README
High Resolution Treatment Effects Estimation: Uncovering Effect Heterogeneities with the Modified Causal Forest

Hugo Bodory, Hannah Busshoff, and Michael Lechner

July 21, 2022

Overview

This replication package reproduces the results of the three empirical applications. Scripts written in Python generate the findings for the tables and figures in the paper and the online supplement. For the empirical applications, we use three data files based on the data sources of (i) Cattaneo (2010), (ii) Keele and Small (2021), and (iii) Ao, Calonico, and Lee (2021). The replicator should expect all programming scripts to run for about 17.5 hours.

Content of the Replication Package

The replication package consists of five folders as listed below:

- Folder `data`
  To replicate the empirical applications, the `data` folder contains the data files `data_rhc.csv` and `data_wia.csv`. The data file for the `Maternal Smoking during Pregnancy` study (`bw_data.csv`) is available upon request.

- Folder `py`
  The `py` folder stores four Python files and one text file, which are required to run the programs for the empirical studies. The file names are: `bw.py`, `rhc.py`, `wia.py`, `config.py`, and `requirements.txt`.

- Folder `txt_files`
  The files `bw_data.txt`, `data_rhc.txt`, and `data_wia.txt` are stored in the `txt_files` folder. These text files show the results of the three empirical applications. They will be generated after running the files saved in the `py` folder.

- Folder `output`
  The `output` folder contains the Python scripts `output.py` and `output_functions.py`. It further stores the sub-folders `output_mcf`, `figures`, and `tables`. The sub-folder `output_mcf` stores the CSV files used for creating the figures and tables in the paper (except the codebooks in Tables S14-S16). These CSV files can be generated by the Python codes in the `py` folder. The plots and tables of the paper are available in the `figures` and `tables` subplots.

- Folder `ReadMe`
  This folder contains the file ReadMe.pdf.

Data Availability and Provenance Statements

The original data for the empirical applications are provided by (i) Cattaneo (2010), (ii) Keele and Small (2021), and (iii) Ao, Calonico, and Lee (2021). The data used by Cattaneo (2010) are not accessible for the general public. In contrast, Keele and Small (2021) employ publicly available health data in their study and the analysis of Ao, Calonico, and Lee (2021) uses public information of the US Department of Labor. For our paper, we slightly process the original data for estimation. Note that the original data sets, as well as the Python files which prepare the data for our applications, are not added to this replication package but are available on request. Copies of all data used in our study are added to this replication package (except the data file `bw_data.csv`, which is available upon request).

Statement about Rights

Subsequent users of the software of this replication package are expected to give proper credit via citation.

Summary of Availability

With the exception of Cattaneo (2010), the empirical data are publicly available.

Further Details on the Data Sources

We create three datasets to replicate the empirical applications. The dataset list below summarizes all file names.

Datasets and Variables

Dataset list

The following dataset list presents the names of the CSV data files used for estimation.

Data file	Source	Empirical Application
`bw_data.csv` (available upon request)	Cattaneo (2010)	Maternal Smoking during Pregnancy
`data_rhc.csv`	Keele and Small (2021)	Right Heart Catheterization
`data_wia.csv`	Ao, Calonico, and Lee (2021)	The Workforce Investment Act Programs

Variables

The codebooks in Tables S14-S16 of the online appendix give information on the variables analyzed in the empirical applications. The variable descriptions in these tables further indicate which of these variables are used as outcomes and treatments.

Computational Requirements

Software Requirements

Python 3.8.8, IDE: Spyder 5.1.5, OS: Microsoft Windows Server 2019.
Python package: `mcf`, version 0.2.4.

Memory and Runtime Requirements

The next table states the compute-time for the Python programs. All programs have been run on an HP Z8 G4 workstation with 2 Intel(R) Xeon(R) Gold 6234 CPUs @ 3.30 GHz, 32 cores (16 physical and 16 virtual cores), and 768 GB RAM.

Program	Wall clock time
`bw.py`	5.3 hours
`rhc.py`	0.5 hours
`wia.py`	11.6 hours
`output.py`	3 minutes

Description of Programs/Code

- The script `config.py` adjusts the default paths for the Python sessions.
- The text file `requirements.txt` installs all dependencies locally. It should be run once before starting the Python codes.
- The Python script `bw.py` imports the data file `bw_data.csv` (available upon request). It creates the textfile `bw_data.txt`, which displays in the rows 3760-3774 the results of Table 1. `bw.py` further generates the CSV files used to create Figures 1+2, as well as Tables S1-S6 in the online appendix.
- The Python script `rhc.py` imports the data file `data_rhc.csv`. It creates the textfile `data_rhc.txt`, which displays in the rows 3468+3480 the results of Table 2. `rhc.py` further generates the CSV files used to create Figures 3-5, as well as Table 3 and Tables S7-S9 in the online appendix.
- The Python script `wia.py` imports the data file `data_wia.csv`. It creates the textfile `data_wia.txt`, which displays in the rows 4395-4400 the results of Table 4. `wia.py` further generates the CSV files used to create Figures 6+7, as well as Tables S10-S13 in the online appendix.
- The Python script `output.py` imports `output_functions.py` and the CSV files saved in the `output/mcf_output` folder. Running `output.py` creates the following output: Table 3, Figures 1-7, and Tables S1-S13 in the online appendix.

Instructions to Replicators

- Download the open source Anaconda Distribution (Anaconda Individual Edition For Windows) here, install the software on your computer, and launch the Spyder IDE.
- Download `config.py`, `requirements.txt`, and the Python scripts `bw.py`, `rhc.py`, `wia.py`, `output.py`, and `output_functions.py` into your working

directory. In addition, download also the folders `data` and `output` into your working directory.

- Open `config.py` in Spyder, edit the file to adjust the default path for your working directory, and run it in Spyder.
- Open `requirements.txt` in Spyder and type the command *`pip install -r py/requirements.txt`* into Spyder's console. Running this command sets up the working environment.
- Launch a new Spyder console (to restart the kernel). Then open and run the python scripts `bw.py`, `rhc.py`, `wia.py` and `output.py` to reproduce the results. All text and CSV files required to generate the tables and figures will be stored in the newly generated folders `data0`, `data1`, `data2`, as well as in the subfolders `figures` and `tables` of the `output` folder. Note that `bw.py` imports the data file `bw_data.csv`, which is available upon request.

List of Tables and Programs

The following table summarizes the program codes written to produce the tables and figures in the paper and online supplement.

Figure/ Table	Script
Table 1	`bw.py`
Table 2	`rhc.py`
Table 4	`wia.py`
Table 3, Figures 1-7, Tables S1-S13	`output.py`

References

Ao, W., S. Calonico, and Y. Lee (2021): "Multivalued Treatments and Decomposition Analysis: An Application to the WIA Program," *Journal of Business & Economic Statistics*, 39(1), 358–371, https://doi.org/10.1080/07350015.2019.1660664.

Cattaneo, M. (2010): "Efficient Semiparametric Estimation of Multi-Valued Treatment Effects under Ignorability," *Journal of Econometrics*, 150, 138–154, https://doi.org/10.1016/j.jeconom.2009.09.023.

Keele, L., and D. Small (2021): "Comparing Covariate Prioritization via Matching to Machine Learning Methods for Causal Inference Using Five Empirical Applications," *The American Statistician*, pp. 1–9, https://doi.org/10.1080/00031305.2020.1867638.