Panel Data Analysis III

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Review

- The previous class started from introducing the DID estimator.
- It is motivated by the TWFE model when the data have a DID structure.
- But the estimator only needs the assumption of parallel trends and is robust to heterogeneous treatment effects.
- This is not true when the data have a more complex structure.
- The within estimator may generate negative weights for individualistic treatment effects.
- We can fix the problem by not using untreated observations when fitting the model.
- This leads to the idea of counterfactual estimation.

The classic factor model

- Remember that in the TWFE model, we assume that $h_t(\mathbf{U}_i) = \mu + \alpha_i + \xi_t$.
- Bai (2003) relaxes this assumption to $h_t(\mathbf{U}_i) = \mathbf{f}_t' \lambda_i$.
- Both $\mathbf{f}_t$ and λ_i are r-dimensional vectors.
- The number of parameters is $(N + T) * r \leq N * T$.
- The former is known as factors, and the latter as factor loadings.
- When $\mathbf{f}_t = (1, \xi_t)'$ and $\lambda_i = (\alpha_i, 1)'$, it boils down to the TWFE model.
- Such a model captures the interaction between variables that vary only over time and only across units.
- E.g., country i's endowments in r different resources and the prices of them in year t.

The classic factor model

- Let's omit the covariates, then

$$Y_{it} = \mathbf{f}_t^{'} \lambda_i + \varepsilon_{it},$$

- We can write the model in matrices:

$$\mathbf{Y} = \mathbf{F}\Lambda' + \varepsilon.$$

where $\mathbf{Y}$ is a $T \times N$ matrix; $\mathbf{F}$ $(T \times r)$ represents factors and Λ $(N \times r)$ represents factor loadings.

- Suppose $\tilde{\mathbf{F}} = \mathbf{F}\mathbf{H}$ and $\tilde{\Lambda}' = \mathbf{H}^{-1}\Lambda'$, then $\mathbf{Y} = \tilde{\mathbf{F}}\tilde{\Lambda}' + \varepsilon$.
- Some assumptions on the structure of $\mathbf{F}$ and Λ are necessary for identification:

$$\frac{\mathbf{F}'\mathbf{F}}{T} = \mathbf{I}_r$$

$$\frac{\Lambda'\Lambda}{N} \text{ is diagonal.}$$

- There are $\frac{r(r-1)}{2} + \frac{r(r+1)}{2} = r^2$ restrictions in total.

The classic factor model (*)

- An intuitive idea is to estimate $\mathbf{F}$ and Λ via OLS and minimize

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(Y_{it}-\mathbf{f}_t^{'}\lambda_i)^2.$$

- We can show that it is equivalent to minimizing

$$tr((\mathbf{Y}-\mathbf{F}\Lambda')(\mathbf{Y}-\mathbf{F}\Lambda')'),$$

  where $tr()$ is the trace of a matrix.
- If we know the value of $\mathbf{F}$, we can estimate Λ with $\Lambda=\mathbf{Y}'\mathbf{F}(\mathbf{F}'\mathbf{F})^{-1}=\frac{\mathbf{Y}'\mathbf{F}}{T}$.
- The second equality uses the identification constraint.
- Plugging the expression into the objective function, it becomes

$$tr(\mathbf{F}'\mathbf{Y}\mathbf{Y}'\mathbf{F})/T$$

- All we need to do is to find $\mathbf{F}$ that minimizes the expression.

The classic factor model (*)

- Bai (2003) shows that $\mathbf{F}$ can be found via principal component analysis (PCA) or singular value decomposition (SVD) of the matrix $\mathbf{YY}'/T$.
- In SVD, we find a matrix $\hat{\mathbf{F}}$ such that

$$\left(\frac{1}{NT} \sum_{i=1}^{N} \mathbf{Y}_i \mathbf{Y}_i^{'} \right) \hat{\mathbf{F}} = \hat{\mathbf{F}} \mathbf{V}_{NT},$$

  where $\mathbf{V}_{NT}$ is the matrix of singular values.
- Bai shows that $\sqrt{N}(\hat{\mathbf{F}}_t - \mathbf{H}'\mathbf{F}_t)$ converges to a normal distribution.
- We are estimating T vectors using $N \times T$ observations, hence $\sqrt{N}$ is the best convergence rate we can get.
- Similarly, $\sqrt{T}(\hat{\Lambda}_i - \mathbf{H}^{-1}\Lambda_i)$ converges to a normal distribution.

The classic factor model

- Bai (2009) considers the DGP with covariates:

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{F}\Lambda' + \varepsilon.$$

- Knowing β, we can calculate residuals and estimate $\mathbf{F}$ and Λ as before; knowing $\mathbf{F}$ and Λ, we can estimate β via

$$\hat{\beta} = \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Q}_{\hat{\mathbf{F}}} \mathbf{X}_i \right)^{-1} \left(\sum_{i=1}^{N} \mathbf{X}_i' \mathbf{Q}_{\hat{\mathbf{F}}} \mathbf{Y}_i \right)$$

  where $\mathbf{Q}_{\hat{\mathbf{F}}}$ is the residual-making matrix of $\hat{\mathbf{F}}$.
- We can start from some initial values and iterate this process until convergence.
- Bai shows that $\sqrt{NT}(\hat{\beta} - \beta)$ converges to a normal distribution that may not concentrate around zero.
- The bias diminishes when heteroskedasticity and autocorrelation are absent in either dimension.

Synthetic control

- Abadie, Diamond, and Hainmueller (2010) propose the synthetic control (SC) method for case studies.
- It assumes that there are N untreated units and a single treated unit ($i = 1$).
- There are T_0 pre-treatment periods and T_1 post-treatment periods.
- The untreated potential outcome obeys the classic factor model as in Bai (2003) or Bai (2009).
- We are interested in is the treatment effect on unit 1 in any period $t > T_0$:

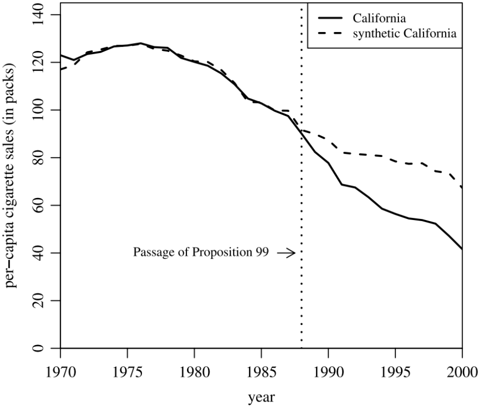
$$\tau_{1t} = Y_{1t}(1) - Y_{1t}(0).$$

Synthetic control

- Abadie, Diamond, and Hainmueller (2010) show that we don't need to estimate factors or factor loadings directly in this case.
- Instead, they propose a weighting algorithm to predict the counterfactual of the treated unit, $Y_{1t}(0)$.
- Intuitively, we weight the untreated units to construct a synthetic treated unit.
- The weights are calculated to balance the pre-treatment trajectory.

Synthetic control

- Ideally, we would like to see a picture like this:



Synthetic control

- We attempt to find a group of weights $\{w_i\}_{i=2}^{N+1}$ which sum to one such that
$$\sum_{i=2}^{N+1} w_i Y_{it} = Y_{1t}$$
for any $1 \leq t \leq T_0$.
- We require the treated unit to lie in the "convex hull" of the untreated ones.
- The predicted counterfactual for unit 1 in period $t > T_0$ is
$$\hat{Y}_{1t}(0) = \sum_{i=2}^{N+1} w_i Y_{it}.$$
- Hence, we can estimate τ_{1t} by $\hat{\tau}_{1t} = Y_{1t}(1) - \hat{Y}_{1t}(0)$.

Synthetic control

- In practice, we search for weights $\{w_i\}_{i=2}^{N+1}$ to minimize the distance between the predictors.
- Denote $\mathbf{Y}_1^0 = (Y_{11}, Y_{12}, \ldots, Y_{1T_0})'$, $\mathbf{W} = (w_2, w_3, \ldots, w_{N+1})'$ and

$$\mathbf{Y}_0^0 = \begin{pmatrix} Y_{21} & Y_{31} & \ldots & Y_{N+1,1} \\ Y_{22} & Y_{32} & \ldots & Y_{N+1,2} \\ \ldots & \ldots & \ddots & \ldots \\ Y_{2T_0} & Y_{3T_0} & \ldots & Y_{N+1,T_0} \end{pmatrix}$$

- We minimize

$$\begin{aligned} & ||\mathbf{Y}_1^0 - \mathbf{Y}_0^0\mathbf{W}||_{\mathbf{V}} \\ = & \sqrt{(\mathbf{Y}_1^0 - \mathbf{Y}_0^0\mathbf{W})'\mathbf{V}(\mathbf{Y}_1^0 - \mathbf{Y}_0^0\mathbf{W})} \end{aligned}$$

Synthetic control

- Note that

$$Y_{1t}(0) - \hat{Y}_{1t}(0) = \mathbf{f}_t^{'}(\lambda_1 - \sum_{i=2}^{N+1} w_i \lambda_i) + \sum_{i=2}^{N+1} w_i(\varepsilon_{1t} - \varepsilon_{it})$$

  for any $t > T_0$.
- Denote $\mathbf{F}_{T_0} = (\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_{T_0})'$, then

$$\begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix} = \begin{pmatrix} Y_{11}(0) - \hat{Y}_{11}(0) \\ Y_{12}(0) - \hat{Y}_{12}(0) \\ \dots \\ Y_{1T_0}(0) - \hat{Y}_{1T_0}(0) \end{pmatrix} = \mathbf{F}_{T_0}(\lambda_1 - \sum_{i=2}^{N+1} w_i \lambda_i)$$

$$+ \sum_{i=2}^{N+1} w_i \left(\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \dots \\ \varepsilon_{1T_0} \end{pmatrix} - \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \dots \\ \varepsilon_{iT_0} \end{pmatrix} \right)$$

Synthetic control

- Multiplying $\mathbf{f}_t'(\mathbf{F}_{T_0}'\mathbf{F}_{T_0})^{-1}\mathbf{F}_{T_0}'$ to both sides of the second equation and subtracting it from the first one, we have

$$
\begin{aligned}
Y_{1t}(0) - \hat{Y}_{1t}(0) &= \sum_{i=2}^{N+1} w_i(\varepsilon_{1t} - \varepsilon_{it}) \\
&- \mathbf{f}_t'(\mathbf{F}_{T_0}'\mathbf{F}_{T_0})^{-1}\mathbf{F}_{T_0}' \sum_{i=2}^{N+1} w_i \left(\begin{pmatrix} \varepsilon_{11} \\ \varepsilon_{12} \\ \dots \\ \varepsilon_{1T_0} \end{pmatrix} - \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \dots \\ \varepsilon_{iT_0} \end{pmatrix} \right)
\end{aligned}
$$

- We rely on Rosenthal's inequality to show that

$$
\mathbf{f}_t'(\mathbf{F}_{T_0}'\mathbf{F}_{T_0})^{-1}\mathbf{F}_{T_0}' \sum_{i=2}^{N+1} w_i \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \dots \\ \varepsilon_{iT_0} \end{pmatrix} \to 0.
$$

Synthetic control

- It is a case-study method and the estimate has no asymptotic distribution.
- We have only one treated observation.
- We rely on a permutation test for statistical inference.
- It is similar to a placebo test: we replace the treated unit with a randomly selected unit from the control group and estimate the effect on it.
- The permutation test differs from Fisher's randomization test since we do not know the treatment assignment algorithm.
- It assumes that each unit has the same probability of being treated.
- There are approaches to construct non-asymptotic confidence intervals but they tend to be wide.

Augmented synthetic control

- The classic synthetic control method is built upon weighting.
- A natural idea is to improve its efficiency by combining weighting with regression.
- We find weights $\{w_i\}_{i=2}^{N+1}$ as before, and fit regression models of Y_{it} $(t > T_0)$ on $(Y_{i1}, Y_{i2}, \ldots, Y_{iT_0})$.
- Denote the estimated coefficients as $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{T_0})$.
- The predicted counterfactual takes the form of

$$\hat{Y}_{1t}(0) = \sum_{i=2}^{N+1} \hat{w}_i Y_{it} + \sum_{s=1}^{T_0} \hat{\beta}_s \left(Y_{1s} - \sum_{i=2}^{N+1} \hat{w}_i Y_{is} \right).$$

- We can fit a ridge regression to further improve precision by allowing for extrapolation.

Generalized synthetic control

- In practice, most data include more than one treated unit.
- In theory, we can find a unique set of weights for each treated unit.
- But it could be more straightforward to estimate factors and factor loadings.
- Xu (2017) considers the following model:

$$Y_{it}(0) = \mathbf{f}_t'\lambda_i + \mathbf{X}_{it}'\beta + \varepsilon_{it},$$
$$Y_{it}(1) = Y_{it}(0) + \tau_{it}.$$

- As in counterfactual estimation, we estimate $\mathbf{f}_t$, λ_i, and β using untreated observations.
- Then, we generate predictions of the counterfactual and estimate the ATT with:

$$\widehat{\tau}_{ATT} = \frac{1}{|\mathcal{M}|}\sum_{(i,t)\in\mathcal{M}}\left(Y_{it} - \hat{Y}_{it}(0)\right).$$

Matrix completion

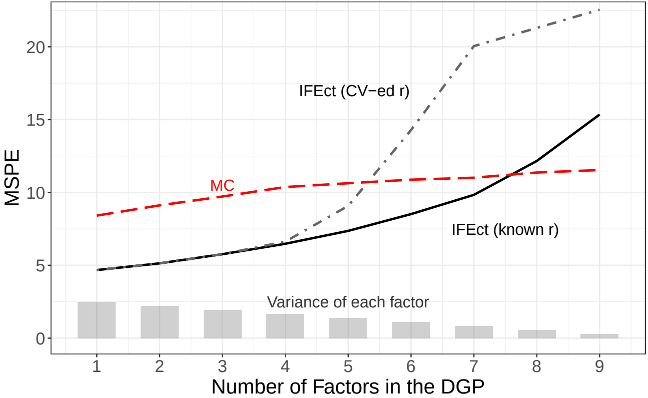
- Athey et al. (2018) suggest that we should directly estimate missing elements in the matrix $\mathbf{L} = \{Y_{it}(0)\}_{N \times T}$.
- We solve the following matrix completion (MC) problem:

$$\widehat{\mathbf{L}} = \arg\min_{\mathbf{L}} \left[\sum_{(i,t)\in\mathcal{O}} \frac{(Y_{it} - L_{it})^2}{|\mathcal{O}|} + \lambda_L \|\mathbf{L}\| \right],$$

- The estimation relies on a different iteration algorithm.
- Athey et al. (2018) prove that the obtained $\hat{\mathbf{L}}$ is asymptotically unbiased for $\mathbf{L}$.

IFE or MC?

- MC penalizes the magnitude rather than the number of singular values (soft impute vs. hard impute).
- Liu, Wang, and Xu (2020) uses simulation to show that matrix completion performs better when the DGP contains many weak factors.



}

SCDID

- Arkhangelsky et al. (2019) propose a different idea to use the factors.
- We solve the following problem:

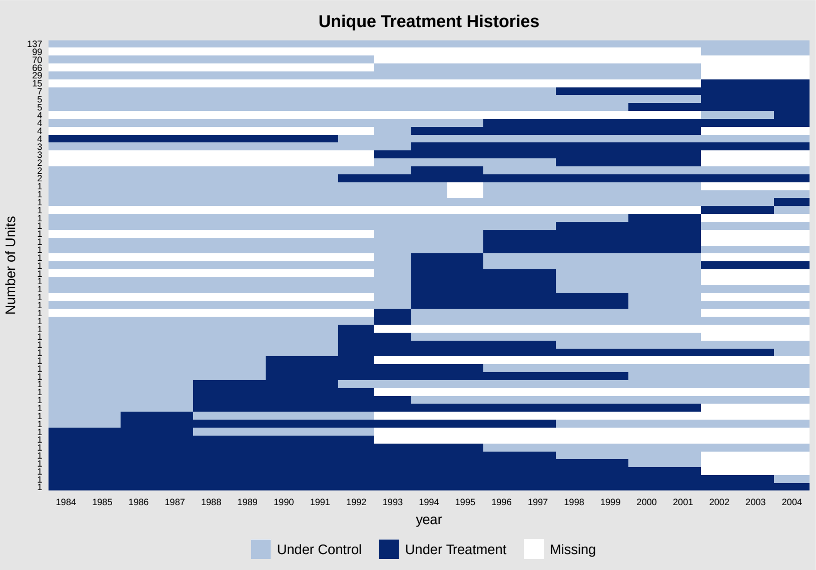
$$\arg\min_{\tau,\mu,\alpha,\xi} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - \mu - \alpha_i - \xi_t - \tau_{it} D_{it}) \mathbf{f}_t' \lambda_i.$$

- They also introduce weighting estimators for $\mathbf{f}_t'$ and λ_i.
- They call this method "synthetic difference-in-differences (SCDID)".
- It works when the data have a structure of staggered adoption.

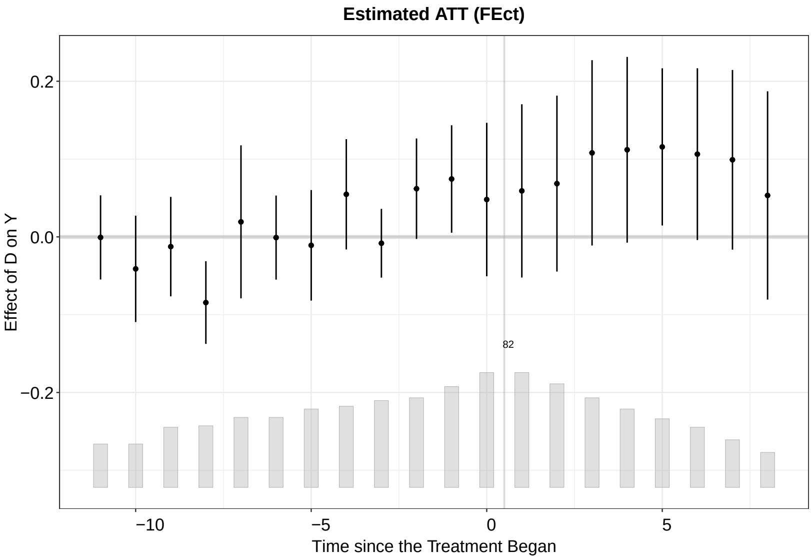
Factor models: application

- Let's examine the study in Anzia and Berry (2011).
- The argument is that female legislators face more structural barriers than their male counterparts.
- Once elected, they will be more competent and perform better.
- The data include 700 districts in the US over 21 years.
- $D_{it} = 1$ if there is a female legislator from district i in year t.
- The outcome is measured by the amount of federal spending a legislator can secure for their district in year t.

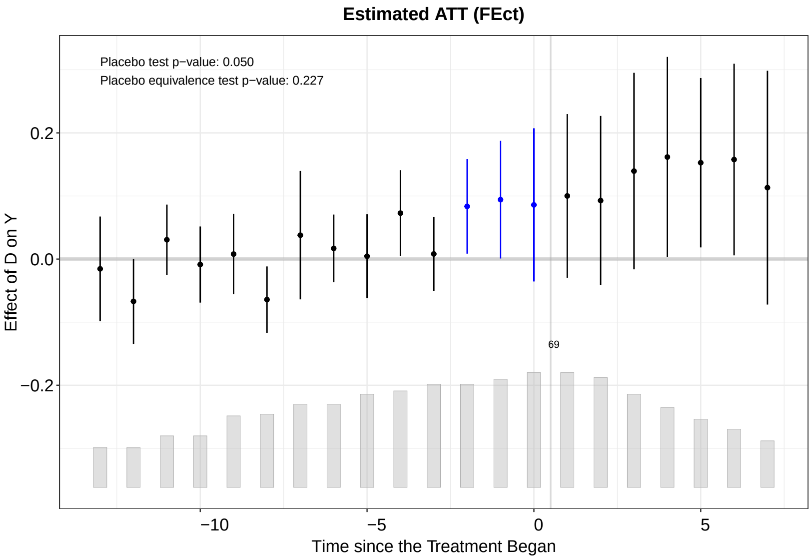
Factor models: application



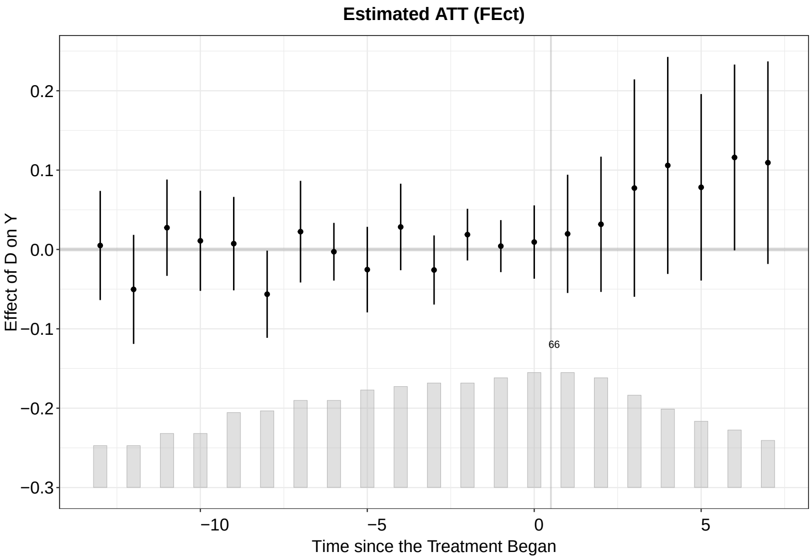
Factor models: application



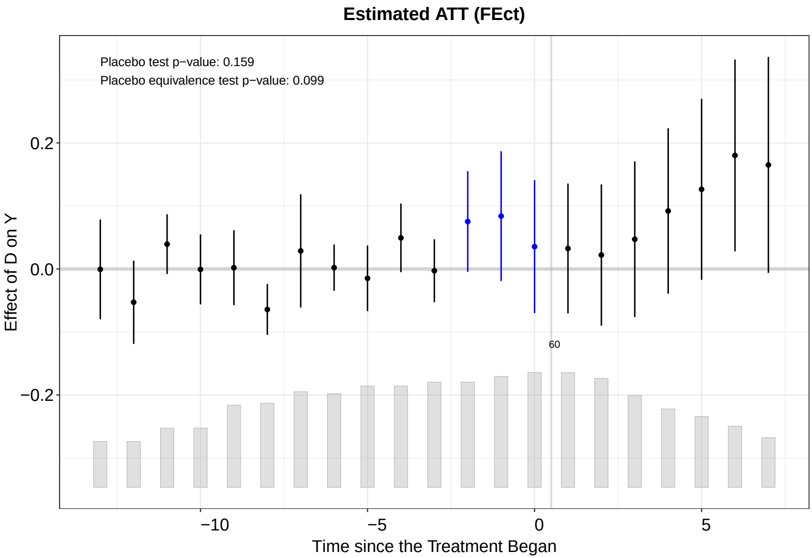
Factor models: application



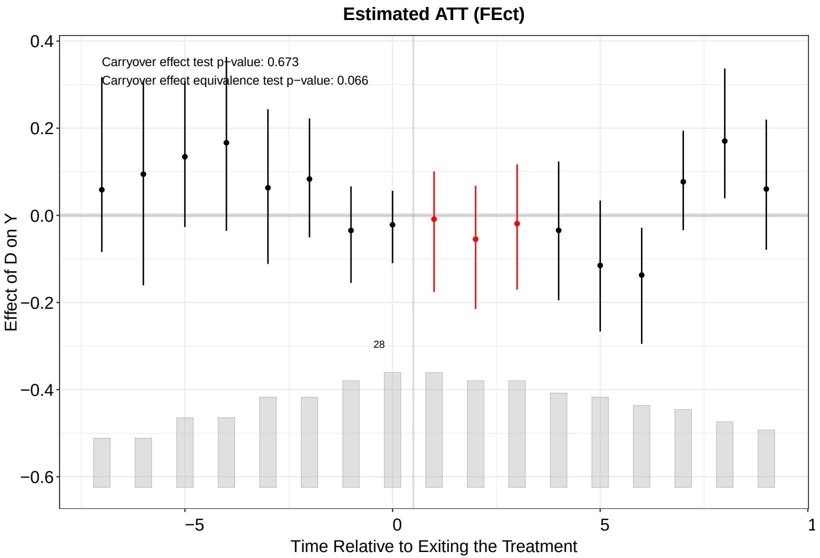
Factor models: application



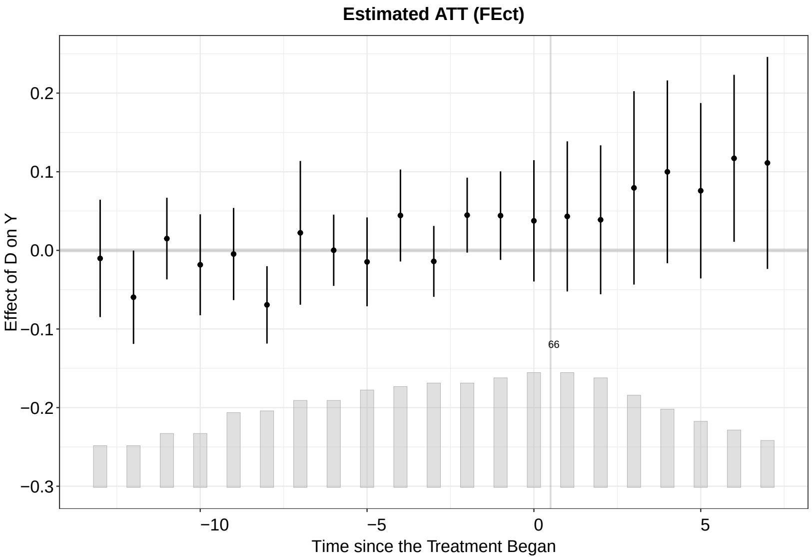
Factor models: application



Factor models: application



Factor models: application



Manifold learning

- In either the TWFE or factor models, we learn the unobservables from observable variables.
- This is the idea of manifold learning: we assume that the unobservable variable is a low-dimensional manifold embedded in a high-dimensional space.
- Feng (2020) formally develops this idea for panel data analysis.
- He assumes that $Y_{it} = f_t(\lambda_i) + \varepsilon_{it}$.
- To estimate λ_i, we first match each unit i to K nearest neighbors.
- Next, we estimate λ_i via SVD on the $K + 1$ outcome histories.
- It is a local factor model.
- Then, we control $\hat{\lambda}_i$ in a doubly robust estimator.

References I

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program." *Journal of the American Statistical Association* 105 (490): 493–505.

Anzia, Sarah F, and Christopher R Berry. 2011. "The Jackie (and Jill) Robinson Effect: Why Do Congresswomen Outperform Congressmen?" *American Journal of Political Science* 55 (3): 478–93.

Arkhangelsky, Dmitry, Susan Athey, David A Hirshberg, Guido W Imbens, and Stefan Wager. 2019. "Synthetic Difference in Differences." National Bureau of Economic Research.

Athey, Susan, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2018. "Matrix Completion Methods for Causal Panel Data Models." National Bureau of Economic Research.

References II

Bai, Jushan. 2003. "Inferential Theory for Factor Models of Large Dimensions." *Econometrica* 71 (1): 135–71.

———. 2009. "Panel Data Models with Interactive Fixed Effects." *Econometrica* 77 (4): 1229–79.

Feng, Yingjie. 2020. "Causal Inference in Possibly Nonlinear Factor Models." *arXiv Preprint arXiv:2008.13651*.

Liu, Licheng, Ye Wang, and Yiqing Xu. 2020. "A Practical Guide to Counterfactual Estimators for Causal Inference with Time-Series Cross-Sectional Data." *Available at SSRN 3555463*.

Xu, Yiqing. 2017. "Generalized Synthetic Control Method: Causal Inference with Interactive Fixed Effects Models." *Political Analysis* 25 (1): 57–76.