Problem set 1

Hyoungchul Kim

2025-02-07

Basic and matching theory

1.

Let $h(X)$ be a function of X. Then expectation of mean squared error is: $\mathbb{E}\left(Y - h(X)\right)^2 = \mathbb{E}(Y - \mathbb{E}(Y|X) + \mathbb{E}(Y|X) + h(X))^2 = \mathbb{E}(Y - \mathbb{E}(Y|X))^2 + \mathbb{E}(\mathbb{E}(Y|X) + h(X))^2 - 2\mathbb{E}\left[(Y - \mathbb{E}(Y|X))\underbrace{(\mathbb{E}(Y|X) - h(X))}_{\text{g(X)}}\right]$.

Then we can show that the last term becomes zero because $\mathbb{E}\left[g(X)Y - g(X)\mathbb{E}(Y|X)\right] = \mathbb{E}\left[g(X)Y - \mathbb{E}(g(X)Y|X)\right] = \mathbb{E}(g(X)Y) - \mathbb{E}(g(X)Y)$, which is due to law of iterated expectation.

Then in order to minimize this error, we need to set $h(X) = \mathbb{E}(Y|X)$.

2. In potential outcomes framework, we can write ATT as $\mathbb{E}(\hat{\tau}_{ATT}) = \mathbb{E}[Y_1 \mid D_i = 1] - \mathbb{E}[Y_0 \mid D_i = 1] = \mathbb{E}[Y_1 \mid D_i = 1] - \mathbb{E}[Y_i \mid D_i = 1]$

Matching empirics

1.

Read the data and do basic data cleaning

First do it for whole sample

```{r}
#| code-fold: true
#| message: false
#| warning: false
library(tidyverse)
library(modelsummary)
nsw = haven::read_dta("data/nsw.dta")
nsw_dw = haven::read_dta("data/nsw_dw.dta")
cps1 = haven::read_dta("data/cps1.dta")
psid1 = haven::read_dta("data/psid1.dta")

# add up by data_id

data = nsw |>
  bind_rows(nsw_dw) |>
  bind_rows(cps1) |>
  bind_rows(psid1)

datasummary(age + education + black + hispanic + married + nodegree + re74 + re75 + re78 ~ data
```

This is for just treatment group

```{r}
#| cold-fold: true
#| message: false
#| warning: false
```

	CPS1		Dehejia-Wahba Sample		Lalonde Sample		PSID	
	mean	sd	mean	sd	mean	sd	mean	sd
age	33.23	11.05	25.37	7.10	24.52	6.63	34.85	10.44
education	12.03	2.87	10.20	1.79	10.27	1.70	12.12	3.08
black	0.07	0.26	0.83	0.37	0.80	0.40	0.25	0.43
hispanic	0.07	0.26	0.09	0.28	0.11	0.31	0.03	0.18
married	0.71	0.45	0.17	0.37	0.16	0.37	0.87	0.34
nodegree	0.30	0.46	0.78	0.41	0.78	0.41	0.31	0.46
re74	14 016.80	9569.80	2102.27	5363.58			19 428.75	13 406.88
re75	13 650.80	9270.40	1377.14	3150.96	3042.90	5066.14	19 063.34	13 596.95
re78	14 846.66	9647.39	5300.76	6631.49	5454.64	6252.94	21 553.92	15 555.35

```
data_treat = data |> filter(treat ==1)

datasummary(age + education + black + hispanic + married + nodegree + re74 + re75 + re78 ~ data
```

	Dehejia-Wahba Sample		Lalonde Sample	
	mean	sd	mean	sd
age	25.82	7.16	24.63	6.69
education	10.35	2.01	10.38	1.82
black	0.84	0.36	0.80	0.40
hispanic	0.06	0.24	0.09	0.29
married	0.19	0.39	0.17	0.37
nodegree	0.71	0.46	0.73	0.44
re74	2095.57	4886.62		
re75	1532.06	3219.25	3066.10	4874.89
re78	6349.14	7867.40	5976.35	6923.80