Applied Statistics and Econometrics

Giuseppe Ragusa
Luiss University

Lecture 14: Panel Data II

The Panel Data regression problem

$$Y_{it} = \underbrace{\beta_0 + \beta_1 X_{1it} + \ldots + \beta_k X_{kit}}_{\text{observables}} + \underbrace{\delta Z_i}_{\text{unobservable}} + u_{it}$$

- Z_i is a factor that does not change over time (density), at least during the years on which we have data.
- Z_i is **not observed**, so its omission could result in **omitted variable bias**.
- Then, we can use the special panel data structure to *eliminate* Z_i

Estimation methods seen so far...

Three estimation methods so far:

1. **If $T=2$, two time periods, T_1 and T_2, run OLS on time demeaned variables**

$$Y_{tT_2} - Y_{tT_1} = \beta_1(X_{1iT_2} - X_{1iT_1}) + \ldots + \beta_k(X_{kiT_2} - X_{kiT_1}) + (u_{iT_2 -} u_{iT_1})$$

2. **OLS adding $n-1$-dummy variables**

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \ldots + \beta_k X_{kit} + \gamma_2 D_{2i} + \ldots + \gamma_n D_{ni} + u_{it}$$

3. **Entity demeaning**

$$\tilde{Y}_{it} = \beta_1 \tilde{X}_{1it} + \ldots + \beta_k \tilde{X}_{kit} + \tilde{u}_{it}$$

where

$$\tilde{Y}_{it} = Y_{it} - \frac{1}{T}\sum_{t=1}^{T} Y_{it},\ \tilde{X}_{it} = X_{it} - \frac{1}{T}\sum_{t=1}^{T} X_{it},\ \tilde{u}_{it} = u_{it} - \frac{1}{T}\sum_{t=1}^{T} u_{it}$$

Equivalences

- If $T = 2$, **all three methods** conduce to the **same estimate** of the parameters of interest $(\beta_1, \ldots, \beta_k)$
- If $T > 2$, $\mathbf{n-1}$ **dummy variables OLS** and entity demeaned OLS are **equivalent**

Regression with Time Fixed Effects

An omitted variable might vary over time but not across states:

- Safer cars (air bags, etc.); changes in national laws
- These produce intercepts that change over time
- Let S_t denote the combined effect of variables which changes over time but not states ("safer cars").

The resulting population regression model is:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + \beta_3 S_t + u_{it}$$

Time effects only

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_3 S_t + u_{it}$$

This model can be recast as having an intercept that varies from one year to the next:

$$\begin{aligned} Y_{i,1998} &= \beta_0 + \beta_1 X_{i,1982} + \beta_3 S_{1982} + u_{i,1982} \\ &= (\beta_0 + \beta_3 S_{1982}) + \beta_1 X_{i,1982} + u_{i,1982} \\ &= \lambda_{1982} + \beta_1 X_{i,1982} + u_{i,1982}, \end{aligned}$$

where $\lambda_{1982} = (\beta_0 + \beta_3 S_{1982})$. Similarly,

$$Y_{i,1983} = \lambda_{1983} + \beta_1 X_{i,1983} + u_{i,1983},$$

where $\lambda_{1983} = (\beta_0 + \beta_3 S_{1983})$.

Two formulations of regression with time fixed effects

1. "T-1 binary regressor" formulation:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_2 B_{2t} + \dots \delta_T B_{Tt} + u_{it}$$

   where

$$B_{2t} = \begin{cases} 1 & \text{when t=2} \\ 0 & \text{otherwise,} \end{cases} \quad B_{3t} = \begin{cases} 1 & \text{when t=3} \\ 0 & \text{otherwise,} \end{cases}, \dots, B_{Tt} = \begin{cases} 1 & \text{when t=T} \\ 0 & \text{otherwise,} \end{cases}$$

2. "Time effects" formulation:

$$Y_{it} = \beta_1 X_{it} + \lambda_t + u_{it}$$

Time fixed effects: estimation methods

1. "T-1 binary regressor" OLS regression

$$Yit = \beta_0 + \beta_1 X_{it} + \delta_2 B_{2it} + \ldots \delta_T B_{Tit} + u_{it}$$

   - Create binary variables $B_2, \ldots, B_T$
   - $B_2 = 1$ if $t = year$ $\#2$, $= 0$ otherwise
   - Regress Y on $X, B_2, \ldots, B_T$ using OLS
   - Where's B_1?

2. "Year-demeaned" OLS regression
   - Deviate Y_{it}, X_{it} from year (not state) averages
   - Estimate by OLS using "year-demeaned" data

Estimation with both entity and time fixed effects

$$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

- When $T = 2$, computing the first difference and including an intercept is equivalent to (gives exactly the same regression as) including entity and time fixed effects.
- When $T > 2$, there are various equivalent ways to incorporate both entity and time fixed effects:
  - entity demeaning & T–1 time indicators
  - time demeaning & n–1 entity indicators
  - T–1 time indicators & n–1 entity indicators
  - entity & time demeaning

Fatality Rate and Beer Tax

```
## State dummies
lm1 <- lm(fatalityrate ~ beertax + state, data = fatalities)
summary_rob(lm1, omit_factor = TRUE)
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.478      0.351    9.91   <2e-16
## beertax       -0.656      0.203   -3.23   0.0013
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 0.19 on 287 degrees of freedom
## Multiple R-squared:  0.905, Adjusted R-squared:  0.889
## F-statistic: 6.06e+03 on 48 and Inf DF,  p-value: <2e-16
## ---
## Factors not reported: state
```

Fatality Rate and Beer Tax

```r
## Year dummies
lm2 <- lm(fatalityrate ~ beertax + year, data = fatalities)
summary_rob(lm2, omit_factor = TRUE)
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8948     0.1060   17.87  < 2e-16
## beertax       0.3663     0.0533    6.87  6.5e-12
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 0.55 on 328 degrees of freedom
## Multiple R-squared:  0.0986, Adjusted R-squared:  0.0794
## F-statistic: 49.4 on 7 and Inf DF,  p-value: <2e-16
## ---
## Factors not reported: year
```

Fatality Rate and Beer Tax

```r
## Both set of dummies dummies
lm3 <- lm(fatalityrate ~ beertax + state + year, data = fatalities)
summary_rob(lm3, omit_factor = TRUE)
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)    3.511      0.447    7.85  4.2e-15
## beertax       -0.640      0.255   -2.51    0.012
## ---
## Heteroskadasticity robust standard errors used
##
## Residual standard error: 0.19 on 281 degrees of freedom
## Multiple R-squared:  0.909, Adjusted R-squared:  0.891
## F-statistic: 6.96e+03 on 54 and Inf DF,  p-value: <2e-16
## ---
## Factors not reported: state year
```

Fatality Rate and Beer Tax

```
plm1 <- plm(fatalityrate ~ beertax + year, data = fatalities, model = "within")
summary_rob(plm1)


## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = fatalityrate ~ beertax + year, data = fatalities,
##     model = "within")
##
## Balanced Panel: n=48, T=7, N=336
##
## Coefficients :
##          Estimate Std. Error t-value Pr(>|t|)
## beertax   -0.6400     0.2354   -2.72   0.0070 **
## year1983  -0.0799     0.0465   -1.72   0.0866 .
## year1984  -0.0724     0.0418   -1.73   0.0844 .
## year1985  -0.1240     0.0425   -2.92   0.0038 **
## year1986  -0.0379     0.0450   -0.84   0.4004
## year1987  -0.0509     0.0477   -1.07   0.2868
## year1988  -0.0518     0.0498   -1.04   0.2990
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Adj. R-Squared :  0.067
```

Fatality Rate and Beer Tax

```r
## Both set of dummies dummies
plm2 <- plm(fatalityrate ~ beertax + state, data = fatalities, model = "within",
    effect = "time")
summary(plm2)
```

```
## Oneway (time) effect Within Model
##
## Call:
## plm(formula = fatalityrate ~ beertax + state, data = fatalities,
##     effect = "time", model = "within")
##
## Balanced Panel: n=48, T=7, N=336
##
## Residuals :
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -0.59600 -0.08100  0.00143  0.08230  0.83900
##
## Coefficients :
##          Estimate Std. Error t-value Pr(>|t|)
## beertax  -0.6400     0.1974   -3.24  0.00133 **
## stateaz  -0.5469     0.2779   -1.97  0.05006 .
## statear  -0.6385     0.2273   -2.81  0.00532 **
## stateca  -1.4852     0.3178   -4.67  4.6e-06 ***
## stateco  -1.4615     0.2998   -4.88  1.8e-06 ***
## statect  -1.8401     0.2926   -6.29  1.2e-09 ***
## statede  -1.2843     0.3068   -4.19  3.8e-05 ***
## statefl  -0.2601     0.1419   -1.83  0.06799 .
## statega   0.5116     0.1899    2.69  0.00749 **
## stateid  -0.6490     0.2687   -2.42  0.01635 *
## stateil  -1.9385     0.3042   -6.37  7.6e-10 ***
## statein  -1.4401     0.2841   -5.07  7.3e-07 ***
```

Fatality Rate and Beer Tax

```
## Both set of dummies dummies
plm3 <- plm(fatalityrate ~ beertax, data = fatalities, model = "within",
    effect = "twoway")
summary_rob(plm3)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = fatalityrate ~ beertax, data = fatalities, effect = "twoway",
##     model = "within")
##
## Balanced Panel: n=48, T=7, N=336
##
## Coefficients :
##         Estimate Std. Error t-value Pr(>|t|)
## beertax   -0.640      0.233   -2.74   0.0065 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Adj. R-Squared :  0.03
## F-statistic: 10.5133 on 1 and 281 DF, p-value: 0.00133
```

The Fixed Effects Regression Assumptions and Standard Errors for Fixed Effects Regression

- Under a panel data version of the least squares assumptions, the OLS fixed effects estimator of β is normally distributed.
- However, a new standard error formula needs to be introduced: the "clustered" standard error formula.
- This new formula is needed because observations for the same entity are not independent (it's the same entity!), even though observations across entities are independent if entities are drawn by simple random sampling.
- Here we consider the case of entity fixed effects. Time fixed effects can simply be included as additional binary regressors.

LS Assumptions for Panel Data

Consider a single X:

$$Y_{it} = \beta_1 X_{it} + \alpha_i + u_{it},\ \ i = 1, \ldots, n,\ \ t = 1, \ldots, T$$

- $E(u_{it}|X_{i1}, \ldots, X_{iT}, \alpha_i) = 0, \quad t = 1, \ldots, T$
- $(X_{i1}, \ldots, X_{iT}, u_{i1}, \ldots, u_{iT})$, $i = 1, \ldots, n$, are i.i.d. draws from their joint distribution.
- (X_{it}, u_{it}) have finite fourth moments.
- There is no perfect multicollinearity (multiple X's)

Assumptions 3&4 are same as least squares assumptions. **Assumptions 1&2 differ**

Assumption #1: $E(u_{it}|X_{i1},\ldots,X_{iT},\alpha_i) = 0, \quad t = 1,\ldots,T$

- u_{it} has mean zero, given the entity fixed effect and the **entire history** of the X's for that entity
- This is an extension of the previous multiple regression Assumption #1
- This means there are no omitted lagged effects (any lagged effects of X must enter explicitly)
- Also, there is not feedback from u to future X:
  - Whether a state has a particularly high fatality rate this year doesn't subsequently affect whether it increases the beer tax.
  - Sometimes this "no feedback" assumption is plausible, sometimes it isn't.

Assumption #2: $(X_{i1}, \ldots, X_{iT}, u_{i1}, \ldots, u_{iT})$, $i = 1, \ldots, n$, are i.i.d. draws from their joint distribution.

- This is an extension of Assumption #2 for multiple regression with cross-section data
- This is satisfied if entities are randomly sampled from their population by simple random sampling.
- This does not require observations to be i.i.d. over time for the same entity – that would be unrealistic. Whether a state has a high beer tax this year is a good predictor of (correlated with) whether it will have a high beer tax next year. Similarly, the error term for an entity in one year is plausibly correlated with its value in the year, that is, $corr(u_{it}, u_{it+1})$ is often plausibly nonzero.

Autocorrelation (serial correlation)

Suppose a variable Z is observed at different dates t, so observations are on Z_t, $t = 1, \ldots, T$. (Think of there being only one entity.) Then Z_t is said to be autocorrelated or serially correlated if

$$corr(Z_t, Z_{t+j}) \neq 0$$

for some dates $j \neq 0$.

- "Autocorrelation" means correlation with itself.
- $cov(Z_t, Z_{t+j})$is called the jth autocovariance of Z_t.
- In the drunk driving example, u_{it} includes the omitted variable of annual weather conditions for state i. If snowy winters come in clusters (one follows another) then u_{it} will be autocorrelated (why?)
- In many panel data applications, u_{it} is plausibly autocorrelated.

Under the LS assumptions for panel data:

- The OLS fixed effect estimator is unbiased, consistent, and asymptotically normally distributed
- However, the usual OLS standard errors (both homoskedasticity-only and heteroskedasticity-robust) will in general be wrong because they assume that u_{it} is serially uncorrelated.
  - In practice, the OLS standard errors often understate the true sampling uncertainty: if u_{it} is correlated over time, you don't have as much information (as much random variation) as you would if u_{it} were uncorrelated.
  - This problem is solved by using "clustered" standard errors.

Clustered standard errors

- If plm is used then you can “simply” do this

```
lm4 <- plm(fatalityrate ~ beertax, data = fatalities, model = "within",
    effect = "twoway")
summary_rob(lm4, cluster = TRUE)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = fatalityrate ~ beertax, data = fatalities, effect = "twoway",
##     model = "within")
##
## Balanced Panel: n=48, T=7, N=336
##
## Coefficients :
##         Estimate Std. Error t-value Pr(>|t|)
## beertax    -0.64       0.35   -1.83    0.069 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```