IEMS 402 Statistical Learning - 2025 Winter
HW4

Tuorui Peng[1]

Exercise 1 Estimating the Derivatives via Kernel Smoothing

- Bias term:

$$\begin{aligned}\left|d_n(x)-p'(x)\right| =&\left|\int_{-1}^{1}\frac{1}{h^2}K\left(\frac{X_i-x}{h}\right)p(x)\,\mathrm{d}x-p'(x)\right|\\ =&\frac{1}{h}\left|\int K(\|v\|)\big(p(x+hv)-vp'(x)\big)\,\mathrm{d}v\right|\\ =&\frac{1}{h}\left|\int K(\|v\|)\big(p(x+hv)-vp'_{x,\beta}(x+hv)+vp'_{x,\beta}(x+hv)-vp'(x)\big)\,\mathrm{d}v\right|\\ \leq&\frac{1}{h}\int K(\|v\|)\left|p(x+hv)-vp'_{x,\beta}(x+hv)\right|\,\mathrm{d}v+\frac{1}{h}\int K(\|v\|)\left|vp'_{x,\beta}(x+hv)-vp'(x)\right|\,\mathrm{d}v\\ \lesssim&Lh^{\beta-1}\int K(\|v\|)\,|v|\,\mathrm{d}v\end{aligned}$$

  in which the last inequality is because $p(\,\cdot\,)-p_{x,\beta}(\,\cdot\,)$ is still a β-Hölder function.

$$\begin{aligned}var(d_n(x))\leq&\frac{1}{nh^4}\int K^2\Big(\frac{X_i-x}{h}\Big)p(x)\,\mathrm{d}x\\ =&\frac{1}{nh^3}\int K^2(v)p(x+hv)\,\mathrm{d}v\\ \lesssim&\frac{1}{nh^3}\sup_{x\in[-1,1]}p(x)\int K^2(v)\,\mathrm{d}v\end{aligned}$$

  where the last inequality is because only the first order term of $p(\,\cdot\,)$ gives non-zero kernel integration.

Put together, we have

$$\mathrm{MSE}\lesssim h^{2(\beta-1)}+\frac{1}{nh^3}$$

Optimal h_n is at $h_n\asymp n^{-\frac{1}{2\beta+1}}$, and the optimal MSE is $n^{-\frac{2(\beta-1)}{2\beta+1}}$.

Exercise 2 An average treatment effect estimator

2.(a)

In completely randomized experiment, we can see that

$$\begin{aligned}\mathbb{E}\left[Y_i(a)1\{A_i=a\}\right]=&\frac{1}{2}\mathbb{E}\left[Y_i(a)1\{A_i=a\}|A_i=0\right]+\frac{1}{2}\mathbb{E}\left[Y_i(a)1\{A_i=a\}|A_i=1\right]\\ =&\begin{cases}\frac{1}{2}\mathbb{E}\left[Y_i(0)|A_i=0\right], & \text{if } a=0\\ \frac{1}{2}\mathbb{E}\left[Y_i(1)|A_i=1\right], & \text{if } a=1\end{cases}\\ =&\frac{1}{2}\mathbb{E}\left[Y_i(a)\right]=\frac{1}{2}\mathbb{E}\left[Y(a)\right]\end{aligned}$$

[1] TuoruiPeng2028@u.northwestern.edu

and we have ATE being

$$\tau = \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right] = 2\mathbb{E}\left[Y(1)1\{A=1\}\right] - 2\mathbb{E}\left[Y(0)1\{A=0\}\right].$$

2.(b)

Note that we can write $\hat{\tau}_n$ as

$$\begin{aligned}
\hat{\tau}_n =& \frac{1}{n}\left[2Y_i(1)1\{A_i=1\} - 2Y_i(0)1\{A_i=0\}\right]\\
\mathbb{E}\left[2Y_i(1)1\{A_i=1\} - 2Y_i(0)1\{A_i=0\}\right] =& \tau\\
var(2Y_i(1)1\{A_i=1\} - 2Y_i(0)1\{A_i=0\}) =& var(\mathbb{E}\left[2Y_i(1)1\{A_i=1\} - 2Y_i(0)1\{A_i=0\}|A\right])\\
&+ \mathbb{E}\left[var(2Y_i(1)1\{A_i=1\} - 2Y_i(0)1\{A_i=0\}|A)\right]\\
=& (\mathbb{E}\left[Y(1)\right] + \mathbb{E}\left[Y(0)\right])^2 + 2\big(var(Y(1)) + var(Y(0))\big)
\end{aligned}$$

By CLT

$$\sqrt{n}(\hat{\tau}_n - \tau) \xrightarrow{\mathrm{d}} N(0, (\mu_1+\mu_0)^2 + 2(\sigma_1^2+\sigma_0^2)), \quad \mu_a = \mathbb{E}\left[Y(a)\right], \sigma_a^2 = var(Y(a))$$

2.(c)

Note that for completely randomized experiment, we have $|S_1| = n - |S_0| \sim \mathrm{Binomial}(n, \frac{1}{2})$. Thus we get (take $a=1$ example)

$$\begin{cases}\sqrt{n}(1 - 2\left|S_1\right|/n) \xrightarrow{\mathrm{d}} N(0,1)\\ 2\left|S_1\right|/n \xrightarrow{\mathrm{p}} 1\end{cases} \overset{\text{Slutsky}}{\Rightarrow} \sqrt{n}(\frac{n}{2\left|S_1\right|} - 1) \xrightarrow{\mathrm{d}} N(0,1)$$

and similarly for $a=0$ we have $\sqrt{n}(\frac{n}{2\left|S_0\right|} - 1) \xrightarrow{\mathrm{d}} N(0,1)$.

2.(d)

We have

$$\begin{aligned}
\sqrt{n}(\hat{\tau}_n^{\text{norm}} - \tau) =& \frac{\sqrt{n}}{\sqrt{|S_1|}}\sqrt{|S_1|}\big(\frac{1}{|S_1|}\sum_i (Y_i(1) - \mathbb{E}\left[Y(1)\right])1\{A_i=1\}\big)\\
&+ \frac{\sqrt{n}}{\sqrt{|S_0|}}\sqrt{|S_0|}\big(\frac{1}{|S_0|}\sum_i (Y_i(0) - \mathbb{E}\left[Y(0)\right])1\{A_i=0\}\big)
\end{aligned}$$

For now we treat A as given, and we have

$$\sqrt{n}(\hat{\tau}_n^{\text{norm}} - \tau)\big|A \xrightarrow{\mathrm{d}} \frac{\sqrt{n}}{\sqrt{|S_1|}} N(0,\sigma_1^2) + \frac{\sqrt{n}}{\sqrt{|S_0|}} N(0,\sigma_0^2)$$

On the other hand we notice that we already have

$$\frac{\sqrt{n}}{\sqrt{|S_1|}} \xrightarrow{\mathrm{p}} \sqrt{2}, \quad \frac{\sqrt{n}}{\sqrt{|S_0|}} \xrightarrow{\mathrm{p}} \sqrt{2}, \quad cov\left(\frac{n}{|S_1|}, \frac{n}{|S_0|}\right) = -1$$

Thus by Slutsky's theorem, we have

$$\sqrt{n}(\hat{\tau}_n^{\text{norm}} - \tau) \xrightarrow{\text{d}} N(0, \sigma^2_{\text{norm}})$$

where

$$\sigma^2_{\text{norm}} = 2\sigma_1^2 + 2\sigma_0^2$$

2.(e)

It seems that based on our results up to now, the conlusion would be that:

$$\begin{aligned}
\sqrt{n}(\hat{\tau}_n - \tau) \xrightarrow{\text{d}} N(0, \sigma^2), \quad \sigma^2 = (\tau_1 + \tau_0)^2 + 2(\sigma_1^2 + \sigma_0^2)\\
\sqrt{n}(\hat{\tau}_n^{\text{norm}} - \tau) \xrightarrow{\text{d}} N(0, \sigma^2_{\text{norm}}), \quad \sigma^2_{\text{norm}} = 2(\sigma_1^2 + \sigma_0^2)
\end{aligned}$$

so we have $\sigma^2 > \sigma^2_{\text{norm}}$ as long as $\tau_1 + \tau_0 > 0$.

Exercise 3 A weighted average treatment effect estimator

3.(a)

With the covariate X involved, we have

$$\begin{aligned}
\tau =& \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right]\\
=& \mathbb{E}\left[\mathbb{E}\left[Y(1)\big(1\{A=1\} + 1\{A=0\}\big)\middle|X=x\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y(0)\big(1\{A=1\} + 1\{A=0\}\big)\middle|X=x\right]\right]
\end{aligned}$$

now note that since $(Y(1), Y(0)) \perp A|X$, we have

$$\begin{aligned}
\mathbb{E}\left[Y(A)1\{A=1\}|X=x\right] =& Y(1)e(x)\\
\mathbb{E}\left[Y(A)1\{A=0\}|X=x\right] =& Y(0)(1-e(x))
\end{aligned}$$

substitute this back to the above equation, we have

$$\begin{aligned}
\tau =& \mathbb{E}\left[\frac{Y(A)1\{A=1\}}{e(x)}\big(1\{A=1\} + 1\{A=0\}\big) - \frac{Y(A)1\{A=0\}}{1-e(x)}\big(1\{A=1\} + 1\{A=0\}\big)\middle|X=x\right]\\
=& \mathbb{E}\left[\frac{Y(A)1\{A=1\}}{e(x)} - \frac{Y(A)1\{A=0\}}{1-e(x)}\middle|X=x\right]\\
=& \mathbb{E}\left[\frac{Y(A)1\{A=1\}}{e(X)}\right] - \mathbb{E}\left[\frac{Y(A)1\{A=0\}}{1-e(X)}\right]
\end{aligned}$$

3.(b)

By CLT, the propensity weighted estimator

$$\frac{1}{n}\sum_{i=1}^n \left[\frac{Y_i 1\{A_i=1\}}{e(X_i)} - \frac{Y_i 1\{A_i=0\}}{1-e(X_i)}\right] \xrightarrow{\text{d}} N(\tau, n\sigma^2_{\text{ps}})$$

where the variance is computed as follows:

- Prepare for the calculation:

$$
\begin{aligned}
\mathbb{E}\left[(1\{A_i=1\}-e(X_i))Y_i|X_i=x\right] =&0\\
var((1\{A_i=1\}-e(X_i))Y_i|X_i=x) =&e(x)(1-e(x))\mathbb{E}\left[Y_i^2|X_i=x\right]\\
=&e(x)(1-e(x))\mathbb{E}\left[(Y(1)1\{A=1\}+Y(0)1\{A=0\})^2|X=x\right]\\
=&e(x)(1-e(x))\big(e(x)v_2(x,1)^2+(1-e(x))v_2(x,0)^2\big)
\end{aligned}
$$

$$
\begin{aligned}
\sigma^2_{\text{ps}} =&var(\frac{Y_i1\{A_i=1\}}{e(X_i)}-\frac{Y_i1\{A_i=0\}}{1-e(X_i)})\\
=&var(\frac{(1\{A_i=1\}-e(X_i))Y_i}{e(X_i)(1-e(X_i))})\\
=&\mathbb{E}\left[var(\frac{(1\{A_i=1\}-e(X_i))Y_i}{e(X_i)(1-e(X_i))}|X_i=x)\right]+var\left(\mathbb{E}\left[\frac{(1\{A_i=1\}-e(X_i))Y_i}{e(X_i)(1-e(X_i))}|X_i=x\right]\right)\\
=&\mathbb{E}\left[\frac{\big(e(X)v_2(X,1)^2+(1-e(X))v_2(X,0)^2\big)}{e(X)(1-e(X))}\right]\\
=&\mathbb{E}\left[\frac{v_2(X,1)^2}{1-e(X)}\right]+\mathbb{E}\left[\frac{v_2(X,0)^2}{e(X)}\right]
\end{aligned}
$$

3.(c)

Note that, given $X=x$, we have

$$
\sigma^2_{\text{ps},x}=\left(\frac{v_2(x,1)^2}{1-e(x)}+\frac{v_2(x,0)^2}{e(x)}\right)\cdot(1-e(x)+e(x))\overset{\text{Cauchy}-\text{Schwarz}}{\geq}(v_2(x,1)+v_2(x,0))^2
$$

which takes the minimum when

$$
\frac{v_2(x,1)}{1-e(x)}=\frac{v_2(x,0)}{e(x)}\Rightarrow e(x)=\frac{v_2(x,0)}{v_2(x,0)+v_2(x,1)}
$$

which is the optimal propensity score chosen to minimize the variance of the propensity score weighted estimator.

One sentence intuition: the choice of propensity score should balance variance within two groups w.r.t. the covariate X, i.e. similar to the idea to avoiding simpson's paradox. Thus this method can lead to significant improvement over the naive randomized experiment estimator when we do have a good covariate X to work with.

Exercise 4 Logistic regression

For logistic regression, we have

$$
\begin{aligned}
\ell(\theta):=\text{ log-likelihood }(\theta)=&\frac{1}{n}\sum_{i=1}^n Y_i\log\pi_\theta(X_i)+(1-Y_i)\log(1-\pi_\theta(X_i)),\quad \pi_\theta(X)=\frac{1}{1+\exp(-\theta'X)}\\
\frac{\partial^2\ell}{\partial\theta\partial\theta'}=&\frac{1}{n}\sum_{i=1}^n\frac{\exp(-\theta'X_i)}{(1+\exp(-\theta'X_i))^2}X_iX_i'\\
I(\theta):=&\mathbb{E}\left[\frac{\exp(-\theta'X_i)}{(1+\exp(-\theta'X_i))^2}X_iX_i'\right]
\end{aligned}
$$

We have convergence of $\hat{\theta}_n = \arg\max_\theta \ell(\theta)$ that:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} N(0, I(\theta)^{-1})$$