Empirical Methods - Exercise 4

Solutions

December 11-12, 2019

Theory Question - Motivating Linear Panel Data

Farmer's production function is:

$$y_{it} = x_{it}\beta + \alpha_i + \epsilon_{it}$$

where: y_{it} is *log(output)*; x_{it} is *log(labour)*; α_i is *log(soil quality)* (an unobservable, time-invariant, farmer-specific characteristic); ϵ_{it} is rainfall (unobservable and iid across farmers).

(a) *Solve the farmer's profit maximization problem assuming he sells output at a common (across farmers) market price P_t and pays common wages W_t. (Hint: It may help to write down the production function in levels instead of logs.) For notational convenience, assume that $E[e^{\epsilon_{it}}] = \lambda$. Does the labor demand depend on α_i? Explain the economic intuition behind the result.*

Answer:

To solve the farmer's profit maximization problem we have to first determine her expected profits. To do so, it helps starting by rewriting the production function in levels, as:

$$\begin{aligned} \log(Y_{it}) &= \beta \log(X_{it}) + \log(A_i) + \varepsilon_{it} \\ &= \log(X_{it}^{\beta} \, A_i \, e^{\varepsilon_{it}}) \end{aligned}$$

(where we define a generic variable Z as $Z = e^z$ so that $z = log(Z)$)

By taking the exponential from both sides, we get:
$Y_{it} = X_{it}^{\beta} \cdot A_i \cdot e^{\varepsilon_{it}}$

$\longrightarrow$ Expected output is:

$$\begin{aligned} E(Y_{it}) &= E(X_{it}^{\beta} \, A_i \, e^{\varepsilon_{it}}) \\ &= X_{it}^{\beta} \, A_i \, E(e^{\varepsilon_{it}}) \\ &= X_{it}^{\beta} \, A_i \, \lambda \end{aligned}$$

We can now plug this result into the profits function:

$$\begin{aligned} E(\Pi_{it}) &= E(P_t\, Y_{it} - W_t\, X_{it}) \\ &= P_t\, X_{it}^{\beta}\, A_i\, \lambda - W_t\, X_{it} \end{aligned}$$

To solve the farmer's (unconstrained) profit maximization problem we hence have to solve the following:

$$\max_{X_{it}} E(\Pi_{it}) = \max_{X_{it}} \{P_t\, X_{it}^{\beta}\, A_i\, \lambda - W_t\, X_{it}\}$$

$\longrightarrow$ *FOC* :

$$\frac{\partial \Pi_{it}}{\partial X_{it}} = 0 \;\longrightarrow\; \beta P_t X_{it}^{\beta-1}\, A_i \lambda - W_{it} = 0$$

So that rearranging terms we obtain farmer's demand for labor:

$$X_{it} = \left(\frac{\beta A_i \lambda P_t}{W_t} \right)^{\frac{1}{1-\beta}}$$

- If $\beta < 1$ (which is the standard assumption in Labor Economics): Recall that $A_i = e^{\alpha_i}$: $\alpha_i \uparrow \rightarrow \ A_i \uparrow \rightarrow X_{it} \uparrow$: good soil quality makes land more productive, and hence more labor is needed to harvest.
- If $\beta = 1$, the maximization problem does not have a maximum: the maximum number of workers the farmer would like to employ could only be determined by a resource constraint (e.g., the total number of people in the economy), otherwise a maximum won't exist.
- if $\beta > 1$: $\alpha_i \uparrow \rightarrow \ A_i \uparrow \rightarrow X_{it} \downarrow$: the better the soil is, the lower is the number of workers needed. *Why?*

(b) *Under what assumption can you recover a consistent estimate for β by running (pooled) OLS? Based on what you found in (a) do you think this assumption is violated in this case? (no proof required)*

Answer:

Pooled OLS requires $Cov(x_{it}, \alpha_i) = 0$ which, as we just saw in part (a), is violated. As a result, our estimator will be inconsistent.

(c) *Suppose you only had $T = 1$ period of data. Propose an estimation strategy that would consistently estimate β. Be careful to explain what assumptions need to hold. Could the variables P_1 and W_1 possibly help? If you would need access to another variable besides $(y_{i1}, x_{i1}, P_1, W_1)$ provide an example of what might work. Discuss what properties your variable must satisfy.*

Answer:

α_i is unobservable, and as we have seen, it is correlated with our regressor. As a result, in a cross-section, we would need to use IV. With only one time period, we won't be able to use P_t or W_t as instruments, because we would only have one data point each, and no cross-sectional variation. An instrument could be *number of children*: since they may well be working in the farm, this variable is likely correlated with our endogenous regressor, and is however uncorrelated with shocks to output (rainfall and soil quality).

Do you believe it? Think about alternative (better) instruments.

(d) *Now suppose that you have access to $T = 10$ periods of data. Which of the following estimators would consistently estimate β: (1) Random Effects, (2) Fixed Effects, or (3) First Differences? Explain your answers by very briefly discussing which assumptions are needed for consistent estimates from each of the three methods, and whether they are likely to hold in this example.*

Answer:

- RE inconsistent as $Cov(x_{it}, \alpha_i) \neq 0$
- FE consistent as $Cov(x_{it}, \varepsilon_{it}) = 0$
- FD consistent as $Cov(\Delta x_{it}, \Delta \varepsilon_{it}) = 0$

(e) *Would you prefer your estimation strategy in (c) or your preferred estimator identified in (d)? Why?*

Answer:

FE and FD better than IV: it may be hard to find a valid instrument in this setting, considering that we have to deal with correlated unobservables. In addition, given that we have quite large number of periods in the analysis ($T = 10$), there should be enough variation in the data to identify our coefficient of interest β reasonably well.

As a result, we prefer the estimator identified in part (d).

(f) *[Extra] Farmers take their harvesting decisions (i.e. how many workers to hire) also based on rain forecasts for the season. This is another variable that the farmers observe. Suppose you are not able to observe this information. Obviously, as for the case of soil quality, rain forecasts are correlated with labor decisions, but they end up in the error term. This means your* ε_{it} *also contains a time-varying shock. You still have* $t = 10$ *periods of data. How would this affect your preferred estimator? How could you fix this?*

Answer:

With time-varying unobservables correlated with our regressor, FE would lead to incosistent estimates. The solution would be to find a valid time-varying instrument. The take-away here is to understand that a Fixed Effects estimator can only work if the unobservable is constant over time (like soil quality).

Empirical Application Question - Basic Panel Data Models

This question uses a panel data set taken from Baltagi and Griffin (1983) "Gasoline Demand in the OECD: An Application of Pooling and Testing Procedures" in the European Economic Review. The data set contains infomation on gasoline consumption in 18 OECD countries (i) over the 19 years from 1960-1978 (t).

(a) *Consider the following specification for a gasoline consumption equation:*

$$\ln\left(\frac{Gas}{Car}\right)_{it} = \beta_0 + \beta_1 \ln\left(\frac{Y}{N}\right)_{it} + \beta_2 \ln\left(\frac{P_{MG}}{P_{GDP}}\right)_{it} + \beta_3 \ln\left(\frac{Car}{N}\right)_{it} + u_{it}$$

where $u_{it} = \alpha_i + \varepsilon_{it}$.

Provide some economic rationale for the regression specification. That is, explain why each variable is included, and the likely sign of the coefficients.

Answer:

- LHS: gasoline consumption per car
- RHS: gasoline consumption is likely to depend on income, prices, and number of cars available
  - per capita income - positive: the more you earn, the more you drive (use less public transport)
  - gasoline price - negative: the higher the price of gas, the lower the demand for gas (find alternative means of transportation)
  - number of cars per person - negative: the more cars you own the less you use each car

(b) *Estimate this specification by a Pooled OLS regression.*

Answer:

```
gasData <- plm.data(gasData, index= c( "co" , "year"))
pols <- plm(c ~ y + p + car + as.factor(year), data=gasData,
            model="pooling")
```

Table 1:

	Dependent variable:
	c
y	0.900^{***}
	(0.037)
p	-0.899^{***}
	(0.031)
car	-0.764^{***}
	(0.019)
Observations	342
R^2	0.857
Note:	*p<0.1; **p<0.05; ***p<0.01

(c) *What are the necessary assumptions for this OLS model to be consistent? Are they likely to be satisfied here? In particular, are there any potential sources of endogeneity that we should worry about? Given your answer to these questions, is OLS the best linear unbiased estimator for this model?*

Answer:

It seems very likely thst our regressors are correlated with the error term, and as such $\mathbb{E}(X_{it}' u_{it}) \neq 0$. In particular, there is a clear simultaneous equation issue between price of gas and consumption of gas. Using a Pooled OLS does not solve the issue.

(d) *Let's focus on β_2, the coefficient on the real price of gasoline. Given your answer to part (c), what is the likely direction of the bias in the Pooled OLS coefficient? Be careful in explaining the economic mechanism driving the bias.*

Answer:

Let's consider two different potential mechanisms.

(1) α_i is unobserved innate propensity to drive (the higher α, the more willing to drive):

- May be related to culture/size of the country
- Think of this as an omitted variable
- Increase in α_i increase gas consumption per-capita ($\beta_2 > 0$)
- At the same time, it may be that places with high innate propensity to drive face also high pressures for keeping gas prices low ($Cov(\alpha_i, p_i) < 0$)

$\longrightarrow$ *negative bias (the true β should hence be closer to zero).*

(2) α_i is quality of transport infrastructure, like roads (the higher the α, the better the quality):

- Think of this as an omitted variable
- In countries with bad transport infrastructure, people may need more time to go from A to B and hence use more gasoline ($\beta_2 < 0$)
- It may also be harder to deliver gasoline to the gas station, or there may be fewer gas stations, which may increase the price of gas in the country ($Cov(\alpha_i, p_i) < 0$)

$\longrightarrow$ *positive bias (the true β should hence be more negative).*

(e) *Estimate the regression specification using:*

(e.i) *The LSDV Estimator. How do the coefficients compare to what you find in (b)? Is this what you expected?*

Answer:

```
lsdv <- plm(c ~ y + p + car + as.factor(co) + as.factor(year),
            data=gasData, model="pooling")
```

Table 2:

	Dependent variable:	
	c	
	POLS	LSDV
	(1)	(2)
y	0.900***	0.051
	(0.037)	(0.091)
p	−0.899***	−0.193***
	(0.031)	(0.043)
car	−0.764***	−0.593***
	(0.019)	(0.028)
Observations	342	342
R^2	0.857	0.981
Note:	*p<0.1; **p<0.05; ***p<0.01	

We can see from comparing the estimated coefficients between the Pooled OLS model and the LSDV model that the LSDV coefficients are lower in magnitude. Is this what you expected? Well, that depends on the story you told before.

(e.ii) *The Within Groups ("Fixed Effect") Estimator. Are the β's different to what you obtained in (b)? Are they different from what you obtained in e.i? Explain.*

Answer:

```
withinGroup <- plm(c ~ y + p + car + as.factor(year),
                   data = gasData, model = "within" )
```

Table 3:

	Dependent variable:		
	POLS	c LSDV	FE
	(1)	(2)	(3)
y	0.900^{***}	0.051	0.051
	(0.037)	(0.091)	(0.091)
p	-0.899^{***}	-0.193^{***}	-0.193^{***}
	(0.031)	(0.043)	(0.043)
car	-0.764^{***}	-0.593^{***}	-0.593^{***}
	(0.019)	(0.028)	(0.028)
Observations	342	342	342
R^2	0.857	0.981	0.883
Note:			*p<0.1; **p<0.05; ***p<0.01

If you look at the estimated coefficients, they are all the same as LSDV: they are indeed analytically equivalent. Standard errors should be somewhat different at some decimal levels, because in one case we are regressing demeaned variables (in the regression without dummy variables) and in ther other we regress our "simple" variables (plus a set of dummies). The R-squared is also different, and this is consistent with the fact that in one specification have way more regressors than in the other.

(e.iii) *The Generalised Least Squares ("Random Effect") Estimator*

Answer:

```
randomEffect <- plm(c ~ y + p + car + as.factor(year),
                    data = gasData, model = "random" )
```

Table 4:

	Dependent variable:
	c
	RE
y	0.204^{***}
	(0.073)
p	-0.287^{***}
	(0.042)
car	-0.606^{***}
	(0.025)
Observations	342
R^2	0.863
Note:	*p<0.1; **p<0.05; ***p<0.01

(f) *Focusing on your estimates of* β_2

(f.i) *What does the OLS estimate of* $\hat{\beta}_2$ *imply about the relationship between gas prices and gas consumption? Does this estimate pass the "smell test", i.e. do you think it's magnitude is likely to be too big, about right, or too small?*

Answer:

The magnitude is pretty big: If there were a 10% increase in gas prices, do we really think people would drive almost 9% less? It should probably be more inelastic.

(f.ii) *Do the results suggest your concerns about endogeneity bias may be reasonable?*

Answer:

If α_1 is unobserved innate propensity to drive, then yes. We indeed thought we would have a negative bias and doing Fixed Effects has made the coefficient much bigger (it is indeed less negative). Note it has also made the economic magnitude more reasonable, i.e. a 10% increase in gas prices is now associated with a 2% reduction in gas consumption, on average, holding all the other factors constant.

(g) *Perform the Classical- and Regression Based Hausman Tests to investigate whether the Random Effects assumption is viable in this setting. What do you find? Does this conclusion surprise you?*

Answer:

"Classical" Hausman Test:

The Hausman Test allows us to test whether $Cov(x_{it}, \alpha_i) = 0$ or not. To do so, we compare the time-varying coefficient in the FE and RE models:

$$H = \frac{\hat{\beta}_{FE} - \hat{\beta}_{RE}}{\sqrt{Var(\hat{\beta}_{FE}) - Var(\hat{\beta}_{RE})}} \tag{1}$$

Recall that, under H_1, RE would be inconsistent, but FE would be consistent; under H_0 instead, RE would be consistent and efficient, and FE would be consistent but inefficient.

The H-test is distributed as χ^2_K (here with 21 degrees of freedom). Given the big change in the coefficients between OLS and FE, we expect to reject the assumption of no unobserved heterogeneity. And indeed, by running the test, we get $p - value < 0.000$, which makes us rejecting the null of no unobserved heterogeneity.

```
phtest(randomEffect , withinGroup)
```

```
Hausman Test
```

data: c ~ y + p + car + as.factor(year) chisq = 141.85, df = 21, p-value < 2.2e-16 alternative hypothesis: one model is inconsistent

"Regression Based" Hausman Test:

We should run an F-test of $H_0 : \xi = 0$ from the unrestricted model $y_{it} = x_{it}'\beta + \bar{w}_i'\xi + \nu_{it}^*$ (cfr. slides 77-78 panel data), where $\bar{w}_i$ are across-time mean values of the subset of regressors we are interested in (recall these cannot include time dummies!).

As a result, we start by computing $\bar{w}_i$:

```
timeAvg <- gasData %>%
                group_by(co) %>%
                summarise(ybar = mean(y, na.rm = T),
                          pbar = mean(p, na.rm = T),
                          carbar = mean(car, na.rm = T))
gasData <- merge(gasData, timeAvg)

hausman.aux <-plm(c ~ ybar + pbar + carbar +
                      y + p + car + as.factor(year),
                  data=gasData, model ="pooling")
```

Table 5:

	Dependent variable:
	c
	Hausman Auxiliary Regression
ybar	0.916^{***}
	(0.224)
pbar	-0.771^{***}
	(0.108)
carbar	-0.202^{***}
	(0.070)
y	0.051
	(0.222)
p	-0.193^{*}
	(0.104)
car	-0.593^{***}
	(0.067)
Observations	342
R^2	0.880
Adjusted R^2	0.871
F Statistic	97.234^{***} (df = 24; 317)
Note:	*p<0.1; **p<0.05; ***p<0.01

The F-test for the null that all three averaged regressor are jointly zero is equal to 20.719 (and we hence reject the null).