Applied Methods - LAB Panel Data

Michael Kummer

Why do we use panel data?

- To eliminate individual unobserved time constant factors.
- To study treatment effects (for instance, policy change).

Exercise 1

Setup:

```
library(data.table)
library(ggplot2)
library(stargazer)
#install.packages("plm")
library(plm)
library(lmtest)
library(stats)
setwd("C:/TopicsDig/Labs/Lab_07W_Panel/")
# change the file's path to your own
```

Exercise 2

```
load("rental.RData")
head(dt.rental)
```

```
##    city year rent    lrent    pop     lpop avginc   lavginc   pctstu y90
## 1:    1   80  197 5.283204  75211 11.22805  11537  9.353314 20.34676   0
## 2:    1   90  342 5.834811  77759 11.26137  19568  9.881651 23.17031   1
## 3:    2   80  323 5.777652 106743 11.57818  19841  9.895506 21.04307   0
## 4:    2   90  496 6.206576 141865 11.86263  31885 10.369891 20.98403   1
## 5:    3   80  216 5.375278  36608 10.50802  11455  9.346182 32.36178   0
## 6:    3   90  351 5.860786  42099 10.64778  21202  9.961851 24.38300   1
```

Goal: Does a stronger presence of students affect rental rates?

- What is the unit of analysis? City.
- What are the time periods?
- How many periods?
- Is there a treatment? No.
- What are the key variables of interest?
- Dependent variable (y)? Rent or log(rent). Which one will we use? lrent for % effects.
- What is the independent variable of interest? pctstu.
- What factors should we control for?

Estimate the equation by pooled OLS.

```
out.ols <- plm(  lrent ~  pctstu + lpop + lavginc + y90
               , c = index("city", "year")
               , model = "pooling"
               , data=dt.rental)
stargazer(out.ols, type = "text")
```

```
##
## ========================================
##                   Dependent variable:
##               ---------------------------
##                           lrent
## ----------------------------------------
## pctstu                 0.005***
##                         (0.001)
##
## lpop                    0.041*
##                         (0.023)
##
## lavginc                0.571***
##                         (0.053)
##
## y90                    0.262***
##                         (0.035)
##
## Constant                -0.569
##                         (0.535)
##
## ----------------------------------------
## Observations              128
## R2                       0.861
## Adjusted R2              0.857
## F Statistic   190.922*** (df = 4; 123)
## ========================================
## Note:        *p<0.1; **p<0.05; ***p<0.01
```

The coefficient on pctstu means that a one percentage point increase in pctstu increases rent by half a percent (.5%). pctstu is very statistically significant.

% variation in y = 100* beta1* variation in x = 0.005 * 100 * variation in x

The positive and very significant coefficient on d90 simply means that, other things in the equation fixed, nominal rents grew by over 26% over the 10 year period.

% variation in y = 100* beta1* variation in x = 0.262 * 100 * variation in x

Test for serial correlation

Serial Correlation

What is serial correlation?

The residuals are correlated over time.

What are the consequences of serial correlation?

Positive Serial Correlation

This means that if you get a positive residual in one period, you are more likely to get a positive residual in the following period. The consequences of this are:

- Standard errors are underestimated
- T-statistics are inflated
- Type-I error increases (false positive, you incorrectly reject the null)

Negative Serial Correlation

This means that if you get a negative residual in a one period, you are more likely to get a positive residual in the following period. The consequences of this are:

- Standard errors are overstated
- F-statistics are understated
- Type-II error increases (false negative, you incorrectly do not reject the null)

How to test for serial correlation?

In order to test for serial correlation you regress the residuals on the residuals from the previous period. lm(u ~ lag(u)). In order to run the regression we need to compute the residuals first. Then we will add these residuals to our panel data frame in order to run the regression.

In this case we ran a pooled OLS, therefore, we did not loose any observations (note the differences in the code from the case presented above.)

```
u <- residuals(out.ols)
length(u)
```

```
## [1] 128
```

```
nrow(dt.rental)
```

```
## [1] 128
```

```
dt.rental[,'u'] <- u
head(dt.rental)
```

```
##    city year rent    lrent    pop     lpop avginc   lavginc   pctstu y90
## 1:    1   80  197 5.283204  75211 11.22805  11537  9.353314 20.34676   0
## 2:    1   90  342 5.834811  77759 11.26137  19568  9.881651 23.17031   1
## 3:    2   80  323 5.777652 106743 11.57818  19841  9.895506 21.04307   0
## 4:    2   90  496 6.206576 141865 11.86263  31885 10.369891 20.98403   1
## 5:    3   80  216 5.375278  36608 10.50802  11455  9.346182 32.36178   0
## 6:    3   90  351 5.860786  42099 10.64778  21202  9.961851 24.38300   1
##                u
## 1: -0.052352509
## 2: -0.080484228
## 3:  0.114505724
## 4: -0.001158549
## 5:  0.012495132
## 6: -0.081490174
```

```
dt.rental$L1_u <- c(NA, dt.rental$u[-nrow(dt.rental)])
summary(dt.rental$L1_u)
```

```
##       Min.    1st Qu.     Median       Mean    3rd Qu.       Max.       NA's
## -0.2423317 -0.0783143 -0.0135525  0.0002261  0.0457265  0.4808174          1
```

```
out.u <- plm( u ~ L1_u, data=dt.rental, model='pooling')
stargazer(out.u, type="text")
```

```
##
## ========================================
##                   Dependent variable:
##               ---------------------------
##                             u
## ----------------------------------------
## L1_u                   0.381***
##                         (0.083)
##
## Constant                0.0003
##                         (0.010)
##
## ----------------------------------------
## Observations              127
## R2                       0.146
## Adjusted R2              0.139
## F Statistic    21.290*** (df = 1; 125)
## ========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

- Why do we have serial correlation? We are not getting rid of the a_i (individuals unobserved time invariante factors).
- Can these a_i also bias our beta estimates? Think whether the percentage of students in one city can be related to some unobservable characteristic that also has an impact on the rent value; what things can you think of?
- What are the consequences of serial correlation to our estimate?
- What model should we use?

Now, estimate the model by first-differences. Compare your estimate of beta pctstu with your previous estimation. Does the relative size of the student population appear to affect rental prices?

Change data do pdata.frame an define the index:

```
pdt.rental   <- pdata.frame( dt.rental,  index = c("city", "year"))
```

Alternatively, you can just use "dt.rental" and define the index inside the plm function.

Create a dummy variable for the year 1990. This is an alternative way to create a dummy:

```
pdt.rental$d2 <- as.double(pdt.rental$year==90)
```

Run fd regression.

```
out.fd <- plm(  lrent ~  pctstu + lpop + lavginc,
              , model = "fd"
              , data=pdt.rental)
stargazer(out.fd, type = "text")
```

```
##
## ========================================
##                   Dependent variable:
##               ---------------------------
##                           lrent
## ----------------------------------------
## pctstu                 0.011***
```

```
##                          (0.004)
##
## lpop                      0.072
##                          (0.088)
##
## lavginc                 0.310***
##                          (0.066)
##
## Constant                0.386***
##                          (0.037)
##
## ---------------------------------------
## Observations               64
## R2                        0.322
## Adjusted R2               0.288
## F Statistic     9.510*** (df = 3; 60)
## =======================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

Interestingly, the effect of pctstu is over twice as large as we estimated in the pooled OLS equation. Now, a one percentage point increase in pctstu is estimated to increase rental rates by about 1.1%. The intercept (constant) gives us the time trend, or the coefficient of the dummy y90. Not surprisingly, we obtain a much less precise estimate when we difference (although the OLS standard errors are likely to be much too small because of the positive serial correlation in the errors within each city). While we have differenced away the a_i, there may be other unobservables that change over time and are correlated with pctstu.

We can test for the presence of heteroskedasticity using the bptest:

```
bptest(out.fd)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  out.fd
## BP = 5.5058, df = 3, p-value = 0.1383
```

We do not reject the null - that we have homoskedasticity.

In the first-differences case, serial correlation is not an issue because we have no time component in the equation.

In any case, when using panel data, we can always use *vcov. = vcovHC(out.fd, method = c("arellano"))* to correct for **both** heteroskedasticity **and** serial correlation. (Note, this correction can only be done for panel data.)

```
stargazer(coeftest(out.fd, vcov. = vcovHC(out.fd, method = c("arellano"))), type = "text")
```

```
##
## ====================================
##               Dependent variable:
##           -------------------------
##
## ------------------------------------
## pctstu             0.011***
##                     (0.003)
##
## lpop                 0.072
##                     (0.067)
```

```
##
## lavginc           0.310***
##                    (0.086)
##
## Constant          0.386***
##                    (0.047)
##
## ===================================
## ===================================
## Note:    *p<0.1; **p<0.05; ***p<0.01
```

Estimate the model by fixed effects to verify that you get identical estimates and standard errors to those in the first-differences model.

```
out.fe <- plm(  lrent ~  0 + pctstu + lpop + lavginc + y90
              , c = index("city", "year")
              , model = "within"
              , data=dt.rental)
stargazer(out.fe, type = "text")
```

```
##
## ========================================
##                  Dependent variable:
##              ---------------------------
##                          lrent
## ----------------------------------------
## pctstu                0.011***
##                        (0.004)
##
## lpop                    0.072
##                        (0.088)
##
## lavginc               0.310***
##                        (0.066)
##
## y90                   0.386***
##                        (0.037)
##
## ----------------------------------------
## Observations             128
## R2                      0.977
## Adjusted R2             0.950
## F Statistic    624.146*** (df = 4; 60)
## ========================================
## Note:        *p<0.1; **p<0.05; ***p<0.01
```

Additional Resources:

https://rstudio-pubs-static.s3.amazonaws.com/372492_3e05f38dd3f248e89cdedd317d603b9a.html#4562_controlling_for_heteroskedasticity:_fixed_effects

Acknowledgements and Thanks:

This lab is based on material by M Godinho de Matos, R Belo and F Reis. Gratefully acknowledged!