Optional Reading: Panel Data DID

(Jing Li, Miami University)

1. The kielmc data in R's Wooldridge package used for garbage incinerator study is pooled cross sections—one cross section in 1978, and the other in 1981. Since we do not observe the same house over time, the houses in the two cross sections in general are different. In short, the kielmc data is not panel data.

2. This note aims to explain how to conduct a DID analysis with panel data.

3. For example, the authors of a famous paper attempt to estimate the causal effect of the increase in minimum wage (treatment) on employment in fast food restaurants. They compare Pennsylvania (PA, state=0, control group), where there is no change in minimum wage, to New Jersey (NJ, state=1, treatment group), where there is change in minimum wage. Their data is a panel data in wide form

```
> d = read.csv("311r_wagewide.csv")
> head(d)
  storeid chain state  ft1  pt1 ma1 wage1 STATUS2  ft2 pt2 ma2 wage2
1      46     1     0 30.0 15.0   3    NA       1  3.5  35   3  4.30
2      49     2     0  6.5  6.5   4    NA       1  0.0  15   4  4.45
3     506     2     0  3.0  7.0   2    NA       1  3.0   7   4  5.00
4      56     4     0 20.0 20.0   4   5.0       1  0.0  36   2  5.25
5      61     4     0  6.0 26.0   5   5.5       1 28.0   3   6  4.75
> d$fte1 = d$ft1+0.5*d$pt1+d$ma1
> d$fte2 = d$ft2+0.5*d$pt2+d$ma2
> d = d[d$storeid!=407&d$STATUS2==1,]
```

   (a) The first column is storeid; the third column state is a treatment dummy that equals one for NJ, and zero for PA.

   (b) ft1 is the number of full time employees before the minimum wage changes, and ft2 is after the change. pt is the number of part time employees, and ma is the number of managers. Outcome variable is full-time-equivalence (fte), a weighted average of ft, pt and ma:

$$fte = ft + 0.5 * pt + ma$$

(c) The STATUS2 variable indicates how a restaurant responds in the two rounds of interviews. Some restaurants are missing (NA) in the second period, maybe because they are out of business or other reasons. Later we will only use restaurants with STATUS2 being one. Also we need to drop the restaurant of storeid 407 because that id is not unique.

(d) There is only one row for each storeid, and in the row we observe ft, pt, ma, and wage twice—one before the treatment and the other after the treatment. Panel data organized in this way is wide form

4. Panel data in wide form is ready for DID analysis—we first compute the after-and-before difference in fte. Then we compute average difference in fte for the untreated and treated groups, respectively. The same DID estimate is reported by regressing that difference in fte onto the treatment dummy state

$$\Delta fte_i = \beta_0 + \beta_1 state_i + u_i \quad \texttt{(DID regression based on wideform differenced data)} \tag{1}$$

The results are

```
> d$dfte = d$fte2-d$fte1
> mean(d$dfte[d$state==0],na.rm=T)
[1] -2.092466
> mean(d$dfte[d$state==1],na.rm=T)
[1] 0.6516502
> 0.6516502-(-2.0924658)
[1] 2.744116

> summary(lm(dfte~factor(state),data=d))$coef
                Estimate Std. Error   t value   Pr(>|t|)
(Intercept)    -2.092466   1.044017 -2.004245 0.04576439
factor(state)1  2.744116   1.163001  2.359513 0.01881213

> t.test(d$dfte[d$state==1],d$dfte[d$state==0],var.equal=T)
Two Sample t-test
data:  d$dfte[d$state == 1] and d$dfte[d$state == 0]
t = 2.3595, df = 374, p-value = 0.01881
```

(a) The average difference in fte in PA is -2.092466, implying that employment in PA fell over time.

(b) Surprisingly, the average difference in fte in NJ is 0.6516502, a positive value. So during the same period the employment rose in NJ. Taken literally, it suggests that an increase in minimum wage be associated with increase in employment. That finding earns the author Nobel prize, because it contradicts with the text-book's prediction that employment would fall when minimum wage rises.

(c) The DID estimate is $(ta - tb) - (ca - cb)$, or

$$0.6516502 - (-2.0924658) = 2.744116$$

and is shown to be significant with t value of 2.359513.

(d) We obtain the same t value with function t.test

5. Panel data in long form is more common in practice. Thus, we transform the wide form (dw) into long form (dl) using the reshape function. Let us focus on the fte and state variables.

```
> dw = d[,c("fte1","fte2","state")]
> head(dw)
   fte1 fte2 state
1 40.50 24.0     0
2 13.75 11.5     0
...
> dl = reshape(dw, varying = c("fte1", "fte2"),v.names = "fte",timevar = "time", ti
> dl = dl[order(dl$id,dl$time),]
> head(dl)
    state time   fte id
1.1     0    1 40.50  1
1.2     0    2 24.00  1
2.1     0    1 13.75  2
2.2     0    2 11.50  2
...
```

Notice that in the long form there are two rows for each restaurant, one is before the treatment (time= 1) and the other is after treatment (time= 2). The variable time

is categorical (factor), but not dummy variable. The variable id is also categorical (factor).

6. The next step is to generate an interaction term of treatment dummy and post-treatment time dummy, and run DVR III

$$y_{it} = \beta_0 + \beta_1 d_i + \beta_2 d_t + \beta_2 d_i d_t + u_{it} \quad \text{(DVR III)} \tag{2}$$

where d_i equals one if a restaurant is in NJ; d_t equals one if time= 2; and $d_i d_t$ is the interaction term.

```
> dl$post = as.integer(dl$time==2)
> dl$inta = dl$state*dl$post
> summary(lm(fte~inta+factor(state)+post,data=dl))$coef
                Estimate Std. Error   t value     Pr(>|t|)
(Intercept)    23.560000   1.070899 22.000215 1.379518e-83
inta            2.738110   1.686981  1.623083 1.049814e-01
factor(state)1 -2.937258   1.193433 -2.461185 1.406650e-02
post           -1.976667   1.514479 -1.305179 1.922223e-01
> mean(dl$fte[dl$time==1&dl$state==0],na.rm=T)
[1] 23.56
> mean(dl$fte[dl$time==1&dl$state==1],na.rm=T)-mean(dl$fte[dl$time==1&dl$state==0],
[1] -2.937258
> mean(dl$fte[dl$time==2&dl$state==0],na.rm=T)-mean(dl$fte[dl$time==1&dl$state==0],
[1] -1.976667
```

The DID estimate is 2.738110 (t value is 1.623083), similar to previous DID estimate of 2.744116.

7. That DID estimate from DVR III can be improved because it fails to account for the special structure of panel data, i.e., the same restaurant has been observed twice. Put differently, the previous regression uses the data effectively as pooled cross sections. A better approach is running a fixed effect DID regression, FEDID with long form data

$$y_{it} = \beta_0 + \beta_1 D_{treatment,post} + \alpha_i + \gamma_t + u_{it} \quad \text{(FEDID)} \tag{3}$$

where

(a) $D_{treatment,post} = 1$ for treated units in post-treatment periods

(b) α_i is a set of dummies for all units except the first one, called unit fixed effect[1]

(c) γ_t is a set of dummies for all except the first period, called time fixed effect

(d) The DID estimate is β_1.

Actually we do not need to create those fixed effect dummies by ourselves. The factor function can be used as a shortcut, see below

```
> summary(lm(fte~inta+factor(time)+factor(id), data = dl))$coef
                    Estimate Std. Error       t value     Pr(>|t|)
(Intercept)     33.296232877   4.490487  7.4148385095 8.224194e-13
inta             2.744115918   1.163001  2.3595126255 1.881213e-02
factor(time)2   -2.092465753   1.044017 -2.0042450489 4.576439e-02
factor(id)2    -19.625000000   6.307452 -3.1113988999 2.005067e-03
factor(id)3    -22.750000000   6.307452 -3.6068445846 3.520073e-04
...
factor(id)397 -11.372057959   6.334201 -1.7953422592 7.340638e-02
```

Notice that the DID estimate 2.744116 and its t value 2.359513 duplicate old results from the wide-form regression.

8. The main benefit of running FEDID regression (3) is to control for unobserved factors that are time-invariant for each unit (such as the location of a restaurant), and factors that are unit-invariant for each period (such as the national trend of using automation to replace labor).

9. Another way to describe the benefit of FEDID regression is that it uses a restaurant as its own control group—a restaurant after the treatment is compared to itself before the treatment. FEDID ensures apple-to-apple comparison.

[1]Dummy for the first unit is excluded to avoid dummy variable trap

10. Alternatively we can use plm package to fit the FEDID regression:

```
> library(plm)
> dl_panel = pdata.frame(dl, index = c("id", "time"))
> summary(plm(fte~inta+factor(time), data = dl_panel, model = "within"))$coef
               Estimate Std. Error   t-value   Pr(>|t|)
inta           2.744116   1.163001  2.359513 0.01881213
factor(time)2 -2.092466   1.044017 -2.004245 0.04576439
>
> # FD estimator
> summary(plm(fte~inta, data = dl_panel, model = "fd"))$coef
             Estimate Std. Error   t-value   Pr(>|t|)
(Intercept) -2.092466   1.044017 -2.004245 0.04576439
inta         2.744116   1.163001  2.359513 0.01881213
```

There is no need to include factor(id) thanks to the "within" option[2]. With "fd" option there is no need to include factor(id) and factor(time)

11. Finally, we can run a robustness check by using the high-end restaurant (with pre-treatment wage>= 5.00) in NJ as control group, provided that the change in minimum wage has no effect on the high-end restaurant. The new DID estimate is 3.728571, see codes below

```
> mean(d$dfte[d$state==1&d$wage1==4.25],na.rm=T)
[1] 1.620879
> mean(d$dfte[d$state==1&d$wage1>=5],na.rm=T)
[1] -2.107692
> mean(d$dfte[d$state==1&d$wage1==4.25],na.rm=T)-mean(d$dfte[d$state==1&d$wage1>=5]
[1] 3.728571
> d$treated = as.integer(d$state==1&d$wage1==4.25)
> summary(lm(dfte~treated,data=d,subset=(state==1&(wage1==4.25|wage1>=5))))$coef
             Estimate Std. Error   t value    Pr(>|t|)
(Intercept) -2.107692   0.990630 -2.127628 0.034960196
treated      3.728571   1.297039  2.874679 0.004616899
```

[2]With "within" option, there is no need to include any regressors that are time-invariant such as id variable and its corresponding dummies

FEDID and Toy Data

12. It is not obvious how to interpret coefficients in FEDID regression (3). So here I design a toy data, and show how to make sense of FEDID via discussion. The toy data look like

```
> toy = dl[1:8,c("fte","id","time")]
> toy$state = c(rep(0,4),rep(1,4))
> toy$post = as.integer(toy$time==2)
> toy$inta = toy$post*toy$state
> toy
      fte id time state post inta
1.1 40.50  1    1     0    0    0
1.2 24.00  1    2     0    1    0
2.1 13.75  2    1     0    0    0
2.2 11.50  2    2     0    1    0
3.1  8.50  3    1     1    0    0
3.2 10.50  3    2     1    1    1
4.1 34.00  4    1     1    0    0
4.2 20.00  4    2     1    1    1
```

There are four units $i = 1, 2, 3, 4$, and two periods $t = 1, 2$. The FEDID is

$$y_{it} = \beta_0 + \beta_1 inta + \alpha_2 D_{i=2} + \alpha_3 D_{i=3} + \alpha_4 D_{i=4} + \gamma D_{t=2} + u_{it} \tag{4}$$

where $D_{i=2} = 1$ when $i = 2$, $D_{t=2} = 1$ when $t = 2$, and so on. Let us discuss, and simplify FEDID for each unit in each period:

$$Discuss \begin{cases} i = 1, t = 1: & y_{11} = \beta_0 + u_{11}, \\ i = 1, t = 2: & y_{12} = \beta_0 + \gamma + u_{12}, \\ i = 2, t = 1: & y_{21} = \beta_0 + \alpha_2 + u_{21}, \\ i = 2, t = 2: & y_{22} = \beta_0 + \alpha_2 + \gamma + u_{22}, \\ i = 3, t = 1: & y_{31} = \beta_0 + \alpha_3 + u_{31}, \\ i = 3, t = 2: & y_{32} = \beta_0 + \beta_1 + \alpha_3 + \gamma + u_{32}, \\ i = 4, t = 1: & y_{41} = \beta_0 + \alpha_4 + u_{41}, \\ i = 4, t = 2: & y_{42} = \beta_0 + \beta_1 + \alpha_4 + \gamma + u_{42} \end{cases} \tag{5}$$

13. Thus, <u>in the absence of error term</u>, interpretation of each coefficient in FEDID is

$$\begin{cases} \beta_0 = y_{11} = \texttt{base group} \\ \alpha_2 = y_{21} - y_{11} \\ \alpha_3 = y_{31} - y_{11} \\ \alpha_4 = y_{41} - y_{11} \\ \gamma = y_{12} - y_{11} = y_{22} - y_{21} = \texttt{ca} - \texttt{cb} \\ \beta_1 = (y_{32} - y_{31}) - \gamma = (y_{42} - y_{41}) - \gamma = (\texttt{ta} - \texttt{tb}) - (\texttt{ca} - \texttt{cb}) \end{cases} \tag{6}$$

Now I can explain why unit fixed effects and time fixed effects are necessary

(a) If we drop unit fixed effects, we effectively assume $\alpha_2 = \alpha_3 = \alpha_4 = 0$, or $y_{21} = y_{11}$, $y_{31} = y_{11}$, and $y_{41} = y_{11}$. That is, each unit is assumed to be the same in the first period, if unit fixed effects are excluded in the regression.

(b) If we drop time fixed effects, we effectively assume $\gamma = 0$, or $y_{12} = y_{11}$ and $y_{22} = y_{21}$. In words, we assume no change over time for the control group, $\texttt{ca} = \texttt{cb}$.

(c) If we drop the interaction term inta, we effectively assume $\beta_1 = 0$, or $(\texttt{ta} - \texttt{tb}) = (\texttt{ca} - \texttt{cb})$. So the DID effect is zero when inta is absent in the regression

14. Next we take error term into account, by incorporating residual, the estimate of error

term, in the above formula. For instance,

$$\hat{\beta}_0 = y_{11} - \hat{u}_{11}$$

$$\hat{\alpha}_2 = y_{21} - y_{11} - (\hat{u}_{21} - \hat{u}_{11})$$

Below are the results of FEDID applied to toy data, and codes to duplicate those coefficients based on formula (6)

```
> summary(lm(fte~inta+post+factor(id), data = toy))$coef
            Estimate Std. Error    t value   Pr(>|t|)
(Intercept)  36.9375   6.560267  5.6304869 0.03012526
inta          3.3750  10.712872  0.3150416 0.78256189
post         -9.3750   7.575144 -1.2376002 0.34144618
factor(id)2 -19.6250   7.575144 -2.5907097 0.12226103
factor(id)3 -24.4375   9.277619 -2.6340270 0.11895587
factor(id)4  -6.9375   9.277619 -0.7477673 0.53256839
>
> uhat = resid(lm(fte~inta+post+factor(id), data = toy))
> b = coef(lm(fte~inta+post+factor(id), data = toy))
> toy$fte[toy$id==1&toy$time==1]-uhat[1]
    1.1
36.9375
> toy$fte[toy$id==2&toy$time==1]-b[1]-uhat[3]
(Intercept)
    -19.625
> gamma = (toy$fte[toy$id==1&toy$time==2]-toy$fte[toy$id==1&toy$time==1])-(uhat[2]-
> gamma
   1.2
-9.375
> did = (toy$fte[toy$id==3&toy$time==2]-toy$fte[toy$id==3&toy$time==1])-(uhat[6]-uh
> did
  3.2
3.375
```

Dynamic FEDID and Toy Data 2

15. It is straightforward to specify a dynamic FEDID regression to allow for multiple periods before and after the treatment

16. Consider toy data 2, which contains four units, two pre-treatment periods, and two post-treatment periods. The data look like

```
> toy = data.frame(dl[1:16,c("fte")])
> toy[is.na(toy)]=0
> toy$id = rep(1:4, each = 4)
> toy$time = rep(1:4, times = 4)
> toy$state = c(rep(0,8),rep(1,8))
> toy$pre1 = as.integer(toy$time==2)
> toy$post1 = as.integer(toy$time==3)
> toy$post2 = as.integer(toy$time==4)
> toy$ipr1 = toy$pre1*toy$state
> toy$ipost1 = toy$post1*toy$state
> toy$ipost2 = toy$post2*toy$state
> names(toy)[1] = "fte"
> toy
     fte id time state pre1 post1 post2 ipr1 ipost1 ipost2
1  40.50  1    1     0    0     0     0    0      0      0
2  24.00  1    2     0    1     0     0    0      0      0
3  13.75  1    3     0    0     1     0    0      0      0
4  11.50  1    4     0    0     0     1    0      0      0
5   8.50  2    1     0    0     0     0    0      0      0
6  10.50  2    2     0    1     0     0    0      0      0
7  34.00  2    3     0    0     1     0    0      0      0
8  20.00  2    4     0    0     0     1    0      0      0
9  24.00  3    1     1    0     0     0    0      0      0
10 35.50  3    2     1    1     0     0    1      0      0
11 20.50  3    3     1    0     1     0    0      1      0
12  0.00  3    4     1    0     0     1    0      0      1
13 70.50  4    1     1    0     0     0    0      0      0
14 29.00  4    2     1    1     0     0    1      0      0
```

```
15 23.50  4    3     1    0     1     0    0     1     0
16 36.50  4    4     1    0     0     1    0     0     1
```

17. For this toy data, the dynamic FEDID regression is

$$y_{it} = \beta_0 + \beta_1 pr1 + \beta_2 ipost1 + \beta_3 ipost2 + \sum_{i=2}^{4} \alpha_i + \sum_{t=2}^{4} \gamma_t + u_{it} \quad (\texttt{dynamic FEDID}) \quad (7)$$

Let us discuss, and simplify dynamic FEDID for each unit in each period (assuming no error term):

$$Discuss \begin{cases} i=1, t=1: & y_{11} = \beta_0, \\ i=1, t=2: & y_{12} = \beta_0 + \gamma_2, \\ i=1, t=3: & y_{13} = \beta_0 + \gamma_3, \\ i=1, t=4: & y_{14} = \beta_0 + \gamma_4, \\ i=2, t=1: & y_{21} = \beta_0 + \alpha_2, \\ i=2, t=2: & y_{22} = \beta_0 + \alpha_2 + \gamma_2, \\ i=2, t=3: & y_{23} = \beta_0 + \alpha_2 + \gamma_3, \\ i=2, t=4: & y_{24} = \beta_0 + \alpha_2 + \gamma_4, \\ i=3, t=1: & y_{31} = \beta_0 + \alpha_3, \\ i=3, t=2: & y_{32} = \beta_0 + \beta_1 + \alpha_3 + \gamma_2, \\ i=3, t=3: & y_{33} = \beta_0 + \beta_2 + \alpha_3 + \gamma_3, \\ i=3, t=4: & y_{34} = \beta_0 + \beta_3 + \alpha_3 + \gamma_4, \\ i=4, t=1: & y_{41} = \beta_0 + \alpha_4, \\ i=4, t=2: & y_{42} = \beta_0 + \beta_1 + \alpha_4 + \gamma_2, \\ i=4, t=3: & y_{43} = \beta_0 + \beta_2 + \alpha_4 + \gamma_3, \\ i=4, t=4: & y_{44} = \beta_0 + \beta_3 + \alpha_4 + \gamma_4, \end{cases} \quad (8)$$

The interpretations of coefficients are

$$\begin{cases} \beta_0 = y_{11} \\ \alpha_2 = y_{21} - y_{11} \\ \alpha_3 = y_{31} - y_{11} \\ \alpha_4 = y_{41} - y_{11} \\ \gamma_2 = y_{12} - y_{11} = y_{22} - y_{21} \\ \gamma_3 = y_{13} - y_{11} = y_{23} - y_{21} \\ \gamma_4 = y_{14} - y_{11} = y_{24} - y_{21} \\ \beta_1 = (y_{32} - y_{31}) - \gamma_2 = (y_{42} - y_{41}) - \gamma_2 = \texttt{pre1 DID} \\ \beta_2 = (y_{33} - y_{31}) - \gamma_3 = (y_{43} - y_{41}) - \gamma_3 = \texttt{post1 DID} \\ \beta_3 = (y_{34} - y_{31}) - \gamma_4 = (y_{44} - y_{41}) - \gamma_4 = \texttt{post2 DID} \end{cases} \tag{9}$$

(a) Most importantly, each β is a DID. For instance, we can write

$$\beta_2 = (y_{33} - y_{31}) - \gamma_3 = (y_{33} - y_{31}) - (y_{13} - y_{11}) = (y_{33} - y_{13}) - (y_{31} - y_{11}) = DID$$

(b) The unit fixed effects are $\sum_{i=2}^{4} \alpha_i$, which capture difference between the first unit and other units in the first period.

(c) The time fixed effects are $\sum_{t=2}^{4} \gamma_t$, which capture temporal change for control group

(d) β_1 is the DID at $t = 2$, or one period before the treatment. Testing Parallel Trend Assumption (PTA) amounts to testing $H_0 : \beta_1 = 0$

(e) β_2 is the DID at $t = 3$, or one period after the treatment

(f) β_3 is the DID at $t = 4$, or two period after the treatment

(g) The model is dynamic in the sense that it allows $\beta_2 \neq \beta_3$

(h) Many researchers report $\beta_1, \beta_2, \beta_3$ with coefficient plot [3]. Usually they hope to find 0 is inside the confidence interval of β_1, but outside the confidence intervals of β_2, β_3.

[3] See `http://www.fsb.miamioh.edu/lij14/311r_or_coefplot.pdf` for a brief introduction

18. For the toy data, the estimated dynamic FEDID is

```
> DFEDID = lm(fte~ipr1+ipost1+ipost2+factor(time)+factor(id), data = toy)
> summary(DFEDID)$coef
               Estimate Std. Error    t value   Pr(>|t|)
(Intercept)    26.59375   13.66281  1.9464335 0.09954998
ipr1           -7.75000   24.44078 -0.3170930 0.76191665
ipost1        -24.62500   24.44078 -1.0075375 0.35256646
ipost2        -20.25000   24.44078 -0.8285334 0.43907629
factor(time)2  -7.25000   17.28224 -0.4195058 0.68945217
factor(time)3  -0.62500   17.28224 -0.0361643 0.97232452
factor(time)4  -8.75000   17.28224 -0.5063001 0.63070692
factor(id)2    -4.18750   12.22039 -0.3426651 0.74353612
factor(id)3    10.71875   19.32213  0.5547396 0.59912667
factor(id)4    30.59375   19.32213  1.5833528 0.16443025
```

19. We can run some version of restricted models.

    (a) Example A: suppose there are only two units, say, we have y_{1t} and y_{3t}. Then we replace $\sum_{i=2}^{4} \alpha_i$ with α_3. The new interpretation can be obtained from (9) after dropping any term associated with y_{2t} and y_{4t}

    (b) Example B: suppose we use state fixed effect other than unit fixed effect. This is essentially same as Example A

    (c) Example C: suppose we impose restriction that $\beta_2 = \beta_3$. Then we need to run

$$y_{it} = \beta_0 + \beta_1 pr1 + \beta_2 (ipost1 + ipost2) + \sum_{i=2}^{4} \alpha_i + \sum_{t=2}^{4} \gamma_t + u_{it} \quad (\texttt{restricted dynamic FEDID})$$

    (d) Example D: How about restriction of $\beta_1 = 0$? Then the restricted regression is

$$y_{it} = \beta_0 + \beta_2 ipost1 + \beta_3 ipost2 + \sum_{i=2}^{4} \alpha_i + \sum_{t=2}^{4} \gamma_t + u_{it} \quad (\texttt{restricted dynamic FEDID})$$

20. The Dynamic FEDID regression can be modified to account for staggered treatment, see `http://www.fsb.miamioh.edu/lij14/311r_or_eventstudy.pdf` for a discussion.