PS690 Introduction to Regression Methods Spring 2026

Lecture 8 — Panel Data

Jiawei Fu, Duke University *Scribe: Jiawei Fu*

1 Overview

Economists traditionally use the term panel data to refer to data structures consisting of observations on individuals for multiple time periods. Other fields such as statistics typically call this structure longitudinal data. Panel data are central in both econometrics and causal inference because repeated observations on the same units allow us to study dynamics and to control for certain kinds of unobserved heterogeneity.

2 Framework

It is standard to index observations by both individual i and time period t, so that Y_{it} denotes the value of a variable for individual i in period t. We index individuals as $i = 1, ..., N$ and time periods as $t = 1, ..., T$. Thus, N is the number of individuals in the panel and T is the number of time periods.

When observations are available for all individuals over the same time periods, we say that the panel is balanced. In that case, each individual contributes exactly T observations and the total number of observations is $n = NT$. When different individuals are observed for different sets of time periods, we say that the panel is unbalanced. Unbalanced panels are very common in practice. They do not create conceptual problems, but they do make the notation more cumbersome.

The set of time periods observed for individual i is denoted by S_i, so individual-specific sums over time are written as $\sum_{t \in S_i}$. We write Y_i for the $T_i \times 1$ vector obtained by stacking the observations Y_{it} for $t \in S_i$ in chronological order. Similarly, X_i denotes the corresponding $T_i \times k$ matrix of regressors. We will also sometimes use full-sample matrix notation, writing $Y = (Y_1', ..., Y_N')'$ as the $n \times 1$ stacked outcome vector and $X = (X_1', ..., X_N')'$ as the full regressor matrix.

3 Pooled Regression

The simplest model in panel regression is pooled regression

$$Y_{it} = X_{it}'\beta + e_{it} \quad \mathbb{E}\left[X_{it}e\right] = 0$$

The standard estimator of β in the pooled regression model is least squares, which can be written as

$$
\begin{aligned}
\hat{\beta}_{pool} &= (\sum_{i=1}^{N}\sum_{t\in S_i} X_{it}X_{it}')^{-1}(\sum_{i=1}^{N}\sum_{t\in S_i} X_{it}Y_{it}') \\
&= (\sum_{i=1}^{N} X_iX_i')^{-1}(\sum_{i=1}^{N} X_iY_i') \\
&= (X'X)^{-1}(X'Y)
\end{aligned}
$$

But pooled regression does not yet exploit the main advantage of panel data. Panel structure becomes especially useful when we want to account for unobserved heterogeneity across units.

4 One-Way Error Component Model

For identification, we often worry about unobserved variables. Here we use c to denote such unobserved heterogeneity. In regression, if any observable regressor X_j is correlated with c, then absorbing c into the error term can create serious omitted-variable bias. Without additional information, we generally cannot consistently estimate β.

Panel data give us an additional dimension—time—that can help us control for certain unobserved confounders. Suppose the unobserved component is individual-specific, or time-invariant, and denote it by c_i. Then the model becomes

$$Y_{it} = X_{it}'\beta + c_i + u_{it}$$

We call c_i a **random effect** when $Cov(X_i, c_i) = 0$, meaning that the unobserved effect is uncorrelated with the observed explanatory variables. When $Cov(X_i, c_i) \neq 0$, we use **fixed-effects** methods. If we explicitly model the dependence between c_i and X_i, then we are working in a **correlated random effects** framework. Practically, the key difference is that the fixed-effects approach leaves the dependence between c_i and X_i unrestricted, whereas the correlated-random-effects approach imposes some structure on it.

4.1 Different Exogeneity Assumptions

There are several different exogeneity assumptions we can impose for panel data. The first one is strict exogeneity assumption: $\mathbb{E}[Y_{it}|X_{i1},...,X_{iT}] = \mathbb{E}[Y_{it}|X_{it}]$. It means that, once X_{it} and c_i are controlled for, X_{is} has no partial effect on Y_{it} for $s \neq t$. This assumption restricts how the expected value of Y_{it} can depend on explanatory variables in other time periods. It excludes lagged dependent variables $Y_{i,t-1}$ from X_{it}; otherwise, u_{it} would be predictable given $X_{i,t+1}$ (because $X_{i,t+1}$ is $Y_{i,t}$ now and it is a function of u_{it}.). Another way is to write $\mathbb{E}[u_{it}|X_{i1},...,X_{iT}] = 0$. u_{it} is uncorrelated with all explanatory variables in all time periods, including future time periods.

This assumption, in turn, implies that explanatory variables in each time period are uncorrelated with the idiosyncratic error in each time period: $\mathbb{E}[X_{is}'u_{it}] = 0, \quad s,t = 1,...,T$. A stronger assumption is $\mathbb{E}[u_{it}|X_i, c_i] = 0$.

The second one is sequential exogeneity: $\mathbb{E}[u_{it}|X_{it}, X_{i,t-1},...,X_{i1}] = 0$. Sequential exogeneity implies that X_{is} is uncorrelated with u_{it} for all $s \leq t$, but it leaves unrestricted the relationship between X_{is} and u_{it} for $s > t$. Intuitively, once we condition on current and past explanatory variables, no additional lags are needed to explain the conditional expectation of Y_{it}.

The last one it the zero contemporaneous correlation: $\mathbb{E}\left[X_{it}'u_{it}\right] = 0, t = 1, ..., T$, or say contemporaneous exogeneity $\mathbb{E}\left[u_{it}|X_{it}\right] = 0$. It only restricts the relationship between the disturbance and explanatory variables in the same time period. Contemporaneous exogeneity says nothing about the relationship between X_{is} and u_{it} for any $s \neq t$

Consider the case $X_t = (1, Y_{t-1})$. Contemporaneous exogeneity only requires the conditional mean $\mathbb{E}\left[Y_t|Y_{t-1}\right]$ to be linear in Y_{t-1}. Sequential exogeneity requires $\mathbb{E}\left[Y_t|Y_{t-1}, Y_{t-2}, ..., Y_1\right] = \mathbb{E}\left[Y_t|Y_{t-1}\right]$, so that one lag is sufficient for the dynamic conditional mean. But strict exogeneity must fail, because future regressors contain future lags of Y, which are mechanically related to the current disturbance.

Strict exogeneity can fail even if X_t does not contain lagged dependent variables. Consider a model relating poverty rates to welfare spending per capita at the city level. A finite distributed lag (FDL) model is

$$poverty_t = \beta_0 + \beta_1 welfare_t + \beta_2 welfare_{t-1} + \beta_3 welfare_{t-2} + u_t$$

Contemporaneous exogeneity holds if we assume that $\mathbb{E}\left[poverty_t|welfare_t, welfare_{t-1}, welfare_{t-2}\right]$ is linear. Whether sequential exogeneity holds or not depends on whether two-year lag is sufficient to capture lagged effects of welfare spending on poverty rates. Strict exogeneity generally fails if welfare spending reacts to past poverty rates. For example,

$$welfare_t = \eta_t + \rho_1 poverty_{t-1} + r_t$$

Generally, strict exogeneity is violated if $\rho_1 \neq 0$, because $welfare_{t+1}$ depends on past poverty and therefore indirectly on u_t, while X_{t+1} contains $welfare_{t+1}$.

4.2 Identification

Now, we want to show β is identified. A natural idea is to remove c_i. We can achieve this by **within transformation**.

Define the mean of a variable for a given individual as

$$\overline{Y}_i = \frac{1}{T_i} \sum_{t \in S_i} Y_i$$

We call this the individual-specific mean because it is computed separately for each individual. Some authors instead call it the time average, since it averages observations over time for a given unit. Subtracting the individual-specific mean from the variable we obtain the deviations

$$\dot{Y}_{it} = Y_{it} - \overline{Y}_i$$

This is known as the within transformation. We also refer to $\dot{Y}_{it}$ as the demeaned values. Note that

$$\dot{Y}_i = Y_i - 1_i \overline{Y}_i = M_i Y_i$$

where $M_i = I_i - 1_i(1_i' 1_i)^{-1} 1' 1_i$ is the residual maker (individual-specific demeaning operator).

Similarly, for the regressors we define

$$\dot{X}_{it} = X_{it} - \overline{X}_i$$

Go back to the regression model. Taking individual-specific averages we obtain

$$\overline{Y}_i = \overline{X}_i'\beta + c_i + \overline{u}_i$$

Subtracting this expression from the original equation, we obtain

$$\dot{Y}_i = \dot{X}_i'\beta + \dot{u}_i$$

The key is that unit fixed effect c_i is eliminated.

Another consequence, however, is that all time-invariant regressors are also eliminated. That is, if the original model had included any regressors $X_{it} = X_i$ which are constant over time for each individual then for these regressors the demeaned values are identically 0. What this means is that it will be impossible to estimate (or identify) a coefficient on any regressor which is time invariant. This is not a consequence of the estimation method but rather a consequence of the model assumptions. In other words, if the individual effect u_i has no known structure then it is impossible to disentangle the effect of any time-invariant regressor X_i. The two have observationally equivalent effects and cannot be separately identified.

Now, we can obtain the fixed-effects or within estimator of β:

$$\begin{aligned}\hat{\beta}_{fe} &= (\sum_{i=1}^{N} \dot{X}_i{}'\dot{X}_i)^{-1}(\sum_{i=1}^{N} \dot{X}_i'\dot{Y}_i) \\ &= (\sum_{i=1}^{N} X_i'M_iX_i)^{-1}(\sum_{i=1}^{N} X_i'M_iY_i)\end{aligned}$$

This estimator implicitly assumes that the matrix $\sum_{i=1}^{N} \dot{X}_i'\dot{X}_i$ is full rank. This requires that each regressor has time variation for at least some individuals in the sample. Under the strict exogeneity condition $\mathbb{E}\left[u_{it}|X_i\right] = 0$, the within estimator is unbiased conditional on the regressors and is consistent as N grows.

The within transformation is not the only transformation which eliminates the individual-specific effect. Another important transformation which does the same is first-differencing. It is defined by $\Delta Y_{it} = Y_{it} - Y_{i,t-1}$. We can also define a matrix differencing operator

$$D_i = \begin{bmatrix} -1 & 1 & 0 & ... & 0 & 0 \\ 0 & -1 & 1 & ... & 0 & 0 \\ \vdots & & & & & \vdots \\ 0 & 0 & 0 & ... & -1 & 1 \end{bmatrix}$$

so that $\Delta Y_i = D_iY_i$. Therefore, we obtain

$$\Delta Y_i = \Delta X_i\beta + \Delta u_i$$

We obtain the differenced estimator.

$$\hat{\beta}_{\Delta} = (\sum_{i=1}^{N} \Delta X_i' \Delta X_i)^{-1} (\sum_{i=1}^{N} \Delta X_i' \Delta Y_i)$$
$$= (\sum_{i=1}^{N} X_i' D_i' D_i X_i)^{-1} (\sum_{i=1}^{N} X_i' D_i' D_i Y_i)$$

For $T = 2$, $\hat{\beta}_{fe} = \hat{\beta}_{\Delta}$.

A third way to estimate the fixed-effects model is to run least squares of Y_{it} on X_{it} and a full set of individual dummies. This dummy-variable approach allows us to recover the individual fixed effects directly. It is algebraically equivalent to the within estimator, so we omit the proof here. Some software reports an intercept; in this setting, it is usually interpreted as the average of the estimated unit effects. When N is large, the within transformation is often preferable because it is much more memory-efficient than including a full set of dummies.

4.3 Asymptotics

The asymptotics requires that $N \to \infty$. This approximation is appropriate for the context of a large number of individuals. We could alternatively derive an approximation for the case where both N and T diverge to infinity but this would not be a stronger result.

Fixed effects regression is effectively estimating $N + k$ coefficients (k slope and N fixed effect) Thus the number of estimated parameters is diverging to infinity at the same rate as sample size yet the the estimator obtains a conventional mean-zero sandwich-form asymptotic distribution.

Also, asymptotics assume that the observations are independent across individuals i. This is commonly used for panel data asymptotic theory. An important implied restriction is that it means that we exclude from the regressors any serially correlated aggregate time series variation.

We focus on the balanced case. For consistency, by WLLN,

$$\frac{1}{N} \sum_{i=1}^{N} \dot{X}_i' \dot{X}_i \to \mathbb{E}[\dot{X}_i' \dot{X}_i]$$

and

$$\mathbb{E}[\dot{X}_i' u_i] = \sum_{t=1}^{T} \mathbb{E}[\dot{X}_{it}' u_{it}] = \sum_{t=1}^{T} \mathbb{E}[X_{it} u_{it}] - \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbb{E}[X_{is} u_{it}] = 0$$

For asymptotic normality, by CLT,

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \dot{X}_i' u_i \to N(0, \Omega_T)$$

where $\Omega_T = \mathbb{E}[\dot{X}_i' u_i u_i' \dot{X}_i]$, and thus

$$\sqrt{N}(\hat{\beta}_{fe} - \beta) = (\frac{1}{N} \sum_{i=1}^{N} \dot{X}_i' \dot{X}_i)^{-1} (\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \dot{X}_i' u_i) \to (\mathbb{E}[\dot{X}_i' \dot{X}_i])^{-1} N(0, V_{\beta})$$

where $V_\beta = \mathbb{E}\,[\dot{X}_i'\dot{X}_i])^{-1}\Omega_T\mathbb{E}\,[\dot{X}_i'\dot{X}_i])^{-1}$.

If we assume idiosyncratic errors are homoskedastic and serially uncorrelated: $\mathbb{E}\,[u_{it}^2|X_i] = \sigma_u^2$ and $\mathbb{E}\,[u_{it}u_{is}|X_i] = 0$, then

$$\hat{V}_{fe}^0 = \hat{\sigma}_u^2(\dot{X}'\dot{X})^{-1}$$

with

$$\hat{\sigma}_u^2 = \frac{1}{n-N-k}\sum_{i=1}^{N}\sum_{t\in S_i}\hat{u}_{it}^2$$

where $\hat{u}_{it} = \dot{Y}_{it} - \dot{X}_{it}\hat{\beta}_{fe}$ A covariance matrix estimator which allows u_{it} to be heteroskedastic and serially correlated across t is the cluster-robust covariance matrix estimator, clustered by individual

$$\hat{V}_{fe}^{cluster} = (\dot{X}'\dot{X})^{-1}(\sum_{i=1}^{N}\dot{X}_i'\hat{u}_i\hat{u}_i'\dot{X}_i)(\dot{X}'\dot{X})^{-1}$$

Some adjustment is multiple $\frac{N}{N-1}\frac{n-1}{n-N-k}$

For the intermediate case where u_{it} is heteroskedastic but serially uncorrelated, we omit here.

For FD estimator, it has similar results. We omit here. FE and FE estimator will be efficient under different homoskedastic error assumption. If we maintain contemporaneous exogeneity, then both of them are inconsistent, the bias of FE estimator shrinks to zero at the rate $\frac{1}{T}$.

5 Two-Way Error Components

In general we expect that economic agents will experience common shocks during the same time period. For example, business cycle fluctuations, inflation, and interest rates affect all agents in the economy. Therefore it is often desirable to include time effects in a panel regression model. We could add time trend

$$Y_{it} = X_{it}'\beta + \gamma t + c_i + u_{it}$$

We can also allow for individual-specific time trends

$$Y_{it} = X_{it}'\beta + \gamma_i t + c_i + u_{it}$$

We could also consider the most flexible specification where the trend is allowed to take any arbitrary shape (previously, we assume linearity), but we require that it is common rather than individual-specific. The model we consider is the two-way error component model

$$Y_{it} = X_{it}'\beta + v_t + c_i + u_{it}$$

where v_t is an unobserved time-specific effect.

Let S_t be the set of individuals i for which the observation t is included in the sample and let N_t be the number of these individuals. Now, to eliminate two fixed effects, we need define time-specific mean at time t

$$\tilde{Y} = \frac{1}{N_t}\sum_{i\in S_i} Y_{it}$$

For the case of balanced panels the two-way within transformation is

$$\ddot{Y}_{it} = Y_{it} - \overline{Y}_i - \tilde{Y}_t + \overline{Y}$$

where $\overline{Y} = \frac{1}{n}\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it}$ is the full-sample mean. Let me explain the intuition. If $Y_{it} = v_t + c_i + u_{it}$, then $\overline{Y}_i = \overline{v} + c_i + \overline{u}_i$, and $\tilde{Y}_t = v_t + \overline{u} + \tilde{u}_t$, and $\overline{Y} = \overline{v} + \overline{c} + \overline{u}$. Hence

$$\begin{aligned}\ddot{Y}_{it} &= v_t + u_i + u_{it} - (\overline{v} + c_i + \overline{u}_i) - (v_t + \overline{u} + \tilde{u}_t) + \overline{v} + \overline{c} + \overline{u} \\ &= u_{it} - \overline{u}_i - \tilde{u}_t + \overline{u} \\ &= \ddot{u}\end{aligned}$$

so the individual and time effects are eliminated.

Thus, we obtain two-way within transformation

$$\ddot{Y}_{it} = \ddot{X}_{it}'\beta + \ddot{u}_{it}$$

If the two-way within estimator is used then the regressors X_{it} cannot include any time-invariant variables X_i or common time series variables X_t. Both are eliminated by the two-way within transformation. Coefficients are only identified for regressors which have variation both across individuals and across time.

6 IVs in Panel Data

For $Y_{it} = X_{it}'\beta + c_i + u_{it}$, $\mathbb{E}[X_{it}u_{it}] \neq 0$, we need IVs. Let Z_{it} be an $l \times 1$ instrumental variable where $l \geq k$. In the first stage,

$$X_{it} = \Gamma Z_{it} + W_i + \xi_{it}$$

Applying the within transformation to the reduced form we find

$$\dot{X}_{it} = \gamma \dot{Z}_{it} + \dot{\xi}_{it}$$

If there is no time-variation in the within-transformed instruments or there is no correlation between the instruments and the regressors after removing the individual-specific means, then the coefficient Γ will be either not identified or singular. In either case the coefficient β will not be identified.

Therefore, we need $\mathbb{E}[\dot{Z}_i'\dot{Z}_i] > 0$ and $rank(\mathbb{E}[\dot{Z}_i'\dot{X}_i]) = k$. In other words, the instruments must exhibit within-individual variation over time.Then,

$$\hat{\beta}_{2sls} = (\dot{X}'\dot{Z}(\dot{Z}'\dot{Z})^{-1}\dot{Z}'\dot{X})^{-1}(\dot{X}'\dot{Z}(\dot{Z}'\dot{Z})^{-1}\dot{Z}'\dot{Y})$$

7 Dynamic Panel

What if only sequential exogeneity holds? One important example is a dynamic panel model with a p^{extth}-order autoregression, additional regressors, and a one-way error-component structure:

$$Y_{it} = \alpha_1 Y_{i,t-1} + ... + \alpha_p Y_{i,t-p} + X_{it}'\beta + c_i + u_{it}$$

For example, in the model

$$Y_{it} = \alpha_1 Y_{i,t-1} + c_i + u_{it}.$$

The condition $\mathbb{E}\left[u_{it}|Y_{i,t-1},...,Y_{i1},c_i\right] = 0$ means that, once we control for c_i, one lag of Y is sufficient to characterize the conditional mean dynamics.

More generally, consider

$$Y_{it} = X_{it}'\beta + c_i + u_{it},$$

with sequential exogeneity. We again eliminate the fixed effect by first differencing:

$$\Delta Y_{it} = \Delta X_{it}'\beta + \Delta u_{it}.$$

Sequential exogeneity implies $\mathbb{E}\left[X_{is}u_{it}\right] = 0$ for $s = 1,...,t$ and $t = 1,...,T$, which in turn implies

$$\mathbb{E}\left[X_{is}\Delta u_{it}\right] = 0$$

for $s = 1,...,t-1$. Therefore, at time t, lagged levels such as $(X_{i1}, X_{i2}, ..., X_{i,t-1})$ can serve as valid instruments for the differenced equation. This is the basic logic behind GMM estimators for dynamic panel data.

8 Panel Data under Causal Inference

In this section, we study event-study designs in which all units begin in the control condition and, once treated, remain treated thereafter.

We observe a panel indexed by $i = 1,...,N$ and $t = 1,...,T$. For each (i,t) pair we observe an outcome Y_{it} and a treatment indicator $D_{it} \in \{0,1\}$. The event-study setup assumes absorbing treatment: once treatment switches on, it never switches off. Formally, $D_{it} \leq D_{it'}$ for all $t \leq t'$.

A **block-adoption design** has a common event time: units either begin treatment right after that common date or never receive treatment. A **staggered-adoption design** allows each unit to have its own treatment date, or to remain untreated forever.

The following materials are from Roth et al. (2023).

8.1 DiD: 2 by 2

Let D_i denote treatment-group status, where $D_i = 1$ for treated units and $D_i = 0$ for untreated units. Consider the simplest 2×2 DiD design with two periods. Let $Y_{it}(0,0)$ be the potential outcome for unit i at time t if it remains untreated in both periods. Similarly, let $Y_{it}(0,1)$ be the potential outcome if the unit is untreated in the first period and treated in the second. Because no one is treated in period 1, the first entry is fixed, and we can simplify the notation by writing only the second argument.

They are potential outcomes; we can never observe them simultaneously. For example, for the untreated units, we never observe $Y_{it}(1)$ because this is the outcome at time t when unit i is treated in the second period. We can only observe the $Y_{it}(0)$.

The observed Y_{it} is represented as

$$Y_{it} = D_i Y_{it}(1) + (1 - D_i) Y_{it}(0).$$

The target parameter is

$$ATT(t) = \mathbb{E}\left[Y_{it}(1) - Y_{it}(0) | D_i = 1\right]$$

It compares, for the treated group, the observed average post-treatment outcome $\mathbb{E}\left[Y_{it}(1)|D_i = 1\right]$ with the counterfactual average outcome that would have been observed had those same units not been treated, $\mathbb{E}\left[Y_{it}(0)|D_i = 1\right]$.

To identify this ATT, we require two assumptions.

Assumption 1 (No-Anticipation)**.** *For all treated unit i and all pre-treatment periods t, $Y_{it}(1) = Y_{it}(0)$.*

This means that unit i does not anticipate treatment and therefore does not change its period 1 outcome in response to future treatment. Thus, for treated units before treatment begins, the observed value equals the untreated potential outcome: $Y_{i1} = Y_{i1}(0)$. With this assumption, $ATT(1) = 0$.

The second assumption is known as parallel trend. It states that in the absence of treatment, the average outcome evolution is the same among treated and comparison groups.

Assumption 2 (2×2 Parallel Trends)**.** *The average change of $Y_{i2}(0)$ from $Y_{i1}(0)$ is the same between treated and untreated groups:*

$$\mathbb{E}\left[Y_{i2}(0)|D_i = 1\right] - \mathbb{E}\left[Y_{i1}(0)|D_i = 1\right] = \mathbb{E}\left[Y_{i2}(0)|D_i = 0\right] - \mathbb{E}\left[Y_{i1}(0)|D_i = 0\right]$$

Under parallel trends, we can construct the missing counterfactual term in $ATT(2)$ as

$$\mathbb{E}\left[Y_{i2}(0)|D_i = 1\right] = \mathbb{E}\left[Y_{i1}(0)|D_i = 1\right] + (\mathbb{E}\left[Y_{i2}(0)|D_i = 0\right] - \mathbb{E}\left[Y_{i1}(0)|D_i = 0\right]).$$

Therefore, ATT(2) is identified:

$$\begin{aligned}
ATT(2) &= \mathbb{E}\left[Y_{i2}(1) - Y_{i2}(0)|D_i = 1\right] \\
&= \mathbb{E}\left[Y_{i2}(1)|D_i = 1\right] - (\mathbb{E}\left[Y_{i1}(0)|D_i = 1\right] + (\mathbb{E}\left[Y_{i2}(0)|D_i = 0\right] - \mathbb{E}\left[Y_{i1}(0)|D_i = 0\right]) \\
&= \underbrace{(\mathbb{E}\left[Y_{i2}(1)|D_i = 1\right] - (\mathbb{E}\left[Y_{i1}(0)|D_i = 1\right])}_{\text{average change for } D_i = 1} - \underbrace{(\mathbb{E}\left[Y_{i2}(0)|D_i = 0\right] - \mathbb{E}\left[Y_{i1}(0)|D_i = 0\right])}_{\text{average change for } D_i = 0} \\
&= (\mathbb{E}\left[Y_{i2}(1)|D_i = 1\right] - (\mathbb{E}\left[Y_{i1}(1)|D_i = 1\right]) - (\mathbb{E}\left[Y_{i2}(0)|D_i = 0\right] - \mathbb{E}\left[Y_{i1}(0)|D_i = 0\right]) \\
&= (\mathbb{E}\left[Y_{i2}|D_i = 1\right] - (\mathbb{E}\left[Y_{i1}|D_i = 1\right]) - (\mathbb{E}\left[Y_{i2}|D_i = 0\right] - \mathbb{E}\left[Y_{i1}|D_i = 0\right])
\end{aligned}$$

In the second to the last row, we apply no anticipation. Therefore, $ATT(2)$ can be uniquely expressed as population moments.

Equivalently, we can also run TWFE model:

$$Y_{it} = c_i + v_t + \beta D_{i,t} + u_{it}$$

or

$$Y_{it} = c_g + v_t + \beta D_{i,t} + u_{it}$$

where g denotes group.

For inference, one often clusters standard errors at the unit level, although the correct choice ultimately depends on the sampling and assignment design. Under regularity conditions, the DiD estimator is asymptotically normal. We can also extend the design to allow for covariates:

Assumption 3 (2 × 2 Conditional Parallel Trends)**.** *The average change of* $Y_{i2}(0)$ *from* $Y_{i1}(0)$ *is the same between treated and untreated groups:*

$$\mathbb{E}\left[Y_{i2}(0) - Y_{i1}(0)|D_i = 1, X_i\right] = \mathbb{E}\left[Y_{i2}(0) - Y_{i1}(0)|D_i = 0, X_i\right]$$

Then, $att(2)$ is identified:

$$\begin{aligned}
ATT(2) &= \mathbb{E}\left[Y_{i2}(1) - Y_{i2}(0)|D_i = 1\right] \\
&= \mathbb{E}\left[Y_{i2}|D_i = 1\right] - \mathbb{E}\left[\mathbb{E}\left[Y_{i2}(0)|X_i, D_i = 1\right]|D_i = 1\right] \\
&= \mathbb{E}\left[Y_{i2}|D_i = 1\right] - \mathbb{E}\left[\mathbb{E}\left[Y_{i1}(0)|X_i, D_i = 1\right] + \mathbb{E}\left[\Delta Y_{i2}(0)|X_i, D_i = 0\right]|D_i = 1\right] \\
&= \mathbb{E}\left[\Delta Y_{i2}|D_i = 1\right] - \mathbb{E}\left[\mathbb{E}\left[\Delta Y_{i2}|X_i, D_i = 0\right]|D_i = 1\right]
\end{aligned}$$

if add X in the TWFE,

$$Y_{it} = c_i + v_t + \beta D_{it} + \gamma X'_{it} + u_{it}$$

Caetano and Callaway (2024) tackle exactly this question. They show that β equals a weighted average of conditional average treatment effects, defined as $ATT_{x_k}(2) = \mathbb{E}\left[Y_{i2}(1) - Y_{i,2}(0)|D_i = 1, X_i = x_k\right]$, with weights that may not be convex, plus three bias terms reflecting misspecification either in the set of control variables or the fact that they are included linearly.

The weighting results suggest that even when the covariates are correctly selected, measured, and added with the correct functional form, β could be negative even when $ATT_{x_k}(2)$ is positive for all values of covariates.

However, if we assume additional assumption that the treatment effects across covarite strata are constant, it works.

8.2 DiD: 2 by T

Now, we generalize the frameworks to multiple time periods, but still two groups. Suppose when $t \geq g$, treated groups starts to be treated. If we want to obtain ATT(t) for $t \geq g$, apparently, we should strength Parallel Trends to any $t \geq g$.

Assumption 4 (2 × T Parallel Trends)**.** *The average change of* $Y_{i,t}(0)$ *from* $Y_{i,g-1}(0)$ *is the same between treated and untreated units for all post-treatment periods* $t \geq g$*:*

$$\mathbb{E}\left[Y_{i,t}(0)|D_i = 1\right] - \mathbb{E}\left[Y_{i,g-1}(0)|D_i = 1\right] = \mathbb{E}\left[Y_{i,t}(0)|D_i = 0\right] - \mathbb{E}\left[Y_{i,g-1}(0)|D_i = 0\right] \; \forall t \geq g$$

Then, each $ATT(t)$ is identified by a DiD comparison between period $g-1$ and t, $\widehat{ATT}(t) = (\overline{Y}_{D=1,t} - \overline{Y}_{D=1,g-1}) - (\overline{Y}_{D=0,t} - \overline{Y}_{D=0,g-1})$

We can use TWFE to estimate $ATT(t)$: Omitting the treatment interaction for $t = g - 1$ avoids multicollinearity and fixes $g - 1$ as the baseline period for all β_t estimates:

$$Y_{it} = c_i + v_t + \sum_{k=1}^{g-2} \beta_k (I[G_i = g] I(t = k)) + \sum_{k=g}^{T} \beta_k (I[G_i = g] I[t = k]) + u_{it}$$

We can show that $\hat{\beta} = \widehat{ATT}(t)$. Once we have those $ATT(t)$ as building blocks, we could construct $ATT = \frac{1}{T-(g-1)} \sum_{t=g}^{T} ATT(t)$, the average treatment effect in the post-period.

Remark 5. *What if $Y_{it} = c_i + v_t + \beta D_{it} + u_{it}$? We can show that*

$$\beta = (\mathbb{E}[Y|G = g, t \geq g] - \mathbb{E}[Y|G = g, t < g]) - (E[Y|G = \infty, t \geq g] - \mathbb{E}[Y|G = \infty, t < g])$$

So, it implicitly uses all available pre-treatment periods. So far, we have assumed parallel trends only for post-treatment periods, $t \geq g$. Thus, at least implicitly, the TWFE regressions rely on a "different" type of PT assumption! A PT version compatible with TWFE and "DiD-by-hand" is that PT holds for all time periods (both pre-and post-treatment). Under PT for all time periods,

$$\beta = \frac{\sum_{s=g}^{T} ATT(g, s)}{T - g + 1} = \frac{\sum_{e=0}^{T-g} ATT(g, g + e)}{T - g + 1}$$

8.3 Staggered treatment adoption $G \times T$

When treatment start dates can vary across units, we need to allow the potential outcomes, and thus the target parameters and identifying assumptions, to reflect this richer notion of treatment.

We therefore index potential outcomes by the time treatment begins, g: $Y_{it}(g)$, and use $Y_{it}(\infty)$ to denote never-treated potential outcomes. We use G_i to denote each unit's treatment date.

$$Y_{it} = \sum_{g} Y_{it}(g) I[G_i = g]$$

Assumption 6 (No-Anticipation with staggered treatment timing)**.** *For all units i that are eventually treated and all pre-treatment periods t, $Y_{it}(g) = Y_{it}(\infty)$*

This assumption imposes that treatment effects are zero in all pretreatment periods as a consequence of units not acting on the potential knowledge of future treatment dates before they are actually exposed to treatment.

We assume that a "never-treated" group always exists in our staggered DiD setup. If all units are eventually treated, we drop all the data from when the last cohort is treated, so the last treated cohort becomes the "never-treated" cohort, and T here denotes the number of available periods in the subset of the data that we will use in our analysis. This is essentially without loss of generality, because under standard DiD assumptions, we cannot identify any ATT for periods where all units are treated. We also dropped data from units treated in the first available period, as such treatment group does not have any pre-treatment data, preventing us from conducting a DiD analysis.

Define group-time average treatment effects:

$$ATT(g, t) = \mathbb{E}[Y_{it}(g) - Y_{it}(\infty)|G_i = g]$$

This is the average treatment effect of starting treatment at period g relative to never starting it, at time period t, among units that actually started treatment in period g. It is simply a set of event study parameters for treatment timing group g. The unobserved counterfactuals are now untreated potential outcome means for each treatment group in each period, $\mathbb{E}\left[Y_{it}(\infty)|G_i = g\right]$.

There are many possible PT assumptions we can impose.

Assumption 7 (Parallel Trends based on never-treated groups)**.** *For every eventually treated group g and post-treatment time period $t \geq g$*

$$\mathbb{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty)|G_i = g\right] = \mathbb{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty)|G_i = \infty\right]$$

This assumption is the analogue of the parallel-trends condition used in the $2 \times T$ design. It uses never-treated units as the comparison group for all eventually treated units and imposes parallel trends only for future untreated potential outcomes. Because the assumption is formulated only in forward-looking form, the farthest we can go back in pre-treatment periods is $t = g-1$, which serves as the only fully justifiable baseline period. It is straightforward to show that for post-treatment periods,

$$ATT(g,t) = \mathbb{E}\left[Y_{it} - Y_{i,g-1}|G_i = g\right] - \mathbb{E}\left[Y_{it} - Y_{i,g-1}|G_i = \infty\right]$$

Note that it highlights that we are essentially back to a 2×2 design when it comes to learning about $ATT(g,t)$: it leverages data from only two periods, t (post) and $g-1$ (pre), and two treatment groups, $G_i = g$ (treated) and $G_i = \infty$ (comparison).

Assumption 8 (Parallel Trends based on not-yet-treated groups)**.** *For every eventually treated group g, not-yet-treated group g' and time periods t such that $t \geq g$ and $g' > t$*

$$\mathbb{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty)|G_i = g\right] = \mathbb{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty)|G_i = g'\right]$$

Under this assumption, one can use not only the never-treated units but any group of units that are not-yet-treated by time t.

$$ATT(g,t) = \mathbb{E}\left[Y_{it} - Y_{i,g-1}|G_i = g\right] - \mathbb{E}\left[Y_{it} - Y_{i,g-1}|G_i > max\{g,t\}\right]$$

this follows a 2×2 set up: it leverages data from only two periods, t (post) and $g-1$ (pre), and two treatment groups, $G_i = g$ (treated), and $G_i > t$ (comparison).

Actually, for any not-yet-group $g' > t$, for $t \geq g$,

$$ATT(g,t) = \mathbb{E}\left[Y_{it} - Y_{i,g-1}|G_i = g\right] - \mathbb{E}\left[Y_{it} - Y_{i,g-1}|G_i = g'\right]$$

One can then flexibly combine the different comparison groups by leveraging user-specified or efficiency-oriented weights.

Assumption 9 (Parallel Trends for every period and group)**.** *For every treatment groups g and g' and time periods t,*

$$\mathbb{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty)|G_i = g\right] = \mathbb{E}\left[Y_{it}(\infty) - Y_{it-1}(\infty)|G_i = g'\right]$$

Under this assumption, one can use of any not-yet-treated units as a comparison group, as well as any pre-treatment period as a baseline period. It impose pre-trends parallel so we can use more information.

It is easy to show that, for any pre-treatment period $t_{pre} < g$ and any not-yet-treated group $g' > t$, we can identify the $ATT(g,t)$ for group $g's$ post-treatment periods $t \geq g$ by

$$ATT(g,t) = \mathbb{E}\left[Y_{it} - Y_{i,t_{pre}}|G_i = g\right] - \mathbb{E}\left[Y_{it} - Y_{i,t_{pre}}|G_i = g'\right]$$

which again maps back to the 2×2 DiD setup, as it leverages two periods, t (post) and t_{pre} (pre), and two groups,$G_i = g$ (treated) and $G_i = g'$ (comparison). This representation makes it clear that, in practice, one can use various pre-treatment periods and not-yet-treated comparison groups to characterize the $ATT(g,t)$ under Assumption. An intuitive estimand is

$$ATT(g,t) = \mathbb{E}\left[Y_{it} - \overline{Y}_{i,t\leq g-1}|G_i = g\right] - \mathbb{E}\left[Y_{it} - \overline{Y}_{i,t\leq g-1}|G_i > max\{g,t\}\right]$$

where $\overline{Y}_{i,t\leq g-1} = \sum_{s=1}^{g-1} \frac{Y_{is}}{g-1}$ is the time average of group g pre-treatment periods for each unit i.

In practice, however, there is no general econometrics guarantee that estimators based on this average will be more precise; see, for instance, Harmon (2024). This arises because it is not always optimal to weigh all pre-treatment periods equally when forming estimators. The optimal way depends on the correlation structure of how the outcome changes over time across different comparison groups.

Instead of reporting every $ATT(g,t)$ separately, we can summarize treatment-effect dynamics by aggregating across cohorts at a given event time. One useful parameter is the average group-time treatment effect by event time:

$$ATT(e) = \mathbb{E}\left[ATT(G, G+e)|G+e \in [1,T], G \leq T\right] = \sum_{g\leq\infty} w_{g,e} ATT(g, g+e)$$

where weight $w_{g,e} = I[g+e \leq T]P[G = g|G+e \leq T, G \leq T]$ gives the share of a group $G = g$ among treated units that have been exposed to treatment for exactly e periods.

8.4 Discussion on TWFE

What if we use $Y_{it} = c_i + v_t + \beta D_{it} + u_{it}$? The static specification yields a sensible estimand when there is no heterogeneity in treatment effects across either time or units. Let $\tau_{i,t}(g) = Y_{it}(g) - Y_{it}(\infty)$. Suppose for all units i,$\tau_{i,t}(g) = \tau$ whenever $t \geq g$. This imposes that (1) all units have the same treatment effect, and (2) the treatment has the same effect regardless of how long it has been since treatment started. Then, under a suitable generalization of the parallel trends assumption (e.g. Assumption 8) and no anticipation assumption , the population regression coefficient β is equal to τ.

Suppose first that there is heterogeneity in time since treatment only. That is $\tau_{it}(g) = \sum_{s\geq 0} \tau_s I[t - g = s]$, so all units have treatment effect τ_s in the s-th period after they receive treatment. In this case β corresponds with a potentially non-convex weighted average of the parameters τ_s:$\beta = \sum_s \omega_s \tau_s$ where the weights sum to 1 but may be negative. The possibility of negative weights is concerning because, for instance, all of the τ_s could be positive and yes the coefficient may be negative! In particular, longer-run treatment effects will often receive negative weights. More generally, if treatment effects vary across both time and units, then $\tau_{it}(g)$ may get negative weight in the TWFE estimand for some combinations of t and g.

Let us see the intuition. From FWL, β is equivalent to the coefficient from a univariate regression of Y_{it} on $D_{it} - \hat{D}_{it}$, where $\hat{D}_{it}$ is the predicted value from a regression of D_{it} on the other right-hand side variables. However, a well-known issue with OLS with binary outcomes is that its predictions may fall outside the unit interval. If the predicted value $\hat{D}_{it}$ is greater than 1, then $D_{it} - \hat{D}_{it}$ will be negative even when a unit is treated, and thus that unit's outcome will get negative weight in $\hat{\beta}$:

$$\hat{\beta} = \frac{\sum_{i,t}(D_{it} - \hat{D}_{it})Y_{it}}{\sum_{i,t}(D_{it} - \hat{D}_{it})^2}$$

Denominator is positive. If $D_{it} = 1$, and $(D_{it} - \hat{D}_{it}) < 0$, then $\hat{\beta}$ will be decreasing in Y_{it} even though unit i is treated at period t. But because $Y_{it} = Y_{it}(\infty) + \tau_{it}(g)$, it follows that $\tau_{it}(g)$ gets negative weight in $\hat{\beta}$.

These negative weights will tend to arise for early-treated units in periods late in the sample. Note $\hat{D}_{it} = \overline{D}_i + \overline{D}_t - \overline{D}$. It follows that if we have a unit that has been treated for almost all periods $\overline{D}_i \approx 1$, and a period in which almost all units have been treated $\overline{D}_t \approx 1$, then $\hat{D}_{it} \approx 2 - \overline{D}$, which will be strictly greater than 1 if there is a nontrivial fraction of non-treated units in some period $\overline{D} < 1$. We thus see that $\hat{\beta}$ will tend to put negative weight on τ_{it} for early-adopters in late periods within the sample.

Remark 10. *Goodman-Bacon (2021) shows that TWFE estimator can be expressed as a weighted average of all possible $2 \times D$ DiD comparisons between pairs of groups and time periods during which one group enters treatment and the other does not. de Chaisemartin and D'Haultfoeuille (2020) provide a related decomposition of the TWFE estimand and emphasize another issue: already-treated units are often used as comparison groups for later-treated units. This can further complicate interpretation. : Treated vs. "Never-treated", Early-treated vs. Later-treated, Later-treated vs. Already-treated.*

How about dynamic TWFE?

$$Y_{it} = \alpha_i + \phi_t + \sum_{r \neq 0} I[t - G_i + 1 = r]\beta + \epsilon_{it}$$

where $t - G_i + 1$ is the time relative to treatment, $t - G_i + 1 = 1$ in the first treated period for unit i, and the summation runs over all possible values except for 0.

Unlike the static specification, the dynamic specification yields a sensible causal estimand when there is heterogeneity only in time since treatment. Borusyak and Jaravel (2018) and Sun and Abraham (2021) imply that if $\tau_{it}(g) = \sum_{s \geq 0} \tau_s I[t - g = s]$ so all units have treatment effect τ_s in the s-th period after treatment, then $\beta_s = \tau_s$ under suitable generalizations of the parallel trends and no anticipation assumptions, Sun and Abraham (2021) show, however, that when dynamic treatment effects differ across adoption cohorts, the coefficients from this specification become difficult to interpret. There are two problems. First, as in the static TWFE regression, the coefficient may place negative weight on the treatment effect r periods after treatment for some units. Second, the coefficient β_r can place nonzero weight on treatment effects at other lags $r' \neq r$, creating cross-lag contamination. For example, β_2 may be influenced by treatment effects that occur three periods after treatment for some cohorts.

8.5 Function Form

In principle, parallel trends is a function-form assumption. First, if the identification assumptions hold for the level of Y, they need not hold for monotone transformations of Y. Consider the following example from Lechner (2011): $\ln Y_{it}(0)|D = d \sim N(0, 2d + 2t)$.

Let us check the parallel-trends assumption. If $\ln Y \sim N(\mu, \sigma^2)$, then the mean of Y depends nonlinearly on σ^2. This means parallel trends in logs need not imply parallel trends in levels.

$$\mathbb{E}\left[lnY_{i2}(0)|D_i = 1\right] - \mathbb{E}\left[lnY_{i1}(0)|D_i = 1\right] = 0 - 0 = 0$$
$$\mathbb{E}\left[lnY_{i2}(0)|D_i = 0\right] - \mathbb{E}\left[lnY_{i1}(0)|D_i = 0\right] = 0 - 0 = 0$$

However,

$$\mathbb{E}\left[Y_{i2}(0)|D_i = 1\right] - \mathbb{E}\left[Y_{i1}(0)|D_i = 1\right] = e^3 - e^2$$
$$\mathbb{E}\left[Y_{i2}(0)|D_i = 0\right] - \mathbb{E}\left[Y_{i1}(0)|D_i = 0\right] = e^2 - e$$

A more general case is

$$E[Y_{i,t}(0)|D = d] = \alpha + t\delta_0 + d\gamma + x\beta + tx\lambda_0 + dx\pi$$
$$E[Y_{i,t}(1)|D = d] = \alpha + t\delta_1 + d\gamma + x\beta + tx\lambda_1 + dx\pi \text{ (can ignore this line)}$$

It satisfies parallel trend:

$$\begin{aligned}
&E[Y_{i,2}(0)|D_i = 0] - E[Y_{i,1}(0)|D_i = 0] \\
&= (\alpha + 2\delta_0 + x\beta + 2x\lambda_0) - (\alpha + \delta_0 + x\beta + x\lambda_0) \\
&= \delta_0 + x\lambda_0
\end{aligned}$$

and

$$\begin{aligned}
&E[Y_{i,2}(0)|D_i = 1] - E[Y_{i,1}(0)|D_i = 1] \\
&= (\alpha + 2\delta_0 + \gamma + x\beta + 2x\lambda_0 + x\pi) - (\alpha + \delta_0 + \gamma + x\beta + x\lambda_0 + x\pi) \\
&= \delta_0 + x\lambda_0
\end{aligned}$$

The key is that it does not include interactions between time and treatment status.

References

Lechner, M. (2011). The estimation of causal effects by difference-in-difference methods. *Foundations and trends in econometrics*, 4(3), 165–224.

Roth, J., Sant'Anna, P. H., Bilinski, A., & Poe, J. (2023). What's trending in difference-in-differences? a synthesis of the recent econometrics literature. *Journal of econometrics*, 235(2), 2218–2244.