ECON60052: Cross Section Econometrics

Lecture 3: Policy Analysis with Panel Data and some Simple Panel Data Methods

Nicky L. Grant

Semester 2

Introduction to Pooling and Panel Data

Panel data has both a time and cross section element. Broadly there are two kinds:

1. **Panel data** or **longitudinal** data: follow the *same* cross-section unit (individual, household, firm, country) through time
2. **Pooled cross sections**: different random samples at different points in time

These slides deal with **short, wide** panels $\Rightarrow$ not many time periods, but lots of individuals...

There are many examples of panel data sets across the world

- GSOEP, PSID, LSMS-ISA, etc.
- Here, we use the British Household Panel Survey (BHPS) (now Understanding Society), which started in 1991

Pooling independent cross sections across time

We may want to pool cross sections:

- ↪ to get bigger sample sizes
- ↪ to investigate the effects of time
- ↪ to investigate whether relationships have changed over time
- ↪ to evaluate policy effects

Data example

Consider 2 cross sections of data observed on a *different* sample of individuals in two time periods labelled $t = 1, 2$

i	t	y_i	1	$d1_i$	$d2_i$	$\boldsymbol{x}_i \rightarrow$		
1	1	4.29	1	1	0	2.42	1	...
2	1	4.86	1	1	0	9.31	0	...
3	1	4.62	1	1	0	6.22	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
	1	5.42	1	1	0	1.33	1	...
	2	5.39	1	0	1	5.92	0	...
	2	5.62	1	0	1	8.11	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
N	2	6.03	1	0	1	9.06	0	...

The vector $\boldsymbol{x}_i$ contains k covariates

The dummy $d1_i$ represents the first time period:

$$d1_i = \begin{cases} 1 & \text{if } t = 1 \\ 0 & \text{if } t = 2, \end{cases} \quad \text{with } d2_i = 1 - d1_i$$

Note instead of pooling data across time, we can think of pooling the data across any dummy variable (eg gender) $\Rightarrow$ essentially the same issues apply

- Are the parameters stable across time (including intercepts)?
- The estimator is called **Pooled OLS**

In the following example, we pool BHPS1996 and BHPS2006 (note they are not random samples in reality but let's pretend they were!)

The variable *male16* is a male dummy variable interacted with the Wave16 dummy variable etc.

A reminder: interaction terms

Let $d2$ denote a dummy variable for period $t = 2$ and m denote a dummy variable for gender

↪ Consider the following model for the pooled cross section with $t = 1, 2$:

$$y_{it} = \alpha_0 + \beta_1 m_{it} + \beta_2 d2_i + \delta m_{it} d2_i + u_{it}$$

↪ How to interpret the parameters?

- $\mathsf{E}(y|m = 0, d2 = 0) = \alpha_0 \Rrightarrow$ average y for women in $t = 1$
- $\mathsf{E}(y|m = 0, d2 = 1) = \alpha_0 + \beta_2 \Rrightarrow$ average y for women in $t = 2$. What is β_2?
- $\mathsf{E}(y|m = 1, d2 = 0) = \alpha_0 + \beta_1 \Rrightarrow$ average y for men in $t = 1$. What is β_1?
- $\mathsf{E}(y|m = 1, d2 = 1) = \alpha_0 + \beta_1 + \beta_2 + \delta \Rrightarrow$ average y for men in $t = 2$.

↪ What is $\widehat{\delta}$?

$\delta = [\mathsf{E}(y|m = 1, d2 = 1) - \mathsf{E}(y|m = 1, d2 = 0)] - [\mathsf{E}(y|m = 0, d2 = 1) - \mathsf{E}(y|m = 0, d2 = 0)]$.
Difference-in-differentials: whether and how much, say, the gender pay gap varied over time

```
. su lrhrpay wave16 southeast age age2 male married manual covered if wave==6

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     lrhrpay |      4485   -2.601872    .5569235  -6.164872      .66056
      wave16 |      4485           0           0          0          0
   southeast |      4485    .2008919     .400712          0          1
         age |      4485    37.50702    11.83322         16         82
        age2 |      4485    1546.771    946.8992        256       6724
-------------+--------------------------------------------------------
        male |      4485    .4894091    .4999436          0          1
     married |      4485    .5823857    .4932209          0          1
      manual |      4485    .3092531    .4622372          0          1
     covered |      4485    .4784838    .4995925          0          1
```

```
. su lrhrpay wave16 southeast age age2 male married manual covered if wave==16

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     lrhrpay |      4092   -2.319596    .5379006  -5.530777    .381414
      wave16 |      4092           1           0          1          1
   southeast |      4088    .1998532    .3999388          0          1
         age |      4092    39.78104     12.3699         16         82
        age2 |      4092    1735.508    1018.982        256       6724
-------------+--------------------------------------------------------
        male |      4092    .4853372     .499846          0          1
     married |      4087     .542941    .4982136          0          1
      manual |      4067    .2505532    .4333849          0          1
     covered |      4092    .4828935    .4997684          0          1
```

```
. reg lrhrpay  southeast  age age2 male married manual covered wave16 male16
> covered16 manual16 if wave==6 | wave==16

      Source |       SS       df       MS              Number of obs =    8543
-------------+------------------------------           F( 11,  8531) =  390.24
       Model |  913.677226    11   83.061566           Prob > F      =  0.0000
    Residual |   1815.7807  8531  .212845001           R-squared     =  0.3347
-------------+------------------------------           Adj R-squared =  0.3339
       Total |  2729.45793  8542  .319533824           Root MSE      =  .46135

------------------------------------------------------------------------------
     lrhrpay |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   southeast |    .090701   .0125359     7.24   0.000     .0661275    .1152744
         age |   .0783872   .0026371    29.72   0.000     .0732178    .0835566
        age2 |  -.0009066   .0000318   -28.50   0.000     -.000969   -.0008443
        male |   .4201241   .0142987    29.38   0.000     .3920951     .448153
     married |    .040595   .0114594     3.54   0.000     .0181318    .0630582
      manual |  -.3514498   .0155264   -22.64   0.000    -.3818852   -.3210143
     covered |   .1961334   .0139355    14.07   0.000     .1688165    .2234503
      wave16 |   .3064201   .0177479    17.27   0.000     .2716298    .3412104
     male16 |  -.0931194   .0209377    -4.45   0.000    -.1341624   -.0520765
  covered16 |  -.0320294   .0200599    -1.60   0.110    -.0713517    .0072928
   manual16 |   .0405851    .023446     1.73   0.083    -.0053748    .0865449
      _cons |  -4.372235   .0503866   -86.77   0.000    -4.471005   -4.273465
------------------------------------------------------------------------------
```

Interpreting the results

Between 1996 and 2006, the proportions covered and male did not change, and the proportion manual fell from 31% to 25%

- But is there a change in the effect of these variables on wages between 1996 and 2006?

Take a look at the interaction terms *covered16*, *manual16* and *male16*:

- the coverage differential has fallen by 0.032 log-points (but is not significant) and the manual differential has risen by 0.041 log-points (and is not significant, just)
- The male differential has fallen by 0.093 log-points (and is significant)
- See Wooldridge Ch7

For the base category (non-covered, non-manual unmarried women), there is an unexplained increase in log real hourly wages of 0.306 log-points

Policy analysis with pooled cross sections

We studied treatment effects in a non-dynamic setting in Lecture 1

Often treatments happen in a dynamic setting, e.g. a tax change is implement in t=2 (that wasn't in place in t=1)

When we observe a cross section over time this allows us to control for some unobserved heterogeneity

A downside is that there may be time trends/dynamic effects we must model also

Difference in Difference Estimator

Ideally we would have random assignment into **control** and **treatment** groups, like in a medical experiment (called groups C and T respectively)

Without random assignment, control for systematic differences between the 2 groups, compare the change in outcomes across the treatment and control groups to estimate the treatment effect

Thus

$$(\overline{y}_{2,T} - \overline{y}_{2,C}) - (\overline{y}_{1,T} - \overline{y}_{1,C})$$

$$\equiv (\overline{y}_{2,T} - \overline{y}_{1,T}) - (\overline{y}_{2,C} - \overline{y}_{1,C})$$

is the **difference-in-differences (DiD) estimator**

Using regression to recover the DiD

This is the same as the following regression but *without* the covariates

$$y = \beta_0 + \delta_0 d2 + \beta_1 dT + \delta_1 d2.dT + u, \tag{1}$$

where $d2$ is a time dummy for after the treatment occurs, and dT is a dummy for whether treated

- The constant is the mean of the dep var for the untreated in period 1, ie $\overline{y}_{1,C}$
- The estimate on *d2* is the differential relative to the base category, ie $\overline{y}_{2,C} - \overline{y}_{1,C}$
- The estimate on *d2.dT* is the difference in differentials $(\overline{y}_{2,T} - \overline{y}_{1,T}) - (\overline{y}_{2,C} - \overline{y}_{1,C})$

Consistent estimation of the treatment effect requires u exogenous so OLS unbiased estimate of δ_1.

The benefit of the regression formulation is it allows for the control of additional xs (selection on observables)

Example: Impact of Increasing Minimum Wages on Employment- Card and Krueger (1994)

New Jersey 1 April 1992: increase in state minimum wage from \$4.25 to \$5.05.

CK (1994) collected data on employment in fast-food restaurants in February 1992 (before) and November 1992 (after)

CK (1994) also analysed data for restaurants in Pennsylvania, where the minimum wage stayed at \$4.25 throughout

Uncovering Treatment Effect Using Panel Data

Here the treatment (D_{st}) is increase in minimum wage by 80c. The independent var. is y_{ist}, employment in rest. *i*, in state *s*, time period *t*.

As in Lecture 1 we must construct an estimator that removes any selection bias

Even if treatment was allocated at random- how to use the panel to construct an estimate of the ATE?

1-3 below are three candidates with some intuitive basis.

1: $E[y_{ist}|\text{s=NJ, t=Nov}] - E[y_{ist}|\text{s=NJ, t=Feb}]$

2: $E[y_{ist}|D_{st} = 1] - E[y_{ist}|D_{st} = 0]$

3: $(E[y_{ist}|\text{s=NJ, t=Nov}] - E[y_{ist}|\text{s=NJ, t=Feb}]) - (E[y_{ist}|\text{s=PA, t=Nov}] - E[y_{ist}|\text{s=PA, t=Feb}])$

Uncovering Treatment Effect Using Panel Data

Assuming there is no selection bias, this differential includes the treatment effect plus any change in time trend. e.g. Suppose $y_{ist} = \alpha + \gamma_s + \lambda_t + \beta D_{st} + u_{ist}$

1. $E[y_{ist}|\text{s=NJ, t=Nov}] - E[y_{ist}|\text{s=NJ, t=Feb}] = \lambda_{Nov} - \lambda_{Feb} + \beta$

2. $E[y_{ist}|D_{st} = 1] - E[y_{ist}|D_{st} = 0] = E[y_{ist}|\text{s=NJ, t=Nov}] - E[y_{ist}|t = \text{Feb}] = \lambda_{Nov} - \lambda_{Feb} + \beta - E[\gamma_s]$

3.$(E[y_{ist}|\text{s=NJ, t=Nov}] - E[y_{ist}|\text{s=NJ, t=Feb}]) - (E[y_{ist}|\text{s=PA, t=Nov}] - E[y_{ist}|\text{s=PA, t=Feb}]) = \beta$

Assuming a common trend λ_t in both states 3. identifies the treatment effect β.

The Difference-in Difference estimator is defined as the sample estimate of 3.

Note: If λ_t varied across states then 3. would also have a bias from improper specification of time trends.

DiD Estimation via Regression

We can uncover the DiD by a regression:

$$y_{ist} = \alpha + \gamma NJ_s + \lambda d_t + \beta(NJ_s \times d_t) + u_{ist}$$

NJ dummy for New Jersey and d_t dummy for post treatment

Interpretation:

1) $E(y|NJ = 0, d_t = 0) = \alpha \Rrightarrow$ PA pre treatment
2) $E(y|NJ = 1, d_t = 0) = \alpha + \gamma \Rrightarrow$ NJ pre treatment
3) $E(y|NJ = 0, d_t = 1) = \alpha + \lambda \Rrightarrow$ PA post treatment
4) $E(y|NJ = 1, d_t = 1) = \alpha + \gamma + \lambda + \beta \Rrightarrow$ NJ post treatment

(4)-(2): $(\alpha + \gamma + \lambda + \beta) - (\alpha + \gamma) \Rrightarrow$ Difference before/after treatment for NJ.

(3)- (1): $(\alpha + \gamma) - (\alpha) \Rrightarrow$ Difference before/after treatment for PA.

$\Rrightarrow$ β Difference-in-difference

Table 5.2.1: Average employment per store before and after the New Jersey minimum wage increase

Variable	PA (i)	NJ (ii)	Difference, NJ-PA (iii)
1. FTE employment before, all available observations	23.33 (1.35)	20.44 (0.51)	-2.89 (1.44)
2. FTE employment after, all available observations	21.17 (0.94)	21.03 (0.52)	-0.14 (1.07)
3. Change in mean FTE employment	-2.16 (1.25)	0.59 (0.54)	2.76 (1.36)

Notes: Adapted from Card and Krueger (1994), Table 3. The table reports average full-time equivalent (FTE) employment at restaurants in Pennsylvania and New Jersey before and after a minimum wage increase in New Jersey. The sample consists of all stores with data on employment. Employment at six closed stores is set to zero. Employment at four temporarily closed stores is treated as missing. Standard errors are reported in parentheses

Caveats on program evaluation

In many cases the agent chooses whether to participate in a program, which may lead to a **self-selection problem**

If we can control for everything that is correlated with both participation and the outcome of interest then there is no problem

If there are unobservables that are correlated with participation $\Rightarrow$ the DiD is biased

If the 'common trends' assumption is not valid $\Rightarrow$ the DiD is biased

Common trends assumption: crucial for identification

Common Trends Assumption: In the absence of treatment, the outcome would follow the same trend for treatment and control groups

We can test for this if have access to pre-treatment data

E.g. Card Krueger (2000) updated their data set with employment at rests. in NJ and PA for more periods, including another increase in minimum wage just for PA to $4.75 in 1996

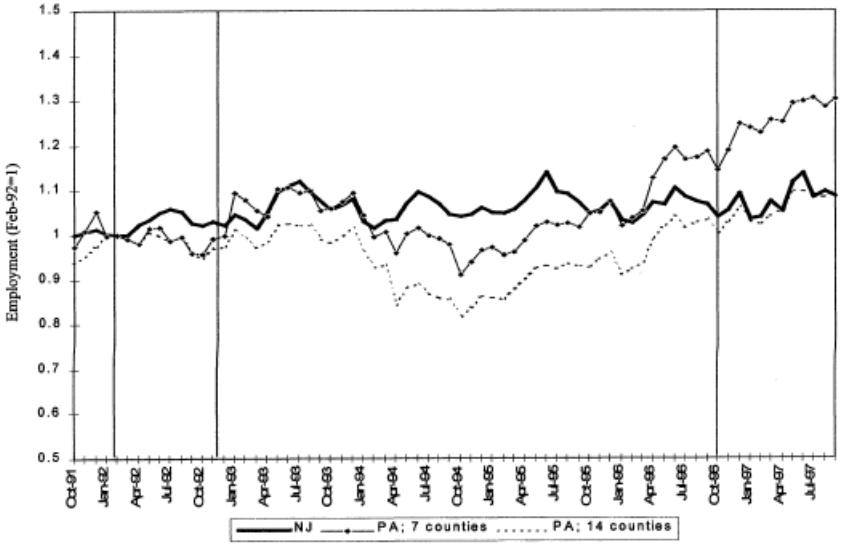

Figure 5.2.2: Employment in New Jersey and Pennsylvania fast-food restaurants, October 1991 to September 1997 (from Card and Krueger 2000). Vertical lines indicate dates of the original Card and Krueger (1994) survey and the October 1996 federal minimum-wage increase.

Two-period panel data analysis

Consider again 2 cross-sections of data observed on the *same* sample of individuals. The periods are labelled $t = 1, 2$

this is called a short, wide panel (lots of n, not much t)

We observe everybody twice ($\Rrightarrow$ the panel is balanced)

Data example

i	t	y_{it}	1	$d1_t$	$d2_t$	$\boldsymbol{x}_{it} \rightarrow$		
1	1	4.29	1	1	0	2.42	1	...
2	1	4.86	1	1	0	9.31	0	...
3	1	4.62	1	1	0	6.22	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
N	1	5.42	1	1	0	1.33	1	...
1	2	5.39	1	0	1	5.92	0	...
2	2	5.62	1	0	1	8.11	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
N	2	6.03	1	0	1	9.06	0	...

Why panel data?

Using panel data methods can be advantageous because it can address some kinds of omitted variable bias

In micro-econometric application, **unobserved heterogeneity** is widespread

These are characteristics of individuals that are not observable, vary between individuals, and cause variations in observed behaviour

For example, there are many empirical models of earnings (wages) that are explained by gender, training, schooling, occupational choice, union status, etc.

Most of these variables are **endogenous**, as they are correlated with factors we fail to control for, such as "motivation" or "ability",which we can never observe except by (poor) proxies

↪ This leads to omitted variables bias

But . . . *if* it can be argued that "motivation" or "ability" are inherently fixed for a given individual ⇒ panel data methods can overcome the **heterogeneity bias**

Two-period panel data model

If it reasonable to assume that the omitted variable is constant over time

$\Rightarrow$ the model is specified as having 2 errors: one fixed over time (a_i) and the other time-varying (u_{it})

Write the population model as

$$y_{it} = \beta_0 + \delta_0 d2_t + \boldsymbol{x}_{it}\boldsymbol{\beta} + a_i + u_{it} \qquad t = 1, 2$$

- a_i is called an **unobserved effect**, a **fixed effect**, or **unobserved heterogeneity**;
- u_{it} is called an **idiosyncratic error**;
- This model is called an **unobserved effects** or **fixed effects** model

We could estimate this model by OLS, replacing $a_i + u_{it}$ by v_{it}, a **composite error**:

$$y_{it} = \beta_0 + \delta_0 d2_t + \boldsymbol{x}_{it}\boldsymbol{\beta} + v_{it}$$

But *if* a_i is correlated with $\boldsymbol{x}_{it}$, pooled OLS will be biased and inconsistent

- since a_i is part of the error term $\Rightarrow$ often called **heterogeneity bias**

With panel data, we can difference-out the unobserved fixed effect

$$\begin{aligned} y_{i1} &= \beta_0 + \boldsymbol{x}_{i1}\boldsymbol{\beta} + a_i + u_{i1} \quad (t = 1) \\ y_{i2} &= \beta_0 + \delta_0 + \boldsymbol{x}_{i2}\boldsymbol{\beta} + a_i + u_{i2} \quad (t = 2) \end{aligned}$$

Subtracting,

$$\Delta y_i = \delta_0 + \Delta \boldsymbol{x}_i \boldsymbol{\beta} + \Delta u_i,$$

where:

$\Delta y_i \equiv y_{i2} - y_{i1}$ is the **difference** of y_i

$\Delta \boldsymbol{x}_i \equiv \boldsymbol{x}_{i2} - \boldsymbol{x}_{i1}$ is a vector of differences of regressors

β_0 and $\beta_0 + \delta_0$ are "macro" effects in that they affect all individuals identically, in the same time period, but only δ_0 is identified

δ_0 is the parameter on $\Delta d2_t$, which is a vector of ones (the new regression constant)

In words, the change in y_{it} between periods 1 and 2 is regressed on the change in the regressors and a constant, using N obs

To see what is happening, reorder the data above, and subtract first row from second for each i:

i	t	y_{it}	1	$d1_t$	$d2_t$	$\boldsymbol{x}_{it}$	Δy_{it}	$\Delta 1$	$\Delta d1_t$	$\Delta d2_t$	$\Delta \boldsymbol{x}_{it}$
1	1	y_{11}	1	1	0	$\boldsymbol{x}_{11}$	.	.	.	.	.
1	2	y_{12}	1	0	1	$\boldsymbol{x}_{12}$	Δy_1	0	–1	1	$\Delta \boldsymbol{x}_1$
2	1	y_{21}	1	1	0	$\boldsymbol{x}_{21}$	.	.	.	. .	.
2	2	y_{22}	1	0	1	$\boldsymbol{x}_{22}$	Δy_2	0	–1	1	$\Delta \boldsymbol{x}_2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	1	y_{N1}	1	1	0	$\boldsymbol{x}_{N1}$	.	.	.	.	.
N	2	y_{N2}	1	0	1	$\boldsymbol{x}_{N2}$	Δy_N	0	–1	1	$\Delta \boldsymbol{x}_N$

First-difference estimator

Note that variables are not defined for period 1 and so dataset comprises just N obs

This is a **first differenced equation**

Are the key assumptions for OLS satisfied?

If yes, OLS is called the **first-differenced estimator** (FD), and is unbiased

Exogeneity. Is Δu_i? uncorrelated $\Delta \boldsymbol{x}_i$

We need

$$E(\Delta \boldsymbol{x}_i \Delta u_i) = E[(\boldsymbol{x}_{i2} - \boldsymbol{x}_{i1})(u_{i2} - u_{i1})] = 0$$

4 sufficient conditions are:

$$E(\boldsymbol{x}_{i1} u_{i1}) = 0$$
$$E(\boldsymbol{x}_{i2} u_{i2}) = 0$$
$$E(\boldsymbol{x}_{i1} u_{i2}) = 0$$
$$E(\boldsymbol{x}_{i2} u_{i1}) = 0$$

Policy Analysis with Two-period Panel Data

i	t	y_{it}	T_{it}	d_{2t}	(T_i)	$\boldsymbol{x}_{it}$	Δy_{it}	ΔT_{it}	Δd_{2t}	$\Delta \boldsymbol{x}_{it}$
1	1	y_{11}	0	0	(0)	$\boldsymbol{x}_{11}$	.	.	.	.
1	2	y_{12}	0	1	(0)	$\boldsymbol{x}_{12}$	Δy_1	0	1	$\Delta \boldsymbol{x}_1$
2	1	y_{21}	0	0	(1)	$\boldsymbol{x}_{21}$	.	.	. .	.
2	2	y_{22}	1	1	(1)	$\boldsymbol{x}_{22}$	Δy_2	1	1	$\Delta \boldsymbol{x}_2$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
N	1	y_{N1}	0	0	(1)	$\boldsymbol{x}_{N1}$	.	.	.	.
N	2	y_{N2}	1	1	(1)	$\boldsymbol{x}_{N2}$	Δy_N	1	1	$\Delta \boldsymbol{x}_N$

Without $\boldsymbol{x}_{it}$ estimate

$$y_{it} = \beta_0 + \gamma_0 d2_t + \delta_1 T_{it} + a_i + u_{it} \tag{2}$$

If program participation takes place only in $t = 2$, then $T_{i1} = 0$

Regress Δy_i on T_{i2} and a constant:

$$\Delta y_i = \gamma_0 + \delta_1 T_{i2} + \Delta u_i$$

Note $\hat{\gamma}_0 = \overline{\Delta y}_C$, and that

$$\hat{\delta}_1 = \overline{\Delta y}_T - \overline{\Delta y}_C \tag{3}$$

$\Rrightarrow$ This is the panel data equivalent of the difference-in-differences (DiD) estimator

- It has the important advantage of being able to control for unobserved heterogeneity

Adding differenced covariates controls for any time-varying variables that might be correlated with program designation

If program participation takes place in both periods, this nice relationship no longer applies, but the interpretation is the same... Regress Δy_{it} on ΔT_{it}, Δ covariates and a constant

Costs of using FD estimators

Each Δx_i must have some variation across $i \Rightarrow$ age, gender, and education are good examples of Δx_i being constant

Even if Δx_i is not constant, it might not vary much, leading to large standard error

Measurement error is attenuated (gets worse) with differencing

↪ See what happens with male dummy m and age a:

i	t	y_{it}	$d2_t$	m	a	Δy_{it}	$\Delta d2_t$	Δm	Δa
1	1	y_{11}	0	1	18	.	.	.	.
1	2	y_{12}	1	1	24	Δy_1	1	0	6
2	1	y_{21}	0	0	28	.	. .	.	.
2	2	y_{22}	1	0	34	Δy_2	1	0	6
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
N	1	y_{N1}	0	1	51	.	.	.	.
N	2	y_{N2}	1	1	57	Δy_N	1	0	6

Readings

↪ Wooldridge, Chapter 7 and 13

↪ Angrist & Pischke, Chapter 5.1, 5.2