Practicals 2

Giovanni Millo

Abstract

Keywords: heterogeneity, fixed effects, first differences, common factors.

This course notes follow the style and typographical conventions of the *Journal of Statistical Software*, particularly as regards `code` and **package names**. R code and output are printed as follows:

```
> print("hello")

[1] "hello"
```

1. Standard panel estimators with plm

1.1. Individual heterogeneity: fixed effects and first differences

Load package **plm** and the dataset `Fatality` from package **Ecdat** *without loading all the other datasets in the package*:

```
> library(plm)
> data(Fatality, package="Ecdat")
```

The Fatality dataset from Stock and Watson, *Introduction to Econometrics*, is a good example of the importance of individual heterogeneity and time effects in a spatially referenced setting.

The research question is whether taxing alcoholics can reduce the road death toll. The basic specification relates the road fatality rate to the (real) beer tax in a classical regression setting:

$$mrall_i = \alpha + \beta beertax_i + \epsilon_i$$

```
> fm <- mrall ~ beertax
```

Data are 1982 to 1988 for each of the continental US States. Most basic step is a cross-sectional analysis for one single year. Subsetting can be done inside the call to `lm`. 1982:

```
> mod82 <- lm(fm, Fatality[Fatality$year==1982, ])
> summary(mod82)
```

```
Call:
lm(formula = fm, data = Fatality[Fatality$year == 1982, ])

Residuals:
    Min      1Q  Median      3Q     Max
-0.9356 -0.4480 -0.1068  0.2295  2.1716

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.0104     0.1391  14.455   <2e-16 ***
beertax       0.1485     0.1884   0.788    0.435
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.6705 on 46 degrees of freedom
Multiple R-squared:  0.01332,	Adjusted R-squared:  -0.008126
F-statistic: 0.6212 on 1 and 46 DF,  p-value: 0.4347
```

The beer tax turns out statistically insignificant. Turning to the last year in the sample:

```
> mod88 <- lm(fm, Fatality[Fatality$year==1988, ])
> summary(mod88)

Call:
lm(formula = fm, data = Fatality[Fatality$year == 1988, ])

Residuals:
     Min       1Q   Median       3Q      Max
-0.72931 -0.36028 -0.07132  0.39938  1.35783

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.8591     0.1060  17.540   <2e-16 ***
beertax       0.4388     0.1645   2.668   0.0105 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.4903 on 46 degrees of freedom
Multiple R-squared:  0.134,	Adjusted R-squared:  0.1152
F-statistic: 7.118 on 1 and 46 DF,  p-value: 0.0105
```

the coefficient is significant *and positive*! Similar results appear for any single year in the sample. Let us loop on years:

```
> library(lmtest)
> for(i in 1982:1988) {
```

```
+   cat(paste("Year", i))
+   print(coeftest(lm(fm, Fatality, subset=(Fatality$year==i))))
+   }

Year 1982
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.01038    0.13908 14.4550   <2e-16 ***
beertax      0.14846    0.18837  0.7881   0.4347
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Year 1983
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.84888    0.12374  14.942  < 2e-16 ***
beertax      0.29858    0.16802   1.777  0.08217 .
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Year 1984
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.80545    0.10964 16.4669   <2e-16 ***
beertax      0.39969    0.15164  2.6358   0.0114 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Year 1985
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.77116    0.10864 16.3036  < 2e-16 ***
beertax      0.39176    0.15467  2.5328  0.01479 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Year 1986
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1.82077    0.11102 16.4001 < 2.2e-16 ***
beertax      0.48029    0.16137  2.9763  0.004639 **
---
```

```
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Year 1987
t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  1.82153    0.11322 16.0880 < 2.2e-16 ***
beertax      0.48304    0.16967  2.8469  0.006575 **
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Year 1988
t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.85907    0.10599  17.540   <2e-16 ***
beertax      0.43875    0.16445   2.668   0.0105 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

More compactly, in **plm**, function `pvcm` estimates separate regressions over individuals (the default) or over time:

```
> pvcm(fm, Fatality, effect="time")

Model Formula: mrall ~ beertax

Coefficients:
     (Intercept) beertax
1982      2.0104 0.14846
1983      1.8489 0.29858
1984      1.8055 0.39969
1985      1.7712 0.39176
1986      1.8208 0.48029
1987      1.8215 0.48304
1988      1.8591 0.43875

> plot(Fatality$beertax, Fatality$mrall, pch=19, col=Fatality$year)
> abline(lm(mrall~beertax, data=Fatality), lty=2, lwd=2)
> unyear <- unique(Fatality$year)
> for(i in 1:length(unyear)) {
+   abline(lm(mrall~ beertax, data=Fatality[Fatality$year==unyear[i], ]),
+          col=unyear[i])
+ }
```

A pooling

$$mrall_{it} = \alpha + \beta beertax_{it} + \epsilon_{it}$$

and a between specification

$$\sum_t mrall_{it} = \alpha + \beta \sum_t beertax_{it} + \epsilon_i$$

do not change the result:

```
> poolmod <- plm(fm, Fatality, model="pooling")
> coeftest(poolmod)

t test of coefficients:

            Estimate Std. Error t value  Pr(>|t|)
(Intercept) 1.853308   0.043567 42.5391 < 2.2e-16 ***
beertax     0.364605   0.062170  5.8647 1.082e-08 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

> bemod <- plm(fm, Fatality, model="between")
> coeftest(bemod)

t test of coefficients:

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.84622    0.11080  16.663   <2e-16 ***
beertax      0.37842    0.15860   2.386   0.0212 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

We suspect the presence of unobserved heterogeneity: in specification terms, we suspect the restriction $\alpha_i = \alpha \forall i$ in the more general model

$$mrall_{it} = \alpha_i + \beta beertax_{it} + \epsilon_{it}$$

to be invalid. A scatterplot with points colored and superposed regression lines *by state* confirms our suspicion:

```
> plot(Fatality$beertax, Fatality$mrall, pch=19, col=Fatality$state)
> abline(lm(mrall~beertax, data=Fatality), lty=2, lwd=2)
> unstate <- unique(Fatality$state)
> for(i in 1:length(unstate)) {
+   abline(lm(mrall~ beertax, data=Fatality[Fatality$state==unstate[i], ]),
+          col=unstate[i])
+ }
```

...too many states, eh? A better way to visualize panel data is through conditional plots in package **lattice**. Let us plot scatterplots of `mrall` vs. `beertax` by state, adding individual regression lines (remember, these are estimated on seven data points only, so do not take them at face value):

```
> library(lattice)
> xyplot(mrall~beertax|state, data=Fatality,
+        panel=function(x,y) {
+            panel.xyplot(x,y)
+            panel.abline(lm(y~x))
+        }
+        )
```

Homogeneity looks unlikely. If intercepts are actually heterogeneous, at a minimum the composite errors in the pooled OLS ($u_{it} = \alpha - \alpha_i + \epsilon_{it}$) are non-spherical and the estimator is inefficient; but if the fixed effects α_i are correlated with the regressor, then the latter is endogenous and pooled OLS is inconsistent. A number of tests are available to check this restriction.

1.2. Testing for individual effects

The testing interface in **plm** generally takes one of two forms (possibly both):

- formula interface
- model interface

with the goal of maximizing flexibility and minimizing computational load, possibly reusing already calculated results.

Testing for intercept homogeneity

The Chow-type pooling test for homogeneity of individual intercepts allows for both: here we use the formula interface. Intercepts are treated as parameters. Null hypothesis is that $\alpha_i = \alpha \forall i$.

```
> pFtest(fm, Fatality)

        F test for individual effects

data:  fm
F = 52.179, df1 = 47, df2 = 287, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Testing for individual effects

The Lagrange multiplier test for individual effects tests the null of spherical errors against the alternative of individual, time-invariant effects:

```
> plmtest(fm, Fatality)
```

```
        Lagrange Multiplier Test - (Honda) for balanced panels

data:  fm
normal = 27.469, p-value < 2.2e-16
alternative hypothesis: significant effects
```

Wooldridge's test for individual effects in the errors uses the formula interface. The null hypothesis here is of the errors' covariance matrix having a diagonal structure, so it has power against a variety of departures from sphericity besides individual effects.

```
> pwtest(fm, Fatality)

        Wooldridge's test for unobserved individual effects

data:  formula
z = 3.2727, p-value = 0.001065
alternative hypothesis: unobserved effect
```

1.3. Fixed effects methods

Both point at the necessity of incorporating (or eliminating!) individual fixed effects. One way is to estimate them explicitly by inclusion of 48 individual intercepts (*LSDV* estimator). This is easily done in R by adding a categorical variable (a `factor`) to the specification (reporting only the relevant coefficient):

```
> lsdvmod <- plm(update(fm, .~.+as.factor(state)), Fatality, model="p")
> coeftest(lsdvmod)["beertax",]

     Estimate    Std. Error       t value      Pr(>|t|)
-0.6558736250  0.1878499921 -3.4914753923  0.0005559707
```

Notice the coefficient! It becomes negative, in line with the theoretical expectations.

A more efficient (but numerically equivalent) way to estimate the same coefficient is by time-demeaning the data (*within* estimator, also fixed effects (FE) estimator). All time-invariants, including the individual intercepts, go to zero. This is the default estimator for function `plm`.

```
> femod <- plm(fm, Fatality)
> summary(femod)

Oneway (individual) effect Within Model

Call:
plm(formula = fm, data = Fatality)

Balanced Panel: n = 48, T = 7, N = 336
```

```
Residuals:
      Min.    1st Qu.     Median    3rd Qu.       Max.
-0.5869619 -0.0828376 -0.0012702  0.0795454  0.8977960

Coefficients:
        Estimate Std. Error t-value Pr(>|t|)
beertax -0.65587    0.18785 -3.4915 0.000556 ***
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Total Sum of Squares:    10.785
Residual Sum of Squares: 10.345
R-Squared:      0.040745
Adj. R-Squared: -0.11969
F-statistic: 12.1904 on 1 and 287 DF, p-value: 0.00055597
```

Notice that the R2 reported is the *within* R2, i.e. that of the regression on demeaned data. The definition of R2 is not obvious in this setting. See `r.squared` for a flexible function calculating the other possible definitions.

1.4. First difference methods

Another way to eliminate time-invariant individual effects is by first differences. Let us take "long" differences over the first and last year in the sample (we do it first the standard, then the "panel way"):

```
> mrall82 <- Fatality[Fatality$year==1982, "mrall"]
> beertax82 <- Fatality[Fatality$year==1982, "beertax"]
> mrall88 <- Fatality[Fatality$year==1988, "mrall"]
> beertax88 <- Fatality[Fatality$year==1988, "beertax"]
> dmrall <- mrall88-mrall82
> dbeertax <- beertax88-beertax82
> summary(lm(dmrall~dbeertax))

Call:
lm(formula = dmrall ~ dbeertax)

Residuals:
     Min       1Q   Median       3Q      Max
-1.22715 -0.09619  0.09212  0.22290  0.67745

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.07204    0.06064  -1.188   0.2410
dbeertax    -1.04097    0.41723  -2.495   0.0162 *
---
```

```
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1

Residual standard error: 0.394 on 46 degrees of freedom
Multiple R-squared:  0.1192,       Adjusted R-squared:    0.1
F-statistic: 6.225 on 1 and 46 DF,  p-value: 0.01625
```

Easier, using panel lagging features:

```
> ldfm.l <- I(mrall-lag(mrall, 6)) ~ I(beertax-lag(beertax, 6))
> coeftest(plm(ldfm.l, Fatality, model="p"))

t test of coefficients:

                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -0.072037   0.060644 -1.1879  0.24098
I(beertax - lag(beertax, 6)) -1.040973   0.417228 -2.4950  0.01625 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

The coefficient estimate is negative and even bigger in absolute value. Alternatively, use the `diff` function:

```
> ldfm.d <- diff(mrall, 6) ~ diff(beertax, 6)
> coeftest(plm(ldfm.d, Fatality, model="p"))

t test of coefficients:

                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.072037   0.060644 -1.1879  0.24098
diff(beertax, 6) -1.040973   0.417228 -2.4950  0.01625 *
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

Taking first differences over successive years:

```
> fdmod <- plm(fm, data=Fatality, model="fd")
> coeftest(fdmod)

t test of coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0031368  0.0119115 -0.2633   0.7925
beertax      0.0136879  0.2852511  0.0480   0.9618
```

significance disappears. To choose between first difference and fixed effects methods, we test the statistical properties of model residuals. If the original errors are stationary, the most appropriate estimator is FE; if they are integrated, then it is FD (whereby the differenced errors, those of the estimated specification, are stationary). We perform Wooldridge's test on original and differenced errors:

```
> pwfdtest(fm, Fatality)

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

data:  plm.model
F = 15.96, df1 = 1, df2 = 238, p-value = 8.629e-05
alternative hypothesis: serial correlation in differenced errors

> pwfdtest(fm, Fatality, h0="fe")

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

data:  plm.model
F = 14.618, df1 = 1, df2 = 238, p-value = 0.0001682
alternative hypothesis: serial correlation in original errors
```

concluding that the truth lies almost exactly in between.

1.5. Common factors and time fixed effects

As Stock and Watson observe, there may be other factors apart from the beer tax influencing the death rate. If these are correlated with taxation, possibly through a common time trend, then omitting them will yield inconsistent estimates. One way of accounting for common factors, varying through time but uniform across states, is to add time fixed effects. Of course we keep individual effects. The specification becomes:

$$mrall_{it} = \alpha_i + d_t + \beta beertax_{it} + \epsilon_{it}$$

Time effects can be included explicitly, as before

```
> coeftest(plm(update(fm, .~.+as.factor(year)), Fatality))

t test of coefficients:

                     Estimate Std. Error t value Pr(>|t|)
beertax             -0.639980   0.197377 -3.2424 0.001328 **
as.factor(year)1983 -0.079903   0.038354 -2.0833 0.038126 *
as.factor(year)1984 -0.072421   0.038352 -1.8883 0.060012 .
as.factor(year)1985 -0.123976   0.038442 -3.2250 0.001408 **
as.factor(year)1986 -0.037864   0.038588 -0.9813 0.327312
as.factor(year)1987 -0.050902   0.038974 -1.3061 0.192600
as.factor(year)1988 -0.051804   0.039623 -1.3074 0.192145
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

or in a *two-way within* specification, which is estimated by OLS on double-demeaned variables:

```
> fe2mod <- plm(fm, Fatality, effect="twoways")
> coeftest(fe2mod)

t test of coefficients:

        Estimate Std. Error t value Pr(>|t|)
beertax -0.63998    0.19738 -3.2424 0.001328 **
---
Signif. codes:  0 âĂŸ***âĂŹ 0.001 âĂŸ**âĂŹ 0.01 âĂŸ*âĂŹ 0.05 âĂŸ.âĂŹ 0.1 âĂŸ âĂŹ 1
```

Let us conclude with the complete model, adding minimum driving age, detention or community service penalties dummies, average miles per driver, unemployment rate and log per-capita real income. Notice how boolean or factor variables are combined through logical operators inside the model formula.

```
> ## modello completo:
> fm <- mrall ~ beertax + I(mlda<=18) + I((mlda>18)&(mlda<=19)) +
+     I((mlda>19)&(mlda<=20)) + I((jaild=="yes")|(comserd=="yes")) +
+     vmiles + unrate + log(perinc)
> fe2mod.c <- plm(fm, Fatality, effect="twoways")
```

`linearHypothesis` in package **car** makes it easy to perform a joint exclusion test on this more complicated specification

```
> #library(car)
> #linearHypothesis(fe2mod.c, c("unrate=0", "log(perinc)=0"), test="F")
```

2. Exercises

2.1. Munnell's productivity model

Munnell (1990), *Public capital productivity*: Does public capital (roads, water facilities, public buildings and structures) help growth? (Example 3 in Baltagi)

48 US states, annual data 1970-1986. Production function:

$$log(gsp) = \alpha + \beta_1 log(pcap) + \beta_2 log(pc) + \beta_3 log(emp) + \beta_4 unemp$$

You are required to:

1. plot `log(gsp)` vs. `log(pcap)`, using colours to mark different states; then plot them conditonally by year
2. estimate the model by cross-sections, then by time series
3. estimate the pooled specification by OLS and by the *between* estimator; pay attention to the coefficient on public capital

4. test for intercept homogeneity and for individual effects
5. determine the most appropriate way to get rid of individual heterogeneity
6. estimate your specification of choice and discuss results
7. assess the need for time fixed effects (*hint: check out ?waldtest*)

Affiliation:
Giovanni Millo
DEAMS, University of Trieste
Piazzale Europa 4
34127 Trieste (Italy)
E-mail: giovanni_millo@deams.units.it