+             random.method="swar",
+                                          effect="individual",
+                                          data=pdata.Wages)
> ht.wage <- plm(lwage~wks+south+smsa+married+
+                  exp+exp2+bluecol+ind+
+                  union+sex+black+ed|bluecol+
+                  south+smsa+ind+sex+black,
+                              model="ht",
+                                          effect="individual",
+                                          data=pdata.Wages)
```

The estimates from fixed effects, random effects using Swamy & Arora (1972) and Hausman & Taylor (1981) appear in Table 3. The estimated effect of education according to the Hausman & Taylor (1981) estimator is 13.8%, which is almost 40% higher than the estimated effect of education found in the random effects framework. Deploying a Hausman test between teh fixed effects framework and the partial correlation setup of Hausman & Taylor (1981) results in a p-value of 0.8113, which fails to reject the partial correlation framework.

TABLE 3.

	Dependent variable:		
	lwage		
	FE	RE	HT
wks	0.001	0.001	0.001
	(0.001)	(0.001)	(0.001)
southyes	-0.002	-0.017	0.007
	(0.034)	(0.027)	(0.032)
smsayes	-0.042^{**}	-0.014	-0.042^{**}
	(0.019)	(0.020)	(0.019)
marriedyes	-0.030	-0.075^{***}	-0.030
	(0.019)	(0.023)	(0.019)
exp	0.113^{***}	0.082^{***}	0.113^{***}
	(0.002)	(0.003)	(0.002)
exp2	-0.0004^{***}	-0.001^{***}	-0.0004^{***}
	(0.0001)	(0.0001)	(0.0001)
bluecolyes	-0.021	-0.050^{***}	-0.021
	(0.014)	(0.017)	(0.014)
ind	0.019	0.004	0.014
	(0.015)	(0.017)	(0.015)
unionyes	0.033^{**}	0.063^{***}	0.033^{**}
	(0.015)	(0.017)	(0.015)
sexmale		0.339^{***}	0.131
		(0.051)	(0.127)
blackyes		-0.210^{***}	-0.286^{*}
		(0.058)	(0.156)
ed		0.100^{***}	0.138^{***}
		(0.006)	(0.021)
Constant		3.924^{***}	2.782^{***}
		(0.103)	(0.308)
Observations	4,165	4,165	4,165
R^2	0.658	0.390	0.363
Adjusted R^2	0.563	0.389	0.362
Note:	*p<0.1; **p<0.05; ***p<0.01		

References

Cornwell, C. & Ruppert, P. (1988), ‘Efficient estimation with panel data: an empirical comparison of instrumental variables estimators’, *Journal of Applied Econometrics* **3**, 149–155.

Cornwell, C. & Trumbull, W. N. (1994), ‘Estimating the economic model of crime with panel data’, *The Review of Economics and Statistics* **76**, 360–366.

Hausman, J. A. & Taylor, W. E. (1981), ‘Panel data and unobservable individual effects’, *Econometrica* **49**, 1377–1398.

Swamy, P. A. V. B. & Arora, S. S. (1972), ‘The exact finite sample properties of the estimators of coefficients in the error components regression models’, *Econometrica* **40**, 261–275.