Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects, American Economic Review, 2020

Wenzhi Wang [*]

August 20, 2024

de Chaisemartin and D'Haultfœuille (2020)

1. Setup

One considers observations that can be divided into G groups and T periods. For every $(g,t) \in \{1,\dots,G\} \times \{1,\dots,T\}$, let $N_{g,t}$ denote the number of observations in group g at period t, and let $N = \sum_{g,t} N_{g,t}$ be the total number of observations. The data may be an individual-level panel or repeated cross-section dataset where groups are, say, individuals' county of birth. The data could also be a cross section where cohort of birth plays the role of time.

One is interested in measuring the effect of a treatment on some outcome. Throughout the paper we assume the treatment is binary, but our results apply to any ordered treatment. Then, for every $(i,g,t) \in \{1,\dots,N_{g,t}\} \times \{1,\dots,G\} \times \{1,\dots,T\}$, let $D_{i,g,t}$ and $(Y_{i,g,t}(0), Y_{i,g,t}(1))$ respectively denote the treatment status and the potential outcomes without and with treatment of observation i in group g at period t.

The outcome of observation i in group g and period t is $Y_{i,g,t} = Y_{i,g,t}(D_{i,g,t})$. For all (g,t), let

$$D_{g,t} = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} D_{i,g,t}, \quad Y_{g,t}(0) = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} Y_{i,g,t}(0)$$
$$Y_{g,t}(1) = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} Y_{i,g,t}(1), \quad \text{and} \quad Y_{g,t} = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} Y_{i,g,t}.$$

Here, $D_{g,t}$ denotes the average treatment in group g at period t, while $Y_{g,t}(0), Y_{g,t}(1)$, and $Y_{g,t}$ respectively denote the average potential outcomes without and with treatment and the average observed outcome in group g at period t.

[*] This note is written in my pre-doc period at the University of Chicago Booth School of Business.

Assumption 1 (Balanced Panel of Groups)**.** For all $(g,t) \in \{1,\dots,G\} \times \{1,\dots,T\}$, $N_{g,t} > 0$.

Assumption 1 requires that no group appears or disappears over time. This assumption is often satisfied. Without it, our results still hold but the notation becomes more complicated as the denominators of some of the fractions below may then be equal to zero.

Assumption 2 (Sharp Design)**.** For all $(g,t) \in \{1,\dots,G\} \times \{1,\dots,T\}$ and $i \in \{1,\dots,N_{g,t}\}$, $D_{i,g,t} = D_{g,t}$.

Assumption 2 requires that units' treatments do not vary within each (g,t) cell, a situation we refer to as a sharp design. This is for instane satisfied when the treatment is a group-level variable, for instance a county or a state law. This is also mechanically satisfied when $N_{g,t} = 1$.

Assumption 3 (Independent Groups)**.** The vectors $(Y_{g,t}(0), Y_{g,t}(1), D_{g,t})_{1 \leq t \leq T}$ are mutually independent.

We consider $D_{g,t}, Y_{g,t}(0), Y_{g,t}(1)$ as random variables. For instance, aggregate randome shocks may affect the average potential outcomes of group g at period t. The treatment status of group g of period t may also be random. The expectations below are taken with respect to the distribution of those random variables. Assumption 3 allows for the possibility that the treatments and potential outcomes of a group may be correlated over time, but it requires that the potential outcomes and treatments of different groups be independent.

Assumption 4 (Strong Exogeneity)**.** For all $(g,t) \in \{1,\dots,G\} \times \{2,\dots,T\}$,

$$\mathbb{E}\left[Y_{g,t}(0) - Y_{g,t-1}(0) \mid D_{g,1}, \dots D_{g,T}\right] = \mathbb{E}\left[Y_{g,t}(0) - Y_{g,t-1}(0)\right].$$

Assumption 4 requires that the shocks affecting a group's $Y_{g,t}(0)$ be mean independent of that group's treatment sequence. This rules out the possibility that a group gets treated because it experiences negative shocks. Assumption 4 is related to the strong exogeneity condition in panel data models, which, as is well known, is necessary to obtain the consistency of the fixed effects estimator.

We now define the FE regression.

Regression 1 (Fixed Effects Regression)**.** Let $\widehat{\beta}_{fe}$ denote the coefficient of $D_{g,t}$ in an OLS regression fo $Y_{i,g,t}$ on group fixed effects, period fixed effects, and $D_{g,t}$. Let $\beta_{fe} = \mathbb{E}\left[\widehat{\beta}_{fe}\right]$.

For all g and t, let $N_{g\cdot} = \sum_{t=1}^{T} N_{g,t}$ and $N_{\cdot t} = \sum_{g=1}^{G} N_{g,t}$ respectively denote the total number of observations in group g and in period t. For any variable $X_{g,t}$ defined in each (g,t) cell, let $X_{g\cdot} = \sum_{t=1}^{T} \left(N_{g,t}/N_{g\cdot}\right) X_{g,t}$ denote the average value of $X_{g,t}$ in group g, let $X_{\cdot t} = \sum_{g=1}^{G} \left(N_{g,t}/N_{\cdot t}\right) X_{g,t}$ denote the average value of $X_{g,t}$ in period t, and let $X_{\cdot\cdot} = \sum_{g,t} \left(N_{g,t}/N\right) X_{g,t}$ denote the average value of $X_{g,t}$. Finally, for any variable $X_{g,t}$, we let $\boldsymbol{X}$ denote the vector $(X_{g,t})_{(g,t) \in \{1,\dots,G\} \times \{1,\dots,T\}}$ collecting the values of that variable in each (g,t) cell. For instance, $\boldsymbol{D}$ is the vector $(D_{g,t})_{(g,t) \in \{1,\dots,G\} \times \{1,\dots,T\}}$ collecting the treatments of all the (g,t) cells.

2. Two-Way Fixed Effects Regressions

2.1. A Decomposition Result

We study the FE regression under the following common trends assumption.

Assumption 5 (Common Trends). For $t \geq 2$, $\mathbb{E}\left[Y_{g,t}(0) - Y_{g,t-1}(0)\right]$ does not vary across g.

Assumption 5 requires that the expectation of the outcome without treatment follow the same evolution over time in every group. When t represents birth cohorts, Assumption 5 requires that the outcome different between consecutive cohorts be the same across groups.

Let $N_1 = \sum_{i,g,t} = D_{i,g,t}$ denote the number of treated units, let

$$\Delta^{TR} = \frac{1}{N_1} \sum_{(i,g,t):D_{g,t}=1} \left[Y_{i,g,t}(1) - Y_{i,g,t}(0)\right]$$

denote the average treatment effect across all treated units, and let $\delta^{TR} = \mathbb{E}\left[\Delta^{TR}\right]$ denote the expectation of that parameter, hereafter referred to as the ATT. For any $(g,t) \in \{1,\ldots,G\} \times \{1,\ldots,T\}$, let

$$\Delta_{g,t} = \frac{1}{N_{g,t}} \sum_{i=1}^{N_{g,t}} \left[Y_{i,g,t}(1) - Y_{i,g,t}(0)\right]$$

denote the ATE in cell (g,t). Note that δ^{TR} is equal to the expectation of a weighted average of the treated cells' $\Delta_{g,t}$:

$$\delta^{TR} = \mathbb{E}\left[\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \Delta_{g,t}\right]. \tag{1}$$

Under the common trends assumption, we show that β_{fe} is also equal to the expectation of a weighted sum of the $\Delta_{g,t}$ terms, with potentially some negative weights.

Let $\varepsilon_{g,t}$ denote the residual of observations in cell (g,t) in the regression of $D_{g,t}$ on group and period fixed effects,

$$D_{g,t} = \alpha + \gamma_g + \lambda_t + \varepsilon_{g,t}.$$

One can show that if the regressors in Regression 1 are not collinear, the average value of $\varepsilon_{g,t}$ across all treated (g,t) cells differs from 0: $\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \varepsilon_{g,t} \neq 0$. Then let $\omega_{g,t}$ denote $\varepsilon_{g,t}$ divided by that average:

$$\omega_{g,t} = \frac{\varepsilon_{g,t}}{\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1} \varepsilon_{g,t}}.$$

Theorem 1. Suppose that Assumptions 1-5 hold. Then

$$\beta_{fe} = \mathbb{E}\left[\sum_{(g,t):D_{g,t}=1} \frac{N_{g,t}}{N_1}\omega_{g,t}\Delta_{g,t}\right].$$

This result implies that in general, $\beta_{fe} \neq \delta^{TR}$, so $\widehat{\beta}_{fe}$ is a biased estimator of the ATT.

Example 1. To illustrate this, we consider a simple example of a staggered adoption design with two groups and three periods, and where the treatments are nonstochastic: group 1 is untreated at periods 1 and 2 and treated at period 3, while group 2 is untreated at period 1 and treated both at periods 2 and 3. We also assume that $N_{g,t}/N_{g,t-1}$ does not vary across g: all groups experience the same growth of their number of observations from $t-1$ to t, a requirement that is for instance satisfied when the data is a balanced panel. Then one can show that

$$\varepsilon_{g,t} = D_{g,t} - D_{g\cdot} - D_{\cdot t} + D_{\cdot\cdot},$$

thus implying that

$$\begin{aligned}
\varepsilon_{1,3} &= 1 - 1/3 - 1 + 1/2 = 1/6 \\
\varepsilon_{2,2} &= 1 - 2/3 - 1/2 + 1/2 = 1/3. \\
\varepsilon_{2,3} &= 1 - 2/3 - 1 + 1/2 = -1/6
\end{aligned}$$

The residual is negative in group 2 and period 3, because the regression predicts a treatment probability larger than one in that cell, a classic extrapolation problem with linear regressions. Then, under the common trends assumption, it follows from Theorem 1 and the fact that the treatments are nonstochastic that

$$\beta_{fe} = 1/2E\left[\Delta_{1,3}\right] + E\left[\Delta_{2,2}\right] - 1/2E\left[\Delta_{2,3}\right].$$

Here, β_{fe} is equal to a weighted sum of the ATEs in group 1 at period 3, group 2 at period 2, and group 2 at period 3, the three treated (g,t) cells. However, the weight assigned to each ATE differs from $1/3$, the proportion that each cell accounts for in the population of treated observations. Therefore, β_{fe} is not equal to δ^{TR}. Perhaps more worryingly, not all the weights are positive. Consequently, β_{fe} may be a very misleading measure of the treatment effect.

We now generalize the previous illustration by characterizing the (g,t) cells whose ATEs are weighted negatively by β_{fe}.

Proposition 1. Suppose that Assumption 1 holds and for all $t \geq 2$, $N_{g,t}/N_{g,,t-1}$ does not vary across g. Then, for all (g,t,t') such that $D_{g,t} = D_{g,t'} = 1$, $D_{\cdot t} > D_{\cdot,t'}$ implies $\omega_{g,t} < \omega_{g,t'}$. Similarly, for all (g,g',t) such that $D_{g,t} = D_{g',t} = 1$, $D_{g\cdot} > D_{g'\cdot}$ implies $\omega_{g,t} < \omega_{g',t}$.

Proposition 1 shows that β_{fe} is more likely to assign a negative weight to periods where a large fraction of groups are treated, and to groups treated for many periods. Then, negative weights are a concern when treatment effects differ between periods with many versus few treated groups, or between groups treated for many versus few periods.

Assumption 6 (Staggered Adoption Designs)**.** For all g, $D_{g,t} \geq D_{g,t-1}$ for all $t \geq 2$.

In staggered adoption designs, $D_{.t}$ is increasing in t, so Proposition 1 implies that $\omega_{g,t}$ is decreasing in t. Proposition 1 also implies that in that design, groups that adopt the treatment earlier are more likely to receive some negative weights.

2.2. Robustness to Heterogeneous Treatment Effects

Theorem 1 shows that in sharp designs with many groups and periods, $\widehat{\beta}_{fe}$ may be a misleading measure of the treatment effect under the standard common trends assumption, if the treatment effect is heterogeneous across groups and time periods. In the corollary below, we propose two robustness measures that can be used to assess how serious that concern is.

2.3. Extension to the First-Different Regression

Regression 2 (First-Difference Regression)**.** Let $\widehat{\beta}_{fd}$ denote the coefficient of $D_{g,t} - D_{g,t-1}$ in an OLS regression of $Y_{g,t} - Y_{g,t-1}$ on period fixed effects and $D_{g,t} - D_{g,t-1}$, among observations for which $t \geq 2$. Let $\beta_{fd} = \mathbb{E}\left[\widehat{\beta}_{fd}\right]$.

When $T = 2$ and $N_{g,2}/N_{g,1}$ does not vary across g, meaning that all groups experience the same growth of their number of units from period 1 to period 2, one can show that $\widehat{\beta}_{fe} = \widehat{\beta}_{fd}$. But, $\widehat{\beta}_{fe}$ differs from $\widehat{\beta}_{fd}$ if $T > 2$ or $N_{g,2}/N_{g,1}$ varies across g.

3. An Alternative Estimator

In this section, we show that it is possible to estimate a well-defined causal effect even if treatment effects are heterogeneous across groups or over time. Let

References

de Chaisemartin, Clément and Xavier D'Haultfœuille (2020) "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects," *American Economic Review*, 110 (9), 2964–2996, 10.1257/aer.20181169.