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

Abstract

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

1 Introduction

Control group selection has become a topic of interest for causal methodologists, where some pre-processing of the control units is done before treatment effects are estimated. For example, Bayani (2021) employ functional PCA and a k-means clustering approach to choose control units for the Robust PCA-Synthetic Control Method. Also for synthetic controls, Yu et al. (2022) experiment with a simple variance minimization method.

Shi and Huang (2023) (henceforth SH) develop the forward-selected panel data approach (fsPDA) for program evaluation as an extension of Hsiao et al. (2012). Exploiting China's Anti-Corruption Campaign (ACC) as the treatment, SH construct a counterfactual for the monthly growth rate of China's import of luxury watches. As the name of fsPDA suggests, SH employ a forward-selection algorithm to choose the control group via a series of least-squares regressions. A key theoretical innovation of fsPDA is to extend the panel data approach to high-dimensions. Li (2024) proposes forward difference-in-differences (fDID), which was recently developed for Stata by Greathouse et al. (2024). fDID also exists in Python. Like SH, Li (2024) also uses forward selection to choose the control group. Obviously however, fDID is based on the difference-in-differences (DID) method.

The task of this paper is to use the fDID method to replicate the original findings presented in SH. The findings replicate fairly successfully, both in terms of the average treatment effect on the treated (ATT) and the standard error. I first detail the control group selection algorithms and subsequent estimation for both approaches. Then, I discuss how the replication was conducted and present the results.

2 The Models

2.1 Notations

I begin by standardizing the notations to be used for both methods. 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 $\mathcal{T}_1 := \{1, 2, \ldots, T_0\}$ and $\mathcal{T}_2 := \{T_0 + 1, \ldots, T\}$ respectively. The full time series is $\mathcal{T} := \mathcal{T}_1 \cup \mathcal{T}_2$. Each of these discrete sets have cardinalities of T_1, T_2, and T. Denote $\mathbf{y}_1 := (y_{1t})_{t \in \mathcal{T}}$ as the $T \times 1$ column vector of outcomes for the treated unit and $\mathbf{Y}_{\mathcal{N}_0} := (\mathbf{y}_j)_{j \in \mathcal{N}_0}$ as the $T \times N_0$ matrix of control unit outcomes. Denote $\mathbf{Y}_U := (\mathbf{y}_j)_{j \in \widehat{U}}$ as the $T \times U$ submatrix of selected controls. Let the treatment effect be $\hat{\Delta}_{\widehat{U}t} := y_{1t} - \hat{y}_{1t}^0(\widehat{U})$ where $\hat{y}_{1t}^0(\widehat{U})$ is the counterfactual based on set $\widehat{U}$. The causal estimand of interest is the ATT, or $\widehat{ATT}_{\widehat{U}} = T_2^{-1} \sum_{t \in \mathcal{T}_2} \hat{\Delta}_{\widehat{U}t}$.

2.2 Forward-Selected Panel Data Approach

Algorithm 1 fsPDA Method

1: $\widehat{U} \leftarrow \emptyset$
2: **while** mBIC not satisfied **do**
3:     **for** each $i \in \mathcal{N}_0 \setminus \widehat{U}$ **do**
4:         Regress $\mathbf{y}_1 = \mathbf{Y}'_{\widehat{U} \cup \{i\}} \boldsymbol{\beta}_{\widehat{U} \cup \{i\}} + \boldsymbol{\beta}_0, \forall t \in \mathcal{T}_1$
5:         Compute $R^2_{\widehat{U} \cup \{i\}}$
6:     **end for**
7:     Select $i^* = \arg\max_{i \in \mathcal{N}_0} R^2_{\widehat{U} \cup \{i\}}$
8:     $\widehat{U} \leftarrow \widehat{U} \cup \{i^*\}$
9: **end while**
10: **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 pre-intervention 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, linear 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 until the modified Bayesian Information Criterion (mBIC) stops the selection process (Wang et al., 2009). The algorithm returns the final selected control group $\widehat{U}$, where ideally $U \ll N_0$. The fsPDA predictions are estimated via OLS

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

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

2.3 Forward Difference-in-Differences

Algorithm 2 Forward Difference-in-Differences (fDID)

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

Next I describe fDID. fDID proceeds iteratively over k total candidate iterations. Like fsPDA, we begin with an empty set of controls. Let $\mathcal{U} := \{\widehat{U}_1, \widehat{U}_2, \ldots, \widehat{U}_{N_0}\}$ represent the set of candidate control groups. For $k = 1$, a one control unit DID model

is estimated $\forall\, i \in \mathcal{N}_0$. For each submodel, the pre-treatment R-squared statistic, R_i^2, is computed. The control unit that maximizes the R-squared is selected as $i_1^* = \arg\max_{i \in \mathcal{N}_0} R_i^2$, forming the first candidate control group $\widehat{U}_1 = \{i_1^*\}$. 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 pre-intervention R-squared statistic when combined with i_1^* is selected as $i_2^* = \arg\max_{i \in \mathcal{N}_0 \setminus \{i_1^*\}} R_{\{i_1^*, i\}}^2$, forming candidate set $\widehat{U}_2 = \{i_1^*, i_2^*\}$. For each k, a DID model is estimated for each remaining control unit $i \in \mathcal{N}_0 \setminus \widehat{U}_{k-1}$. Each iteration combines the previously selected controls $\widehat{U}_{k-1}$ with one additional control to form a new candidate set $\widehat{U}_k$. These candidate sets are subsequently added to $\mathcal{U}$. This process continues until $k = N_0$. The control group returned by fDID is the one that maximizes the R-squared statistic across all candidate iterations, $\widehat{U} := \operatorname{argmax}_{\widehat{U}_k \in \mathcal{U}} R^2(\widehat{U}_k)$. The fDID predictions take the following form

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

where we estimate DID except only using the selected control group $\widehat{U}$. Here is the formula for fDID's standard error:

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

which is then used to construct confidence intervals and t-statistics.

A few comments are in order: firstly as Li (2024) notes, fsPDA may possibly 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 $\widehat{\boldsymbol{\beta}}_0$, or the time-fixed effect from the standard two-way fixed effects setup. Additionally, fDID constrains $\widehat{\boldsymbol{\beta}}_{\widehat{U}}$ to be proportional, whereas fsPDA's coefficients are unconstrained. Practically, this means it is likely the case that the fsPDA method may, on average, have better in-sample fit compared to fDID, making it a more flexible estimator in the case that fDID's parallel pre-intervention trends assumption does not hold.

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 exported from 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 had one observation per time period. Treatment begins January of 2013, where $t = 36$, giving us 35 pre-intervention periods and 36 post-intervention periods.

4   Results of Replication

Figure 1 presents the results of the fDID method. The returned average treatment 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, or roughly a 3.1% decrease in the import of luxury watches. The calculated 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 very similar 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
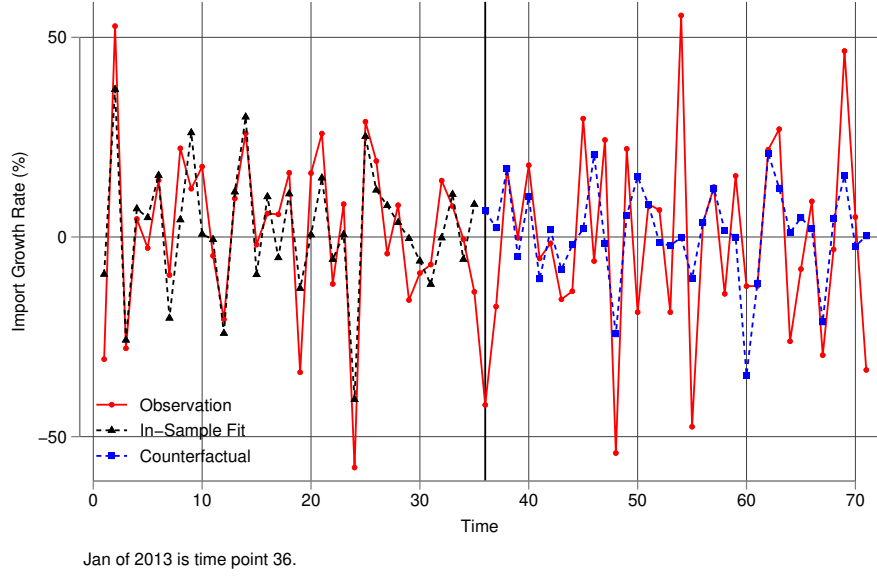
Jan of 2013 is time point 36.

Figure 1: Estimated via fDID

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, as we noted above, is to be expected since the unconstrained regression fsPDA is based on tends to obtain better in-sample fit (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 actual names of the control units 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

Bayani, M. (2021). Robust pca synthetic control. https://doi.org/10.48550/ARXIV. 2108.12542

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

Yu, L., Tran, T., & Lee, W.-S. (2022). Revitalising the silk road: Evidence from railway infrastructure investments in northwest china.