Take-Home Midterm: LaLonde (1986) and Nonexperimental ATT

Due: Friday, February 13 (11:59pm)

Overview

In the last homework you used the experimental NSW data to estimate the ATE. In this take-home midterm you will use nonexperimental data to estimate the ATT. You will work with the Dehejia–Wahba (1999) subset of the experimental sample, as well as observational control data, and will then evaluate estimators.

Reference tutorial: https://yiqingxu.org/tutorials/lalonde/

We follow many of the procedures used in the tutorial, but they have some built in functions and use some machine learning estimators that we haven't covered; we are going to put together the pieces using different estimation techniques. You are allowed to use the tutorial for reference, for understanding the data and context, but their code will not carry over directly.

Data and setup

In our analysis, as in the tutorial, will use three samples built around the LaLonde–Dehejia–Wahba (LDW) data: (1) LDW-Experimental, which is the experimental subset of the original experimental data; (2) LDW-CPS1, which pairs the LDW-Experimental treated units with observational controls from the Current Population Survey (CPS-SSA-1) ; and (3) LDW-PSID1, which pairs the same LDW treated units with observational controls from the Panel Study of Income Dynamics (PSID-1).

`ldw_cps` and `ldw_psid` already *include* the LDW treated observations. The experimental control group is stored separately as `ldw_co`.

```r
library(Matching)
```

```
## Loading required package: MASS

## ##
## ##  Matching (Version 4.10-15, Build Date: 2024-10-14)
## ##  See https://www.jsekhon.com for additional documentation.
## ##  Please cite software as:
## ##   Jasjeet S. Sekhon. 2011. ``Multivariate and Propensity Score Matching
## ##   Software with Automated Balance Optimization: The Matching package for R.''
## ##   Journal of Statistical Software, 42(7): 1-52.
## ##
```

```r
library(estimatr)

# Load prepared objects (nsw, ldw, ldw_cps, ldw_psid, ldw_co, etc.)
lalonde_url <- "https://raw.githubusercontent.com/xuyiqing/lalonde/master/data/lalonde.RData"
load(url(lalonde_url))

# define variables
Y <- "re78"
treat <- "treat"
```

```
covar <- c("age", "education", "black", "hispanic", "married", "nodegree",
           "re74", "re75", "u74", "u75")
```

Problems

Part 1: Constructing the Comparison Samples

1.1 Re-estimate propensity scores (LDW-Experimental, CPS1, PSID1)

Estimate the treatment propensity score for each of these datasets: LDW-Experimental (`ldw`), LDW-CPS1 (`ldw_cps`), and LDW-PSID1 (`ldw_psid`). Use a logit model with `glm(..., family = binomial())` and the covariate set `covar`.

Create and use the following object names:

- `exp_df` for LDW-Experimental
- `cps_untrim` for LDW-CPS1 (untrimmed)
- `psid_untrim` for LDW-PSID1 (untrimmed)

```
exp_df <- ldw
cps_untrim <- ldw_cps
psid_untrim <- ldw_psid
```

Deliverable: redefined datasets, no results reporting needed for this problem.

1.2 Propensity score overlap plots

Using `ggplot2` and `geom_histogram`, plot propensity score overlap for each merged dataset and for the true experimental population. Use different fill colors for treated vs control and label each plot clearly (LDW-CPS1 vs LDW-PSID1 vs LDW-Experimental). Use `aes(y = after_stat(density))` so the y-axis is density rather than counts. Also report the number of treated and control observations for each dataset you plot.

Deliverable: three overlap plots each accompanied by treated/control counts.

1.3 Add experimental controls + re-estimate + trim

For each of the two combined datasets, create a new data set that merges in the experimental controls from `ldw_co` to expand the control group. Then, on these new expanded data sets, estimate each unit's propensity **of being included in the experiment** using `covar`.

We are going to use this experimental propensity score to implement trimming. On these new data sets, drop units with extreme propensity scores as follows: - LDW-CPS1: drop units with propensity score > 0.9 - LDW-PSID1: drop units with propensity score > 0.8

Create and use these object names:

- `ldw_cps.plus` and `ldw_psid.plus` for the expanded datasets
- `ldw_cps.plus_trim` and `ldw_psid.plus_trim` after trimming
- `cps_trim` and `psid_trim` for ATT analysis (samples 1&3, 1&4)

Deliverable: redefined datasets, no results reporting needed for this problem.

Note: Before `rbind`, make sure the two data frames have the same columns and create a new indicator (e.g., `experimental`) for the experiment status. Then estimate `experimental ~ covar` by logit and trim on that fitted probability.

1.4 Propensity score matching (1:1)

Using the trimmed datasets from Problem 1.3, perform 1:1 nearest-neighbor propensity score matching **without replacement**. Steps:

1) Keep only treated LDW units and nonexperimental controls:
   - CPS: `sample %in% c(1, 3)`
   - PSID: `sample %in% c(1, 4)`
2) Re-estimate the propensity score (logit `glm`) on the filtered sample.
3) Match treated to controls on the propensity score with `Matching::Match`, using `M = 1`, `replace = FALSE`, `ties = FALSE`, and `estimand = "ATT"`. The key arguments are:
   - `Y`: outcome (`re78`)
   - `Tr`: treatment indicator (`treat`)
   - `X`: the propensity score vector
4) Extract the matched rows using `index.treated` and `index.control`, then bind those rows to form a matched dataset for later steps.

Create and use: - `ldw_cps_matched` and `ldw_psid_matched`

Note: `Matching::Match` returns indices. You can recover the matched data with something like `matched_df <- rbind(df[m$index.treated, ], df[m$index.control, ])`.

1.5 Re-assess overlap after matching

Plot propensity score overlap for the matched CPS and PSID datasets. Use `ggplot2` and `geom_histogram` with `aes(y = after_stat(density))`. Report the number of treated and control observations for each matched dataset.

Deliverable: two overlap plots each accompanied by treated/control counts

1.6 Covariate balance (by hand)

Check covariate balance. For each covariate in `covar`, compute standardized mean differences (treated - control) pre-matching and post-matching. Use the original LDW-CPS1 / LDW-PSID1 samples for pre-matching and the matched datasets for post-matching.

For each dataset (CPS and PSID), make one plot where: - x-axis = standardized mean difference (SMD) - y-axis = covariate name - show matched vs unmatched SMDs in different colors on the same plot

Comment on what you see.

Deliverable: two balance plots + a brief interpretation.

Note: A simple SMD for covariate `x` is `(mean(x_t) - mean(x_c)) / sd(x_all)`, where `x_t` and `x_c` are treated and control values and `x_all` is the pooled sample. You can wrap this in a small function and apply it across `covar`.

Part 2: ATT Estimation

Estimate the ATT using the five samples you created in Part 1: `exp_df`, `cps_untrim`, `psid_untrim`, `cps_trim`, and `psid_trim`.

For each sample, compute the following estimators:

- difference in means,
- Lin (`estimatr::lm_lin`),
- IPW,
- stabilized IPW, and

- doubly robust.

Use bootstrap confidence intervals for all estimates and report a single table with estimates and 95% CIs. Stratify bootstraps so that each bootstrap sample has the same number of treated and control units as the original sample (i.e., resample treated and control units separately). Use at least 500 bootstrap replications. Re-estimate the propensity scores on each bootstrap.

Deliverable: one table with estimates + bootstrap 95% CIs. Optional: plot estimates.

Part 3: Placebo Analysis

Repeat the analysis using **placebo outcomes** to assess unconfoundedness. Set the outcome to `re75`, and remove `re75` and `u75` from the conditioning set. Create **new trimmed samples** based on 1:1 matching of propensity scores estimated via logit (without using `re75`/`u75`), and then estimate the ATT **by hand** using the same estimators as Part 2 (diff, Lin, IPW, stabilized IPW, doubly robust), with bootstrap CIs.

For the placebo, use five datasets: `exp_df`, `cps_untrim`, `psid_untrim`, and `cps_placebo_trim` (new placebo-trimmed CPS from matching), `psid_placebo_trim` (new placebo-trimmed PSID from matching).

Deliverable: one table with estimates + bootstrap 95% CIs. Optional: plot estimates.