Forward and Onward? Replicating a Research Claim 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 study the impact of China's Anti-Corruption Campaign on luxury watch imports. My results largely agree with their treatment effects and uncertainty estimation.

1 Introduction

Lan and Li (2018) find luxury watch imports covary with changes in Chinese leadership, proposing the import of expensive watches as a proxy for corruption. 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. Their case study is 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 the forward difference-in-differences (fDID) method for program evaluation. Like SH, Li (2024) also advocates for forward selection to choose the control group for the treated unit. Once either method selects the control group, their next step is to conduct inference for the treatment effect. Greathouse et al. (2024) develops fDID for Stata. The goal of this paper is to replicate SH using the fDID method. To fix ideas, I begin by explaining the basic estimation and inference for both algorithms. After, I detail the ground rules for the replication process. I then present the results and conclude.

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 $\mathbf{Y}$. 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

$\mathcal{N}_0 = \mathcal{N} \setminus \{1\}$. Time is indexed by t. Our outcomes, thus, are y_{jt}. Denote pre-post-policy periods as $\mathcal{T}_1 = \{1, 2, \ldots, T_0\}$ and $\mathcal{T}_2 = \{T_0 + 1, \ldots, T\}$, where $\mathcal{T} = \mathcal{T}_1 \cup \mathcal{T}_2$. Both algorithms select some subset of controls, $\widehat{U} \subset \mathcal{N}_0$, to impute the counterfactual. Let the treatment effect for either algorithm 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

fsPDA begins with an empty set of control units. The first step is to run a linear regression of the pre-intervention outcomes of the treated unit upon the pre-intervention outcomes of each control unit, giving us N_0 submodels. We add the control unit which maximizes the R-squared statistic to set $\widehat{U}$, removing it from set $\mathcal{N}_0$. We then, including the first selected unit as a predictor, use the remaining $N_0 - 1$ control units in linear regression submodels, giving us $N_0 - 1$ 2-unit models. Of these submodels, we choose the control unit that, in conjunction with the first selected control unit, maximizes the pre-intervention R-squared statistic. This is the new candidate control group. The algorithm proceeds for the rest of the control units, adding control units until a modified Bayesian information criterion stops selection (Wang et al., 2009). This returns the selected control group. The fsPDA predictions are calculated via the following linear regression model

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

where $\mathbf{Y}'_{\widehat{U}t}$ is the submatrix of control unit outcomes based on the selected set $\widehat{U}$ and $\widehat{\boldsymbol{\beta}}_{\widehat{U}}$ corresponds to the least-squares coefficients for those selected control units. To conduct inference, SH use a heteroskedasticity and autocorrelation consistent estimator of the long run variance via $\hat{\rho}^2_{\tau\widehat{U}} = T_2^{-1} \sum_{ts \in T_2} (\hat{\Delta}_{\widehat{U}t} - \bar{\Delta}_{\widehat{U}})(\hat{\Delta}_{U_s} - \bar{\Delta}_U) \cdot \mathbf{1}\{|t-s| \leq \tau\}$ 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

Parallel trends is the identifying assumption of difference-in-differences designs, and forward DID is no exception to this rule. Parallel trends, at the simplest level, says

that *absent* the treatment, the difference between the counterfactual and the average of the control group would be constant. However, parallel trends is a statement about the counterfactual, meaning we in principle cannot test it. Thus, we focus on validating parallel pre-intervention trends. The basic idea of fDID is that if parallel pre-intervention trends do not hold with all controls, it may still hold with some controls, $\widehat{U} \subset \mathcal{N}_0$. Below, I describe the selection method for fDID.

Much like fsPDA, fDID begins with an empty set of controls, which we iteratively build upon. To build the first candidate model, we estimate the following regression submodels N_0 times

$$\mathbf{y}_{1t} = \mathbf{Y}'_{\mathcal{N}_0 t}\widehat{\boldsymbol{\beta}}_{\mathcal{N}_0} + \widehat{\boldsymbol{\beta}}_0, \; \forall\, t \in \mathcal{T}_1 \quad \text{s.t. } \widehat{\boldsymbol{\beta}}_{\mathcal{N}_0} = \frac{1}{N_0} \tag{2}$$

where each control is used per estimation plus a constant. We calculate the R-squared statistic, which corresponds to each submodel. We then choose the submodel which maximizes the R-squared statistic, and we add that corresponding control unit to set $\widehat{U}$. This model is the first candidate DID model. The second iteration proceeds similarly where, like SH, we include the first selected control unit along with the remaining $N_0 - 1$ controls in two-control unit DID models. We calculate the R-squared statistic for each submodel, returning $N_0 - 1$ R-squared statistics. We find the submodel with the maximum R-squared of these iterations, and add the control to set $\widehat{U}$. This model becomes the second candidate DID model. We continue adding controls until there are as many candidate DID models as there are controls. The final control group selected by fDID is the the candidate model with the highest R-squared statistic. With the control group selected, the final DID model is

$$\hat{\mathbf{y}}^0_{1t}(\widehat{U}) = \mathbf{Y}'_{\widehat{U}t}\widehat{\boldsymbol{\beta}}_{\widehat{U}} + \widehat{\boldsymbol{\beta}}_0 \quad \text{s.t. } \widehat{\boldsymbol{\beta}}_{\widehat{U}} = \frac{1}{|\widehat{U}|}, \tag{3}$$

Standard errors for fDID are computed like

$$\hat{\Omega} = \left[\left(\frac{T_2}{T_1}\right) \cdot T_1^{-1} \sum_{t \in \mathcal{T}_1} \hat{v}^2_{1t}\right]^{0.5}, \quad \hat{v} = \mathbf{y}_{1t} - (\mathbf{Y}'_{\widehat{U}t}\widehat{\boldsymbol{\beta}}_{\widehat{U}} + \widehat{\boldsymbol{\beta}}_0) \tag{4}$$

which we then use to construct 95% confidence intervals.

There are a few distinctions between the approaches: 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, $\widehat{\beta}_0$ specifically. Additionally, fDID constrains $\widehat{\boldsymbol{\beta}}_{\widehat{U}}$ to be proportional, whereas fsPDA's coefficients may both vary and be positive or negative. This means that the fsPDA method is more flexible than fDID, and can be used in situations where no subset of controls would, by this algorithm, satisfy the parallel pre-trends assumption.

3 Replication Process

Now I describe the replication process. I downloaded the fsPDA package for R from Zhentao Shi's GitHub, which contains the package as well as the treatment vector and the control group matrix. The outcome data, as described in SH, come from the United Nations. 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. I then concatenated the treatment vector and the control group matrix together in R. I exported this dataframe as a comma separated value file, so that it could be used in Stata. This concluded the work in R.

The analysis in Stata proceeded similarly: firstly, I installed the fdid package as described in Greathouse et al. (2024). I began by importing the wide-form dataset 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 (Febraury 2010 to December of 2015). Treatment begins January of 2013, giving us 35 pre-intervention periods and 36 post-intervention periods. The treatment is equal to 1 for $t > 35$ and if $j = 1$, else 0. 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.

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 percent decrease in the monthly import of watches in the post-intervention period. For fsPDA, the ATT was -0.0308. 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. In terms of point estimates, both estimators agree that the ACC had a small impact on the import of luxury

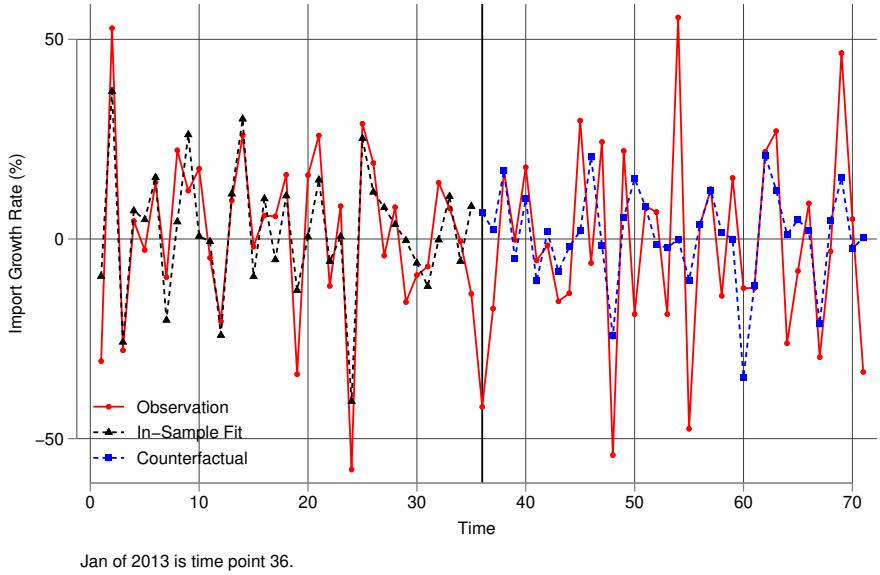

Figure 1: Estimated via fDID

watches. Their standard errors are also quite close, which suggests they quantify the precision of the treatment effect similarly in this instance.

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. In terms of pretreatment fit, 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 differnce 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 beyond their numeric ids. 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/https://doi.org/10.1002/jae.1230

Lan, X., & Li, W. (2018). Swiss watch cycles: Evidence of corruption during leadership transition in china. *Journal of Comparative Economics*, *46*(4), 1234–1252. https://doi.org/https://doi.org/10.1016/j.jce.2018.07.019

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