Readme to Replication Data for: Finitely Heterogeneous Treatment Effect in Event-study

Myungkou Shin

January 6, 2025

This document describes R program files to replicate all tables and figures in the manuscript and the Supplementary Appendix for "Finitely Heterogeneous Treatment Effect in Event-study."

1 Data

Tables 5-6 and Figure 1 of the manuscript and Tables 1-12 and Figures 1-3 of the Supplementary Appendix use `ros_ccd_aejp.dta` from Lutz, Byron. Replication data for: The End of Court-Ordered Desegregation. Nashville, TN: American Economic Association [publisher], 2011. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-10-13. `https://doi.org/10.3886/E114756V1`.

2 Program files

The program files are follows:

- `main.Rmd`: combines `classification.R`, `summary_stat.R`, `catt_estimation.R`, and `cross_validation.R`.
- `functions.R`: contains functions used in `main.Rmd`, `classification.R`, `summary_stat.R`, `catt_estimation.R`, and `cross_validation.R`.
- `simulation.R`: produces Tables 1-4 of the manuscript.
- `simulation_function.R`: contains functions used in `simulation.R`.
- `data_cleaning.R`: cleans the dataset from Lutz (2011).
- `classification.R`: produces type classification results to be used in `summary_stat.R`, `catt_estimation.R` and Tables 2-4 and 12 of the Supplementary Appendix.

- `summary_stat.R`: produces Tables 5-6 of the manuscript and Table 5-10 of the Supplementary Appendix.
- `catt_estimation.R`: produces Figure 1 of the manuscript and Figures 1-3 of the Supplementary Appendix.
- `cross_validation.R`: produces Table 1 and Table 11 of the Supplementary Appendix.

3 Functions

This section describes select functions from functions.R which can be used to compute the type-specific DiD estimator.

3.1 kmeans_fun

This function takes the pretreatment subset of the panel data and estimates latent type structure across units. The function has six arguments. Below is the list of the arguments:

- `n_of_initial_values`: A positive integer for the number of initial type assignments to be drawn.
- `e_data`: A $(n \cdot T_0)$-by-$(p+5)$ matrix that contains the pretreatment panel data. The columns of the matrix need to be sorted in the following order: first-differenced outcome variable, unit index, time index, treatment timing, weight, p control covariates. The unit indices and the time indices need to be positive integers: $1, \ldots, n$ and $1, \ldots, T_0$.
- `K`: A positive integer for the number of types.
- `eps`: A positive real number for a stopping criterion; the iteration stops when the biggest increment in the parameter is smaller than eps.
- `maxiter`: A positive integer for a stopping criterion; the iteration stops when the number of iteration passes maxiter.
- `spline`: A function that takes an integer $1, \ldots, T_0$ and produces regressors to model the (type-specific) time trends. For example, the most flexible specification of type-and-time

fixed-effect can be constructed by letting spline function produce a T_0-by-1 vector where only one element is one and the rest are zeros: $\big(\mathbf{1}\{t=1\},\cdots,\mathbf{1}\{t=T_0\}\big)^\intercal$.

The output of `kmeans_fun` contains coefficients on the control covariates, coefficients for the type-specific trends and type assignment.

3.2 estFullCATT

This function takes the full panel data and estimates the type-cohort-and-time specific average treatment effect in treated units as defined in (3) of the manuscript. The function has seven arguments. Below is the list of the arguments:

- `gamma`: A n-by-1 vector for the type assignment. The elements of the vector ranges from $1,\ldots,K$.

- `Pi`: A n-by-$(p_\pi+1)$ matrix that contains propensity score estimates. p_π is the dimension of the model parameter in the assignment model. The first column of the matrix contains the propensity score ratio given the arguments `t`, `k` and `e` and an estimate on the assignment model parameter $\hat{\xi}$:

$$\frac{\pi_e(X_i,k,\hat{\xi})}{\pi_{t+}(X_i,k,\hat{\xi})}.$$

  The last p_π columns contain the plug-in estimate on the first derivative of the propensity score ratio:

$$\frac{\partial}{\partial\xi}\frac{\pi_e(X_i,k,\xi)}{\pi_{t+}(X_i,k,\xi)}\bigg|_{\xi=\hat{\xi}}.$$

  π_e is defined in page 37 of the manuscript. For more discussion, see Section C.4 of the manuscript and Section D of the Supplementary Appendix. When there is no control covariates X_i to be used, a placeholder for `Pi` has ones in its first column and zeros in the rest.

- `L`: A n-by-p_π matrix that contains the plug-in estimates on the scores in the linear approximation of the assignment model parameter ξ:

$$l^\pi(X_i,\hat{k}_i,E_i).$$

l^π is defined in page 39 of the manuscript. For more discussion, see Section C.4 of the manuscript and Section D of the Supplementary Appendix. When there is no control covariates X_i to be used, a placeholder for `L` has zeros everywhere.

- `e_data`: A $(n \cdot T)$-by-$(p+5)$ matrix that contains the panel data. The columns of the matrix need to be sorted in the following order: outcome variable in level, unit index, time index, treatment timing, weight, p control covariates. The unit indices and the time indices need to be positive integers: $1, \ldots, n$ and $1, \ldots, T$.

- `t`: time index.

- `k`: type.

- `e`: treatment timing.

The last three arguments specifies the treatment effect parameter $CATT_t(k, e)$.

3.3 estCATT

This function averages the estimates from estFullCATT and estimates the dynamic type-specific average treatment effect in treated units as defined in (4) of the manuscript. The function has six arguments: the first four arguments are identical with those for `estFullCATT` and

- `r`: relative treatment timing.

- `k`: type.

The last two arguments specifies the treatment effect parameter $\beta_r(k)$.