On the Use of Linear Fixed Effects Regression Models for Causal Inference

Kosuke Imai

Department of Politics
Princeton University

Joint work with In Song Kim

Washington University in St. Louis
April 27, 2012
Motivation

- Fixed effects models are a primary workhorse for causal inference.
- Used for stratified experimental and observational data.
- Also used to adjust for unobservables in observational studies:
  - “Good instruments are hard to find ... , so we’d like to have other tools to deal with unobserved confounders. This chapter considers ... strategies that use data with a time or cohort dimension to control for unobserved but fixed omitted variables” (Angrist & Pischke, *Mostly Harmless Econometrics*).
  - “fixed effects regression can scarcely be faulted for being the bearer of bad tidings” (Green *et al.*, *Dirty Pool*).

- Common claim: Fixed effects models are superior to matching estimators because the latter can only adjust for observables.

- **Question:** What are the exact causal assumptions underlying fixed effects regression models?
Main Methodological Results

1. Standard (one-way and two-way) FE estimators are equivalent to particular matching estimators

2. Common belief that FE models adjust for unobservables but matching does not is wrong

3. Identify the information used implicitly to estimate counterfactual outcomes under FE models

4. Identify potential sources of bias and inefficiency in FE estimators

5. Propose simple ways to improve FE estimators using weighted FE regression

6. Within-unit matching, first differencing, propensity score weighting, difference-in-differences are all equivalent to weighted FE model with different regression weights

7. Offer a specification test for the standard FE model

8. Develop easy-to-use software
Matching and Regression in Cross-Section Settings

<table>
<thead>
<tr>
<th>Units</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Treatment status</td>
<td>T</td>
<td>T</td>
<td>C</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>Outcome</td>
<td>$Y_1$</td>
<td>$Y_2$</td>
<td>$Y_3$</td>
<td>$Y_4$</td>
<td>$Y_5$</td>
</tr>
</tbody>
</table>

Estimating the Average Treatment Effect (ATE) via matching:

$$Y_1 - \frac{1}{2} (Y_3 + Y_4)$$
$$Y_2 - \frac{1}{2} (Y_3 + Y_4)$$
$$\frac{1}{3} (Y_1 + Y_2 + Y_5) - Y_3$$
$$\frac{1}{3} (Y_1 + Y_2 + Y_5) - Y_4$$
$$Y_5 - \frac{1}{2} (Y_3 + Y_4)$$
Matching Representation of Simple Regression

- Cross-section simple linear regression model:
  \[ Y_i = \alpha + \beta X_i + \epsilon_i \]

- Binary treatment: \( X_i \in \{0, 1\} \)

- Equivalent matching estimator:
  \[
  \hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} \left( \hat{Y}_i(1) - \hat{Y}_i(0) \right)
  \]

  where
  \[
  \hat{Y}_i(1) = \begin{cases} 
  \frac{1}{\sum_{i'=1}^{N} X_{i'}} \sum_{i'=1}^{N} X_{i'} Y_{i'} & \text{if } X_i = 1 \\
  \frac{1}{\sum_{i'=1}^{N} (1-X_{i'})} \sum_{i'=1}^{N} (1 - X_{i'}) Y_{i'} & \text{if } X_i = 0
  \end{cases}
  \]
  \[
  \hat{Y}_i(0) = \begin{cases} 
  \frac{1}{\sum_{i'=1}^{N} X_{i'}} \sum_{i'=1}^{N} X_{i'} Y_{i'} & \text{if } X_i = 0 \\
  \frac{1}{\sum_{i'=1}^{N} (1-X_{i'})} \sum_{i'=1}^{N} (1 - X_{i'}) Y_{i'} & \text{if } X_i = 1
  \end{cases}
  \]

- Treated units matched with the average of non-treated units
One-Way Fixed Effects Regression

- Simple (one-way) FE model:

\[ Y_{it} = \alpha_i + \beta X_{it} + \epsilon_{it} \]

- Commonly used by applied researchers:
  - Stratified randomized experiments (Duflo et al. 2007)
  - Stratification and matching in observational studies
  - Panel data, both experimental and observational

- \( \hat{\beta}_{FE} \) may be biased for the ATE even if \( X_{it} \) is exogenous within each unit

- It converges to the weighted average of conditional ATEs:

\[
\hat{\beta}_{FE} \xrightarrow{p} \frac{\mathbb{E}\{\text{ATE}_i \sigma_i^2\}}{\mathbb{E}(\sigma_i^2)}
\]

where \( \sigma_i^2 = \sum_{t=1}^{T} (X_{it} - \overline{X}_i)^2 / T \)

- How are counterfactual outcomes estimated under the FE model?
- Unit fixed effects \( \rightarrow \) within-unit comparison
Mismatches in One-Way Fixed Effects Model

- **Circles**: Proper matches
- **Triangles**: “Mismatches” $\implies$ attenuation bias

- **T**: treated observations
- **C**: control observations
Matching Representation of Fixed Effects Regression

Proposition 1

\[ \hat{\beta}^{FE} = \frac{1}{K} \left\{ \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left( \hat{Y}_{it}(1) - \hat{Y}_{it}(0) \right) \right\}, \]

\[ \hat{Y}_{it}(x) = \begin{cases} Y_{it} & \text{if } X_{it} = x \\ \frac{1}{T-1} \sum_{t' \neq t} Y_{it'} & \text{if } X_{it} = 1 - x \end{cases} \text{ for } x = 0, 1 \]

\[ K = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \left\{ X_{it} \cdot \frac{1}{T-1} \sum_{t' \neq t} (1 - X_{it'}) + (1 - X_{it}) \cdot \frac{1}{T-1} \sum_{t' \neq t} X_{it'} \right\} \cdot \]

- \( K \): average proportion of proper matches across all observations
- More mismatches \( \implies \) larger adjustment
- Adjustment is required except very special cases
- “Fixes” attenuation bias but this adjustment is not sufficient
- Fixed effects estimator is a special case of matching estimators
**Unadjusted Matching Estimator**

- Consistent if the treatment is exogenous within each unit
- Only equal to fixed effects estimator if heterogeneity in either treatment assignment or treatment effect is non-existent

![Diagram showing matching estimator with time periods and units labeled as C and T.](image-url)
Unadjusted Matching = **Weighted FE Estimator**

**Proposition 2**

The unadjusted matching estimator

\[
\hat{\beta}^M = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} (\hat{Y}_{it}(1) - \hat{Y}_{it}(0))
\]

where

\[
\hat{Y}_{it}(1) = \begin{cases} 
Y_{it} & \text{if } X_{it} = 1 \\
\frac{\sum_{t'=1}^{T} X_{it'} Y_{it'}}{\sum_{t'=1}^{T} X_{it'}} & \text{if } X_{it} = 0
\end{cases}
\]

\[
\hat{Y}_{it}(0) = \begin{cases} 
\frac{\sum_{t'=1}^{T} (1-X_{it'}) Y_{it'}}{\sum_{t'=1}^{T} (1-X_{it'})} & \text{if } X_{it} = 1 \\
Y_{it} & \text{if } X_{it} = 0
\end{cases}
\]

is equivalent to the weighted fixed effects model

\[
(\hat{\alpha}^M, \hat{\beta}^M) = \arg \min_{(\alpha, \beta)} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} (Y_{it} - \alpha_i - \beta X_{it})^2
\]

\[
W_{it} = \begin{cases} 
\frac{\sum_{t'=1}^{T} X_{it'}}{\sum_{t'=1}^{T} X_{it'}} & \text{if } X_{it} = 1, \\
\frac{\sum_{t'=1}^{T} (1-X_{it'})}{\sum_{t'=1}^{T} (1-X_{it'})} & \text{if } X_{it} = 0.
\end{cases}
\]
Equal Weights

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>0</td>
</tr>
<tr>
<td>T</td>
<td>1/2</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1/2</td>
</tr>
<tr>
<td>T</td>
<td>0</td>
</tr>
</tbody>
</table>

Different Weights

- Any within-unit matching estimator leads to weighted fixed effects regression with particular weights
- We derive regression weights given a matching estimator for various quantities (ATE, ATT, etc.)
**Theorem:** General Equivalence between Weighted One-Way FE and Matching Estimators

General matching estimator

\[
\hat{\beta}^M = \frac{1}{\sum_{i=1}^{N} \sum_{t=1}^{T} C_{it}} \sum_{i=1}^{N} \sum_{t=1}^{T} C_{it} \left( \overline{Y}_{it}(1) - \overline{Y}_{it}(0) \right)
\]

where \(0 \leq C_{it} < \infty, \sum_{t=1}^{T} \sum_{i=1}^{N} C_{it} > 0\),

\[
\overline{Y}_{it}(1) = \begin{cases} 
Y_{it} & \text{if } X_{it} = 1 \\
\sum_{t'=1}^{T} v_{it} X_{it'} Y_{it'} & \text{if } X_{it} = 0 
\end{cases}
\]

\[
\overline{Y}_{it}(0) = \begin{cases} 
\sum_{t'=1}^{T} v_{it} (1 - X_{it'}) Y_{it'} & \text{if } X_{it} = 1 \\
Y_{it} & \text{if } X_{it} = 0 
\end{cases}
\]

\[
\sum_{t'=1}^{T} v_{it} X_{it'} = \sum_{t'=1}^{T} v_{it} (1 - X_{it'}) = 1
\]

is equivalent to the weighted one-way fixed effects estimator

\[
W_{it} = \sum_{i'=1}^{N} \sum_{t'=1}^{T} w_{it'} t' \quad \text{and} \quad w_{it'} t' = \begin{cases} 
C_{it} & \text{if } (i, t) = (i', t') \\
v_{it} & \text{if } (i, t) \in \mathcal{M}_{i', t'} \\
0 & \text{otherwise.}
\end{cases}
\]
First Difference = Matching = Weighted One-Way FE

\[ \Delta Y_{it} = \beta \Delta X_{it} + \epsilon_{it} \] where \( \Delta Y_{it} = Y_{it} - Y_{i,t-1}, \Delta X_{it} = X_{it} - X_{i,t-1} \)

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>C T C C T</td>
<td>0 1 0 0</td>
</tr>
<tr>
<td>T C T C C</td>
<td>0 1 0 0</td>
</tr>
<tr>
<td>C C T C C</td>
<td>0 0 0 0</td>
</tr>
<tr>
<td>T T T C T</td>
<td>0 0 0 0</td>
</tr>
</tbody>
</table>
Implications for Time-Varying Confounders $Z_{it}$

1. **Discrete covariates**: the saturated model

2. **Regression-adjusted matching**:
   \[
   Y_{it} - \hat{g}(Z_{it}) \quad \text{where} \quad g(z) = \mathbb{E}(Y_{it} \mid X_{it} = 0, Z_{it} = z)
   \]

3. **Direct regression adjustment**:
   \[
   \arg \min_{(\alpha, \beta, \delta)} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} \left( Y_{it} - \alpha_i - \beta X_{it} - \delta^\top Z_{it} \right)^2
   \]
   - **Ex post interpretation**: $Y_{it} - \hat{\delta}^\top Z_{it} = \alpha_i + \beta X_{it} + \epsilon_{it}$

4. **Inverse-propensity score weighting** with normalized weights
   \[
   \hat{\beta}^W = \frac{1}{N} \sum_{i=1}^{N} \left\{ \sum_{t=1}^{T} \frac{X_{it} Y_{it}}{\hat{\pi}(Z_{it})} / \sum_{t=1}^{T} \frac{X_{it}}{\hat{\pi}(Z_{it})} - \sum_{t=1}^{T} \frac{(1 - X_{it}) Y_{it}}{1 - \hat{\pi}(Z_{it})} / \sum_{t=1}^{T} \frac{(1 - X_{it})}{1 - \hat{\pi}(Z_{it})} \right\}
   \]
   where $\pi(Z_{it}) = \Pr(X_{it} = 1 \mid Z_{it})$ is the propensity score
   - within-unit weighting followed by across-units averaging
Prop. Score Weighting = Transformed Weighted FE

**Proposition 3**

\[
(\hat{\alpha}^W, \hat{\beta}^W) = \arg \min_{(\alpha, \beta)} \sum_{i=1}^N \sum_{t=1}^T W_{it} (Y_{it}^* - \alpha_i - \beta X_{it})^2
\]

where the transformed outcome \( Y_{it}^* \) is,

\[
Y_{it}^* = \begin{cases} 
\frac{\left( \sum_{t'=1}^T X_{it'} \right) Y_{it}}{\hat{\pi}(Z_{it})} / \sum_{t'=1}^T \frac{X_{it'}}{\hat{\pi}(Z_{it'})} & \text{if } X_{it} = 1 \\
\frac{\left\{ \sum_{t'=1}^T (1-X_{it'}) \right\} Y_{it}}{1-\hat{\pi}(Z_{it})} / \sum_{t'=1}^T \frac{(1-X_{it'})}{1-\pi(Z_{it'})} & \text{if } X_{it} = 0
\end{cases}
\]

and the weights are the same as before

\[
W_{it} = \begin{cases} 
\frac{\sum_{t'=1}^T X_{it'}}{T} & \text{if } X_{it} = 1, \\
\frac{\sum_{t'=1}^T (1-X_{it'})}{\sum_{t'=1}^T (1-X_{it'})} & \text{if } X_{it} = 0.
\end{cases}
\]
Fast Computation and Standard Error Calculation

- **Standard FE estimator:**
  - “demean” both \( Y \) and \( X \)
  - regress demeaned \( Y \) on demeaned \( X \)

- **Weighted FE estimator:**
  - “weighted-demean” both \( Y \) and \( X \)
  - regress weighted-demeaned \( Y \) on weighted-demeaned \( X \)

- **Model-based standard error calculation**
  - Various robust sandwich estimators
  - Easy standard error calculation for matching estimators
Specification Test

- Should we use standard or weighted FE models?
- Standard FE estimator is more efficient if its assumption is correct
- Weighted FE estimator is consistent under the same assumption
- White (1980) shows that any misspecified least squares estimator converges to the weighted least squares that minimizes mean squared prediction error

Specification test:
- Null hypothesis: standard FE model is correct
- Does the difference between standard and weighted FE estimators arise by chance?
Mismatches in Two-Way FE Model

\[ Y_{it} = \alpha_i + \gamma_t + \beta X_{it} + \epsilon_{it} \]

**Units**

**Units**

**Time periods**

**Units**

- **Triangles**: Two kinds of mismatches
  - Same treatment status
  - Neither same unit nor same time
Some mismatches can be eliminated

You can NEVER eliminate them all
Weighted Two-Way FE Estimator

Proposition 4

The adjusted matching estimator

\[ \hat{\beta}^{M^*} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{1}{K_{it}} (\hat{Y}_{it}(1) - \hat{Y}_{it}(0)) \]

\[ \hat{Y}_{it}(x) = \begin{cases} \frac{1}{m_{it}} \sum_{(i', t') \in M_{it}} Y_{it'} + \frac{1}{n_{it}} \sum_{(i', t) \in N_{it}} Y_{i't} - \frac{1}{m_{it} n_{it}} \sum_{(i', t') \in A_{it}} Y_{i't'} & \text{if } X_{it} = x \\ \frac{m_{it} n_{it}}{m_{it} n_{it} + a_{it}} & \text{if } X_{it} = 1 - x \end{cases} \]

\[ A_{it} = \{(i', t') : i' \neq i, t' \neq t, X_{it'} = 1 - X_{it}, X_{i't} = 1 - X_{it}\} \]

\[ K_{it} = \frac{m_{it} n_{it}}{m_{it} n_{it} + a_{it}} \]

and \( m_{it} = |M_{it}|, n_{it} = |N_{it}|, \) and \( a_{it} = |A_{it} \cap \{(i', t') : X_{i't'} = X_{it}\}|. \)

is equivalent to the following weighted two-way fixed effects estimator,

\[ (\hat{\alpha}^{M^*}, \hat{\gamma}^{M^*}, \hat{\beta}^{M^*}) = \arg \min_{(\alpha, \beta, \gamma)} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} (Y_{it} - \alpha_{i} - \gamma_{t} - \beta X_{it})^2 \]
Weighted Two-way Fixed Effects Model

\[ \hat{\beta}^{M*} = \frac{1}{NT} \sum_{i=1}^{N} \sum_{t=1}^{T} \frac{1}{K_{it}} \left( \hat{Y}_{it}(1) - \hat{Y}_{it}(0) \right) \]
Proof by Picture: \( \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(2X_{it} - 1) \alpha_i^* = \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it}(2X_{it} - 1) \gamma_t^* = 0 \)

\[
W_{it} = \sum_{i' = 1}^{N} \sum_{t' = 1}^{T} w_{it'}^{i'} \quad \text{and} \quad w_{it'}^{i'} = \left\{
\begin{array}{ll}
\frac{m_{i' t'} n_{i' t'}}{m_{i' t'} n_{i' t'} + a_{i' t'}} & \text{if } (i, t) = (i', t') \\
\frac{m_{i' t'} n_{i' t'} + a_{i' t'}}{m_{i' t'} n_{i' t'} + a_{i' t'}} & \text{if } (i, t) \in M_{i' t'} \\
\frac{m_{i' t'} n_{i' t'} + a_{i' t'}}{(2X_{it} - 1)(2X_{it'} - 1)} & \text{if } (i, t) \in N_{i' t'} \\
0 & \text{if } (i, t) \in A_{i' t'} \\
\end{array}
\right.
\]

**Treatment**

C  T  C  C  T  0  1/6  0  -1/3  1/6

T  C  T  T  C  0  -1/2  0  1  -1/2

C  C  T  C  C  0  1/6  0  -1/3  1/6

T  T  T  C  T  0  1/6  0  -1/3  1/6

**Weights**
Cross Section Analysis = Weighted Time FE Model

- Treatment group
- Control group
- Counterfactual

Graph showing changes in average outcome over time between treatment and control groups.
First Difference = Weighted Unit FE Model

- **Average Outcome**
  - Treatment group
  - Control group
  - Counterfactual

- **Time Points**
  - Time t
  - Time t+1

- **Graph**
  - Y-axis: Average Outcome
  - X-axis: Time (t to t+1)
What about Difference-in-Differences (DiD)?

Average Outcome

- treatment group
- counterfactual
- control group

Please refer to the slide for a detailed explanation.
General DiD $=$ Weighted Two-Way (Unit and Time) FE

- $2 \times 2$: standard two-way fixed effects
- General setting: Multiple time periods, repeated treatments

Weights can be negative $\implies$ the method of moments estimator
Fast computation is not available
Controversy

- Rose (2004): No effect of GATT membership on trade
- Tomz et al. (2007): Significant effect with non-member participants

The central role of fixed effects models:

- Rose (2004): one-way (year) fixed effects for dyadic data
- Tomz et al. (2007): two-way (year and dyad) fixed effects
- Rose (2005): “I follow the profession in placing most confidence in the fixed effects estimators; I have no clear ranking between country-specific and country pair-specific effects.”
- Tomz et al. (2007): “We, too, prefer FE estimates over OLS on both theoretical and statistical ground”
Data and Methods

1. Data
   - Data set from Tomz et al. (2007)
   - 162 countries, and 196,207 (dyad-year) observations

2. Year fixed effects model:
   \[
   \ln Y_{it} = \alpha_t + \beta X_{it} + \delta \top Z_{it} + \epsilon_{it}
   \]
   - \(X_{it}\): Formal membership (1) Both vs. One, (2) One vs. None
   - \(Z_{it}\): 15 dyad-varying covariates (e.g., log product GDP)

3. Weighted one-way fixed effects model:
   \[
   \arg \min_{(\alpha, \beta, \delta)} \sum_{i=1}^{N} \sum_{t=1}^{T} W_{it} (\ln Y_{it} - \alpha_t - \beta X_{it} - \delta \top Z_{it})^2
   \]
   - Equal weights
   - Inverse-propensity score weighting
   - With and without restriction (one country shared)
Preliminary Empirical Results

(a) Dyad with Both GATT Members vs. One GATT Member

Estimated Effects (log of trade)

- Year Fixed Effects
- Equal Weights
- Propensity Score Weights

Average Treatment Effect (ATE)

Without Restriction

With Restriction

(b) Dyad with One GATT Member vs. No GATT Member

Estimated Effects (log of trade)

- Year Fixed Effects
- Equal Weights
- Propensity Score Weights

Average Treatment Effect (ATE)

Without Restriction

With Restriction
Concluding Remarks

- Standard one-way and two-way FE estimators are adjusted matching estimators
- FE models are not a magic bullet solution to endogeneity
- In many cases, adjustment is not sufficient for removing bias
- Key Question: “Where are the counterfactuals coming from?”
- Different causal assumptions yield different weighted FE models
- Weighted FE models encompass a large class of causal assumptions: stratification, first difference, propensity score weighting, difference-in-differences
- Model-based standard error, specification test
- Easy-to-use software, R package \texttt{wfe}, available
Send comments and suggestions to:

kimai@Princeton.Edu

More information about this and other research:

http://imai.princeton.edu