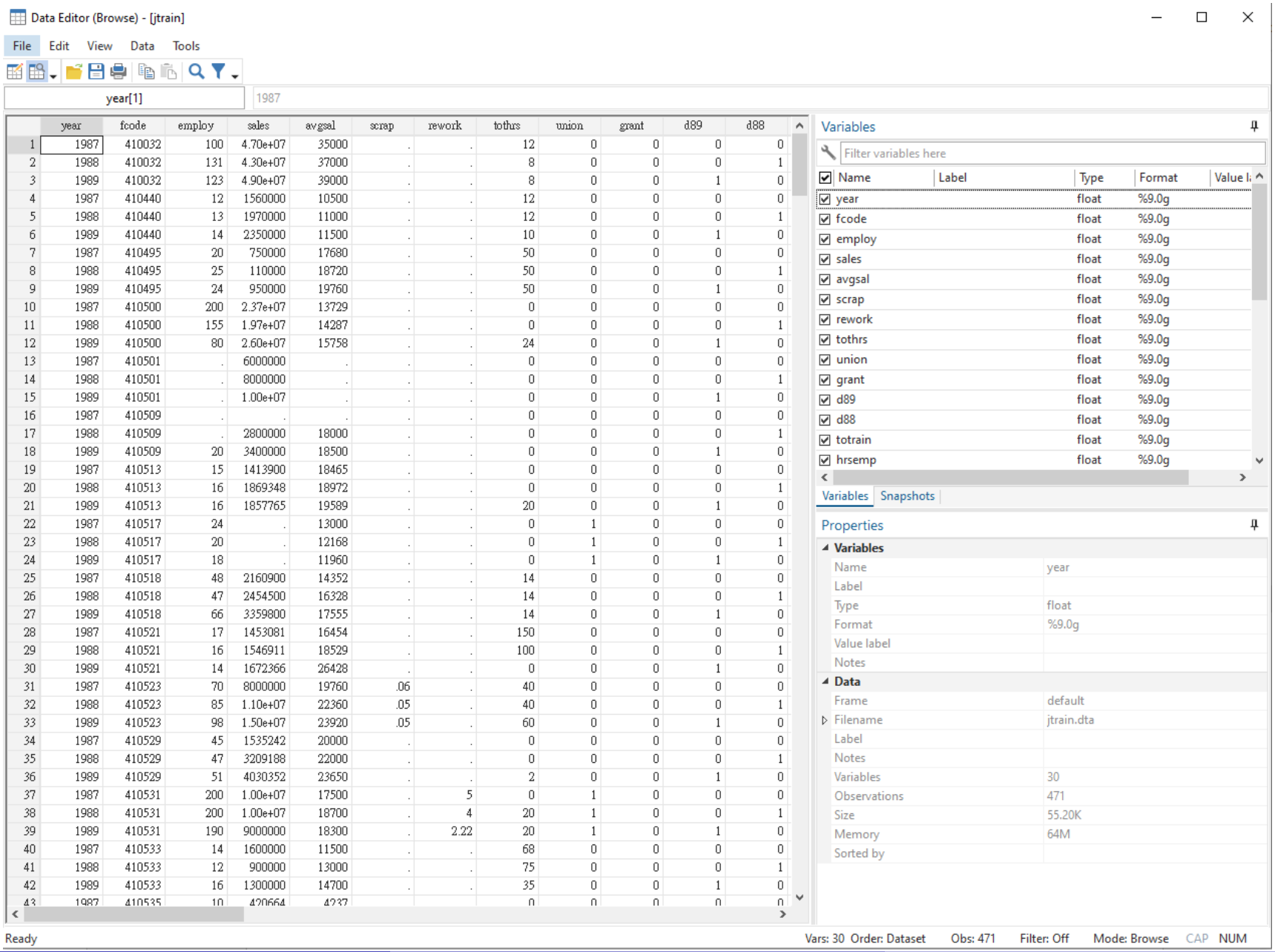
Lecture 5: Introduction to Panel Data Models

Yu Bai

City University of Macau

How does a panel data set look like?



	year	fcode	employ	sales	avgsal	scrap	rework	tothrs	union	grant	d89	d88
1	1987	410032	100	4.70e+07	35000	.	.	12	0	0	0	0
2	1988	410032	131	4.30e+07	37000	.	.	8	0	0	0	1
3	1989	410032	123	4.90e+07	39000	.	.	8	0	0	1	0
4	1987	410440	12	1560000	10500	.	.	12	0	0	0	0
5	1988	410440	13	1970000	11000	.	.	12	0	0	0	1
6	1989	410440	14	2350000	11500	.	.	10	0	0	1	0
7	1987	410495	20	750000	17680	.	.	50	0	0	0	0
8	1988	410495	25	110000	18720	.	.	50	0	0	0	1
9	1989	410495	24	950000	19760	.	.	50	0	0	1	0
10	1987	410500	200	2.37e+07	13729	.	.	0	0	0	0	0
11	1988	410500	155	1.97e+07	14287	.	.	0	0	0	0	1
12	1989	410500	80	2.60e+07	15758	.	.	24	0	0	1	0
13	1987	410501	.	6000000	.	.	.	0	0	0	0	0
14	1988	410501	.	8000000	.	.	.	0	0	0	0	1
15	1989	410501	.	1.00e+07	.	.	.	0	0	0	1	0
16	1987	410509	.	.	.	.	.	0	0	0	0	0
17	1988	410509	.	2800000	18000	.	.	0	0	0	0	1
18	1989	410509	20	3400000	18500	.	.	0	0	0	1	0
19	1987	410513	15	1413900	18465	.	.	0	0	0	0	0
20	1988	410513	16	1869348	18972	.	.	0	0	0	0	1
21	1989	410513	16	1857765	19589	.	.	20	0	0	1	0
22	1987	410517	24	.	13000	.	.	0	1	0	0	0
23	1988	410517	20	.	12168	.	.	0	1	0	0	1
24	1989	410517	18	.	11960	.	.	0	1	0	1	0
25	1987	410518	48	2160900	14352	.	.	14	0	0	0	0
26	1988	410518	47	2454500	16328	.	.	14	0	0	0	1
27	1989	410518	66	3359800	17555	.	.	14	0	0	1	0
28	1987	410521	17	1453081	16454	.	.	150	0	0	0	0
29	1988	410521	16	1546911	18529	.	.	100	0	0	0	1
30	1989	410521	14	1672366	26428	.	.	0	0	0	1	0
31	1987	410523	70	8000000	19760	.06	.	40	0	0	0	0
32	1988	410523	85	1.10e+07	22360	.05	.	40	0	0	0	1
33	1989	410523	98	1.50e+07	23920	.05	.	60	0	0	1	0
34	1987	410529	45	1535242	20000	.	.	0	0	0	0	0
35	1988	410529	47	3209188	22000	.	.	0	0	0	0	1
36	1989	410529	51	4030352	23650	.	.	2	0	0	1	0
37	1987	410531	200	1.00e+07	17500	.	5	0	1	0	0	0
38	1988	410531	200	1.00e+07	18700	.	4	20	1	0	0	1
39	1989	410531	190	9000000	18300	.	2.22	20	1	0	1	0
40	1987	410533	14	1600000	11500	.	.	68	0	0	0	0
41	1988	410533	12	900000	13000	.	.	75	0	0	0	1
42	1989	410533	16	1300000	14700	.	.	35	0	0	1	0
43	1987	410535	10	420664	4237			0	0	0	0	0

Pooling Independent Cross Sections across Time

- Many surveys of individuals, families, and firms are repeated at regular intervals, often each year.
- One reason for using independently pooled cross sections is to increase the sample size.
- Typically, to reflect the fact that the population may have different distributions in different time periods, we allow the intercept to differ across periods, which can be easily accomplished by including dummy variables.

Return to Education and the Gender Wage Gap

- Suppose you have data collected from the year 1978 and estimate the model:

$$\log(\text{wage}_i) = \beta_0 + \beta_1 \text{female}_i + \boldsymbol{\gamma}' \mathbf{Z}_i + u_i.$$

- Now you have data from the year 1985. How would you specify the model when data from both years are available?
- Easy! Just add dummies:

$$\log(\text{wage}_{it}) = \beta_0 + \delta_0 d_t + \beta_1 \text{female}_i + \delta_1 d_t \text{female}_i + \boldsymbol{\gamma}' \mathbf{Z}_{it} + u_{it}.$$

The Chow test for structural change across time

- Two time periods?
- More than two time periods?
  1. Estimate restricted model (pooled OLS) and obtain SSR_r
  2. Estimate unrestricted model: cross-sectional regression for each period to obtain $\text{SSR}_\text{ur} = \text{SSR}_1 + \text{SSR}_2 + \cdots + \text{SSR}_{n-1} + \text{SSR}_n$
  3. There are $(T-1)k$ restrictions, with total # of $(T+1)k$ parameters estimated.
  4. Construct F-test as usual:

$$F_{\mathcal{H}_0} = \frac{\text{SSR}_r - \text{SSR}_{ur}}{\text{SSR}_{ur}} \frac{n-(T+1)k}{(T-1)k} \overset{d}{\sim} F\left((T-1)k, n-(T+1)k\right),$$

  where $n = n_1 + n_2 + \cdots + n_{T-1} + n_T$.

  5. How to make the test robust to heteroskedasticity?

Example: Effect of a Garbage Incinerator's Location on Housing Prices

- Some background:
  - The rumor that a new incinerator would be built in North Andover began after 1978, and construction began in 1981.
  - The incinerator was expected to be in operation soon after the start of construction, but actually began operating in 1985.
- The hypothesis is that the price of houses located near the incinerator would fall relative to the price of more distant houses.
- Data: prices of houses that sold in 1978 and another sample on those that sold in 1981, in North Andover
- A house is defined to be near the incinerator if it is within three miles.

Example: Effect of a Garbage Incinerator's Location on Housing Prices

- If we just use the 1981 data and estimate the simple model:

$$rprice_i = \gamma_0 + \gamma_1 nearinc_i + u_i,$$

  can $\hat{\gamma}_{1n}$ have any causal interpretation?
- OK if $\mathbb{E}[u_i | nearinc_i] = 0$. rather unlikely...
- Note that

$$\hat{\gamma}_{1n} = \overline{rprice}_{81,nr} - \overline{rprice}_{81,fr}$$

Example: Effect of a Garbage Incinerator's Location on Housing Prices

- What about only consider houses located near the incinerator and do a polled regression?

$$rprice_{it} = \beta_0 + \delta_0 y_{t,81} + u_{it}$$

- Note that

$$\hat{\delta}_{0n} = \overline{rprice}_{81,nr} - \overline{rprice}_{78,nr}$$

- Works if the trend of house prices remains the same in these two years.

Example: Effect of a Garbage Incinerator's Location on Housing Prices

- **DID** (difference-in-differences) estimator:

$$\hat{\delta}_{1n} = \left(\overline{rprice}_{81,nr} - \overline{rprice}_{81,fr}\right) - \left(\overline{rprice}_{78,nr} - \overline{rprice}_{78,fr}\right), \tag{1}$$

  which can be obtained by estimating

$$rprice_{it} = \beta_0 + \delta_0 y_{t,81} + \beta_1 nearinc_i + \delta_1 y_{t,81} \cdot nearinc_i + u_{it}.$$

Example: Effect of a Garbage Incinerator's Location on Housing Prices

```
. reg rprice nearinc y81 y81nrinc
```

Source	SS	df	MS		
				Number of obs	= 321
				F(3, 317)	= 22.25
Model	6.1055e+10	3	2.0352e+10	Prob > F	= 0.0000
Residual	2.8994e+11	317	914632749	R-squared	= 0.1739
				Adj R-squared	= 0.1661
Total	3.5099e+11	320	1.0969e+09	Root MSE	= 30243

rprice	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
nearinc	-18824.37	4875.322	-3.86	0.000	-28416.45	-9232.293
y81	18790.29	4050.065	4.64	0.000	10821.88	26758.69
y81nrinc	-11863.9	7456.646	-1.59	0.113	-26534.67	2806.866
_cons	82517.23	2726.91	30.26	0.000	77152.1	87882.36

Policy analysis with pooled cross sections

- Why DID? It is used when the data arise from a **natural-experiment** (or a **quasi-experiment**).
- In a natural experiment, the **control and treatment groups** arise from the particular policy change.
- The equation of interest is

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2 \cdot dT + other\ factors$$

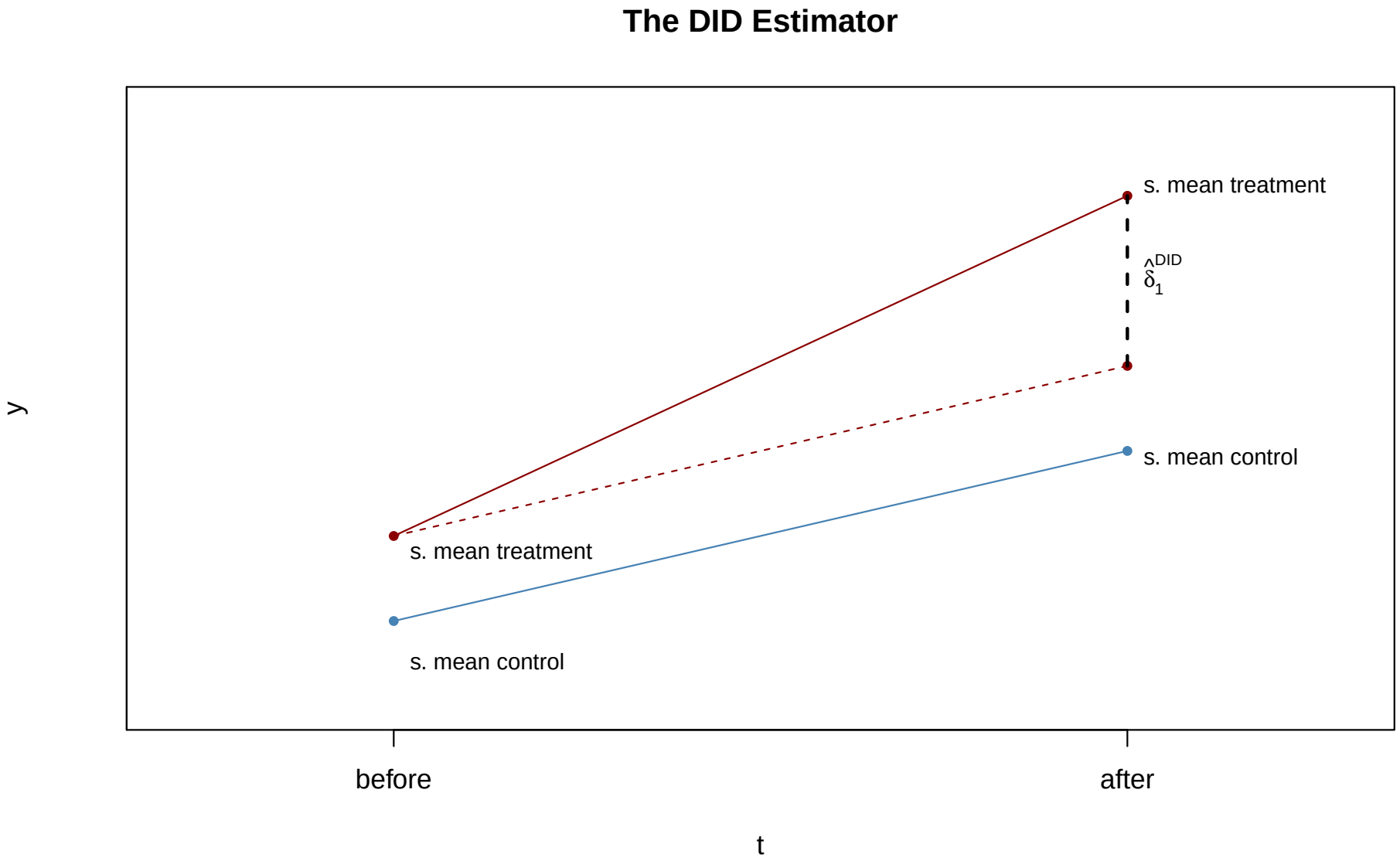
  where
  - $dT = 1$ if in the treatment group
  - $d2 = 1$ if in the (post-policy change) time period

Policy analysis with pooled cross sections

	Before	After	After-Before
Control	β_0	$\beta_0 + \delta_0$	δ_0
Treatment	$\beta_0 + \beta_1$	$\beta_0 + \delta_0 + \beta_1 + \delta_1$	$\delta_0 + \delta_1$
Treatment-Control	β_1	$\beta_1 + \delta_1$	δ_1

- δ_1 is interpreted as an **average treatment effect (ATE)**.

Policy analysis with pooled cross sections



Two-period panel data analysis

- (1) is nice, but the treatment structure in the general setting may not be as simple as it is.
- Needs a formal treatment on panel data models
- Will start with a model with two time periods

Two-period panel data analysis

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}, \tag{2}$$

where

- $d2_t = 1$ if $t = 2$
- x_{it} is a variable of interest
- a_i: fixed effect
- $v_{it} = a_i + u_{it}$: component error

Two-period panel data analysis

- How to estimate β_1?
- Pooled OLS? Hard due to the presence of a_i
- For $t = 2$, we have
$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2}.$$
- For $t = 1$, we have
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1}.$$
- Differencing gives
$$\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$$
- Can do OLS (**first-differenced estimator**), but needs two conditions:
  - Strictly exogeneity
  - Δx_i must have some variation across i

Still remembered this?

Problem 1

Suppose that the equation

$$y_t = \alpha + \delta t + \beta_1 x_{t1} + \cdots + \beta_k x_{tk} + u_t,$$

satisfies the sequential exogeneity assumption.

- Suppose you difference the equation to obtain

$$\Delta y_t = \delta + \beta_1 \Delta x_{t1} + \cdots + \beta_k \Delta x_{tk} + \Delta u_t.$$

  Why does applying OLS on the differenced equation not generally result in consistent estimator of the β_j?

Program evaluation model in general

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

where

- y_{it}: an outcome variable
- $prog_{it}$: a program participation dummy variable
- a_i: fixed effect

$$\Delta y_{it} = \beta_1 \Delta prog_{it} + \Delta u_{it}$$
$$\hat{\beta}_{1n} = \overline{\Delta y}_{treat} - \overline{\Delta y}_{control}$$

- panel version of DID estimator
- allows us to control for person-, firm-, or city-specific effects

Policy analysis with pooled cross sections

- Relies on the **parallel trends assumption**:

parallel trends assumption

The time trend for the treatment group would be the same as control group in the absence of the intervention.

Example: Effect of a Michigan job training program on worker productivity of manufacturing firms

$$scrap_{it} = \beta_0 + \delta_0 y88_t + \beta_1 grant_{it} + a_i + u_{it}, \quad t = 1, 2,$$

where

- $scrap_{it}$: scrap rate of firm i during year t
- $grant_{it}$: 1 if firm i in year t received a job training grant
- $y88$: a dummy variable for 1988
- a_i: firm fixed effect

Example: Effect of a Michigan job training program on worker productivity of manufacturing firms

```
. reg Dscrap Dgrant
```

Source	SS	df	MS	Number of obs	=	54
				F(1, 52)	=	1.17
Model	6.73345587	1	6.73345587	Prob > F	=	0.2837
Residual	298.400031	52	5.73846213	R-squared	=	0.0221
				Adj R-squared	=	0.0033
Total	305.133487	53	5.7572356	Root MSE	=	2.3955

Dscrap	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
Dgrant	-.7394436	.6826276	-1.08	0.284	-2.109236	.6303488
_cons	-.5637143	.4049149	-1.39	0.170	-1.376235	.2488069

Example: Effect of a Michigan job training program on worker productivity of manufacturing firms

```
. reg lDscrap Dgrant
```

Source	SS	df	MS	Number of obs	=	54
				F(1, 52)	=	3.74
Model	1.23795567	1	1.23795567	Prob > F	=	0.0585
Residual	17.1971851	52	.330715099	R-squared	=	0.0672
				Adj R-squared	=	0.0492
Total	18.4351408	53	.347832845	Root MSE	=	.57508

lDscrap	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
Dgrant	-.3170579	.1638751	-1.93	0.058	-.6458974	.0117816
_cons	-.0574357	.097206	-0.59	0.557	-.2524938	.1376224

- Having a job training grant is estimated to lower the scrap rate by about 27.2%: $\exp(-.317) - 1$.

Differencing with more than two time periods

- With $T = 3$, we need to have an additional dummy variable

$$y_{it} = \delta_1 + \delta_2 d2_t + \delta_3 d3_t + \beta_1 x_{it} + a_i + u_{it},$$

  where $d2_t = 1$ if in period 2 and $d3_t = 1$ if in period 3.
- Differencing gives

$$\Delta y_{it} = \delta_2 \Delta d2_t + \delta_3 \Delta d3_t + \beta_1 \Delta x_{it} + \Delta u_{it}$$

- Since we have $\Delta d2_t = 1$, $\Delta d3_t = 0$ for $t = 2$ and $\Delta d2_t = -1$, $\Delta d3_t = 1$ for $t = 3$, we can estimate a model with an intercept

$$\Delta y_{it} = \alpha_0 + \alpha_3 d3_t + \beta_1 \Delta x_{it} + \Delta u_{it}.$$

- Serial correlation in Δu_{it} might be an issue.

Potential Pitfalls in First Differencing Panel Data

- Potential problems with the method when the key explanatory variables do not vary much over time
- May have serious biases if strictly exogeneity assumption fails

Assumption FD.4

For each t, the expected value of the idiosyncratic error given the explanatory variables in *all* time periods and the unobserved effect is zero: $\mathbb{E}(u_{it}|\mathbf{X}_i, a_i) = 0$.

- Can be worse than pooled OLS if explanatory variable contains measurement error

Fixed Effects Estimation

- Consider a model with a single explanatory variable:

$$y_{it} = \beta_1 x_{it} + a_i + u_{it}, \quad i = 1, 2, \cdots, N, \quad t = 1, 2, \cdots, T,$$

  where $\text{Cov}(x_{it}, a_i) \neq 0$.
- When $\text{Cov}(x_{it}, a_i) \neq 0$, we call it **fixed-effects** model.
- Averaging the above equation over time, we get

$$\overline{y}_i = \beta_1 \overline{x}_i + a_i + \overline{u}_i.$$

- We then have

$$y_{it} - \overline{y}_i = \beta_1 (x_{it} - \overline{x}_i) + u_{it} - \overline{u}_i, \tag{3}$$

  which can be estimated using OLS. Why?

Fixed Effects Estimation

- On degrees of freedom:
  - It is $NT - N - k$, not $NT - k$.
- How to compute R^2?
  - Based on (3): the amount of time variation in the y_{it} that is explained by the time variation in the explanatory variables

The Dummy Variable Regression

- Recall that our model with a single explanatory variable is

$$y_{it} = \beta_1 x_{it} + a_i + u_{it}, \quad i = 1, 2, \cdots, N, \quad t = 1, 2, \cdots, T,$$

  where $\text{Cov}(x_{it}, a_i) \neq 0$.
- Can estimate those a_is using dummies:

$$y_{it} = \beta_1 x_{it} + \sum_{i=2}^{N} \lambda_i d_i + u_{it}$$

- $\hat{\beta}_1$ can be obtained using LS and we call this least square dummy variable (**LSDV**) approach.
- $\hat{a}_i$ can be computed by

$$\hat{a}_i = \overline{y}_i - \hat{\beta}_1 \overline{x}_{i1}, \quad i = 1, \cdots, N. \tag{4}$$

The Dummy Variable Regression

- In microeconometrics applications, N is generally large so there are many parameters to be estimated if we use LSDV.
- $\hat{a}_i$ obtained from (4) is unbiased, but not consistent with a fixed T as $N \to \infty$.
  - $\hat{\beta}_{1n}$ is of interest, a_is are just nuisance parameters.

FE or FD?

- Identical when $T = 2$
- When $T = 3$, both FE and FD are unbiased, but efficiency depends on the properties of u_{it}.
  - FE is more efficient than FD if u_{it} are serially correlated.
  - FD can eliminate concerns on spurious regressions particularly with large T small N panel.
- FE is also sensitive to classical measurement error problem.
- Make sense to report both in practice.
- Both FE and FD cannot recover coefficients attached to time-constant variables.
  - As FE also eliminates overall intercept, we need a year dummy in FE if FD also includes intercept to make them identical when $T = 2$.

Example cont.

```
. xtreg logscrap grant y88 if inrange(year,1987,1988), fe

Fixed-effects (within) regression               Number of obs      =       108
Group variable: fcode                           Number of groups   =        54

R-squared:                                      Obs per group:
     Within  = 0.1392                                         min =         2
     Between = 0.0049                                         avg =       2.0
     Overall = 0.0006                                         max =         2

                                                F(2, 52)           =      4.20
corr(u_i, Xb) = -0.0674                         Prob > F           =    0.0203

------------------------------------------------------------------------------
    logscrap | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       grant |  -.3170579   .1638751    -1.93   0.058    -.6458975    .0117816
         y88 |  -.0574357    .097206    -0.59   0.557    -.2524938    .1376224
       _cons |   .5974341   .0553369    10.80   0.000     .4863924    .7084757
-------------+----------------------------------------------------------------
     sigma_u |  1.4833025
     sigma_e |   .4066418
         rho |   .93009745   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(53, 52) = 26.42                     Prob > F = 0.0000
```

FE with unbalanced panels

- For each $i = 1, \cdots, N$, # of available T_i can be different.
- Easy in terms of formula:

$$\hat{\beta}_{1n}^{\text{FE}} = \frac{\sum_i \sum_t \mathbb{1}_{it} \left(x_{it} - \overline{x}_i\right) \left(y_{it} - \overline{y}_i\right)}{\sum_i \sum_t \mathbb{1}_{it} \left(x_{it} - \overline{x}_i\right)^2}$$

- Units with $T_i = 1$ play no role in FE.
- May be more efficient compared to FD under general missing patterns.

Random Effects Models

- Repeat again our model with a single explanatory variable:

$$y_{it} = \beta_1 x_{it} + a_i + u_{it}, \quad i = 1, 2, \cdots, N, \;\; t = 1, 2, \cdots, T. \tag{5}$$

- If we assume that a_i is not correlated with explanatory variable:

$$\text{Cov}\left(x_{it}, a_i\right) = 0, \;\; t = 1, 2, \cdots, T.$$

- Then, we call (5) a **random effect model**.

Random Effects Models

- Estimation can be done using pooled OLS, but is unlikely to be efficient.
- If we define $v_{it} = a_i + u_{it}$, under random effects assumptions,

$$\text{Corr}\left(v_{it}, v_{is}\right) = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_u^2}, \quad t \neq s,$$

  where $\sigma_a^2 = \mathbb{V}(a_i)$ and $\sigma_u^2 = \mathbb{V}(u_{it})$.
- Error term has serial correlation, so standard inference based on no serial correlation is also not valid.

Random Effects Models

- Deriving the GLS transformation that eliminates serial correlation in the errors requires sophisticated matrix algebra.
- It can be shown that the transformed equation takes the form

$$y_{it} - \theta \overline{y}_i = \beta_0(1-\theta) + \beta_1 \left(x_{it} - \theta \overline{x}_i\right) + \left(v_{it} - \theta \overline{v}_i\right), \tag{6}$$

  which involves **quasi-demeaned data** on the variable.
- Time-constant explanatory variables can be included in the model. Why?

Random Effects Models

- Feasible GLS requires the knowledge of θ, which can be obtained by replacing σ_a^2, σ_u^2 with estimated counterparts.
- Clearly, RE becomes pooled OLS when $\theta = 0$, and FE is obtained when $\theta = 1$.
  - As σ_a^2 is often large than σ_u^2, FE and RE estimates can be very similar in finite sample (only in numerical sense).
- If we look at the error term in (6):

$$v_{it} - \theta\overline{v}_i = (1-\theta)a_i + u_{it} - \theta\overline{u}_i,$$

  a_i is now weighted by $1-\theta$ and disappears when $\theta = 1$.

RE or pooled OLS?

- Breusch and Pagan (1980) test on $\mathcal{H}_0 : \sigma_a^2 = 0$?
- Not recommended (from a modern perspective) for the following reasons:
  - (i) has nothing to say about whether pooled OLS is consistent;
  - (ii) may pick up the serial correlation in u_{it} itself;
  - (iii) assumes $\mathcal{N}$ in deriving asymptotic distributions.

RE or FE?

- The null hypothesis is

$$\mathcal{H}_0 : \text{Cov}\left(x_{it}, a_i\right) = 0.$$

  - Under $\mathcal{H}_0$, both FE and RE are consistent, but RE is efficient.
  - Under $\mathcal{H}_A$, FE is consistent but RE becomes invalid.

- Hausman (1978) suggests the following Wald-type test statistic

$$H = \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right)' \left(\mathbb{V}(\hat{\beta}_{FE}) - \mathbb{V}(\hat{\beta}_{RE})\right)^{-1} \left(\hat{\beta}_{FE} - \hat{\beta}_{RE}\right) \overset{d}{\sim} \chi^2_k,$$

  where k is the # of regressors.

RE or FE?: A modern perspective

- The situation under $\mathcal{H}_0$ should be considered the exception rather than the rule, unless the key policy variable is set experimentally—say, each year, children are randomly assigned to classes of different sizes.
- Failure to reject $\mathcal{H}_0$ means either that
  - the RE and FE estimates are sufficiently close so that it does not matter which is used; OR
  - the sampling variation is so large in the FE estimates that one cannot conclude practically significant differences are statistically significant.

General Policy Analysis with Panel Data

- A general equation is

$$y_{it} = \eta_1 + \alpha_2 d2_t + \beta w_{it} + a_i + u_{it}, \quad t = 1, 2,$$

  where w_{it} is the binary intervention indicator and β is our interest.
- FE and FD should be identical. They all produce the DD estimator, $\hat{\beta}_{DD}$.
- If $w_{it} = prog_i \cdot d2_t$, we have the usual DID estimator.

Example: A Wage Equation Using Panel Data

TABLE 14.2 Three Different Estimators of a Wage Equation

Dependent Variable: log(*wage*)

Independent Variables	Pooled OLS	Random Effects	Fixed Effects
educ	.091 (.005)	.092 (.011)	———
black	−.139 (.024)	−.139 (.048)	———
hispan	.016 (.021)	.022 (.043)	———
exper	.067 (.014)	.106 (.015)	———
$exper^2$	−.0024 (.0008)	−.0047 (.0007)	−.0052 (.0007)
married	.108 (.016)	.064 (.017)	.047 (.018)
union	.182 (.017)	.106 (.018)	.080 (.019)

Going further 14.3

The union premium estimated by fixed effects is about 10 percentage points lower than the OLS estimate. What does this strongly suggest about the correlation between union and the unobserved effect?