Panel data methods

Ozan Bakış[1]

[1]Bahcesehir University, Department of Economics and BETAM

Outline

1. Panel data methods

Definitions I

- A panel (longitudinal) data set consists of observations of one or several variables over time. We can think it as many time series data set One ofr each cross-sectional unit).

- 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.

- We can use both **plm** and **fixest** packages to estimate panel data methods. The first thing to do is to convert the usual data (**data.frame** in R language) into panel data (**pdata.frame**)

```r
library("plm")
pdat = pdata.frame(my_data, index=c("id_var","time_var"))
library("fixest")
feols( ..., panel.id =c("id_var","time_var"))
```

Definitions II

where **`id_var`** and **`time_var`** are, respectively, unit and time identifiers.

Unobserved effects model I

- An unobserved effects model (or a fixed effects model) can be written as:

  $$y_{it} = \beta_0 + \lambda_t + \beta_1 x_{1it} + \beta_2 x_{2it} + \dots + \underbrace{a_i + u_{it}}_{v_{it}}$$

  where λ_t is a shorthand for all year dummies: $\delta_0 year_2 + \delta_1 year_3 \dots$.

- v_{it} is the composite error term consisting of a time-constant (a_i)and a time-variable component (u_{it}). Usually, the last one is known as the "idiosyncratic shock".

- the unobserved effect (a_i) is the main reason of unobserved heterogeneity. It represents all factors that are constant but unobservable to the econometrician.

Two-period panel data analyis I

- Consider the following example where we have unemployment and crime rates data for two years (1982, 1987) for 46 cities in USA (`crime2.rds`).

- We want to know whether an increase in unemployment rate in a city causes an increase in crime rates. The unemployment rate is in percent and crime rate is measured as the number of crimes per 1000 people.

```r
url = "https://github.com/obakis/econ_data/raw/master/crime2.rds"
download.file(url, "crime.rds", mode ="wb")
crime = readRDS("crime.rds")
```

- A basic model that can be used is

$$\mathrm{crmrte}_{it} = \beta_0 + \delta_0 \mathrm{d87}_t + \beta_1 \mathrm{unem}_{it} + v_{it}, \quad t = 1, 2.$$

Two-period panel data analyis II

- The above model is called a pooled OLS (POLS) model. The variable $d87_t$ is a constructed time dummy for the second time period: $d87_t = 1$ if $year = 1987$ and $d87_t = 0$ if $year = 1982$. This model does not use the ”panel data” characteristics.

```
reg1 = lm(crmrte ~ unem+d87, data=crime)
coef(summary(reg1))
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   93.420      12.74   7.333 9.92e-11
## unem           0.427       1.19   0.359 7.20e-01
## d87            7.940       7.98   0.996 3.22e-01
```

- The coefficient on unemployment is positive but not significant. When we have multiple years, we need to use year dummies to allow changing intercepts (control for common macro shocks) to get a good estimate of a causal effect.

Two-period panel data analyis III

- It is likely that $\widehat{\beta}_1$ is still biased because of unobserved city characteristics that are correlated with unemployment (such as migration, temperature etc.) so that $\text{Cov}(unem_{it}, v_{it}) \neq 0$.

- If there are unobserved factors making $\text{Cov}(unem_{it}, v_{it}) \neq 0$, then we have omitted variables problem, and OLS estimates are biased!

- Let a_i be a city specific unobserved factor that is time constant. And assume that it is correlated both with unemployment and crime rates. Given that in the above regression, naturally a_i is part of the error term, we expect $\text{Cov}(x_{it}, v_{it}) \neq 0$.

Two-period panel data analyis IV

- So, it would be better to decompose the composite error into two components $v_{it} = a_i + u_{it}$ (a_i: unobserved fixed effect; u_{it}: idiosyncratic error) and specify an **unobserved (or fixed) effects model** :

  $$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + \underbrace{a_i + u_{it}}_{v_{it}}$$

  The estimate in the POLS model is likely to be biased because for unbiasedness we need $Cov(x_{it}, v_{it}) = 0$ which requires

  $$Cov(x_{it}, a_i) = 0, \quad \text{and} \quad Cov(x_{it}, u_{it}) = 0$$

- Even if the second one is respected, it is usually unreasonable to assume the first one. Because, usually we have city fixed unobserved effects that affect crime rates and are correlated with unemployment rate. In such situations, we say that POLS suffers from heterogeneity bias.

Two-period panel data analyis V

- A first option to eliminate heterogeneity bias is estimate a_i along with other parameters. This is the same thing as estimating an intercept for each i, and is called **the dummy variable regression**:

  $$crmrte_{it} = \underbrace{\theta_i}_{\beta_0+a_i} + \delta_0 d87_t + \beta_1 unem_{it} + u_{it}$$

  In practice, for n cross-sectional units we need to use $n-1$ dummy variables to estimate θ_i.

- **Important:** Since a_i is modeled as a covariate we need only $\mathrm{Cov}(x_{it}, u_{it}) = 0$ to get unbiased estimates. $\mathrm{Cov}(x_{it}, a_i) \neq 0$ is allowed in this model!

```r
crime$id = rep(1:46, each=2) # create panel id which is missing
dvr = lm(crmrte ~ factor(id) + d87 + unem, data=crime)
summary(dvr)$coef[c(1:3,45:48),]
```

Two-period panel data analyis VI

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     51.49     12.346   4.171  0.00014
## factor(id)2     17.29     14.195   1.218  0.22967
## factor(id)3      4.69     14.239   0.329  0.74351
## factor(id)45   -17.75     14.178  -1.252  0.21713
## factor(id)46    -3.28     14.517  -0.226  0.82244
## d87             15.40      4.702   3.276  0.00206
## unem             2.22      0.878   2.527  0.01519
```

```r
library("fixest")
dvr2 = feols(crmrte ~ unem + i(year) | id,
             panel.id = c("id","year"),
             data=crime, vcov="iid")
summary(dvr2)
```

Two-period panel data analyis VII

```
## OLS estimation, Dep. Var.: crmrte
## Observations: 92
## Fixed-effects: id: 46
## Standard-errors: IID
##           Estimate Std. Error t value  Pr(>|t|)
## unem          2.22      0.878    2.53 0.0151893 *
## year::87     15.40      4.702    3.28 0.0020605 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 9.80504     Adj. R2: 0.774292
##                 Within R2: 0.196062
```

```
dvr3 = feols(crmrte ~ unem | id+year,
             panel.id = c("id","year"),
             data=crime, vcov="cluster")
summary(dvr3)
```

Two-period panel data analyis VIII

```
## OLS estimation, Dep. Var.: crmrte
## Observations: 92
## Fixed-effects: id: 46,  year: 2
## Standard-errors: Clustered (id)
##      Estimate Std. Error t value  Pr(>|t|)
## unem     2.22      0.815    2.72 0.0092419 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 9.80504     Adj. R2: 0.774292
##                 Within R2: 0.1267
```

- If the explanatory variable changes over time, instead of estimating a_i directly using the model, $y_{it} = \beta_0 + a_i + \delta_0 d2_t + \beta_1 x_{it} + u_{it}$, a second option is to eliminate it:

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

Two-period panel data analyis IX

Or in first differenced form, also called estimating equation:

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

where $\Delta = change$ (or $difference$). Estimating the above yields the first-difference (FD) estimator. Differencing away a_i, is a powerful way of isolating causal effects. The key assumption for unbiasedness is $\text{Cov}(\Delta u, \Delta x) = 0$ (strict exogeneity). Again $\text{Cov}(x_{it}, a_i) \neq 0$ is allowed

1. We can estimate the "estimating equation" directly by using `lm` function (by manually creating $ccrmrte = \Delta crmrte, cunem = \Delta unem$)

   ```
   reg3 = lm(ccrmrte ~ cunem, data=crime) # ccrmrte = crmrte_2 - crmrte_1, etc.
   coef(summary(reg3))
   ```

   ```
   ##             Estimate Std. Error t value Pr(>|t|)
   ## (Intercept)    15.40      4.702    3.28  0.00206
   ## cunem           2.22      0.878    2.53  0.01519
   ```

2. Or alternatively using `plm` or `fixest` packages:

Two-period panel data analyis X

```
# library("plm")
# pdat = pdata.frame(crime, index=c("id","year"))
# reg4 = plm(crmrte ~ d87 + unem, data=pdat, model="fd")
# summary(reg4)$coef
reg5 = feols(d(crmrte) ~ d(unem) | year,
             panel.id = c("id","year"),
             data=crime, vcov="iid")
```

```
## NOTE: unitary time step taken: 5.
## NOTE: 46 observations removed because of NA values (LHS: 46, RHS: 46).
```

```
summary(reg5)
```

```
## OLS estimation, Dep. Var.: d(crmrte, 1)
## Observations: 46
## Fixed-effects: year: 1
## Standard-errors: IID
##             Estimate Std. Error t value Pr(>|t|)
## d(unem, 1)     2.22      0.878    2.53 0.015189 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 19.6     Adj. R2: 0.106852
##              Within R2: 0.1267
```

Two-period panel data analyis XI

- Now, we have a positive and statistically significant relationship between the crime and unemployment rates. Thus, differencing to eliminate time-constant effects makes a big difference in this example.

- When unemployment rate goes up by 1 percentage point, the number of crimes per 1000 people increases by 2.2.

Returns to union membership and marriage I

- `wagepan.rds`: Working men from 1980 to 1987, so eight years. $N = 545$. Use $lwage$ as the dependent variable and standard panel data methods.

  $$lwage_{it} = \beta_0 + \beta_1 educ_{it} + \beta_2 exper_{it} + \beta_3 married_{it} + \beta_4 union_{it} + \delta_t + a_i + u_{it}$$

  δ_t refers to **set of year dummies**. Union status and marital status change over time. Education does not. Experience does, but if we know the experience in 1980 we know it in any other year (increases by one each year).

- While experience, union status and marital status change over time, race and education stay constant.

Returns to union membership and marriage II

- We are likely to have two problems. First, unobserved effects are probably endogenous: $\mathrm{Cov}(\mathbf{X}_{it}, a_i) \neq 0$. This implies that people that are married and have union membeship are likely to differ from others. Second, strict exogeneity is likely to fail: $\mathrm{Cov}(\mathbf{X}_{it}, u_{it}) \neq 0$. Why are union status and marital status changing over time? Do shocks to wages contribute?

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

wagepan$f_year = factor(wagepan$year)
library("plm")
pdat = pdata.frame(wagepan, index=c("nr","year"))
library("fixest")
```

- One can use POLS for computing returns to marraige and union membership.

Returns to union membership and marriage III

```r
# reg_po = plm(lwage ~ educ + exper + black + union + married + f_year,
#              data=pdat,
#              model="pooling")
# summary(reg_po)$coef
reg_po = feols(lwage ~ educ + exper + black + union + married | year,
               data=wagepan,
               panel.id = c("nr","year"), vcov="iid")
summary(reg_po)
```

```
## OLS estimation, Dep. Var.: lwage
## Observations: 4,360
## Fixed-effects: year: 8
## Standard-errors: IID
##          Estimate Std. Error t value   Pr(>|t|)
## educ       0.0923    0.00515   17.94  < 2.2e-16 ***
## exper      0.0304    0.00549    5.53 3.2982e-08 ***
## black     -0.1400    0.02324   -6.02 1.8504e-09 ***
## union      0.1869    0.01710   10.93  < 2.2e-16 ***
## married    0.1109    0.01567    7.07 1.7557e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Returns to union membership and marriage IV

```
## RMSE: 0.480008     Adj. R2: 0.185341
##                  Within R2: 0.121537
```

- But the estimate is likely to be biased. What about FD or FE?

```
# reg_fd = plm(lwage ~ educ + exper + black + union + married + f_year,
#              data=pdat,
#              model="fd")
# summary(reg_fd)$coef
reg_fd = feols(d(lwage) ~ d(educ)+d(exper)+d(black)+d(union)+d(married) | year,
               data=wagepan, panel.id = c("nr","year"),
               vcov="cluster")
```

```
## NOTE: 545 observations removed because of NA values (LHS: 545, RHS: 545).
## The variable 'd(exper, 1)' has been removed because of collinearity (see
$collin.var).
```

```
summary(reg_fd)
```

Returns to union membership and marriage V

```
## OLS estimation, Dep. Var.: d(lwage, 1)
## Observations: 3,815
## Fixed-effects: year: 7
## Standard-errors: Clustered (nr)
##                Estimate Std. Error t value Pr(>|t|)
## d(union, 1)     0.0419     0.0219    1.91 0.056375 .
## d(married, 1)   0.0403     0.0242    1.67 0.096484 .
## ... 1 variable was removed because of collinearity (d(exper, 1))
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.442721     Adj. R2: 0.002494
##                  Within R2: 0.002004
```

```
# reg_fe = plm(lwage ~ educ + exper + black + union + married + f_year,
#              data=pdat,
#              model="within")
# summary(reg_fe)$coef
# reg_fe2 = feols(lwage ~ educ+exper+black+union+married+f_year|nr, data=wagepan,

#                 panel.id = c("nr","year"))
# summary(reg_fe2)
```

Returns to union membership and marriage VI

```
reg_fe = feols(lwage ~ educ + exper + black + union + married | nr + year,
               data=wagepan, panel.id = c("nr","year"),
               vcov="cluster")
```

```
## The variables 'educ', 'exper' and 'black' have been removed because of
collinearity (see $collin.var).
```

```
summary(reg_fe)
```

```
## OLS estimation, Dep. Var.: lwage
## Observations: 4,360
## Fixed-effects: nr: 545,  year: 8
## Standard-errors: Clustered (nr)
##         Estimate Std. Error t value  Pr(>|t|)
## union     0.0834     0.0231    3.62 0.0003279 ***
## married   0.0583     0.0213    2.73 0.0064602 **
## ... 3 variables were removed because of collinearity (educ, exper and black)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.330217     Adj. R2: 0.559649
##                  Within R2: 0.007635
```

Returns to union membership and marriage VII

- Observable time-constant variables's coefficients are dropped (education and black). So we can not compute the level of the returns to education.

- `exper` is dropped for another reason. The reason is we can not separately identify / estimate the effect of `exper` and years in estimating equation of FD or FE. Both increases by one for every year in the sample. The starting points for exper are different across people but this is not distinguishable from the fixed effect, a_i.

Random effects estimator (RE) I

- What to do when x_{it} is constant such as education or race dummies?
  $\Rightarrow$ If $x_{it} = \bar{x}$ for each individual i and time t, we cannot estimate its coefficient using any of estimators we know: POLS, FD, FE. Actually, ”no perfect collinearity” assumption (one of Gauss-Markov assumptions) rules out this.
  $\Rightarrow$ If $x_{it} = x_i$ for each time t this means that x differs across people but does not change over time for the same individual. 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 (x_i) and as a result we can not rely on FE or FD estimation.

- Any solution?
  $\Rightarrow$ No perfect solution: an improved version of pooled OLS: random effects estimator

Random effects estimator (RE) II

- We already know that when $\text{Cov}(x_{it}, a_i) = 0$, we can use pooled OLS estimator. But, because of the presence of a_i in the composite error term $v_{it} = a_i + u_{it}$, we need to deal with serial correlation in the error term.

- To deal with the serial correlation described above, one solution is using generalized least squares (GLS). This is known as ”random effects” (RE) estimation.

- Even if RE allows time constant covariates, it is crucial to keep in mind that RE rely on not only strict exogeneity with respect to u_{it}: $\text{Cov}(\mathbf{X}_{is}, u_{it}) = 0$ for $s \neq t$ but also exogeneity of fixed effects a_i, $\text{Cov}(\mathbf{X}_{it}, a_i) = 0$. This is likely to be wrong !

- For policy analysis, RE is typically less convincing than FD or FE: we want to allow $\mathbf{X}_{it}$ to be correlated with the time-constant factors in a_i.

Random effects estimator (RE) III

- However, with good time-constant controls, RE may be convincing. This is because more is taken out of a_i as we add time-constant variables.

- The serial correlation in $\{v_{it}\}$ (ρ) is also known as intraclass correlation. It is the proportion of total variance in dependent variable that is due to a_is.

$$\text{Corr}(v_{it}, v_{is}) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2} \equiv \rho$$

- A useful characterization of RE is in terms of a ”partially-time-demeaned” equation. It can be shown that for a given $\theta(T, \rho)$, that is between zero and one and increases with both ρ and T, Then RE estimate is obtained by regressing

$$y_{it} - \hat{\theta}\bar{y}_i \text{ on } \mathbf{X}_{it} - \hat{\theta}\bar{\mathbf{X}}_i$$

$y_{it} - \hat{\theta}\bar{y}_i$ and $\mathbf{X}_{it} - \hat{\theta}\bar{\mathbf{X}}_i$ are ”partially-time-demeaned” variables.

Random effects estimator (RE) IV

- We have the following

  $$\begin{aligned} \hat{\theta} &\rightarrow 0 \Rightarrow \hat{\beta}_{RE} \approx \hat{\beta}_{POLS} \\ \hat{\theta} &\rightarrow 1 \Rightarrow \hat{\beta}_{RE} \approx \hat{\beta}_{FE} \end{aligned}$$

  When θ approaches 1, and RE converges to FE !

- Actually, $\hat{\theta}$ is a measure for how much of the unobserved effect a_i we are extracting from the error term, during the estimation. If θ goes to zero we are leaving a large part in the error term and as a result the bias of the RE estimator will be larger.

- While POLS leaves a_i entirely in the error term; FE (or FD) remove it completely and RE leaves it partially in the error term.

```r
reg_po = feols(lwage ~ educ + exper + black + union + married | year,
                data=wagepan,
                panel.id = c("nr","year"), vcov="iid")
summary(reg_po)
```

Random effects estimator (RE) V

```
## OLS estimation, Dep. Var.: lwage
## Observations: 4,360
## Fixed-effects: year: 8
## Standard-errors: IID
##          Estimate Std. Error t value   Pr(>|t|)
## educ       0.0923    0.00515   17.94  < 2.2e-16 ***
## exper      0.0304    0.00549    5.53 3.2982e-08 ***
## black     -0.1400    0.02324   -6.02 1.8504e-09 ***
## union      0.1869    0.01710   10.93  < 2.2e-16 ***
## married    0.1109    0.01567    7.07 1.7557e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.480008     Adj. R2: 0.185341
##                  Within R2: 0.121537
```

```
# reg_p = plm(lwage ~ educ+exper+married+union+f_year, data=pdat, model="pooling")

# summary(reg_p)$coef
```

Random effects estimator (RE) VI

- With POLS everything is statistically significant. Returns to union wage premium is about 17.5%. Marriage premium is about 12.5%. Is this causal, or is their a matching story?

- When a_i is in the error term (case of POLS) the marriage and union premiums are relatively higher. If married or unionized workers are also more able, then, our results will be biased.

- When we eliminate a_i from the error term (FD or FE), then they go down significantly which suggests that POLS estimates are biased. Otherwise, the estimate for marriage and union premiums would not change significantly.

```r
summary(reg_fe2)
```

```
## Error in h(simpleError(msg, call)): error in evaluating the argument
'object' in selecting a method for function 'summary': object 'reg_fe2' not
found
```

Random effects estimator (RE) VII

- RE estimates of union membership and marriage premiums are much smaller compared POLS.

```
reg_re = plm(lwage ~ educ + exper + black + union + married + f_year,
             data=pdat,
             model="random")
summary(reg_re)
```

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = lwage ~ educ + exper + black + union + married +
##     f_year, data = pdat, model = "random")
##
## Balanced Panel: n = 545, T = 8, N = 4360
##
## Effects:
##                  var std.dev share
## idiosyncratic 0.125   0.353  0.54
## individual    0.105   0.324  0.46
## theta: 0.641
```

Random effects estimator (RE) VIII

```
##
## Residuals:
##    Min. 1st Qu.  Median 3rd Qu.    Max.
## -4.5657 -0.1441  0.0268  0.1916  1.5422
##
## Coefficients:
##               Estimate Std. Error z-value Pr(>|z|)
## (Intercept)     0.1608     0.1469    1.09   0.2738
## educ            0.0940     0.0105    8.95  < 2e-16 ***
## exper           0.0333     0.0111    2.99   0.0028 **
## black          -0.1376     0.0470   -2.93   0.0034 **
## union           0.1108     0.0179    6.18  6.4e-10 ***
## married         0.0733     0.0168    4.36  1.3e-05 ***
## f_year1981      0.0787     0.0242    3.26   0.0011 **
## f_year1982      0.0983     0.0309    3.18   0.0015 **
## f_year1983      0.1071     0.0396    2.70   0.0069 **
## f_year1984      0.1403     0.0493    2.84   0.0045 **
## f_year1985      0.1562     0.0595    2.62   0.0087 **
## f_year1986      0.1820     0.0700    2.60   0.0093 **
## f_year1987      0.2069     0.0806    2.57   0.0103 *
## ---
```

Random effects estimator (RE) IX

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    658
## Residual Sum of Squares: 545
## R-Squared:      0.172
## Adj. R-Squared: 0.169
## Chisq: 901.556 on 12 DF, p-value: <2e-16
```

```
# library("lme4") # for random effects estimator
# reg_re = lmer(lwage ~ educ + exper + black + union + married + f_year + (1|nr),
#               data = pdat)
# summary(reg_re)$coef
```

- The partial-time-demeaning parameter, $\widehat{\theta} = 0.64$, so not close to zero and pretty far from one. But, we expect them to be closer to the FE estimates than to the POLS estimates.

Random effects estimator (RE) X

- The remaining positive returns to marriage or union can be interpreted in two ways and unfortunately we cannot distinguish between these alternative hypothesis.
  $\Rightarrow$ marriage or union really makes men (reminder: we have only data on men) more productive
  $\Rightarrow$ firms judge married and unionized men as more stable or reliable and pay them higher wages.

Application I

Effect of Direct Sale Points (”tanzim satış”) on inflation rate in Turkey

- On 11 February 2019, the government introduced Direct Sale Points (DSP). DSP are municipality-led vegetable stalls which sell products at discounted prices (called “tanzim satış” in Turkish).
- Only 9 products have been sold at these DSP (tomatoes, potatoes, onions, green peppers and other vegetables). The data (”tanzim_pooled.rds”) contains monthly inflation rate in February for these 9 products in all regions of Turkey in 2018 (before DSP) and in 2019 (after DSP).
- In February 2019 DSP were operated only in Istanbul and Ankara (called “tanzim regions”), later they were operated in other cities as well. Since there were only 65 DSP in Tanzim regions (50 in Istanbul and 15 in Ankara) but a large population (15 million in Istanbul and almost 6 million in Ankara) it is not obvious that there is an effect on inflation rates.

Application II

```r
furl = "https://github.com/obakis/econ_data/raw/master/tanzim_panel.rds"
download.file(furl, destfile = "tanzim_panel.rds", mode="wb")
tanzim = readRDS("tanzim_panel.rds")
```

- To calculate the effect of DSP we can use simple averages

```r
reg_po1 = lm(inf ~ dsp+ y19 + nuts2, data=tanzim)
summary(reg_po1)
```

```
##
## Call:
## lm(formula = inf ~ dsp + y19 + nuts2, data = tanzim)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -22.20  -9.59  -1.85   8.30  38.62
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   5.4434     3.4979    1.56     0.12
## dsp          -6.4570     4.2039   -1.54     0.13
## y19          -5.6480     1.1660   -4.84  1.8e-06 ***
```

Application III

```
## nuts2TR21    -2.3831     4.5532   -0.52     0.60
## nuts2TR22     3.1321     4.5532    0.69     0.49
## nuts2TR31     0.4946     4.5532    0.11     0.91
## nuts2TR32     1.4981     4.5532    0.33     0.74
## nuts2TR33     2.2672     4.5532    0.50     0.62
## nuts2TR41    -0.9140     4.5532   -0.20     0.84
## nuts2TR42    -1.3841     4.5532   -0.30     0.76
## nuts2TR51     0.4576     4.0390    0.11     0.91
## nuts2TR52     0.1799     4.5532    0.04     0.97
## nuts2TR61    -0.0512     4.5532   -0.01     0.99
## nuts2TR62     2.5462     4.5532    0.56     0.58
## nuts2TR63     2.5034     4.5532    0.55     0.58
## nuts2TR71     0.8141     4.5532    0.18     0.86
## nuts2TR72     1.2443     4.5532    0.27     0.78
## nuts2TR81     1.2519     4.5532    0.27     0.78
## nuts2TR82     2.1824     4.5532    0.48     0.63
## nuts2TR83     0.3363     4.5532    0.07     0.94
## nuts2TR90     4.8643     4.5532    1.07     0.29
## nuts2TRA1     2.9662     4.5532    0.65     0.52
## nuts2TRA2     0.2680     4.5532    0.06     0.95
## nuts2TRB1     2.5411     4.5532    0.56     0.58
```

Application IV

```
## nuts2TRB2     5.0159     4.5532    1.10     0.27
## nuts2TRC1     1.2519     4.5532    0.27     0.78
## nuts2TRC2    -0.9974     4.5532   -0.22     0.83
## nuts2TRC3    -2.0754     4.5532   -0.46     0.65
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.1 on 440 degrees of freedom
## Multiple R-squared:  0.0965,Adjusted R-squared:  0.041
## F-statistic: 1.74 on 27 and 440 DF,  p-value: 0.0131
```

```
# library(plm)
# pdat = pdata.frame(tanzim, index = c("id","year"))
#reg_fd = plm(inf ~ y19 + nuts2 + dsp, model="fd", data=pdat)
#reg_re = plm(inf ~ y19 + nuts2 + dsp, model="random", data=pdat)
#summary(reg_re)
#summary(reg_fd)
# reg_po = plm(inf ~ y19 + nuts2 + dsp, model="pooling", data=pdat)
# reg_fe = plm(inf ~ y19 + nuts2 + dsp, model="within", data=pdat)
reg_po = feols(inf ~ dsp + nuts2 | y19,
                data=tanzim,
```

Application V

```
                panel.id = c("id","year"), vcov="iid")
summary(reg_po)
```

```
## OLS estimation, Dep. Var.: inf
## Observations: 468
## Fixed-effects: y19: 2
## Standard-errors: IID
##           Estimate Std. Error t value Pr(>|t|)
## dsp        -6.4570       4.20 -1.5360  0.12527
## nuts2TR21  -2.3831       4.55 -0.5234  0.60097
## nuts2TR22   3.1321       4.55  0.6879  0.49188
## nuts2TR31   0.4946       4.55  0.1086  0.91355
## nuts2TR32   1.4981       4.55  0.3290  0.74231
## nuts2TR33   2.2672       4.55  0.4979  0.61878
## nuts2TR41  -0.9140       4.55 -0.2007  0.84100
## nuts2TR42  -1.3841       4.55 -0.3040  0.76128
## nuts2TR51   0.4576       4.04  0.1133  0.90984
## nuts2TR52   0.1799       4.55  0.0395  0.96851
## nuts2TR61  -0.0512       4.55 -0.0112  0.99103
## nuts2TR62   2.5462       4.55  0.5592  0.57631
## nuts2TR63   2.5034       4.55  0.5498  0.58273
```

Application VI

```
## nuts2TR71   0.8141       4.55  0.1788  0.85819
## nuts2TR72   1.2443       4.55  0.2733  0.78477
## nuts2TR81   1.2519       4.55  0.2749  0.78349
## nuts2TR82   2.1824       4.55  0.4793  0.63196
## nuts2TR83   0.3363       4.55  0.0739  0.94116
## nuts2TR90   4.8643       4.55  1.0683  0.28597
## nuts2TRA1   2.9662       4.55  0.6514  0.51510
## nuts2TRA2   0.2680       4.55  0.0588  0.95310
## nuts2TRB1   2.5411       4.55  0.5581  0.57707
## nuts2TRB2   5.0159       4.55  1.1016  0.27123
## nuts2TRC1   1.2519       4.55  0.2749  0.78349
## nuts2TRC2  -0.9974       4.55 -0.2191  0.82671
## nuts2TRC3  -2.0753       4.55 -0.4558  0.64876
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 11.7     Adj. R2: 0.041033
##              Within R2: 0.036976
```

Application VII

```r
reg_fe = feols(inf ~ dsp + nuts2 | id + year,
               data=tanzim,
               panel.id = c("id","year"),
            cluster = "nuts2") # vcov = "cluster" is not automatically implied !
```

```
## The variables 'nuts2TR21', 'nuts2TR22' and twenty-three others have been
removed because of collinearity (see $collin.var).
```

```r
summary(reg_fe)
```

```
## OLS estimation, Dep. Var.: inf
## Observations: 468
## Fixed-effects: id: 234,  year: 2
## Standard-errors: Clustered (nuts2)
##     Estimate Std. Error t value  Pr(>|t|)
## dsp    -6.46       1.38   -4.67 8.687e-05 ***
## ... 25 variables were removed because of collinearity (nuts2TR21, nuts2TR22 and 23 oth
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 8.42908     Adj. R2: 0.063888
##                 Within R2: 0.01031
```