Control Group Selection in High Dimensions: A Narrow Replication Using Panel Data Approach and Difference-in-Differences*

October 11, 2024

Abstract

I use the forward difference-in-differences method of Li (2024) and Stata's regression control method by Yan and Chen (2022) 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.

Key words: Greedy algorithm, causal inference, synthetic control method

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, clustering approaches are sometimes invoked. Bayani (2021) employs functional PCA and a k-means clustering approach to choose control units for the Robust PCA-Synthetic Control Method. Dynamic time warping has also been used (Biazoli et al., 2024). 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), extending the original formulation by Hsiao et al. (2012) to when there are very many controls relative to pre-treatment periods. The fsPDA algorithm chooses the control group for a treated unit via a series of least-squares regressions. Empirically, SH exploit China's January 2013 Anti-Corruption Campaign (ACC) as the treatment. Since luxury watches are a bribe good

*We thank Zhentao Shi, Damian Clarke, Eric Sevigny, and Jason Coupet for helpful comments.

(Lan & Li, 2018), SH construct a counterfactual for the monthly growth rate of China's import of luxury watches absent the ACC. Their data have 35 pre-intervention periods and 87 candidate control units. Extending SH, Li (2024) proposes forward difference-in-differences (fDID). Li (2024) also uses forward selection to choose the control group for a treated unit. However instead of OLS, Li's forward selection method uses a series of difference-in-differences (DID) regressions to select the control group. fDID was recently developed for Stata by Greathouse et al. (2024). fsPDA also has an existing Stata implementation via the *Regression Control Method* (RCM). Yan and Chen (2022) explicitly citing SH's fsPDA as one of the available models that the RCM can implement, in addition to Hsiao et al. (2012).

Organization The task of this paper is to use Stata's fDID and RCM to replicate the empirical findings presented in SH. fDID replicates SH fairly successfully, both in terms of the average treatment effect on the treated (ATT) and the standard error. The Stata RCM does not replicate, however. 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 Estimation

Notations I begin by standardizing the notations to be used for both methods. Let ":=" be a definition. 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 $\mathcal{N} = \{1, 2, \dots, N\}$ 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, with cardinalities $T_1 = |\mathcal{T}_1|$ and $T_2 = |\mathcal{T}_2|$, are $\mathcal{T}_1 \coloneqq \{1, 2, \dots, T_0\}$ and $\mathcal{T}_2 \coloneqq \{T_0 + 1, \dots, T\}$ respectively. Let the time series be $\mathcal{T} \coloneqq \mathcal{T}_1 \cup \mathcal{T}_2$, with cardinality $T = |\mathcal{T}|$. Let $\mathbf{y}_j \coloneqq [y_{jt}, \dots y_{NT}]^\top \in \mathbb{R}^T$ be a generic vector of outcomes. Denote $\mathbf{y}_1 \coloneqq (y_{1t})_{t \in \mathcal{T}}$ as the $T \times 1$ column vector of outcomes for the treated unit and $\mathbf{Y}_{\mathcal{N}_0} \coloneqq (\mathbf{y}_j)_{j \in \mathcal{N}_0}$ as the $T \times N_0$ matrix of control unit outcomes. Denote $\mathbf{Y}_U \coloneqq (\mathbf{y}_j)_{j \in \widehat{U}}$ as the $T \times U$ submatrix of selected controls. Let the treatment effect be $\widehat{\Delta}_{\widehat{U}t} \coloneqq y_{1t} - \hat{y}^0_{1t}(\widehat{U})$ where $\hat{y}^0_{1t}(\widehat{U})$ is the counterfactual based on set $\widehat{U}$. The causal estimand of interest is the ATT, or $\widehat{ATT}_{\widehat{U}} = T_2^{-1} \sum_{t \in \mathcal{T}_2} \widehat{\Delta}_{\widehat{U}t}$.

2.1 Forward-Selected Panel Data Approach

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. We find the control unit that maximizes the R-squared statistic $i_1^* = \arg\max_{i \in \mathcal{N}_0} R^2_{\widehat{U} \cup \{i\}}$. 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 stops the selection process (Wang et al., 2009). After, fsPDA predicts the counterfactual via OLS

$$\hat{\mathbf{y}}_1^0(\widehat{U}) = \mathbf{Y}'_{\widehat{U}} \widehat{\boldsymbol{\beta}}_{\widehat{U}} + \widehat{\boldsymbol{\beta}}_0 \, \forall t \in \mathcal{T} \tag{2.1}$$

where $\hat{\mathbf{y}}_1^0(\widehat{U})$ corresponds to our predictions, $\widehat{\boldsymbol{\beta}}_{\widehat{U}}$ corresponds to the coefficients, and $\widehat{\boldsymbol{\beta}}_0$ is the intercept.

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: $\widehat{U}$

To conduct inference, SH use a heteroskedasticity and autocorrelation consistent estimator of the long run variance via $\widehat{\rho}^2_{\tau \widehat{U}} = T_2^{-1} \sum_{ts \in T_2} (\widehat{\Delta}_{\widehat{U}t} - \bar{\Delta}_{\widehat{U}})(\widehat{\Delta}_{Us} - \bar{\Delta}_U) \cdot \mathbf{1}\{|t-s| \leq \tau\}$ 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

Next, I describe fDID. The procedure iterates over k total candidate iterations, 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 $k = 1$, we estimate a one-control DID model for each unit $i \in \mathcal{N}_0$. The control unit that maximizes the pre-treatment R-squared, R_i^2, is selected as $i_1^* = \arg\max_{i \in \mathcal{N}_0} R_i^2$, forming the first candidate control group, $\widehat{U}_1 = \{i_1^*\}$. The set $\mathcal{U}$ is updated accordingly, $\mathcal{U} \cup \widehat{U}_1 = \{\widehat{U}_1\}$. For $k = 2$, a DID model is estimated for each remaining control $i \in \mathcal{N}_0 \setminus \{i_1^*\}$, where each model includes i_1^* and one additional control. The control that maximizes R^2 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 the next candidate control group, $\widehat{U}_2 = \{i_1^*, i_2^*\}$. $\mathcal{U}$ is updated accordingly $\mathcal{U} \cup \widehat{U}_2 = \{\widehat{U}_1, \widehat{U}_2\}$. At each iteration k, a DID model is estimated for each remaining control $i \in \mathcal{N}_0 \setminus \widehat{U}_{k-1}$, combining the previously selected controls $\widehat{U}_{k-1}$ with one additional unit. The process continues until $k = N_0$, with $\mathcal{U}$ being updated at each step, $\mathcal{U} \cup \widehat{U}_k$. The final control group returned by fDID is the one that maximizes the R-squared statistic across all candidate sets: $\widehat{U} \coloneqq \text{argmax}_{\widehat{U}_k \in \mathcal{U}} R^2(\widehat{U}_k)$.

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: $R_k^2(\widehat{U}_{k-1} \cup \{i\})$
  end for
  Update $\mathcal{U} \cup \widehat{U}_k \leftarrow \mathcal{U} \cup \left(\widehat{U}_{k-1} \cup \left\{\text{argmax}_{i \in \mathcal{N}_0 \setminus \widehat{U}_{k-1}} R_k^2(\widehat{U}_{k-1} \cup \{i\})\right\}\right)$
end for
Return: $\widehat{U} \coloneqq \text{argmax}_{\widehat{U}_k \in \mathcal{U}} R^2(\widehat{U}_k)$

The fDID predictions take the following form

$$\hat{\mathbf{y}}_1^0(\widehat{U}) = \mathbf{Y}'_{\widehat{U}} \widehat{\boldsymbol{\beta}}_{\widehat{U}} + \widehat{\boldsymbol{\beta}}_0 \quad \text{s.t. } \widehat{\boldsymbol{\beta}}_{\widehat{U}} = \frac{1}{U}, \tag{2.2}$$

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

$$\hat{\Omega} = \left[\left(\frac{T_2}{T_1}\right) \cdot T_1^{-1} \sum_{t \in \mathcal{T}_1} \hat{v}_{1t}^2\right]^{0.5}, \quad \hat{v}_{1t} = \mathbf{y}_1 - \hat{\mathbf{y}}_1^0(\widehat{U}) \tag{2.3}$$

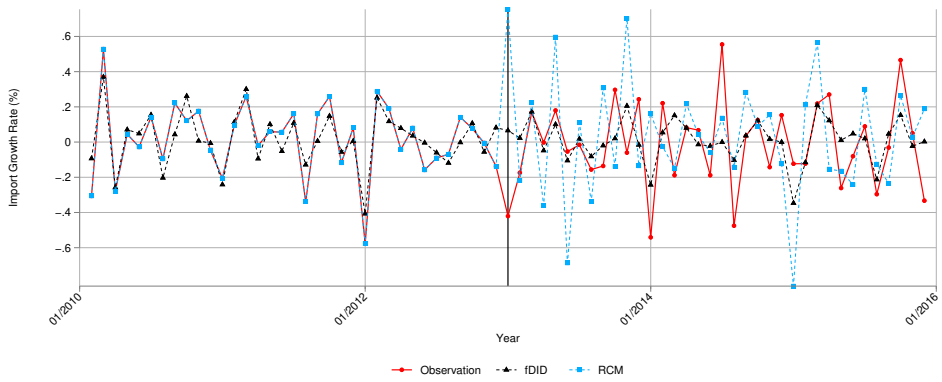
which is then used to construct confidence intervals and t-statistics.

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 $\widehat{\beta}_0$. Additionally, fDID constrains $\widehat{\beta}_{\widehat{U}}$ 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.

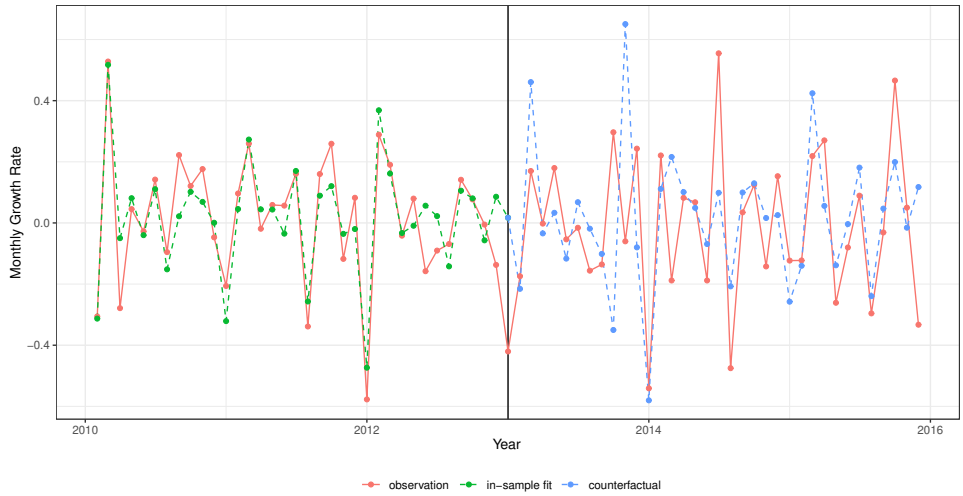
3 Replication Process

R Workflow 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.

Stata Workflow In Stata 17, I installed the fdid package as well as rcm as described in Greathouse et al. (2024) and Yan and Chen (2022). 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. I estimated fdid with the default settings, `fdid import, tr(treat) unitnames(unit)`. For RCM, the specification was `rcm import, trperiod(36) trunit(1) method(forward) criterion(mbic)`.



(a) Stata: fDID and RCM with Forward Selection and MBIC



(b) R: fsPDA

Figure 1: Comparison of Methods

4 Results of Replication

Point Estimates and Uncertainty In panels 1a and 1b respectively of Figure 1, I present the results of the fDID, RCM and fsPDA methods. 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. For RCM, the ATT was -0.0501. The standard error for fDID is 0.0265. For fsPDA, the standard error is 0.0274. RCM reports neither a standard error for the ATT or a t-statistic. fDID's t-statistic is 0.95 and fsPDA t-statistic is 1.12. Both fDID and fsPDA have similar ATTs and very similar standard errors.

The Selected Controls 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 RCM selects, for some reason, 33 controls, almost as many as we have pre-treatment time periods. fDID selects 14 controls, specifically: 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. For RCM, the model overfits with an R-squared of 100%. The difference of R-squared statistics, regarding fDID and fsPDA, 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 differences between fDID and fsPDA are 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: the most concerning of them is that the results of RCM do not come close to matching fDID's results or the results of SH. The in-sample predictions overfit drastically. It is not apparent why this would be, as Yan and Chen (2022) note that their package allows for use of the modified Bayesian criterion as SH use s well as the forward selection method proposed by SH. I should note that when we use forward-selection and the AICc for RCM, we then get 14 control units selected as well as the exact same ATT as fsPDA reports. However, the stopping rule for SH is specifically the modified Bayesian Criterion, so I report the RCM closest to SH's specification.

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" as the selected controls. 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, especially since 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 compared to fDID.

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

Biazoli, L., de Ávila, E. S., & de Oliveira, I. R. C. (2024). Combining cluster analysis with synthetic control for evaluating economic impacts of the dam breach in mariana, brazil. *Empirical Economics*, 1–21. https://doi.org/10.1007/s00181-024-02627-7

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

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/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

Yan, G., & Chen, Q. (2022). Rcm: A command for the regression control method. *The Stata Journal*, *22*(4), 842–883. https://doi.org/10.1177/1536867X221140960

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