Introduction to R Programming

Ozan Bakış[1]
[1]Bahcesehir University, Department of Economics and BETAM

Outline

Storing panel data I

- The best way to store panel data is to stack the time periods for each i on top of each other. In particular, the time periods for each unit should be adjacent, and stored in chronological order (from earliest period to the most recent). This is sometimes called the "long" storage format. It is by far the most common.

```
f_url = "https://github.com/obakis/econ_data/raw/master/wagepan.rds"
download.file(url = f_url, destfile = "wagepan.rds", mode="wb")
dat = readRDS("wagepan.rds")
```

- While not absolutely necessary for some procedures, it is best to tell R that you have a panel data set. We can convert a regular data frame into a panel data frame using

```
pdat = pdata.frame(dat, index=c("i_var","t_var"))
```

In our example

Storing panel data II

```
head(dat,3)
```

```
##   nr year agric black bus construc ent exper fin hisp poorhlth hours
## 1 13 1980     0     0   1        0   0     1   0    0         0  2672
## 2 13 1981     0     0   0        0   0     2   0    0         0  2320
## 3 13 1982     0     0   1        0   0     3   0    0         0  2940
##   manuf married min nrthcen nrtheast occ1 occ2 occ3 occ4 occ5 occ6
## 1     0       0   0       0        1    0    0    0    0    0    0
## 2     0       0   0       0        1    0    0    0    0    0    0
## 3     0       0   0       0        1    0    0    0    0    0    0
##   occ7 occ8 occ9 per pro pub rur south educ tra trad union lwage d81
## 1    0    0    1   0   0   0   0     0   14   0    0     0  1.20   0
## 2    0    0    1   1   0   0   0     0   14   0    0     1  1.85   1
## 3    0    0    1   0   0   0   0     0   14   0    0     0  1.34   0
##   d82 d83 d84 d85 d86 d87 expersq
## 1   0   0   0   0   0   0       1
## 2   0   0   0   0   0   0       4
## 3   1   0   0   0   0   0       9
```

Storing panel data III

```
vars = c("nr", "year", "exper", "hours", "educ", "lwage")
#install.packages("plm")
library(plm)
```

```
## Loading required package: Formula
```

```
pdat = pdata.frame(dat, index=c("nr","year"))
head(pdat[,vars])
```

```
##          nr year exper hours educ lwage
## 13-1980 13 1980     1  2672   14  1.20
## 13-1981 13 1981     2  2320   14  1.85
## 13-1982 13 1982     3  2940   14  1.34
## 13-1983 13 1983     4  2960   14  1.43
## 13-1984 13 1984     5  3071   14  1.57
## 13-1985 13 1985     6  2864   14  1.70
```

```
pdim(pdat)
```

```
## Balanced Panel: n = 545, T = 8, N = 4360
```

Estimating fixed effects model: FD I

- If the explanatory variable changes over time – at least for some units in the population – heterogeneity bias can be solved by differencing away a_i in the model. For this, write the time periods in reverse order for any unit i:

$$\begin{aligned} y_{i2} &= (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2} \\ y_{i1} &= \beta_0 + \beta_1 x_{i1} + a_i + u_{i1} \end{aligned}$$

  Subtract time period one from time period two to get

$$y_{i2} - y_{i1} = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

Estimating fixed effects model: FD II

- If we define $\Delta y_i = y_{i2} - y_{i1}$, where $\Delta = \textit{change}$ (or *difference*) (similarly for Δx_i and Δu_i) we get the following estimating equation

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

  from which we derive the *first-difference estimator*. (With more than two time periods, other orders of differencing are possible; hence the qualifier "first" .) We will refer to the *FD estimator*.

- Differencing away the unobserved effect, a_i, is simple but can be very powerful for isolating causal effects. In estimating equation we need, among other Gauss-Markov assumptions, that Δu being uncorrelated with Δx (strict exogeneity).

Estimating fixed effects model: FD III

- Important: β_1 is the original coefficient we are interested in. We compute it using the estimating equation

$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$

but interpret it as if we have estimated it in levels equation (FE model)

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}$$

and controlled for a_i.

- Notice that the intercept in the differenced equation, is the *change* in the intercept over the two time periods. It is sometimes interesting to study this change.

- If $\Delta x_i = 0$ for all i, or even if Δx_i is the same nonzero constant, this strategy does not work. We need some variation in Δx_i across i.

Estimating fixed effects model: FD IV

- **Example:** Effects of trade on employment in Turkey
- Once we create a panel data set, it is easy to create variables for changes, growth rates, lagged values etc.

```
f_url = "https://github.com/obakis/econ_data/raw/master/tur_xmlab2y.rds"
download.file(url = f_url, destfile = "tur_xmlab2y.rds", mode="wb")
dat2y = readRDS("tur_xmlab2y.rds")
f_url = "https://github.com/obakis/econ_data/raw/master/tur_xmlab.rds"
download.file(url = f_url, destfile = "tur_xmlab.rds", mode="wb")
dat = readRDS("tur_xmlab.rds")

library(plm)
pdat2y = pdata.frame(dat2y, index=c("nuts3","year"))
head(pdat2y)
```

Estimating fixed effects model: FD V

```
##            pr_no lfp_rate un_rate emp_rate year nuts3   export
## TR100-2009    34     46.5   11.20     41.3 2009 TR100 55539993
## TR100-2010    34     47.8   14.28     41.0 2010 TR100 53149408
## TR211-2009    59     57.3    9.10     52.1 2009 TR211   483240
## TR211-2010    59     55.1    9.58     49.8 2010 TR211   546332
## TR212-2009    22     47.0   14.30     40.3 2009 TR212    93267
## TR212-2010    22     55.2    9.17     50.1 2010 TR212    79095
##              import ihs_x ihs_m   gr_x gr_m
## TR100-2009 78756263  18.5  18.9     NA   NA
## TR100-2010 98454135  18.5  19.1 -0.237 1.18
## TR211-2009   473826  13.8  13.8     NA   NA
## TR211-2010   603210  13.9  14.0  0.890 1.75
## TR212-2009    83882  12.1  12.0     NA   NA
## TR212-2010   198437  12.0  12.9 -1.358 7.16
```

Estimating fixed effects model: FD VI

```
# ihs_x = ln(x) (approximately, except that it is defined at 0)
pdat2y$d_x = diff(pdat2y$ihs_x)
pdat2y$d_emp = diff(pdat2y$emp_rate)
pdat2y$y10 = pdat2y$year==2010 # y10=0 represents 2009, y10=1 is for 2010
fd1 = lm(d_emp ~ d_x+y10, data=pdat2y)
coef(summary(fd1))
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    1.515      0.510    2.97  0.00395
## d_x            0.811      0.668    1.22  0.22779
```

```
fd2 = plm(emp_rate ~ ihs_x + year, data=pdat2y,  model = "within")
coef(summary(fd2))
```

```
##          Estimate Std. Error t-value Pr(>|t|)
## ihs_x       0.811      0.668    1.22  0.22779
## year2010    1.515      0.510    2.97  0.00395
```

Estimating fixed effects model: FD VII

- Note that $y10$ is dropped in the FD regression and the intercept is interpreted as the change in the intercept over the two time periods.
- The intercept is equal to 1.74 which means that, if export level did not change, the employment rate would be predicted to increase by about 1.74 percentage points from 2009 to 2010.
- The coefficient on `ihs_x` is statistically significant and of the expected sign: a fifty percent ($50/100$) increase in the exports increases the employment rate by about $0.4 = 0.81/2$ percentage points (almost half points).

Estimating fixed effects model: FD VIII

- First Differencing can be used with more than two years of panel data, but we must be careful to account for serial correlation (and, as usual, possibly heteroskedasticity) in the FD equation. This is because the FD equation is no longer just a single cross section.
- Generally, we should also include a full set of time dummies for a convincing analysis. With $T = 3$ time periods we can write

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \mathbf{x}_{it}\beta + a_i + u_{it}$$

- First differentiation yields ($d1$ and a_i are time constant)

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \Delta \mathbf{x}_{it}\beta + \Delta u_{it}$$

Estimating fixed effects model: FD IX

- When we difference we lose the first time period, as before, but we are left with a panel data "in difference form" when $T \geq 3$:
- For $t = 2$, we have $\Delta d2_t = 1 - 0 = 1$ and $\Delta d3_t = 0$, but for $t = 3$, we have $\Delta d2_t = 0 - 1 = -1$ and $\Delta d3_t = 1 - 0 = 1$ so that there is no intercept for $t = 3$.
- To avoid this problem, unless the time intercepts in the original model (δ_ts), are of direct interest (usually they are not), it is easier to include an overall intercept and a time dummy for the third period. So, we have only 2 time dummies when $T = 3$.
- When $T > 3$

$$\Delta y_{it} = \alpha + \gamma_3 d3_t + \cdots + \gamma_T d3_T + \Delta \mathbf{x}_{it}\beta + \Delta u_{it}$$

Estimating fixed effects model: FD X

- We follow the same idea for $T = 3$ above. We put an intercept and a time dummy for periods starting from period 3. This works because we have $T - 1$ time periods on each unit i for the first-differenced equation. So, in total we have $T - 1$ time dummies.

```
pdat = pdata.frame(dat, index=c("nuts3","year"))
fd3 = plm(emp_rate ~ ihs_x + year, data=pdat, model = "fd")
coef(summary(fd3))
```

```
##           Estimate Std. Error t-value Pr(>|t|)
## ihs_x        0.267      0.233    1.15 2.52e-01
## year2009    -0.455      0.333   -1.36 1.73e-01
## year2010     1.235      0.476    2.60 9.76e-03
## year2011     3.191      0.583    5.47 7.85e-08
## year2012     3.007      0.672    4.47 1.01e-05
## year2013     3.052      0.755    4.04 6.39e-05
```

Estimating fixed effects model: FE I

- When we believe that $Cov(x_{it}, a_i) \neq 0$, differencing is only one method of eliminating a_i (which itself is sometimes called a *fixed effect*).
- Alternatively, can use the "fixed effects" or "within" transformation: remove the within i time averages.
- Consider the simple model

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}$$

Average this equation across t to get $\bar{y}_i = \beta_0 + \beta_1 \bar{x}_i + a_i + \bar{u}_i$ where $\bar{y}_i = T^{-1} \sum_{t=1}^{T} y_{it}$ is a "time average" for unit i. Similarly for $\bar{x}_i$ and $\bar{u}_i$.

Estimating fixed effects model: FE II

- Subtract the time-averaged equation – sometimes called the "between equation" – from other time periods:

$$y_{it} - \bar{y}_i = \beta_1(x_{it} - \bar{x}_i) + (u_{it} - \bar{u}_i)$$

  As with the FD equation, this equation is free of a_i.

- As in the case of the FD estimator, the "time-demeaned" (or "within") equation is used as an estimating equation to get estimates (β_1). But these estimates are interpreted as as if we have estimated it in levels and controlled for a_i, as

$$y_{it} = \beta_0 + \beta_1 x_{it} + a_i + u_{it}$$

- β_1 is called the "fixed effects (FE) estimator" or the "within estimator."

Estimating fixed effects model: FE III

- There is yet another way to compute $\hat{\beta}_1$. Keep the original data, y_{it} and x_{it}, and run a regression of y_{it} on $n-1$ dummy variables (one for each i) and x_{it}. Called the "dummy variable regression". Often not practical when n is large.

```
## FE estimator using plm package.
fe1 = plm(emp_rate ~ ihs_x + year, data=pdat, model = "within")
coef(summary(fe1))
```

```
##           Estimate Std. Error t-value Pr(>|t|)
## ihs_x        0.197      0.283   0.696 4.87e-01
## year2009    -0.458      0.503  -0.910 3.64e-01
## year2010     1.254      0.509   2.463 1.42e-02
## year2011     3.216      0.513   6.266 9.62e-10
## year2012     3.034      0.515   5.891 8.14e-09
## year2013     3.089      0.525   5.882 8.60e-09
```

Estimating fixed effects model: RE I

- What if x_{it} is (almost) constant?
  $\Rightarrow$ A crucial condition in order to be able to apply FD estimator is to have enough variation in Δx_{it}. Actually, this is required by "no perfect collinearity" of Gauss-Markov assumptions.

  $\Rightarrow$ When x_{it} does not change over time (or change very little), we would expect x_{it} to be constant. Ex. gender or years of schooling for people who have completed their schooling. We will be limited in what we can learn in that case. In such cases we can not separate the effect of a_i on y_{it} from the effect of time-constant factors (such as education) and as a result we can not rely on FE or FD estimation.

- Any solution?

Estimating fixed effects model: RE II

- We already saw that when $Cov(x_{it}, a_i) = 0$ is the case, we can use POLS. But, then we said that because of the presence of a_i in the error term in each period we need to deal with serial correlation. And using heteroskedasticity-robust standard errors does not solve the problem.
- To deal with the serial correlation described above, one solution is using generalized least squares (GLS). This is known as "random effects" (RE) estimation.
- For program evaluation, typically RE would be less convincing because one would like to participation to be correlated with the time-constant factors in a_i.
- One case where RE may be of interest is using good time-constant controls in which case more is taken out of a_i.

Estimating fixed effects model: RE III

```
## RE estimator using plm package. Note: we use same covariates as pooling
# fd1 = plm(emp_rate ~ ihs_x + year, data=pdat, model = "fd")
# fe1 = plm(emp_rate ~ ihs_x + year, data=pdat, model = "within")
re1 = plm(emp_rate ~ ihs_x + year, data=pdat, model = "random")
coef(summary(re1))
```

```
##             Estimate Std. Error t-value Pr(>|t|)
## (Intercept)  42.8859      2.617 16.3857 3.12e-48
## ihs_x         0.0115      0.206  0.0559 9.55e-01
## year2009     -0.4662      0.503 -0.9264 3.55e-01
## year2010      1.3056      0.506  2.5786 1.02e-02
## year2011      3.2823      0.508  6.4556 2.65e-10
## year2012      3.1054      0.509  6.0970 2.23e-09
## year2013      3.1877      0.515  6.1919 1.28e-09
```

The coefficient on unemployment rate is now significant, which suggests that an increase in unemployment affects crime rates.