Fast Forward: Replicating a Panel Data Approach Using Forward Difference-in-Differences

Abstract

I use forward difference-in-differences method of Li (2024) to narrowly replicate Shi and Huang (2023), who study the impact of China's Anti-Corruption Campaign on luxury watch imports using a panel data approach.

1 Introduction

As an extension of Hsiao et al. (2012), Shi and Huang (2023) (henceforth SH) develop the forward-selected panel data approach, so called because forward-selection is used to choose the control group for the treated unit. SH examine the causal impact of China's Anti-Corruption Campaign (ACC) on the growth rate of China's import of luxury watches. Recently, Li (2024) proposes forward difference-in-differences (fDID). Like SH, Li (2024) also uses forward selection to choose the control group. Greathouse et al. (2024) develops fDID for Stata, which I use to replicate SH.

2 The Models

A scalar is an italicized lowercase letter, y. A vector is a bold lowercase letter $\mathbf{y}$. Let a matrix be a bold uppercase letter $\mathbf{Y}$. Indexed by j, we observe $\mathcal{N} = \{1, 2, \ldots, N\}$ units where the set $\mathcal{N}$ has cardinality $N = |\mathcal{N}|$. $j = 1$ is the treated unit with the controls being $\mathcal{N}_0 = \mathcal{N} \setminus \{1\}$ whose cardinality is $N_0 = |\mathcal{N}_0|$. Let $\widehat{U} \subset \mathcal{N}_0$ be a subset of controls, with cardinality $U = |\widehat{U}|$. Indexed by t, pre and post-intervention periods are $\mathcal{T}_1 \coloneqq \{1, 2, \ldots, T_0\}$ and $\mathcal{T}_2 \coloneqq \{T_0 + 1, \ldots, T\}$. The full time series is $\mathcal{T} \coloneqq \mathcal{T}_1 \cup \mathcal{T}_2$, and let each of these sets have cardinalities of T_1, T_2, and T. Denote $\mathbf{y}_1 \coloneqq (y_{1t})_{t \in \mathcal{T}}$ as the column vector of outcomes for the treated unit and $\mathbf{Y}_{\mathcal{N}_0} \coloneqq (\mathbf{y}_j)_{j \in \mathcal{N}_0}$ as the $T \times N_0$ matrix of control unit outcomes. Denote $\mathbf{Y}_{\widehat{U}} \coloneqq (\mathbf{y}_j)_{j \in \widehat{U}}$ as the $T \times |\widehat{U}|$ submatrix of selected controls. Let the treatment effect be $\hat{\Delta}_{\widehat{U}t} \coloneqq y_{1t} - \hat{y}^0_{1t}(\widehat{U})$ where $\hat{y}^0_{1t}(\widehat{U})$ is the counterfactual based on set $\widehat{U}$. The causal estimand of interest is the average treatment effect on the treated, or $\widehat{ATT}_{\widehat{U}} = T_2^{-1} \sum_{t \in \mathcal{T}_2} \hat{\Delta}_{\widehat{U}t}$.

2.1 Forward-Selected Panel Data Approach

Algorithm 1 fsPDA Method

$\widehat{U} \leftarrow \emptyset$
while mBIC not satisfied **do**
  for each $i \in \mathcal{N}_0 \setminus \widehat{U}$ **do**
    Regress $\mathbf{y}_1 = \mathbf{Y}'_{\widehat{U}\cup\{i\}}\boldsymbol{\beta}_{\widehat{U}\cup\{i\}} + \boldsymbol{\beta}_0,\ \forall t \in \mathcal{T}_1$
    Compute $R^2_{\widehat{U}\cup\{i\}}$
  end for
  Select $i^* = \arg\max_{i\in\mathcal{N}_0} R^2_{\widehat{U}\cup\{i\}}$
  $\widehat{U} \leftarrow \widehat{U} \cup \{i^*\}$
end while
Return: $\widehat{U}$

fsPDA begins with an empty set of control units. In the first iteration, we run a linear regression of the treated unit's pre-treatment outcomes on the outcomes of each control unit. This produces N_0 submodels, each with an associated R-squared statistic. The control unit that maximizes the R-squared statistic in this iteration is denoted $i_1^* = \arg\max_{i\in\mathcal{N}_0} R^2_{\widehat{U}\cup\{i\}}$. We then add i_1^* to the set $\widehat{U}$, and remove it from $\mathcal{N}_0$. In the second iteration, regressions are run with the remaining $N_0 - 1$ control units, where each model now includes two control units: i_1^* and one from the remaining set. The control unit that maximizes the R-squared statistic in this iteration is selected as $i_2^* = \arg\max_{i\in\mathcal{N}_0} R^2_{\widehat{U}\cup\{i\}}$, and added to $\widehat{U}$, forming the updated candidate control group. The algorithm continues selecting control units iteratively until the modified Bayesian Information Criterion (mBIC) stops the selection process (Wang et al., 2009). The result is the final selected control group $\widehat{U}$. The fsPDA predictions are estimated via OLS

$$\hat{\mathbf{y}}_1^0(\widehat{U}) = \mathbf{Y}'_{\widehat{U}}\widehat{\boldsymbol{\beta}}_{\widehat{U}} + \widehat{\boldsymbol{\beta}}_0\ \forall t \in \mathcal{T} \tag{1}$$

where $\hat{\mathbf{y}}_1^0(\widehat{U})$ corresponds to our predictions and $\widehat{\boldsymbol{\beta}}_{\widehat{U}}$ corresponds to the least-squares coefficients. To conduct inference, SH use a heteroskedasticity and autocorrelation consistent estimator of the long run variance via $\widehat{\rho}^2_{\tau\widehat{U}} = T_2^{-1}\sum_{ts\in T_2}(\hat{\Delta}_{\widehat{U}t} - \bar{\Delta}_{\widehat{U}})(\hat{\Delta}_{Us} - \bar{\Delta}_U)\cdot\mathbf{1}\{|t-s| \leq \tau\}$ where τ is the number of lags and $\bar{\Delta}_{\widehat{U}}$ is the variance of the estimated treatment effect. The formula for the t-statistic is $t_{\widehat{U}} = \frac{\sqrt{T_2}\cdot\bar{\Delta}_{\widehat{U}}}{\hat{\rho}_{\tau\widehat{U}}}$.

2.2 Forward Difference-in-Differences

Algorithm 2 Forward Difference-in-Differences (fDID)

Initialize: $\widehat{U}_0 \leftarrow \emptyset$
for $k = 1$ **to** N_0 **do**
  for each $i \in \mathcal{N}_0 \setminus \widehat{U}_{k-1}$ **do**
    Estimate: $\mathbf{y}_1 = \widehat{\boldsymbol{\beta}}_0 + \mathbf{Y}'_{\widehat{U}_{k-1}\cup\{i\}} \widehat{\boldsymbol{\beta}}_{\widehat{U}_{k-1}\cup\{i\}}$ s.t. $\widehat{\boldsymbol{\beta}}_{\widehat{U}_{k-1}} = \frac{1}{U_{k-1}}, \quad \forall t \in \mathcal{T}_1$
    Compute: $R^2_k(\widehat{U}_{k-1} \cup \{i\})$
  end for
  Update $\mathcal{U} \cup \widehat{U}_k \leftarrow \mathcal{U} \cup \left(\widehat{U}_{k-1} \cup \left\{\text{argmax}_{i\in\mathcal{N}_0\setminus\widehat{U}_{k-1}} R^2_k(\widehat{U}_{k-1}\cup\{i\})\right\}\right)$
end for
Return: $\widehat{U} \coloneqq \text{argmax}_{\widehat{U}_k\in\mathcal{U}} R^2(\widehat{U}_k)$

2.3 Forward Difference-in-Differences

fDID proceeds iteratively over k, starting with an empty set of controls. Let $\mathcal{U} \coloneqq \{\widehat{U}_1, \widehat{U}_2, \ldots, \widehat{U}_{N_0}\}$ represent the set of candidate control groups. For $k = 1$, a one unit DID model is estimated N_0 times. For each control $i \in \mathcal{N}_0$, the pre-treatment R-squared statistic R^2_i is computed. The control unit that maximizes the R-squared is selected as $i^*_1 = \arg\max_{i\in\mathcal{N}_0} R^2_i$, forming the first candidate DID modell= $\widehat{U}_1 = \{i^*_1\}$. For $k = 2$, a DID model is estimated for each remaining $N_0 - 1$ control units, where each model includes i^*_1 from the previous iteration and one additional control. The control that maximizes the R-squared statistic when combined with i^*_1 is selected as $i^*_2 = \arg\max_{i\in\mathcal{N}_0\setminus\{i^*_1\}} R^2_{\{i^*_1,i\}}$, forming candidate set $\widehat{U}_2 = \{i^*_1, i^*_2\}$. For each k, a DID model is estimated for each remaining control unit $i \in \mathcal{N}_0 \setminus \widehat{U}_{k-1}$, combining the previously selected controls $\widehat{U}_{k-1}$ with one additional control. The control that maximizes the R-squared in each iteration is added to the candidate set $\widehat{U}_k$. This process continues until $k = N_0$. The control group returned by fDID is the one that maximizes the R-squared statistic across all iterations, $\widehat{U} \coloneqq \text{argmax}_{\widehat{U}_k\in\mathcal{U}} R^2(\widehat{U}_k)$. The final DID model is

$$\widehat{\mathbf{y}}^0_1(\widehat{U}) = \mathbf{Y}'_{\widehat{U}} \widehat{\boldsymbol{\beta}}_{\widehat{U}} + \widehat{\boldsymbol{\beta}}_0 \quad \text{s.t.} \quad \widehat{\boldsymbol{\beta}}_{\widehat{U}} = \frac{1}{U}. \tag{2}$$

Here is the formula for fDID's standard error:

$$\hat{\Omega} = \left[\left(\frac{T_2}{T_1} \right) \cdot T_1^{-1} \sum_{t \in \mathcal{T}_1} \hat{v}_{1t}^2 \right]^{0.5}, \quad \hat{v}_{1t} = \mathbf{y}_1 - (\mathbf{Y}'_{\widehat{U}} \widehat{\boldsymbol{\beta}}_{\widehat{U}} + \widehat{\boldsymbol{\beta}}_0) \tag{3}$$

A few comments: firstly as Li (2024) notes, fsPDA may overfit the pre-intervention data if fsPDA selects too many controls (see the San Diego or Atlanta examples from Li (2024)). In contrast, fDID cannot overfit since it only estimates a single parameter, $\widehat{\boldsymbol{\beta}}_0$ specifically. Additionally, fDID constrains $\widehat{\boldsymbol{\beta}}_{\widehat{U}}$ to be proportional, whereas fsPDA's coefficients are unconstrained.

3 Replication Process

I first downloaded the fsPDA package for R from Zhentao Shi's GitHub. This package contains the dataset as well as the corresponding R function. The outcome data were originally in dollars, coming from from the United Nations COMTRADE database. As described in SH, this value is transformed to (what appears to be since they do not specify) the month over month growth rate of the import values of 88 goods. I then used the provided R code in the fsPDA vignette to estimate the effect of the ACC on the growth rate of the import of luxury watches, where this good is compared to 87 controls that also were not bribe goods. I then concatenated the treatment vector and the control group matrix together in R into a single dataframe. I exported this dataframe as a comma separated value file, so that it could be used for analysis Stata. This concluded the work in R.

In Stata 17, I firstly, I installed the fdid package as described in Greathouse et al. (2024). I then imported the wide-form dataset, the one that was created in R, into Stata. I then generated a time variable from $t = \{1, 2, \ldots, 71\}$, which would represent the time periods of interest (February 2010 to December of 2015). I then reshaped this dataset to long format, such that each of the 88 total import goods (the unit of analysis) had one observation per time period. Treatment begins January of 2013, giving us 35 pre-intervention periods and 36 post-intervention periods. The treatment variable was equal to 1 for $t > 35$ and if $j = 1$, else 0.

4 Results of Replication

Figure 1 presents the results of the fDID method. The returned average treatment

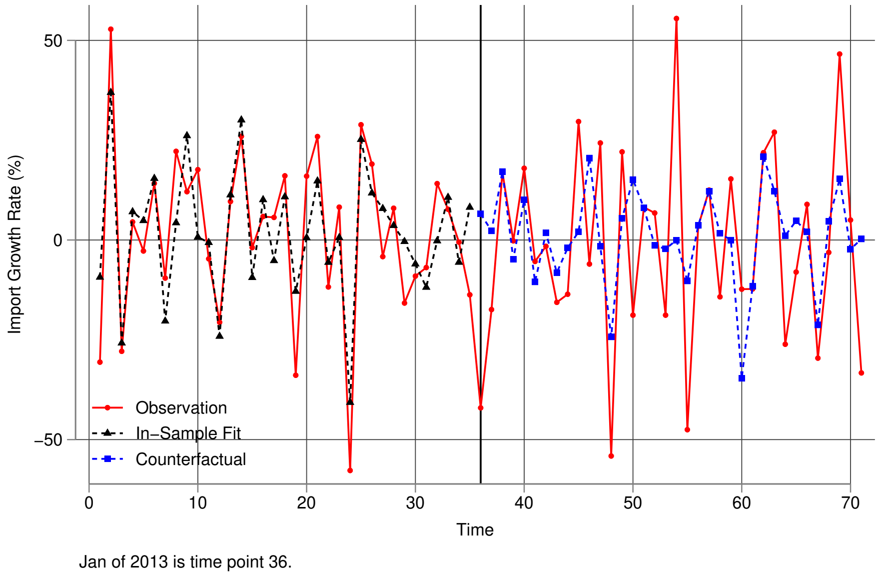

Figure 1: Estimated via fDID

effect on the treated unit using fDID is −0.0252, or a 2.52% decrease in the monthly import of watches in the post-intervention period. For fsPDA, the ATT was −0.0308, a 3.1% decrease. The standard error for fDID is 0.0265, and its t-statistic is 0.95. For fsPDA, the standard error is 0.0274 and its t-statistic −1.12. Both estimators have similar ATTs and standard errors.

Now I discuss the selected control units. Since the innovation of both algorithms are predicated on their ability to choose the right control group among a pool of many potentially irrelevant units, listing the selected controls makes sense. The fsPDA method chooses control units with the numeric ids 60, 45, and 25. The control units selected by fDID are: Control 60, Control 82, Control 26, Control 58, Control 45, Control 50, Control 15, Control 19, Control 21, Control 54, Control 81, Control 31, Control 85, and Control 51. The pre-intervention R-squared from fsPDA is 0.777, and the analogous metric for fDID is 0.707, or a difference of around 7 percentage points. This suggests that the pre-intervention parallel trend assumption seems to hold for the fDID method. The difference of R-squared statistics is to be expected since the unconstrained regression fsPDA is based on is more likely to obtain better fit anyways

(Gardeazabal & Vega-Bayo, 2017). Overall though, the difference is marginal: both methods select two of the same control units and have good in-sample fit and produce similar out-of-sample predictions.

There are a few caveats to this replication, however: firstly, the original R data do not index the names of the controls to the control units. SH list “knitted or crocheted fabric”, “cork and articles of cork”, and “salt, sulfur, earth, stone, plaster, lime and cement”. However, the control untis are not named in the R data. Otherwise, it would be interesting to see which specific controls were selected by fDID. SH also write in the published paper “the t-statistic is -2.457, with a p-value 1.40%”. However, when the code is ran (as of October 5th 2024), we get the t-statistic I list above. It is not obvious why this discrepancy exists, as a t-statistic of 2.457 would reject the null hypothesis of 0 ATT at the 5% size. However, this aside, the results are quite consistent using the currently public version of the R code.

5 Conclusion

Broadly, I succeeded in replicating the empirical findings from Shi and Huang (2023). In terms of future research, it would be interesting to more formally compare both fDID and fsPDA's selection methods to different synthetic control methods. The idea would be to theoretically explain why the algorithms select the control units they do. A corollary to this would be to use finely tuned and realistic synthetic studies to measure the degree to which each method selects the proper set of control units.

References

Gardeazabal, J., & Vega-Bayo, A. (2017). An empirical comparison between the synthetic control method and hsiao et al.'s panel data approach to program evaluation. *Journal of Applied Econometrics*, *32*(5), 983–1002. https://doi.org/10.1002/jae.2557

Greathouse, J., Coupet, J., & Sevigny, E. (2024, August). Greed is good: Estimating forward difference-in-differences in stata. https://jgreathouse9.github.io/publications/FDIDSJ.pdf

Hsiao, C., Steve Ching, H., & Ki Wan, S. (2012). A panel data approach for program evaluation: Measuring the benefits of political and economic integration of hong kong with mainland china. *Journal of Applied Econometrics*, *27*(5), 705–740. https://doi.org/10.1002/jae.1230

Li, K. T. (2024). Frontiers: A simple forward difference-in-differences method. *Marketing Science*, *43*(2), 239–468. https://doi.org/10.1287/mksc.2022.0212

Shi, Z., & Huang, J. (2023). Forward-selected panel data approach for program evaluation. *Journal of Econometrics*, *234*(2), 512–535. https://doi.org/10.1016/j.jeconom.2021.04.009

Wang, H., Li, B., & Leng, C. (2009). Shrinkage Tuning Parameter Selection with a Diverging number of Parameters. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *71*(3), 671–683. https://doi.org/10.1111/j.1467-9868.2008.00693.x