Prof. Yiping Lu
IEMS 402 Statistical Learning
November 21, 2024

Homework 4: Asymptotic Theory

Question 1. Estimating the Derivatives via Kernel Smoothing Given a scalar $\beta > 1$, let p be a probability density function on $\mathbb{R}$ such that $p \in \Sigma(\beta)$ (i.e., β-th order Hölder class). We are interested in nonparametric estimation of the derivative p'.

Given a kernel function $K : \mathbb{R} \to \mathbb{R}$ supported on $[-1, 1]$ satisfying the conditions

$$\int_{\mathbb{R}} u^j K(u)\, du = \begin{cases} 1 & j = 1, \\ 0 & j = 0, 2, \cdots, \lfloor \beta \rfloor. \end{cases}$$

Let $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} p$. Given bandwidth $h > 0$, consider the kernel-based estimator

$$\hat{d}_n(x) := \frac{1}{nh^2} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

For any x_0, and prove the MSE bound

$$\mathbb{E}\left[|\hat{d}_n(x_0) - p'(x_0)|^2\right] \leq n^{-\frac{2(\beta-1)}{1+2\beta}}.$$

Question 2. (An average treatment effect estimator) In the Neyman-Rubin (potential outcomes) approach to causal estimation, one treats estimation as a missing data problem. Let $A \in \{0, 1\}$ be an action (often called the treatment or intervention). The *potential outcomes* are the pair $(Y(0), Y(1)) \in \mathbb{R}$, where $Y(0)$ is the response when action $A = 0$ is chosen and $Y(1)$ the response when $A = 1$ is chosen. Thus, for any individual, we observe a single response: under action $A = a$, we observe $Y(a)$ but never $Y(1 - a)$. The *average treatment effect* is the difference

$$\tau := \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)],$$

where the expectation is taken over the population of individuals we might intervene on. Here, $A = 1$ is the treatment, while $A = 0$ indicates the control (untreated) action, and we may use the notation $Y1\{A = a\} := Y(a)1\{A = a\}$.

The "gold standard" approach is a randomized experiment, where for individuals $i = 1, 2, \ldots, n$, one chooses $A_i \in \{0, 1\}$ uniformly and observes $Y_i(A_i) \in \mathbb{R}$. We assume that individuals are i.i.d.

(a) Show that for $a \in \{0, 1\}$, we have $\mathbb{E}[Y_i(a)1\{A_i = a\}] = \frac{1}{2}\mathbb{E}[Y(a)]$ in the randomized experiment setting, and hence that $\tau = 2\mathbb{E}[Y(1)1\{A = 1\}] - \mathbb{E}[Y(0)1\{A = 0\}]$.

We consider two mean-based estimators. For $a \in \{0, 1\}$, define the sets $S_a = \{i \in [n] \mid A_i = a\}$ (i.e., the treatment and control groups). The *basic estimator* is

$$\hat{\tau}_n := \frac{1}{n} \sum_{i \in S_1} 2Y_i - \frac{1}{n} \sum_{i \in S_0} 2Y_i.$$

(b) Give the asymptotic distribution of $\hat{\tau}_n$. (That is, give the limit distribution of $\sqrt{n}(\hat{\tau}_n - \tau)$.)

We also consider the slightly more nuanced mean-based estimator, which normalizes by the sample sizes,

$$\hat{\tau}_n^{\text{norm}} := \frac{1}{|S_1|} \sum_{i \in S_1} Y_i - \frac{1}{|S_0|} \sum_{i \in S_0} Y_i.$$

(c) For $a \in \{0, 1\}$, give the asymptotic distribution of $\sqrt{n}\left(\frac{n}{2|S_a|} - 1\right)$.

(d) Give the asymptotic distribution of the mean-based estimator $\hat{\tau}_n^{\text{norm}}$.

Hint:

- It may be useful to split the quantities by considering the means $\tau_a = \mathbb{E}[Y(a)]$ for $a \in \{0, 1\}$ separately.
- Using delta method

(e) **(Extra Credits)** In the preceding parts, you have shown that

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

Show that if the means $\tau_a = \mathbb{E}[Y(a)]$ satisfy $\tau_0 \neq \tau_1$, then $\sigma^2 > \sigma^2_{\text{norm}}$.

Hint: The new estimator rolls out the variance of sampling treatments.

Question 3. (A weighted average treatment effect estimator) We consider the same setting as in problem 3.5, but take an alternative approach, where we may differentially sample individuals based on their covariates X. To that end, consider a *propensity score* (the propensity for being treated)

$$e(x) := \mathbb{P}(A = 1 \mid X = x).$$

Now, we assume that given an individual with covariates $X = x$, we assign treatment A conditionally according to the propensity score (3.2), that is, $\mathbb{P}(A = a \mid X = x) = e(x)$, so that $(Y(0), Y(1)) \perp A \mid X$, that is, the potential responses $(Y(0), Y(1))$ are independent of A given X.

(a) Show that the average treatment $\tau = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ also equals

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

(b) Define the conditional second moments $v_2(x, a) := \sqrt{\mathbb{E}[Y(a)^2 \mid X = x]}$, and consider the propensity weighted estimator

$$\hat{\tau}_n^{\text{ps}} := \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i 1\{A_i = 1\}}{e(X_i)} - \frac{Y_i 1\{A_i = 0\}}{1 - e(X_i)}\right].$$

Compute the asymptotic variance σ^2_{ps} in

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

as a function (with appropriate expectations) of $v_2(x, a)$ and $e(x)$.

(c) Which choice of propensity score $e(x)$ minimizes the asymptotic variance σ^2_{ps}? Give a one-sentence (heuristic) intuition for this choice. When does this improve over the "gold standard" approach of the pure randomized experiment in part (b) in Q. 3.5?

Question 4. (Logistic regression) Consider d-dimensional random vectors $X_1, X_2, \cdots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}$, with $\mathbb{E}\left[\|X\|_2^2\right] < +\infty$. Let the binary labels be generated as

$$Y_i | X_i \sim \text{Bernoulli}(\pi_{\theta^*}(X_i)), \quad \text{for } i = 1, 2, \cdots, n,$$

where we define

$$\pi_\theta(x) := \frac{1}{1 + \exp(-x^\top \theta)}.$$

Let Θ be a compact set such that θ^* lies in the interior of Θ. Consider the maximal likelihood estimator

$$\hat{\theta}_n := \arg\max_{\theta \in \Theta} \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i \log \pi_\theta(X_i) + (1 - Y_i)\log\left(1 - \pi_\theta(X_i)\right)\right\}.$$

Let $n \to +\infty$ with everything else fixed. Assume that the Fisher information is non-singular, derive and prove the convergence rate and asymptotic distribution for $\hat{\theta}_n$.

Northwestern University