Difference-in-Differences

Alessandro Caggia

June 2025

Abstract

1. PANEL DATA

1.1. Idea

Now we have panel data, so we can have repeated measures for a given individual.

- **Issue**: If unobserved variables (selection on observables not possible) are correlated with treatment (D) and potential outcomes (Y_{0i}, Y_{1i})[1], unconfoundedness will not hold:

$$(Y_{0i}, Y_{1i}) \not\perp\!\!\!\perp D_i$$

  e.g., smarter people (unobserved IQ, higher (Y_{0i}, Y_{1i})) self-select into a supplementary math course: T will have a higher score than C in the math exam, but this is not driven entirely by the supplementary course but by intrinsic ability.
- **Solution**: The goal of Difference-in-Differences (DID) and panel data models is to leverage panel data to control for permanent unobserved factors (e.g., individual fixed effects): if we have more than one observation per individual, we can account for time-invariant unobserved variables.
  - In a cross section: if someone has a high outcome, is it due to smartness or the treatment? We cannot tell as there are no past observations. Since smartness is unobserved, we cannot control it directly.
  - With repeated observations, we can distinguish the endogenous time-invariant smartness component (pre-existing higher outcomes) from the treatment effect, and use the between-group comparison to disentangle the treatment effect from time effects.
- Go from the individual to the group level: imagine you have two groups. First, you remove the endogenous time-invariant component in the two groups exploiting within-group variation. Second, you exploit between-group variation to estimate the TE.

[1] The problem is not heterogeneous TEs, the problem is selection into treatment. If individuals have heterogeneous TEs, and this heterogeneity is equally represented across treated and control groups, it does not bias the estimatein fact, it is exactly the TE we want to identify.

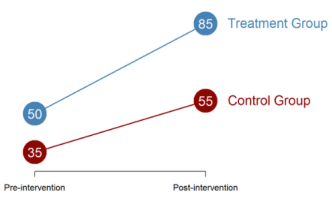


Figure 1. 2 time periods

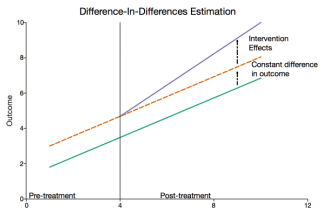


Figure 2. n time periods

1.2. Unconfoundedness

We now assume unconfoundedness holds if we control for a time trend, observed time-varying covariates, and a vector of time-invariant unobserved variables A_i (expansions with respect to our standard model):

$$E(Y_{0it} \mid A_i, X_{it}, t, D_{it}) = E(Y_{0it} \mid A_i, X_{it}, t)$$

1.3. Functional form

- We impose a **linear, additive functional** form for potential outcomes:

$$E(Y_{0it} \mid A_i, X_{it}, t) = \alpha + \lambda_t + A_i'\gamma + X_{it}'\beta$$
$$E(Y_{1it} \mid A_i, X_{it}, t) = E(Y_{0it} \mid A_i, X_{it}, t) + \rho_{it}$$

- Assume that **Treatment Effects are homogeneous and constant over time**: $\rho_{it} = \rho$.
- **The switching equation**:

$$E(Y_{it} \mid A_i, X_{it}, t, D_{it}) \underbrace{=}_{\text{CMI: D out}} E(Y_{0it} \mid A_i, X_{it}, t) + D_{it}E(Y_{1it} - Y_{0it} \mid A_i, X_{it}, t)$$
$$= \alpha + \lambda_t + A_i'\gamma + X_{it}'\beta + \rho D_{it}$$

- The regression equation at the individual level is then (above is Expected Val):

$$Y_{it} = \alpha_i + \lambda_t + X_{it}'\beta + \rho D_{it} + \varepsilon_{it}$$

- where $\alpha_i \equiv \alpha + A_i'\gamma$ is the individual fixed effect (time-invariant heterogeneity (e.g., ability, preferences)). If not controlled for, it creates bias:
  - **Composition**:
    - α: common intercept across individuals.
    - $A_i'\gamma$: a vector of time-invariant unobserved variables.
  - **Identification**: α and $A_i'\gamma$ not separately identified.
    - One intercept absorbs both α and $A_i'\gamma$.
    - Need to impose restrictions (eg. if Ai is normally distributed one could say that alpha is the mean of α_i)
  - **Autocorrelation**: this structure induces serial correlation in $u_{it} = \alpha_i + \varepsilon_{it}$ for a given individual, since α_i is shared across all t. The issue is generated by the $A_i'\gamma$ component, not by the general constant, as the general constant shifts equally all individuals.
- ε_{it} **is the idiosyncratic error**, varying over time and individuals.

1.4. How to get rid of α_i

1.4.1. Random Effects

- **Random Effects**: α_i (unobserved) is moved into the error term, assuming it is uncorrelated with the regressors.
  - We can use pooled OLS, pooling all time periods. Better: for efficiency, use GLS, which accounts for correlation in errors (again, $u_{it} = \alpha_i + \varepsilon_{it}$).
- If errors are correlated with regressors, you will have bias.
- REs are no longer used; they were more common in cases where you have low observations or low computing power.

1.4.2. Fixed Effects and First Differences

- **Fixed Effects**: α_i and λ_t are treated as parameters to estimate (via individual and time dummies). FE allows α_i to correlate with regressors and controls for it;

- Using individual dummies generates a demeaned model (**you are controlling at the individual level, exploiting within-individual (over-time) variation, by removing individual-specific means**):

$$Y_{it} - \overline{Y}_i = \lambda_t - \overline{\lambda} + (X'_{it} - \overline{X}'_i)\beta + \rho(D_{it} - \overline{D}_i) + \varepsilon_{it} - \overline{\varepsilon}_i$$

- The rationale is: you remove the average, so necessarily you remove the unobserved fixed components.
- By the Frisch-Waugh-Lovell Theorem, this is equivalent to:
  1. Regress Y, time FE, X, and D on full individual dummies and get residuals.
  2. Regress Y residuals on X, time FE, and D residuals.
- Alternatively, use First Differences (FD):

$$Y_{it} - Y_{it-1} = \lambda_t - \lambda_{t-1} + (X'_{it} - X'_{it-1})\beta + \rho(D_{it} - D_{it-1}) + \varepsilon_{it} - \varepsilon_{it-1}$$

- **Key facts:**
  - Both FE and FD are unbiased estimators of the treatment effect.
  - When $T = 2$, they yield the same result.
  - With $T > 2$, efficiency differs depending on serial correlation in ε_{it}.
  - with FD you reduce sample size by 50% in FE you lose dfs (1 df for each individual, N dfs lost! use `xtreg ..., fe` to adjust standard errors for lost degrees of freedom.)

1.4.3. Incidental Parameters Problem

- When $N \to \infty$ and T is fixed, it is impossible to consistently estimate the N individual fixed effects as the number of parameters grows with N. Truly, the problem is solved just by large T.
- Incidental problem because it is emerging form thousands of parameters that are not what you want to itentify but are necessary to identify
- When differencing from individual means, we avoid estimating fixed effects directly, thus solving the problem.

1.4.4. FE and Measurement Error

- Fixed effects estimates are highly sensitive, more than OLS, to measurement error, which can lead to strong attenuation bias (maybe try IV).
- **Proof**: Say that $x = x^* + \nu$ and assume that:

$$\text{cov}(x^*_{it}, \varepsilon_{it}) = 0, \quad \text{cov}(\nu_{it}, \nu_{it-1}) = 0$$

  the first assumption rules out endogeneity between the true regressor (hence also the proxy as the proxy is the true plus white noise) and the error, the second assumption is that measurement error is uncorrelated over time (there is no autocorrelation in the measurement error).
- **OLS attenuation bias:**[1]

$$\widehat{\beta}^{OLS} = \beta \cdot \frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_u} = \beta\left(1 - \frac{\sigma^2_\nu}{\sigma^2_{x^*} + \sigma^2_\nu}\right)$$

- **Fixed Effects attenuation bias (with serial correlation):**

$$\widehat{\beta}^{FE} = \beta\left(1 - \frac{\sigma^2_\nu}{(1-\rho_x)\sigma^2_{x^*} + \sigma^2_\nu}\right), \quad \rho_x = \frac{\text{cov}(x^*_{it}, x^*_{it-1})}{\text{var}(x^*_{it})}$$

- ρ_x is the autocorrelation of the true regressor x^*_{it} across time for the same individual. When $\rho_x = 0$, FE and OLS yield the same bias. When $\rho_x \approx 1$, the difference $x^*_{it} - x^*_{it-1}$ is small and the bias from measurement error increases. Why? As you can see ρ_x shrinks the true variability in $\sigma^2_{x^*}$, hence shrinks the signal to noise ratio (there is less signal, that is, information, around! you need more observations!).

2. Difference-in-Differences (DID)

2.1. Basic Setup

- DID is a special case with two time periods ($t = 0$ pre-treatment, $t = 1$ post-treatment) and two groups ($g = 0$ control, $g = 1$ treatment). We have a **within group variation** in treatment status (in the treated group).
- Does not require panel data, repeated cross-sections are sufficient.

2.2. Assumptions

Strong Assumption (which implies treatment effect is homogeneous and constant over time):

$$\mathbb{E}(Y_{0igt} \mid g, t) = \gamma_g + \lambda_t, \quad \mathbb{E}(Y_{1igt} \mid g, t) = \mathbb{E}(Y_{0igt} \mid g, t) + \rho$$

where:

- γ_g: **group fixed effect** (the group version of α_i, we will remove the time varying unobservabels proper of the group!). Captures time-invariant differences between groups (e.g., region, demographics). Visually: the initial vertical gap between the lines.
- λ_t: **time fixed effect**. Captures all time-varying shocks that affect all groups equally. Visually: a shift or slope change that applies equally to all groups over time.

2.2.1. Unconfoudedness and Parallel Trends

- Such a strong assumption **implies unconfoundednes, look at the two expected values above**, the treated outcome is just as the control outcome plus the treatment effect! That is, **In absence of treatment, outcomes would evolve similarly across groups, ie the pre-post change in the treatment group would have followed the same trend as the pre-post change in the control group if the treatment group had not been treated** :

$$\mathbb{E}[Y_{0it} - Y_{0it-1} \mid D_{it} = 1] = \mathbb{E}[Y_{0it} - Y_{0it-1} \mid D_{it} = 0]$$

  Note: the assumption above about parallel trends refers to both the pre-treatment trend (the famous pre-trend test) and the post-treatment trend (look at the t subscript). However, only the pre-trend can be tested empirically.

[1] Proof: attenuation bias for OLS. Suppose the true model is:

$$y = \beta x^* + \varepsilon$$

but the true regressor x^* is not observed. Instead, we observe:

$$x = x^* + u$$

We estimate the regression:

$$y = \gamma x + v$$

and want to know whether $\gamma = \beta$. OLS gives:

$$\gamma = \frac{\text{Cov}(y, x)}{\text{Var}(x)}$$

You don't have the true model $y = \beta x^* + \varepsilon$, fill in $x = x^* + u$:

$$\gamma = \frac{\text{Cov}(\beta x^* + \varepsilon,\ x^* + u)}{\text{Var}(x^* + u)}$$

Compute the numerator:

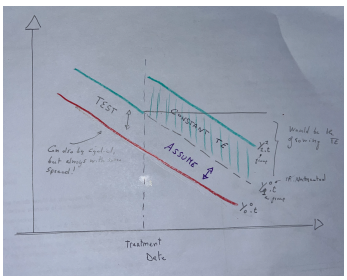
$$\text{Cov}(\beta x^* + \varepsilon,\ x^* + u) = \beta\,\text{Cov}(x^*, x^*) + \beta\,\text{Cov}(x^*, u) + \text{Cov}(\varepsilon, x^*) + \text{Cov}(\varepsilon, u)$$
$$= \beta\,\text{Var}(x^*) \quad \text{(since all cross-covariances are zero)}$$

Compute the denominator:

$$\text{Var}(x^* + u) = \text{Var}(x^*) + \text{Var}(u)$$

Final expression:

$$\gamma = \beta \cdot \frac{\text{Var}(x^*)}{\text{Var}(x^*) + \text{Var}(u)} = \beta \cdot \frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_u}$$



- **Common violations of pre-trend**:
  - **Only one group has a trend change due to external factor** (eg another treatment polic). Violations can come from either group treatment or control.
  - **Group composition changes over time**. Violations can come from either group treatment or control.
  - **Ashenfelter dip**: pre-treatment drop for treated group.
- **Robustness Checks**:
  - **Visually:** Graph trends before treatment for the two groups, check they are parallel (we need more than two periods). Note: If covariates affect the outcome trends, it may be better to condition on them (e.g., plot residualized trends after regressing out covariates).

    $$\mathbb{E}[Y_{0it} - Y_{0it-1} \mid D_{it} = 1, X_{it}] = \mathbb{E}[Y_{0it} - Y_{0it-1} \mid D_{it} = 0, X_{it}]$$

  - **In the regression:** Control for group-specific trends, results should not change. This is because they can have different pretrends, eg T is growing fastly than C, and we need assume it would have continued to grow faster. more infra
  - **Discussion:** Add a discussion to Justify parallel trends assumption. Justify the absence of any other simultaneous treatment

Remark 1: Remark: The SUTVA is implicit in the unconfoundedness assumption: a violation of the SUTVA would lead to an increase in the outcome of the control group after treatment (rhs of the equation, see figure). In the plot, you have that with treatment not only the outcome of T shifts but also the outcome of control. Plus, to be precise, we would need to change our potential outcome framework!

Remark 2 Under the parallel trends assumption, we are essentially saying that we want the treatment and control groups to be as similar as possible (they have to follow the same evolution). The goal is to ensure that assignment to treatment is not driven by intrinsic characteristics, but rather that treatment was exogenously or randomly assigned (ideally due to some factor that cannot be manipulated, such as random border assignment in Africa). However, the more similar the two groups are, the higher the risk of spillovers. Trade-off: The potential violation of SUTVA is the cost of ensuring that the treatment and control groups are highly comparable. Therefore, you must provide a convincing argument for why SUTVA holds in your setting.

2.2.2. Regression Specification (Fully Saturated)

Now we will use a fully saturated regression. Silly remark: in the regression above you had aggregated the constant in the group dummy.

- Model:

  $$Y_{igt} = c_0 + c_1 D_g + c_2 D_t + c_3 D_g D_t + \varepsilon_{igt}$$

- We can estimate the treatment effect by taking the difference between the two differences in means:

  $$\begin{aligned}\widehat{\rho} &= \left(\overline{Y}_{t=1,g=1} - \overline{Y}_{t=0,g=1}\right) - \left(\overline{Y}_{t=1,g=0} - \overline{Y}_{t=0,g=0}\right) \\ &= \text{(Tr Post - Pre)} - \text{(Cr Post - Pre)}\end{aligned}$$

- Math:

  $$\begin{aligned}&\mathbb{E}(Y_{igt} \mid g=1, t=1) - \mathbb{E}(Y_{igt} \mid g=1, t=0) \\ &-\left[\mathbb{E}(Y_{igt} \mid g=0, t=1) - \mathbb{E}(Y_{igt} \mid g=0, t=0)\right] \\ &= [(c_0 + c_1 + c_2 + c_3) - (c_0 + c_1)] - [(c_0 + c_2) - c_0] \\ &= (c_2 + c_3) - c_2 = c_3\end{aligned}$$

- in the table below the horizontal rows remove the constant and the group effect (again, differencing removes unobserved components.
- the vertical line removes the time trend

	t=1	t=0	DIF
g=1	$c_0 + c_1 + c_2 + c_3$	$c_0 + c_1$	$c_2 + c_3$
g=0	$c_0 + c_2$	c_0	c_2
DIF	$c_1 + c_3$	c_1	c_3

Treatment effect: c_3

2.2.3. Regression specification

The regression specification is:

$$Y_{igt} = \gamma_g + \lambda_t + \rho D_{gt} + \varepsilon_{igt}$$

$D_{gt} = 1$ only for $g = 1$ and $t = 1$.

2.2.4. Extensions

- If you have **Repeated Cross-section:** you can pool the data and use (trivial)

  $$Y_{igt} = c_0 + c_1 D_g + c_2 D_t + c_3 D_g D_t + \varepsilon_{igt}$$

- If you have **panel data**. Use FE, see below the regression at the individual level (with individual level partialling out)

  $$Y_{it} = \alpha_i + \lambda_t + X'_{it}\beta + \rho D_{it} + \varepsilon_{it}$$

- you could also use the FD model

3. Leads and Lags of Treatment (now we have multiple t)

- Estimate dynamic effects:

  $$Y_{igt} = \alpha_g + \lambda_t + \sum_{\tau=0}^{m} \rho_{-\tau} D_{g,t-\tau} + \sum_{\tau=1}^{q} \rho_{+\tau} D_{g,t+\tau} + X'_{igt}\beta + \varepsilon_{igt}$$

  note τ makes you go back in time!
- Purpose:
  - Detect pre-trends (check if $\rho_{+\tau} = 0$ before treatment)
  - Observe treatment effect over time.

3.1. Controlling for Covariates

General Principle: We include covariates to avoid bias only when they affect the untreated potential outcome Y^0. VERY TRIVIAL:

- **Time-invariant or independent of treatment:** say 1) race is time-invariant, 2) educ varies but does not affect treatment
  - If effect on outcome Y^0 is constant: dont need to control (there is the individual level fe)
  - If effect on outcome Y^0 varies over time: include interaction with t (i.e., $t \times x$).
- **Time-varying and correlated with treatment:** say race changes and it changes with treatment
  - If effect on the outcome Y^0 is constant: include x_t in levels.
  - If time-varying effect on the outcome Y^0: include $t \times x_t$ interaction.

3.2. Group-Specific Trends

Suppose the treatment group was already growing faster than the control group before treatment. Then, any post-treatment difference in outcomes could be due to this ongoing trend, not the treatment itself. By adding group-specific trends to the regression, we are asking whether *there is still a treatment effect after accounting for the fact that the treatment group was already growing faster?*. Just make the plot; the gap is shrinking/expanding![2]. Hence:

- Allow differential trends across groups (linear/polynomial):

$$Y_{igt} = \gamma_{0g} + \gamma_{1g}t + \lambda_t + X'_{igt}\beta + \rho D_{igt} + \varepsilon_{igt}$$

- Problem: we cannot control for non-linear trends differential across groups easily. Use triple differences!

3.3. Triple Differences

3.3.1. Idea

- **Motivation:** Difference-in-Differences (DID) may fail if group-specific trends are nonlinear or not parallel. Triple Differences (DDD) adds a third dimension to difference out these biases.
- **Setup and Example:** Medicaid introduced in some states (g), and only for families with young children (a).
  - **Setup:** We observe:
    - Two groups (g): a *treated region* and an *untreated region*.
    - Two time periods (t): before and after a policy intervention.
    - Two subgroups (a):
      - **Treated subgroup:** families with children.
      - **Untreated subgroup:** families without children.
  - **Why Triple Differences (DDD) helps:**
    - You would have compared families with child in the treated state with families without child in the treatment state.
    - what if these two types of families intrinsiclaly had different trends? (families without child are younger, and earning of young people are declining relatively more than earnings of reach people?)
    - To correct for this differntial trend, you partial out from those two groups the equivalent groups in another state.

3.3.2. Regression Specification:

$$Y_{iagt} = \gamma_{gt} + \lambda_{at} + \theta_{ag} + \rho D_{agt} + \varepsilon_{iagt}$$

where:

- D_{agt} is a dummy for treated group (state g) × treated subgroup a (families with child) × post period.
- γ_{gt}: group-time fixed effects (controls for group-specific trends).
- λ_{at}: subgroup-time fixed effects.
- θ_{ag}: group-subgroup fixed effects (e.g., in the treated state, families without children may systematically earn less). being in that subgroup in that state creates an advanteage per se, baeline.

Alternative Saturated Specification

$$Y_{iagt} = c_0 + c_1 D_g + c_2 D_t + c_3 D_g D_t + c_4 D_a + c_5 D_t D_a + c_6 D_g D_a + c_7 D_g D_t D_a + \varepsilon_{iagt}$$

- c_7 identifies the DDD estimate

[2] Note that this formula no longer works as it is based on the parallel trend assumption

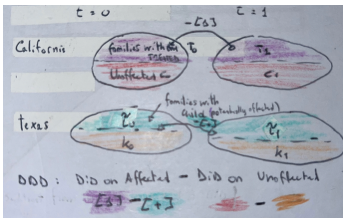
$\hat{\rho} = \left(\overline{Y}_{t=1,g=1} - \overline{Y}_{t=0,g=1}\right) - \left(\overline{Y}_{t=1,g=0} - \overline{Y}_{t=0,g=0}\right)$ = (Treated Post - Pre) – (Control Post - Pre)

3.3.3. Non-Parametric Specification

Color Legend:

- Treated region, families with children (Group A)
- Treated region, families without children (Group B)
- Untreated region, families with children (Group A)
- Untreated region, families without children (Group B)

$$\mathbb{E}(Y_{igt} \mid g{=}1, t{=}1, a{=}1) - \mathbb{E}(Y_{igt} \mid g{=}1, t{=}0, a{=}1) - \mathbb{E}(Y_{igt} \mid g{=}0, t{=}1, a{=}1) + \mathbb{E}(Y_{igt} \mid g{=}0, t{=}0, a$$
$$- \Big[\mathbb{E}(Y_{igt} \mid g{=}1, t{=}1, a{=}0) - \mathbb{E}(Y_{igt} \mid g{=}1, t{=}0, a{=}0) - \mathbb{E}(Y_{igt} \mid g{=}0, t{=}1, a{=}0) + \mathbb{E}(Y_{igt} \mid g{=}0,$$



- you can control both for the differential trends bewteen states (obvious baseline) and the differential trends between groups.

3.4. Standard Errors

- **Problem:** DID estimates can severely underestimate standard errors due to serial correlation.
- **Background:**
  - Downards bias arising from time series structure in the data (seria correlation).
  - Persistency reduces the amount of information contained in teh data (lowe signal to noise ratio). Also, I guess, lower effective sample size.
- **Solutions and Alternatives**:
  1. **Collapse data into pre- and post-treatment periods:** Aggregating eliminates serial correlation within units over time, making standard errors more reliable.
  2. **Specify the correlation structure directly:** Use autoregressive models (e.g., AR(1)) for the error term.
  3. **Block bootstrap:** Resample at the group level (e.g., states or clusters). Valid when the number of blocks is large.
  4. **Clustered Standard Errors:** Allows for arbitrary serial correlation within each group (e.g., state, firm). Consistent only when the number of clusters is large.

3.5. DID with Treatment at Different Times (Staggered Design)

3.5.1. Fraiming the Issue

- **Canonical Model (TWFE):**

$$Y_{igt} = \gamma_g + \lambda_t + \rho D_{gt} + \nu_{igt}$$

  - γ_g: group fixed effects
  - λ_t: time fixed effects
  - D_{gt}: treatment indicator for group g at time t
  - ρ: estimated average treatment effect
- **Interpretation of ρ in staggered adoption:**

$$\rho_{\text{fe}} = \mathbb{E}\left(\sum_{(g,t):D_{gt}=1} W_{g,t}\Delta_{g,t}\right)$$

where:

  - $\Delta_{g,t}$: treatment effect in group g at time t
  - $W_{g,t}$: weight assigned to cell (g,t), with $\sum W_{g,t} = 1$
- **Problem:** some $W_{g,t}$ may be negative.

- Even if all $\Delta_{g,t} > 0$, negative weights can make $\rho_{\text{fe}} < 0$.
- Arises due to comparisons across differentially treated groups in TWFE.
  - note treatment timing may be endogenous: groups with higher treatment benefit are treated before

- **Weight decomposition (de Chaisemartin):** We will build the ATE as a weigheted average of the two by two comparisons in treatemnt effects. The weigth sin such weighted average are determined by the leverage of the treatment dummy on the regression:

$$D_{gt} = \alpha+\gamma_g+\lambda_t+\varepsilon_{g,t} \Rightarrow \varepsilon_{g,t} = D_{gt}-\alpha-\gamma_g-\lambda_t = D_{g,t}+\underbrace{D_{\cdot,\cdot} - D_{g,\cdot} - D_{\cdot,t}}_{D^{pred}}$$

We need the above because the weights are driven by the errors e_{gt}! These residuals measure how much the treatment dummy $D_{g,t}$ deviates from what would be expected based on group and time averages.

$$\rho_{\text{FE}} = \mathbb{E}\left(\sum_{(g,t):D_{gt}=1} \frac{N_{g,t}}{N_1} w_{g,t}\Delta_{g,t}\right), \quad \text{where } w_{g,t} = \frac{\varepsilon_{g,t}}{\sum_{(g,t):D_{gt}=1} \frac{N_{g,t}}{N_1}\varepsilon_{g,t}}$$

- so if $\varepsilon_{g,t}$ is negative the weight is negative!
- Later periods of early adopters are very exposed to negative weights because: 1) the group is usually treated, as it was treated from earlier stages, and 2) in later stages many groups have been treated. Early adopters in late periods face high expected treatment! *My Idea:* The weight is negative if $D - D^{\text{pred}}$ is negative, so if the predicted level of treatment is greater than the actual one that is, if we are greatly expecting treatment (consider that the linear regression does not know that there is a 01 bound, so it is possibly giving to an early adopter a predicted value > 1).
- With homogeneous treatment effects, you still have negative weights, but since the weights sum to one, negative weights are compensated by positive ones, and both positive and negative weights are multiplying the same TE, so they cancel out .The issue is that if those groups with negative weights have higher treatment effects, the ATE is negative!

- **Illustrative example:**

  - Group 1: treated only in period 3
  - Group 2: treated in periods 2 and 3
  - Same group sizes
  - **Step-by-step computation of residuals** $\varepsilon_{g,t} = D_{g,t} - D_{g,\cdot} - D_{\cdot,t} + D_{\cdot,\cdot}$
    - **Step 1: Treatment status** $D_{g,t}$

	$t=1$	$t=2$	$t=3$
$g=1$	0	0	1
$g=2$	0	1	1

The dummies represent the probability of being treated

    - **Step 2: Row averages**: $D_{1,\cdot}$ represents the probability of being treated of group 1 across all t.

$$D_{1,\cdot} = \frac{0+0+1}{3} = \frac{1}{3}, \quad D_{2,\cdot} = \frac{0+1+1}{3} = \frac{2}{3}$$

    - **Step 3: Column averages**: $D_{\cdot,1}$ represents the probability of being trated in time t

$$D_{\cdot,1} = \frac{0+0}{2} = 0, \quad D_{\cdot,2} = \frac{0+1}{2} = \frac{1}{2}, \quad D_{\cdot,3} = \frac{1+1}{2} = 1$$

    - **Step 4: Grand average**

$$D_{\cdot,\cdot} = \frac{0+0+1+0+1+1}{6} = \frac{3}{6} = \frac{1}{2}$$

    - **Step 5: Compute residuals**

$$\varepsilon_{1,3} = D_{1,3} - D_{1,\cdot} - D_{\cdot,3} + D_{\cdot,\cdot} = 1 - \frac{1}{3} - 1 + \frac{1}{2} = \frac{1}{6}$$
$$\varepsilon_{2,2} = D_{2,2} - D_{2,\cdot} - D_{\cdot,2} + D_{\cdot,\cdot} = 1 - \frac{2}{3} - \frac{1}{2} + \frac{1}{2} = \frac{1}{3}$$
$$\varepsilon_{2,3} = D_{2,3} - D_{2,\cdot} - D_{\cdot,3} + D_{\cdot,\cdot} = 1 - \frac{2}{3} - 1 + \frac{1}{2} = -\frac{1}{6}$$

  - $\varepsilon_{1,3} = 1/6, \varepsilon_{2,1} = 1/3, \varepsilon_{2,3} = -1/6$
  - Then:
$$\rho_{\text{fe}} = \frac{1}{2}\mathbb{E}[\Delta_{1,3}] + \mathbb{E}[\Delta_{2,2}] - \frac{1}{2}\mathbb{E}[\Delta_{2,3}]$$
  - Suppose:

$$\mathbb{E}[\Delta_{1,3}] = \mathbb{E}[\Delta_{2,2}] = 1, \quad \mathbb{E}[\Delta_{2,3}] = 4 \Rightarrow \rho_{\text{fe}} = \frac{1}{2}(1)+1-\frac{1}{2}(4) = -0.5$$

  - Despite all ATEs being positive, the estimated ρ_{fe} is negative.

This is connected to the cross-comparison interpretation of the problem (beacon): The issue arises when later-treated groups are compared to earlier-treated groups, and the earlier-treated group is incorrectly used as a control, even though it's already treated. So if you have 1) a group that has been treated for long, this will be a poor control for the other recently treated compared units and will create a negative ATE, 2) if you have many treated groups at t, then it is more likely to get a poor comparison.

the fun part is that these poor counterfactuals are likely exactly those with high ate. they are terrible counterfactuals and they are drawing down the ate!

3.6. Solution

- how to notice there is a problem? compute standard deviation of the ATEs across the treated (g,t) cell (where negative weights may come in) and compare it with the absolute value of the DID estimate (e.g., using not yet treated as control!).
- **Solution:** Estimate treatment effects by comparing treated units only to *not-yet-treated* units. Similar to matching vs OLS: drop poor controls!

4. DID Advanced Topics

4.1. Lagged Dependent Variable

- Sometimes we **must include the lagged outcome** Y_{t-1} as a covariate (control for it).
- **Why include the lagged outcome (Y_{t-1})? Ashenfelter's Dip and Pre-trends Bias**
  If treatment D_t depends on the past outcome Y_{t-1}, and Y_{t-1} is correlated with the current outcome Y_t, then D_t is endogenous: it is correlated with Y_t through Y_{t-1}. This leads to omitted variable bias if Y_{t-1} is not controlled for. Note: this is, first and foremost, a violation of parallel trends. Also, recognize that this is selection biasselection bias is what causes the violation of parallel trends. This is self-selection because the treated group shares something in commonsomething specific that makes them more likely to enter treatment. Most likely, they were chosen to be treated **because** they experienced a drop in income. There is no internal validity: pre-trends are violated, and the treatment group is not comparable to the control group. **Be careful not to misunderstand: you may think "had the control group not received treatment, their Y_0 would have been the same."*

No: the control group experienced a drop in income, and that's why they were selected. This clearly creates positive bias!
Graph intuition: Initially, treated and control groups have parallel trends. A dip in Y_{t-1} for treated units breaks this trend. You want to control for Y_{t-1} that is, absorb the dip. Including Y_{t-1} restores the trend. Imagine the "smoothing" effect a time dummy would have: a time dummy is common to all individuals, but Y_{t-1} provides an individual-specific intensity.

- Y_{t-1} is clearly not eliminated by individual fixed effects.
- In OLS, including Y_{t-1} is straightforward, but in Fixed Effects (FE) models, this becomes more complex (see subsection).

4.1.1. OLS with Lagged Dependent Variable

- One possibiliyy is to **assume unconfoundedness** given lagged outcomes even without controlling for fixed effects (basically say you don't need FE):

$$\mathbb{E}[Y_{0it} \mid Y_{it-k}, X_{it}, D_{it}] = \mathbb{E}[Y_{0it} \mid Y_{it-k}, X_{it}]$$

- Estimate via OLS or Random Effects (RE)

$$Y_{it} = \alpha + \theta Y_{it-k} + \lambda_t + \rho D_{it} + X'_{it}\beta + \varepsilon_{it}$$

no nickel bias (more infra) in OLS adn RE (no differencing trasformation)

4.2. Lagged Dependent Variable with Fixed Effects

- When we assume unconfoundedness given lagged outcomes and fixed effects:

$$\mathbb{E}[Y_{0it} \mid Y_{it-k}, \alpha_i, X_{it}, D_{it}] = \mathbb{E}[Y_{0it} \mid Y_{it-k}, \alpha_i, X_{it}]$$

the treatment status is independent of the potential outcome.
- Estimation equation:

$$Y_{it} = \alpha_i + \theta Y_{it-k} + \lambda_t + \rho D_{it} + X'_{it}\beta + \varepsilon_{it}$$

- Problem: **Nickell bias** (see below).
  - Consider first-differencing (fixed effects disappear):

$$\Delta Y_{it} = \theta \Delta Y_{it-1} + \Delta\lambda_t + \rho \Delta D_{it} + \Delta X'_{it}\beta + \Delta\varepsilon_{it}$$

  - Problem! $\Delta\varepsilon_{it} = \varepsilon_{it} - \varepsilon_{it-1}$ is correlated with ΔY_{it-1}.
  - ΔY_{it-1} contains ε_{it-1}, which is in $\Delta\varepsilon_{it}$.
  - $\Rightarrow$ correlation between regressor and error term $\Rightarrow$ bias.
  - Nickell (1981) proved that OLS is inconsistent in short panels.
  - Solution:
    * If ε_{it} is serially correlated ($\text{corr}(\varepsilon_{it}, \varepsilon_{it-k}) \neq 0$). Then there might not be any solution.
    * If the correlation does not hold up to period k: use Y_{it-2-k} as IV for Y_{it-1} (or First-differences: $Y_{it-2} - Y_{it-3}, X_{it-2} - X_{it-3}$, etc. valid instrument for an AR(1) process). They become uncorrelated (trivial). if there is maximum AR(k) correlation the instrument is valid, but we **will also need to check the first stage** (not trivial).
    * Empirically we can bound the effect of interest:
      · If we include fixed effects and no lagged outcome, we expect upwards biased estimates (we have a dip and then up: Dip in Y_{it-1} leads to mechanical rebound; FE attributes this to treatment → overestimates effect).
      · If we do not include fixed effects but we include lagged outcomes, we expect downwards biased estimates (boh)

5. Staggered Did Rho itnerpretation

Weighted average of all possible two-by-two DID estimators. Only recovers the ATE when treatment effects are homogeneous

6. DID in levls vs DID in logs

- **Key assumption**: **common trends**, which is *not* invariant to monotonic transformations of the outcome (parallel trend in levels are not parallel trends in logs). Thsi is obvious bc in levels you have the difference between the outcomes t obe fixed, in logs you have the raito to be fixed
- **Treatment Effects found using DID can differ greatly depending on whether levels or logs are used** as the outcome variable.