Forward and Onward? Replication of a Research Claim Using A Novel Difference-in-Differences Method

Jared Amani Greathouse
jgreathouse3@student.gsu.edu
Policy Data Analysis
Andrew Young School of Policy Studies, Georgia State University

October 6, 2024

Abstract

Shi and Huang (2023) propose a forward-selected approach to program evaluation as an extension of Hsiao et al. (2012). In particular, the central case study of Shi and Huang (2023) is the Chinese Anti-Corruption campaign. The goal was to see how the campaign affected the import of luxury watches. Shi and Huang (2023) develop and justify a forward-selection algorithm to choose the ideal control group from a pool of 87 candidate control units. Using OLS, they estimate the out of sample counterfactual for watch imports had the anti-corruption campaign never happened. Li (2024) proposes a forward-selected difference-in-differences method. I use this method to narrowly replicate the results of Shi and Huang (2023). Fortunately, my results largely agree with theirs, both in terms of treatment effects as well as uncertainty estimation.

1 Introduction

Shi and Huang (2023) (henceforth SH) study the causal impact of China's 2012 Anti-Corruption Campaign (ACC) on the import of luxury watches. Beginning in January of 2013, SH note "the campaign aimed at cracking down graft and power abuse in all party apparatus, government bureaucracies and military departments." Corruption via direct bribery can be hard to measure empirically. Lan and Li (2018) find luxury watch imports covary with changes in Chinese leadership, offering that metric as a proxy for political corruption. SH use a forward-selected panel data approach (fsPDA) to study the impact of the anti-corruption campaign, estimating a counterfactual for how the growth rate of luxury watch imports would have looked absent the ACC.

Li (2024) proposes forward difference-in-differences (fDID), based on the difference-in-differences (DID) method. A software package to implement fDID was provided by Greathouse et al. (2024) in Stata 16. Like SH, a forward selection algorithm is used in fDID to select the control group for the treated unit. My results using fDID are broadly consistent with SH in terms of the treatment effect as well as the uncertainty underlying the ATT.

The paper is organized as follows: section 2 explains fsPDA by SH and fDID by Li (2024). Section 3 explains the replication process, and section 4 presents these results. Finally, section 5 concludes.

2 The Models

I briefly introduce notations. A scalar is an italicized lowercase letter, y, vector is a bold lowercase letter $\mathbf{y}$, and a matrix is 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. The causal estimand of interest is the ATT, or $\widehat{ATT}_{\widehat{U}} = \frac{1}{T_2} \sum_{t \in \mathcal{T}_2} \left(y_{1t} - \hat{y}^0_{1t}(\widehat{U}) \right)$.

2.1 Forward-Selected Panel Data Approach

Algorithm 1 Forward-Selected Panel Data Approach

Initialize $\widehat{U}_0 \leftarrow \emptyset$
for $k = 1$ to N_0 **do**
  for $i \in \mathcal{N}_0 \setminus \widehat{U}_{k-1}$ **do**
    Estimate $\mathbf{y}_{1t} = \mathbf{Y}'_{\widehat{U}_{k-1} \cup \{i\}, t} \widehat{\boldsymbol{\beta}}_{\widehat{U}_{k-1} \cup \{i\}} + \widehat{\boldsymbol{\beta}}_0, \quad \forall t \in \mathcal{T}_1$
    Compute R^2_i, the R^2 statistic for the regression model
  end for
  Select $i^* = \arg\max_{i \in \mathcal{N}_0} R^2_i$
  Update $\widehat{U}_k \leftarrow \widehat{U}_{k-1} \cup \{i^*\}$
  if $k > \widehat{R} = \arg\min_{k \in \{1,2,\ldots N_0\}} \log\left(\sigma^2_{\widehat{U}k} + \log\log N_0 \cdot k \left(\log T_1 \right) / T_1 \right)$ **then**
    Break
  end if
end for
Return $\widehat{U}_k$

I summarize the key points of the fsPDA selection method, following SH. We begin with an empty set of control units. We estimate 1 to N_0 candidate iterations, adding a new control upon each iteration. The forward selection algorithm takes the most aggressive step in each iteration to maximize the pre-intervention R-squared statistic. 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, as we did in iteration 1, giving us $N_0 - 1$ 2-unit models. We add to set $\widehat{U}$ the new control unit that maximizes the pre-intervention R-squared statistic, making this the new candidate control group/regression model. The algorithm adds control units until a modified Bayesian information criterion stops selection (Wang et al., 2009), returning us the final pool of selected units. The fsPDA predictions are calculated via the following linear regression model

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

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. SH use a heteroskedasticity and autocorrelation consistent estimator of the long run variance to compute the t-statistic, standard error, and p-value.

2.2 Forward Difference-in-Differences

Algorithm 2 Forward Difference-in-Differences

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 $\mathbf{y}_{1t} = \widehat{\boldsymbol{\beta}}_0 + \mathbf{Y}'_{\widehat{U}_{k-1}\cup\{i\},t}\widehat{\boldsymbol{\beta}}_{\widehat{U}_{k-1}\cup\{i\}}$ s.t. $\widehat{\boldsymbol{\beta}}_{\widehat{U}_{k-1}} = \frac{1}{|\widehat{U}_{k-1}|}, \quad \forall t \in \mathcal{T}_1$
    Compute the R-squared statistic: $R_k^2(\widehat{U}_{k-1} \cup \{i\})$
  end for
  $\widehat{U}_k \leftarrow \widehat{U}_{k-1} \cup \left\{\text{argmax}_{i \in \mathcal{N}_0 \setminus \widehat{U}_{k-1}} R_k^2(\widehat{U}_{k-1} \cup \{i\})\right\}$
end for
Return $\widehat{U}^* \leftarrow \text{argmax}_{k \in \{1,2,\ldots,N_0\}} R_k^2(\widehat{U}_k)$

fDID is simply the standard two-way fixed effects DID model with a selected control group which is collected by a related forward-selection algorithm. Because it is a DID method, fDID is based on a parallel trends assumption. Parallel trends says that *absent* the treatment, the difference between the counterfactual and the average of the control group would be constant, expressed in scalar form like $\hat{y}_{1t}^0 - \bar{y}_{\mathcal{N}_0 t} = \hat{\beta}_0 + \epsilon$ where $\bar{y}_{\mathcal{N}_0 t} \coloneqq \frac{1}{N_0}\sum_{j \in \mathcal{N}_0} y_{jt}$. However, parallel trends is a statement about the counterfactual, meaning we in principle cannot test it in the post-intervention period. Thus, we focus on validating this in the pre-intervention period.

The basic idea of fDID is that if parallel pre-intervention trends do not hold with all controls, they may hold with some controls, $\widehat{U} \subset \mathcal{N}_0$. As described in Doudchenko and Imbens (2016), the proportionality of the betas means $\mathbf{Y}'_{\widehat{U}t}\widehat{\boldsymbol{\beta}}_{\widehat{U}}$ is the average of control units. The only estimated coefficient $\boldsymbol{\beta}_0$ is the constant. So, if the selected control group is parallel to the pre-intervention trends of the treated unit, it should be well approximated by the β_0 constant added to the control group mean. Much like fsPDA, we begin with an empty set of controls. Then, we estimate N_0 one control unit DID models, without covariates, using each control unit. We add the unit which maximizes the pre-intervention R-squared 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 (in this case 87) two-control unit DID models. The best two unit DID model becomes the second candidate DID model. Unlike fsPDA, we continue adding controls until there are as many candidate DID models as there are controls. The final model selected by fDID is the the candidate model with the highest pre-treatment R-squared statistic. The final regression model is

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

Standard errors for fDID are computed like

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

which we then use to construct 95% confidence intervals.

The key estimation difference between both algorithms, as highlighted in Li (2024), is that fsPDA may overfit the pre-intervention data if fsPDA selects too many controls (see the San Diego or Atlanta examples from Li (2024)). Overfitting to the in-sample period may harm out of sample predictions, which are the ones we care most about. In contrast, fDID cannot overfit since it only estimates a single parameter, $\boldsymbol{\beta}_0$ specifically. Additionally, fDID constrains $\widehat{\boldsymbol{\beta}}_{\widehat{U}}$ to be proportional, whereas fsPDA's coefficients may both vary and be positive or negative.

3 Replicating Shi and Huang (2023)

To replicate the results, I downloaded the fsPDA package for R from Zhentao Shi's github. I then used the provided R code to estimate the effect as it is written in the vignette for the fsPDA package. I then concatenated the treatment vector and the control group matrix together in R, exporting this dataframe as a comma separated value file. This concluded the work in R. I had to take a few more steps for the Stata analysis: firstly, I installed the fdid package as described in Greathouse et al. (2024). In Stata, I imported the wide-form dataset that was created in R. I then generated a time variable from $t = \{1, 2, \ldots, 71\}$. I then reshaped this dataset to long format. For fDID's purposes, the treatment is equal to 1 for $t > 35$ and if the unit is luxury watches, else 0.

4 Results of Replication

Figure 1 presents the results of the fDID method, using the exact same data as provided by Shi and Huang (2023). The returned average treatment effect on the treated

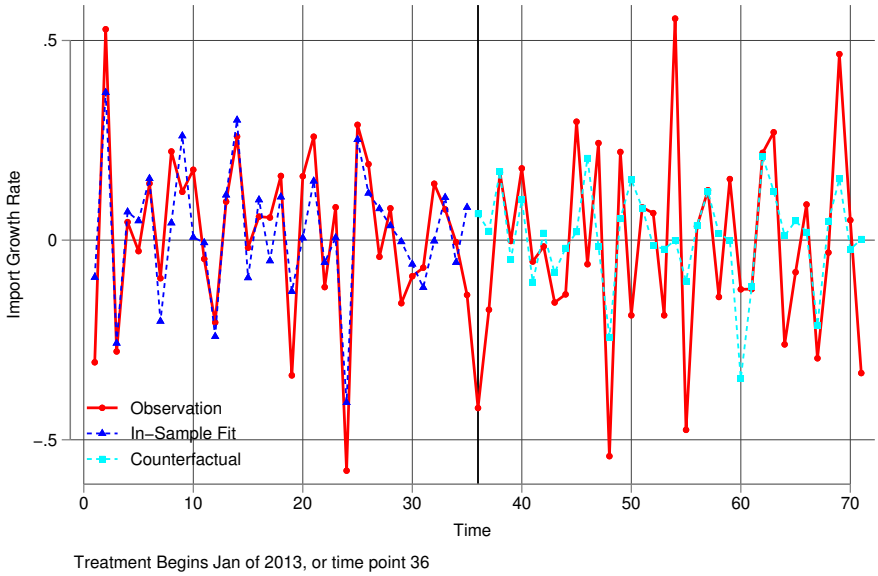

Figure 1: Estimated via fDID

unit using fDID is -0.02523, or a 2.52 percent decrease in the monthly import of watches in the post-intervention period. For fsPDA, the ATT was -0.03089581. The

standard error for fDID is 0.02653, and its t-statistic is 0.95. For fsPDA, the standard error is 0.02743310 and its t-statistic -1.12622398. In terms of point estimates, both estimators agree that the ACC had a small impact on the import of luxury watches.

Now I discuss the selected control units. Since 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 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. Overall, both methods select two of the same control units and have good in-sample fit, thus producing similar out-of-sample counterfactual 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. In Shi and Huang (2023), the authors list "knitted or crocheted fabric", "cork and articles of cork", and "salt, sulfur, earth, stone, plaster, lime and cement". However, in the R data, we only see C1, C2... and so on. 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 and 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. The ATTs are very similar, as are the estimated uncertainty statistics. In terms of future research, it would be interesting to more formally compare both fDID and fsPDA's selection methods to other methods which pre-process the control group before conducting estimation, such as the synthetic control method. 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

Doudchenko, N., & Imbens, G. W. (2016, October). Balancing, regression, difference-in-differences and synthetic control methods: A synthesis. https://doi.org/10.3386/w22791

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