Why No Control Group Causes Collinearity in Event Study Designs

Bas Machielsen

Introduction

When conducting event study analyses with firm and year fixed effects, including a control group is crucial to avoid perfect multicollinearity. This document explains why omitting a control group leads to collinearity issues, using a detailed mathematical derivation.

The variable that becomes redundant is **any one of the cohort-specific event-time dummies**. For simplicity, let's focus on the dummy for being treated for the first time: the $k = 0$ dummy.

We will show that the $k = 0$ dummy variable can be perfectly constructed using:

1. The full set of **firm fixed effects** (dummies for each firm).
2. The full set of **year fixed effects** (dummies for each year).
3. The full set of **other relative-time dummies** ($k \neq 0$).

The Setup

Let's define our dummy variables:

- F_{ij}: A dummy variable that is 1 if the observation is for firm `i` and `j=i`, 0 otherwise. The coefficients on these are the α_i.
- T_{it}: A dummy variable that is 1 if the observation is in year `t` and `s=t`, 0 otherwise. The coefficients on these are the λ_t.
- D^k_{ik}: A dummy variable that is 1 if the observation for firm `i` is `k` periods from its event. The coefficients on these are the δ_k.

The regression model is:

$Y_{it} = \sum_{j=2}^{N} \alpha_j F_{ij} + \sum_{s=t_1}^{t_{max}} \lambda_s T_{is} + \sum_{k=m,k\neq -1,0}^{M} \delta_k D^k_{ik} + \delta_0 D^0_{i0} + \varepsilon_{it}$

(I've isolated the D^0_{i0} term, which corresponds to `k=0`, for clarity).

The collinearity problem means we can write D_{i0}^0 as a linear combination of the other regressors.

The Linear Combination

Let E_i be the year that firm `i` gets treated.

The dummy variable for $k = 0$ is defined as:

$D_{i0}^0 = 1$ if $t = E_i$, and 0 otherwise.

Now, consider the following sum for any observation (i, t).

$\sum_{j=2}^{N} F_{ij} \times T_{j,E_j}$

Let's unpack this:

- F_{ij}: This is 1 only if we are looking at an observation for firm `j`.
- T_{j,E_j}: This is a dummy variable that is 1 *only if the current year is the event year for firm j.*
- The product $F_{ij} \times T_{j,E_j}$ is 1 only for firm `j` in its specific event year E_j.
- The summation $\sum_{j=2}^{N}$ goes through every firm in the dataset.

Let's trace what this summation equals for a given observation $((i, t))$:

- The term F_{ij} will be 0 for all `j` except for $j = i$.
- Therefore, the entire sum collapses to just one term: $F_{ii} \times T_{i,E_i}$.
- F_{ii} is always 1 for firm i.
- So, the sum equals T_{i,E_i}. This is a dummy that is 1 if the current year t is the event year for the current firm i.

But that is **exactly the definition of the $k = 0$ dummy variable!**

So, we have shown that:

$D_{i0}^0 = \sum_{j=2}^{N} (F_{ij} \times T_{j,E_j})$

This looks complex, so let's simplify. For each firm j, let's define a new variable $C_j = T_{j,E_j}$. This is just the dummy for the specific calendar year when firm j was treated. Our equation becomes:

$D_{i0}^0 = \sum_{j=2}^{N} F_{ij} \times C_j$

This is a linear combination of the interaction between firm dummies and year dummies. Since both firm dummies (for the α_i) and year dummies (for the λ_t) are already in the model, this relationship creates perfect multicollinearity. The regression cannot estimate a unique coefficient for D_{i0}^0 because its effect is indistinguishable from this combination of firm and year effects.

A Concrete Example

Let's use the example from the slide:

- Firm 2 treated in 2023 ($E_2 = 2023$)
- Firm 3 treated in 2024 ($E_3 = 2024$)
- Firm 4 treated in 2025 ($E_4 = 2025$)
- No control group.

The regressors in the model are:

- Firm Dummies: $F2, F3, F4$.
- Year Dummies: $T2022, T2023, T2024, T2025, ...$
- Relative Time Dummies: $D(k = -2), D(k = -1)(\text{omitted}), D(k = 0), D(k = 1), ...$

Let's construct the $D(k = 0)$ dummy using the others:

- The interaction $F2 \cdot T2023$ is a dummy that is 1 only for Firm 2 in 2023.
- The interaction $F3 \cdot T2024$ is a dummy that is 1 only for Firm 3 in 2024.
- The interaction $F4 \cdot T2025$ is a dummy that is 1 only for Firm 4 in 2025.

If you add these three interaction terms together, what do you get?

$(F2 \cdot T2023) + (F3 \cdot T2024) + (F4 \cdot T2025)$

This new variable is equal to 1 whenever *any* firm is in its first treatment year, and 0 otherwise. This is **precisely the $D(k = 0)$ dummy variable**.

$D(k = 0) = F2 \cdot T2023 + F3 \cdot T2024 + F4 \cdot T2025$

Since the $D(k = 0)$ column in the regression matrix can be perfectly constructed by adding other columns that are already present (interactions of the included fixed effects), the matrix is not full rank, and the model is not identified.