Overview

This file describes the replication material for Leavitt, T. and L. A. Hatfield. (2025). Averaged Prediction Models (APM): Identifying Causal Effects in Controlled Pre-Post Settings with Application to Gun Policy. *The Annals of Applied Statistics*, 19(3), 1826-1846. DOI: 10.1214/25-AOAS2011

Computational Requirements

I ran the code on a 2023 MacBook Pro with an Apple M3 Max chip and 36 GB memory. The operating system is Sonoma 14.7.1. The R version is 4.4.2 (2024-10-31) and the RStudio version is 2024.12.1+563.

All simulations for this study were performed using a high-performance virtual machine on Amazon Web Services (AWS). This cloud computing environment was used for only the simulation study, enabling efficient parallel computation across a large number of replications. All post-simulation analysis and visualization were performed locally.

AWS Instance Specifications

- Cloud Provider: Amazon Web Services (AWS)
- Instance Type: `c6a.16xlarge`
  - vCPUs: 64
  - RAM: 128 GiB
- Operating System: Ubuntu Server 22.04 LTS (64-bit)
- Region: US East (Ohio) (`us-east-2b`)
- Storage: 100 GiB SSD (gp3 EBS volume)
- RStudio Server: Version 2023.09.1+494 (accessed via web browser)

Software Stack

- R Version: 4.4.2 (2024-10-31)
- RStudio Version: 2023.09.1+494
- Parallelization framework:
  - `parallel` package with 120 worker threads via `parLapply()`

R Packages Used in Simulations

- `parallel` (version 4.4.2)
- `dplyr` (version 1.1.4)
- `magrittr` (version 2.0.3)
- `apm` (version 0.1.0)
- `stats` (base R)
- `utils` (base R)

R Packages Used in Data Analysis and Plot Generation

- `dplyr` (version 1.1.4)
- `magrittr` (version 2.0.3)
- `ggplot2` (version 3.5.1)
- `ggrepel` (version 0.9.6)
- `patchwork` (version 1.3.0)
- `ggExtra` (version 0.10.1)
- `tidyr` (version 1.3.1)
- `apm` (version 0.1.0)

Data

Files

- `data/data_construction.R`
- `data/ptpdata.RData`
- `data/schell_et_al_sim_data.RData`

Description

- `data/data_construction.R` constructs the dataset used for the analysis
- `data/ptpdata.RData` contains state-year homicide data for Missouri and neighboring states
- `data/schell_et_al_sim_data.RData` is the full dataset used for the simulation

Code and Output

To reproduce results, clone this repository and run `master.R` from the project root. All file paths are relative to the repository root, so there is no need to change the working directory manually.

Files

- `install_packages.R`
- `master.R`
- `code_and_output/fig_1_code.R`
- `code_and_output/fig_2_code.R`
- `code_and_output/fig_3_code.R`
- `code_and_output/fig_4_code.R`
- `code_and_output/analysis.R`
- `code_and_output/fits.RData`
- `code_and_output/ests.RData`
- `code_and_output/simulation/apm_simulation.R`
- `code_and_output/simulation/sim_res.RData`
- `code_and_output/simulation/sim_plots/apm_simulation_plots.R`
- Several .pdf output figures used in the manuscript and supplement
- Several .RData files containing main analysis output from `apm` functions, as well as simulation results

Description

- `master.R` runs the full replication workflow in order
- `install_packages.R` installs all required R packages for replication (including specific versions via source `.tar.gz` files)
- `analysis.R` runs the primary causal analysis, generating `code_and_output/fits.RData` and `code_and_output/ests.RData`
- `fig_1_code.R` through `fig_4_code.R` generate the four main figures in the manuscript
- `apm_simulation.R` runs simulations over different sample sizes
- `apm_simulation_plots.R` generates all plots included in the online supplement
- The .pdf files in `code_and_output/` and `code_and_output/simulation/sim_plots/` contain the figures referenced in the manuscript and supplement

Simulation Details

The simulation data (`data/schell_et_al_sim_data.RData`) consist of state-level crude death rates from 1994 to 2008, with 2008 designated as the post-treatment year. In each simulation, five states are randomly selected as treated states, while the remaining 45 states serve as controls. The pre-treatment period is divided into training years (1994 - 1998) and validation years (1999 - 2007), mirroring the structure of our empirical application in Section 5 of the manuscript.

- Each simulation scenario fixed the ratio of treated to control units at approximately 1:8, with total sample sizes ranging from 9 to 18,000 units (e.g., 1 treated and 8 controls, up to 2,000 treated and 16,000 controls). Seven such scenarios were defined, each with 1,000 simulation replications.
- Each simulation iteration involved:
  - Treated and control units sampled with replacement from the population.
  - Set of fixed candidate models fit via `apm_pre()` with 1,000 posterior draws.
  - Treatment point and bound estimation with `M = 1` using `apm_est()` with 1,000 bootstrap replications, which were used to compute standard errors and confidence intervals.
  - For each replication, the simulation stored the following
    - posterior model probabilities
    - estimates of the ATT and its bounds under each candidate model
    - Bayesian model averaging (BMA) estimates of the ATT and its bounds
    - estimates of variance over the model posterior (holding the sample fixed)
    - estimates of variance over bootstrapped samples (holding the model posterior fixed)
    - overall BMA variance estimates
    - confidence interval coverage indicators
- In the final (largest sample size) scenario – comprising 2,000 treated and 16,000 control units – the simulation was restricted to the two models with the highest average posterior probability from the preceding scenario. This decision was made for computational feasibility, as fitting all candidate models at that scale was prohibitively slow and memory-intensive. The restriction is methodologically sound, as the remaining models exhibited near-zero posterior probability across all prior simulations and did not contribute meaningfully to inference.
- Simulations were executed in parallel using `parLapply()` and `makeCluster()`, with reproducibility ensured via `set.seed(..., kind = "L'Ecuyer-CMRG")`, which enables consistent and independent random number streams across parallel workers.
- All simulation outputs were saved to `code_and_output/simulation/sim_res.RData` for further summary and visualization.