Econometrics II

Lecture 5
Panel Data Models II, Clustering, and DiD

Marie Paul

University of Duisburg-Essen
Ruhr Graduate School in Economics

Summer Semester 2020

Outline of lecture

4. Clustering and DiD
   - Applying Panel Data Methods to Other Data Structures
   - Difference-in-differences (DiD)

Self-study plan for Lecture V

- This lecture discusses the application of panel data methods on other data structures and clustering.
- For our next virtual meeting (on May 28th), please work through all slides of Lecture V. Readings:
- Wooldridge 2019: p. 480-483 (this is on other data structures and on whether to cluster)
- If you want to go deeper into the topic of whether to cluster or not, I suggest to read the paper by Abadie, Athey, Imbens, and Wooldridge (2017).
- Please read Cameron and Miller (2015) attached to my email.
- Please read an introduction to DID, e.g. Wooldridge (2019), p.427-439.
- If you are particularly interested in specific topics I suggest to read the cited papers on these topics, but this is not required.
- As usual, please prepare the "your own research" questions and collect questions and topics to discuss them in our virtual meeting.

Outline

4. Clustering and DiD
   - Applying Panel Data Methods to Other Data Structures
   - Difference-in-differences (DiD)

Other Data Structures

Based on Wooldridge 2019:

Panel data methods can be applied on **certain data structures** that do not involve time:

Example 4.1 (Effects on teen childbearing on future economic outcomes)

Geronimus and Korenman (1992) use pairs of sisters in a fixed-effects approach (cross-sectional data, i is family and t is the women). Teenbirth may be endogenous, in particular correlated with an unobserved family effect. FE may fix this problem and a CRE estimation is a good way to test whether averaging over regressors is needed.

More generally, **panel data models** can be applied to a **cluster data**: clusters of units are sampled from a **population of clusters** rather than from the population of individuals (e.g. each family is sampled from the population of families and a family is then a cluster.)
Need many clusters, as the **number of clusters** is **what N is in panel data**.

Cluster Data

Example 4.2 (Linked-employer-employee data (LIAB) of IAB)

Sampling occurs at the firm level (first sample firms) and worker information is added. Some variables vary only on the firm level. In addition unobserved firm characteristics within firms must be accounted for. If these are correlated with regressors FE may help, but (in a cross section) we can then only estimate effects of variables that vary within firms. But the LIAB is in addition a panel data set. Nevertheless some firm variables do not vary over time.

People often say "*I included firm fixed effects in the analysis.*" Maybe they also use several time periods, but do not use FE in the time (panel) dimension.

Other Data Structures

- Further examples: 1) National Education Panel Study (NEPS), sampling based on **schools and classes**. 2) **Regional** data if sampling occurs e.g. on **county levels**.
- In a **true cluster data**, the clusters are **first drawn** from a **population of clusters** and then individuals are drawn from the clusters.
- Using **OLS** "clustered SE" should be used: see below.
- **RE**, or more generally, FGLS, can be applied if unobserved cluster effects are uncorrelated with regressors to gain **efficieny**. Assumption that the **specified error structure** (e.g. RE) really holds is strong.
- The RE approach used to be the **standard approach**, but "clustering" and using OLS has become very popular with **clustered SE** implemented in Stata....

Clustering: when?

Traditional view: if **errors for units within clusters are correlated**.
New paper by **Abadie, Athey, Imbens, and Wooldridge (2017, NBER)**:

- In a **true clustered sample** (*"sampling design problem"*).
- Not in a random sample from the population which has identifiers on e.g. state. Then one could also cluster on education, gender...
- Unless **in a random sample the key policy variable** (or variable of interest in general) is applied at a **group level** (is invariant at a group level) and these groups then become the clusters. If we do not cluster here the SE are only valid for the treatment effect that conditions on the **particular situation** we observe.
- This *"experimental design problem"* occurs *"when clusters of units, rather than units, are assigned to a treatment"*.
- Researcher should assess whether **the assignment mechanism is clustered**. Otherwise do not cluster. The fact that clustering would change the SE is not an argument in favor of clustering (in the view of Abadie et al.).
- Clustering or not is an important decision because it is not uncommon that clustered SE are **several times larger** than normal SE!

Clustering

The following slides are based on Cameron and Miller (JHR, 2015): "A Practitioner's Guide to Cluster-Robust Inference".
Includes much more information than presented here and highly recommanded when decisions how to cluster comes up.

To start with assume: **no fixed-effects**, it is **clear how** to form the clusters, we have **many clusters**, and **one regressor**. All results extend to multiple regressor.

Cluster Robust Variance Estimate (CRVE)

Recall estimator for the **heteroscedastic variance matrix** (White, 1980):

$$Var_{het}[\hat{\beta}] = (\frac{\sum_i x_i^2 \hat{u_i}^2}{\sum_i x_i^2})^2 \tag{4.1}$$

If **all errors** would be **correlated** with each other the variance the idea of White would correspond to:

$$Var_{cor}[\hat{\beta}] = (\frac{\sum_i \sum_j x_i x_j \hat{u}_i \hat{u}_j}{\sum_i x_i^2})^2 \tag{4.2}$$

but this equals zero since $\sum_i x_i \hat{u}_i = 0$.
If we consider **clustered errors**, with $E[u_i, u_j] = 0$ unless observations i and j are in the same cluster, we obtain the **heteroscedasticity and cluster-robust estimate**:

$$Var_{clu}[\hat{\beta}] = (\frac{\sum_i \sum_j x_i x_j \hat{u}_i \hat{u}_j \mathbb{1}[i, j, \text{in same cluster}]}{\sum_i x_i^2})^2 \tag{4.3}$$

Using **multiple regressors** and matrix notation this gives the same formula as the **panel robust variance matrix**.

Cluster Robust Variance Estimate (CRVE)

In the special case that there is only one observation in each cluster $Var_{clu} = Var_{het}$. But generally $Var_{clu} \geqslant Var_{het}$ due to the addition of terms when $i \neq j$, but they are in the same cluster.
The amount of increase is larger, the:

1. more positively correlated the **regressors** are across observations in the same cluster (via $x_i x_j$)
2. more positively correlated are the **errors** in the same clusters (via $u_i u_j$)
3. more **observations** are in the **same cluster**.

When we have an **aggregated regressor** the cluster-robust SE can be much larger even if there is low within-cluster error correlation, because the regressor of interest is perfectly correlated within clusters.

Note that when estimating the variance for one particular cluster there is **no averaging going one**, but the **averaging occurs over all cluster** and this makes the variance estimate consistent. Thus, for the **asymptotics** to work we need **many clusters**. But we should still use a **finite sample correction**.
In **Stata** use: regress y x, vce (cluster id-clu)

Cluster Specific Fixed Effects

- Includes a separate intercept for each cluster: **add dummies (LSDV) or use FE estimator**.
- Not possible if regressor is **invariant** in cluster (e.g. policy variable in cross-section or not varying over time if several years available).
- Cluster robust variance matrix formula carries over.
- FE will generally not **completely control** for within-cluster error correlation, so still use CRVE. Because FE may take out correlation e.g. within schools, but additional correlation within classes possible. Or serial correlation in the error within cluster (e.g. same persons).
- **FE estimation and LSDV** lead to the same coefficients but due to different finite sample degrees of freedom correction they may lead to **different SE**. Use FE: xtreg y x, ge vce(robust). (Or do a correction on the LSDV estimators).

What to cluster over?

- **Larger** and fewer clusters have **less bias** at the cost of more variablility. Rather avoid bias → cluster at the state level and not at the county level if you are uncertain. Provided there are enough states to avoid the problem of few clusters.
- So if clusters are **nested** rather use the **broader cluster**.
- If the clusters are **not nested**, a **multi-way cluster-robust** variance matrix may be a (complicated) solution.
- Or **cluster in one dimension** and attempt to eliminate concerns about clustering using **appropriate controls** in the other dimension.
- It is **wrong** to cluster at the **intersection**, e.g. state-year level. This would assume that observations are independent if they are in the same state but in different years.
- For **spatial data** it is assumed that model erros are more correlated when the distance between observations is smaller. This can be accounted for with **weights getting smaller** with distance larger.

Few Clusters

- If there are many observations within each cluster the **coefficients** may still be well estimated if there are **few clusters**. But the **variance matrix** can be **biased** because asymptotics do not work.
- At least one should use a finite sample correction (as Stata does).
- According to some studies depending on the situation **few** may range from 20 to 50 states in a state-year data set. More is better.
- Otherwise there exist **bias corrections** and **bootstrap methods**.

Clustering

Example 4.3

Dehos and Paul (2020): The Effect of After-School Programs on Maternal Employment

Table 1: OLS Results

	(1)	(2)	(3)	(4)
		OLS *(SOEP)*		
Hours worked				
ASP attendance	6.905 (0.928)***	5.408 (0.833)***	5.612 (0.835)***	2.154 (0.745)***
Employment				
ASP attendance	0.191 (0.030)***	0.154 (0.027)***	0.160 (0.027)***	0.058 (0.028)**
Observations	5 103	5 103	5 103	4 005
Time dummies	yes	yes	yes	yes
State dummies	yes	yes		
Individual covariates		yes	yes	yes
County covariates		yes	yes	yes
County fixed effects			yes	yes
Past labor market attach.				yes

Clustering

Example 4.4 (Your own research)

Do cases for clustering occur in your research and how do people deal with them?

Outline

4. Clustering and DiD
   - Applying Panel Data Methods to Other Data Structures
   - Difference-in-differences (DiD)

Introduction

Difference-in-Differences (DID) is an old and well established technique. Motivated as FE on averaged data or before-after combined with treated-control identification strategy.

It can be applied with panel data or with pooled cross-sections.

Simplest setting: **two groups** and **two time periods** where none exposed to treatment in **first period**; the treatment group exposed in **second period**.

Average gain **over time** in the non-exposed group is **subtracted** from the gain in the exposed group.

Double differencing removes bias in second period comparisons and biases in comparisons over time within the treatment group.

The basic 2×2 model can be **extended** in various directions:

- Multiple groups, multiple periods.
- Models with covariates.
- Additional dimensions (triple difference etc.).

Basic Setup

$$\hat{\tau}_{DID} = \underbrace{\left[\overline{Y}_1^T - \overline{Y}_0^T\right]}_{\Delta_t^T} - \underbrace{\left[\overline{Y}_1^C - \overline{Y}_0^C\right]}_{\Delta_t^C} \tag{4.4}$$

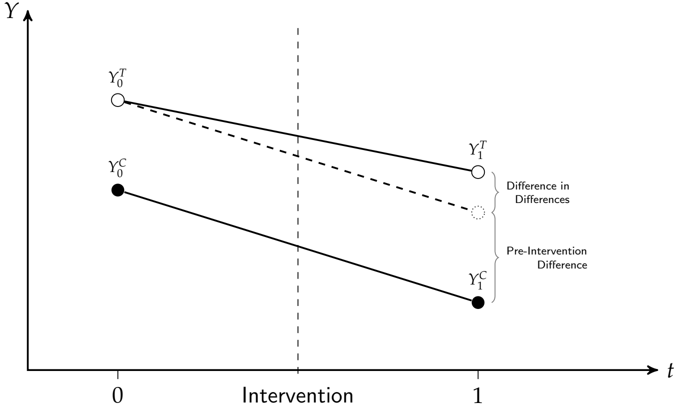

Figure 4.1. Difference in Differences.

Definitions

Observations of outcome variable Y from time period $T_i \in \{0, 1\}$. Binary **treatment variable** $D_i \in \{0, 1\}$.

Bear in mind, D signifies **group membership** and not necessarily treatment (e.g. minimum wage not necessarily binding). In what follows, suppress individual index i.

The parameter of interest is usually the **average treatment effect of the treated** where the untreated group is used as a counterfactual.

$$
\begin{aligned}
ATT_t &= \mathbb{E}\left[Y_t^1 - Y_t^0 \middle| D = 1 \right] \\
&= \mathbb{E}\left[\underbrace{\mathbb{E}\left[Y_t^1 - Y_t^0 \middle| X = x, D = 1 \right]}_{\theta_t(x)} \middle| D = 1 \right] \\
&= \mathbb{E}_{X|D=1}\left[\theta_t(x) \right].
\end{aligned} \tag{4.5}
$$

DID Assumptions

Assumption DID.1 (Stable Unit Treatment Value Assumption - SUTVA)

$$Y_t = DY_t^1 + (1 - D)\, Y_t^0 \tag{4.6}$$

$\Rightarrow$ One and only one **potential outcome** observed. No peer or GE effects.

Assumption DID.2 (Exogeneity - EXOG)

$$X^1 = X^0 = X,\ \forall x \in \chi. \tag{4.7}$$

$\Rightarrow$ Conditioning variables are independent of treatment. **Bad controls** will partially capture effect.

Assumption DID.3 (No Effect Prior to Treatment - NEPT)

$$\theta_0\,(x) = 0,\ \forall x \in \chi. \tag{4.8}$$

$\Rightarrow$ No **anticipation** effects from being treated.

DID Assumptions II

Assumption DID.4 (Common Trend - CT)

$$\begin{aligned}
&\mathbb{E}\left[Y_1^0|X=x,D=1\right]-\mathbb{E}\left[Y_0^0|X=x,D=1\right]\\
&=\mathbb{E}\left[Y_1^0|X=x,D=0\right]-\mathbb{E}\left[Y_0^0|X=x,D=0\right]\\
&=\mathbb{E}\left[Y_1^0|X=x\right]-\mathbb{E}\left[Y_0^0|X=x\right],\ \forall x\in\chi.
\end{aligned} \tag{4.9}$$

⇒ Control group reflects **counterfactual trend** of treatment group in absence of treatment.

Assumption DID.5 (Common Support - CS)

$$\begin{aligned}
&\Pr\left[T\cdot D=1|X=x,(T,D)\in\{(t,d),(1,1)\}\right]<1\\
&\qquad\forall\,(t,d)\in(0,1),(0,0),(1,0);\quad \forall x\in\chi.
\end{aligned} \tag{4.10}$$

⇒ There must exist an **overlap** between all four group/time subsamples for all values of X.

Identification

Let's prove that the DID model is identified:

$$\theta_1(x) = \mathbb{E}\left[Y_1^1 - Y_1^0 | X = x, D = 1\right]$$

$$\overset{SUTVA}{=} \underbrace{\mathbb{E}\left[Y_1 | X = x, D = 1\right]}_{\text{identified}} - \mathbb{E}\left[Y_1^0 | X = x, D = 1\right].$$

$$\begin{aligned}
\mathbb{E}\left[Y_1^0 | X = x, D = 1\right] &\overset{CT}{=} \mathbb{E}\left[Y_1^0 | X = x, D = 0\right] - \mathbb{E}\left[Y_0^0 | X = x, D = 0\right] \\
&+ \mathbb{E}\left[Y_0^0 | X = x, D = 1\right]. \\
&\overset{SUTVA}{=} \underbrace{\mathbb{E}\left[Y_1 | X = x, D = 0\right] - \mathbb{E}\left[Y_0 | X = x, D = 0\right]}_{\text{identified}} + \mathbb{E}\left[Y_0^0 | X = x, D = 1\right].
\end{aligned} \tag{4.11}$$

$$\mathbb{E}\left[Y_0^0 | X = x, D = 1\right] \overset{NEPT}{=} \mathbb{E}\left[Y_0^1 | X = x, D = 1\right] \overset{SUTVA}{=} \underbrace{\mathbb{E}\left[Y_0 | X = x, D = 1\right].}_{\text{identified}}$$

Thus,

$$\begin{aligned}
\theta_1(x) = &\left\{\mathbb{E}\left[Y_1 | X = x, D = 1\right] - \mathbb{E}\left[Y_0 | X = x, D = 1\right]\right\} \\
&- \left\{\mathbb{E}\left[Y_1 | X = x, D = 0\right] - \mathbb{E}\left[Y_0 | X = x, D = 0\right]\right\}.
\end{aligned}$$

Finally, using the common support assumption we obtain

$$ATT_1 = \mathbb{E}_{X|D=1}\left[\theta_1(x)\right]. \tag{4.12}$$

A Parametric Specification

The ATT is often estimated using **linear regression**:

$$Y_i = \alpha + \gamma_t \cdot T_i + \gamma_d \cdot D_i + \tau_{DID} \cdot T_i D_i + \epsilon_i \tag{4.13}$$

We can improve the performance of DID by controlling for **covariates**:

$$Y_i = \alpha + \gamma_t \cdot T_i + \gamma_d \cdot D_i + \tau_{DID} \cdot T_i D_i + \mathbf{X}_{it}\beta + \epsilon_i \tag{4.14}$$

This modification has two advantages:

1. Increases the **efficiency** of the estimated τ.
2. Makes the **common time trend** assumption more credible.

Taking expectations over group and time indicators in (4.14) above yields

$$\begin{aligned}\theta_1(x) =& \mathbb{E}\left[Y_1|X=x, D=1\right] - \mathbb{E}\left[Y_0|X=x, D=1\right] \\ & - \left[\mathbb{E}\left[Y_1|X=x, D=0\right] - \mathbb{E}\left[Y_0|X=x, D=0\right]\right] \\ & = (\alpha + \gamma_t + \gamma_d + \tau_{DID} + x\beta) - (\alpha - \gamma_d - x\beta) \\ & - \left[(\alpha + \gamma_t + x\beta) - (\alpha + x\beta)\right] \\ & = \gamma_t + \tau_{DID} - \gamma_t = \tau_{DID}.\end{aligned} \tag{4.15}$$

Violations of the CT Assumption

In applied work with multiple time periods, it is common to do **placebo tests** or impose **group-specific time trends** as a specification test:

$$Y_i = \alpha + \gamma_t \cdot T_i + \gamma_d \cdot D_i + \tau_{DID} \cdot T_i D_i + \gamma_{dt} \cdot time_t D_i + \epsilon_i. \tag{4.16}$$

This implicitly changes identifying assumption. Instead of **common trend**, we now get **parallel growth**:

$$\mathbb{E}\left[\Delta\left(Y^0_{t*+1} - Y^0_{t*}\right) \mid X, D = 1\right] = \mathbb{E}\left[\Delta\left(Y^0_{t*+1} - Y^0_{t*}\right) \mid X, D = 0\right]. \tag{4.17}$$

One better alternative is to estimate a fully flexible model to derive appropriate identifying assumption:

$$Y_i = \alpha + \sum_{v=t_2}^{T} \gamma_v \cdot \mathbb{1}\left(t = v\right) + \gamma_d \cdot D + \sum_{v=t_2}^{T} \gamma_{dv} \cdot \mathbb{1}\left(t = v\right) D + \epsilon_i, \tag{4.18}$$

where tests on γ_{dv} can be used to assess the common trend assumption.

Problems with Standard Errors

Recall : Grouped residuals inflate **standard errors**.

Consider the simple bivariate case

$$Y_{ig} = \alpha + \beta x_{ig} + e_{ig}, \tag{4.19}$$

where there are G groups and common group errors:

$$e_{ig} = v_g + \eta_{ig}. \tag{4.20}$$

Component v_g captures that group members are exposed to the same **environment**: classroom, teacher, weather...

The **intraclass correlation coefficient** is given by

$$\rho_e = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2}, \tag{4.21}$$

...but the OLS estimator assumes iid residuals ($v_g = 0 \rightarrow \rho_e = 0$).

The Moulton Factor

The Moulton factor: ratio between clustering and iid variance.

$$\frac{\mathbb{V}_{cluster}\left(\hat{\beta}\right)}{\mathbb{V}_{iid}\left(\hat{\beta}\right)} = 1 + \left[\frac{\mathbb{V}\left(n_g\right)}{\bar{n}} + \bar{n} - 1\right]\rho_x\rho_e \tag{4.22}$$

where

$$\rho_x = \frac{\sum_g \sum_j \sum_{i \neq j}\left(x_{ig} - \bar{x}\right)\left(x_{jg} - \bar{x}\right)}{\mathbb{V}\left(x_{ig}\right)\sum_n n_g\left(n_g - 1\right)}. \tag{4.23}$$

Hence, the standard errors get inflated whenever

- Intraclass correlation is high (ρ_e).
- Average group size is large ($\bar{n}$).
- High intraclass correlation also in x_{ig} (ρ_x).

Recall also the subsection on clustering.

By design these scenarios are often appear in a **DID setting**. See *Bertrand, Duflo and Mullainnathan (QJE, 2004)* for an investigation of the problem based on surveying published papers and a Monte Carlo study. Not correcting standard errors many **placebo treatment effects are significant**.

Standard Errors in DID

According to **Bertrand et al. (2004)**, three factors make serial correlation an especially important issue in the DID context: often **long-time series** used, **dependent variable** often highly positively serially correlated, treatment variables **changes very little within a state** over time (often intervention stays on).

Proposed solutions:

- Some papers cluster at the year-state level or aggregate on a year-state level. But this does not help in the Monte Carlo study of Bertrand et al.
- **Block bootstrap**: performs well when number of states are large enough, difficult to implement.
- **Aggregate data into pre- and postintervention** (adaption if they occur at different times) and adjust t-statistics for small number of observations (works well even with small number of states, e.g. 10).
- Allow for an **unrestricted covariance over time within states** (works well when number of states is large, e.g. 50). That means the **clustering** should occur on the **state** assuming error independence across state. Not on state-year interaction since e-g- the error for California in 2010 is likely correlated with error for California in 2009.)

DiD Example

Example 4.5 (Gathmann and Sass, JOLE, 2018: Taxing Childcare: Effects on Childcare Choices, Family Labor Supply and Children)

Table 3: Effect of Home Care Subsidy on Female Labor Supply

	Labor Force Participation (Year of Eligibility)		Labor Force Participation (Year after Eligibility)		Labor Force Participation (2 Years after Eligibility)		Hours Worked (Year of Eligibility)		Full-time Employment (Year of Eligibility)		In School (Year of Eligibility)	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment Dummy	0.010	-0.006	-0.022	-0.044*	-0.041	-0.048*	2.935	3.215	0.151	0.164	-0.054***	-0.056***
	[0.041]	[0.047]	[0.021]	[0.017]	[0.025]	[0.020]	[1.827]	[1.868]	[0.079]	[0.083]	[0.009]	[0.009]
Observations	2,660	2,660	2,616	2,616	2,583	2,583	1,839	1,839	1,793	1,793	2,657	2,657
R Squared	0.118	0.133	0.116	0.177	0.128	0.195	0.048	0.053	0.025	0.030	0.105	0.107
Implied Elasticity				-0.14		-0.15						

SOURCE. – Micro Census (2005-2010).

NOTE. – The table reports the effect of the home care subsidy on female labor supply at the extensive and intensive margin as well as the probability of attending a (general or vocational) school or university. The sample is restricte to women aged 18-45 living in East Germany between 2005 and 2010 with at least one 2 years-old child in the household who have worked some time during their life. The treatment variable is the interaction effect of a parent with a two year-old child living in Thuringia (where the home care subsidy was introduced) and an indicator for the post-policy period (after July of 2006). All specifications include age, marital status, foreign citizenship and education of the parent as well as household size and number of children in the household. To control for aggregate economic conditions, we further include state unemployment and GDP growth rates (linear and squared terms), state and year fixed effects as well as state-specific trends. Even columns adjust for the number of preschool children and newborn children in the household. All standard errors are clustered at the state level.

* $p<0.1$.
** $p<0.05$.
*** $p<0.01$.

DiD Example

Example 4.6 (Gathmann and Sass, JOLE, 2018)

7.4 Estimation of Standard Errors

Our empirical strategy relies on policy changes in a single state which raises the question of how to compute correct standard errors. Our main results are based on standard errors that are clustered at the state level. However, this approach does not account for the small number of clusters. Table A7 reports alternative approaches to calculate standard errors. We rerun variants of equation (1) with standard errors clustered at the state-year level. Further, we include separate state clusters for the pre- and post-policy period to allow for breaks in the temporal dependence of the error term over time. Finally, clustering at the state level might also be too coarse to capture different local market conditions. We therefore use geographic information in the Micro Census to cluster roughly at the level of the county. The standard errors in table A7 are often larger than in our baseline such that some coefficients lose statistical significance. Finally, we implement a wild bootstrap procedure to estimate standard errors with state-dependent errors and a small number of clusters (Miller, Cameron, and Gelbach 2008). This procedure generates p values similar to the baseline and hence confirms our qualitative conclusions.

Difference-in-differences

Example 4.7 (Your own research)

Have you ever estimated a DiD-model? Which problems did you worry about?