Advanced Models B 52805 (Midterm quiz, 2021)

Prof. P. Chigansky

Consider the statistical functional[1]

$$T(F) = \begin{cases} \dfrac{1}{F(b)-F(a)} \displaystyle\int_a^b x dF(x), & \text{if } F(b)-F(a) > 0, \\ 0, & \text{otherwise,} \end{cases}$$

where $a < b$ are fixed real numbers.

(1) Specify the plug-in estimator $T(\widehat{F}_n)$ and find its value at the data set

$$(-1, 0, 1, \tfrac{1}{2}, \tfrac{1}{2}),$$

when $a = 0$ and $b = 1$.

(2) Prove that $T(\widehat{F}_n)$ is consistent at any F.

(3) Find the influence function of $T(F)$ and the corresponding limit variance V_F.

(4) Argue that the plug-in estimator of V_F is consistent and specify the asymptotic $1-\alpha$ confidence interval for $T(F)$, obtained by means of nonparametric Delta-method.

(5) Show that the limit minimax risk of $T(\widehat{F}_n)$ on the set of all distributions $\mathcal{F}$ is infinite,

$$\lim_{n\to\infty} \sup_{F\in\mathcal{F}} \mathbb{E}_F \left|\sqrt{n}\big(T(\widehat{F}_n) - T(F)\big)\right|^p = \infty,$$

for any $p > 0$.

Hint: show that there exists a positive constant a_p such that

$$\sup_{c\in\mathbb{R}} \mathbb{E}_{F_c} \left|T(\widehat{F}_n) - T(F_c)\right|^p \ge a_p, \quad \forall n,$$

where F_c is the Laplace distribution with the density

$$f_c(x) = \frac{1}{2} e^{-|x-c|}, \quad x \in \mathbb{R}, \quad c \in \mathbb{R}.$$

[1] This is the conditional mean $\mathbb{E}_F(X|X \in (a,b])$ for $X \sim F$

Advanced Models B 52805 (Midterm quiz, 2021)

Prof. P. Chigansky

- Quiz duration is 90 minutes
- Open material
- All questions have the same weight

Consider the statistical functional[1]

$$T(F) = \begin{cases} \dfrac{1}{F(b)-F(a)} \displaystyle\int_a^b x dF(x), & \text{if } F(b)-F(a)>0, \\ 0, & \text{otherwise,} \end{cases}$$

where $a<b$ are fixed real numbers.

(1) Specify the plug-in estimator $T(\widehat{F}_n)$ and find its value at the data set

$$(-1, 0, 1, \tfrac{1}{2}, \tfrac{1}{2}),$$

when $a=0$ and $b=1$.

Adopting the convention $0/0=0$, the plug-in estimator is

$$T(\widehat{F}_n) = \frac{\sum_{m=1}^n X_m \mathbf{1}_{\{a<X_m\le b\}}}{\sum_{m=1}^n \mathbf{1}_{\{a<X_m\le b\}}}.$$

For the data at hand, the value of $T(\widehat{F}_n)$ is

$$T(\widehat{F}_n) = \frac{1+\frac{1}{2}+\frac{1}{2}}{3} = \frac{2}{3}.$$

(2) Prove that $T(\widehat{F}_n)$ is consistent at any F.

[1]This is the conditional mean $\mathbb{E}_F(X|X\in[a,b])$ for $X\sim F$

If F is such that $F(b) = F(a)$, then $\mathbb{P}_F(X_j \in (a,b]) = 0$ and hence $T(\widehat{F}_n) = 0 = T(F)$, $\mathbb{P}_F$-a.s. due to the above convention. Otherwise, if $F(b) > F(a)$, by the LLN

$$T(\widehat{F}_n) = \frac{\frac{1}{n}\sum_{m=1}^n X_m \mathbf{1}_{\{a<X_m\le b\}}}{\frac{1}{n}\sum_{m=1}^n \mathbf{1}_{\{a<X_m\le b\}}} \xrightarrow[n\to\infty]{\mathbb{P}_F-a.s.} \frac{\mathbb{E}_F X_1 \mathbf{1}_{\{X_1\in(a,b]\}}}{\mathbb{P}_F(X_1\in(a,b])} = T(F).$$

This proves consistency at any F.

(3) Find the influence function of $T(F)$ and the corresponding limit variance V_F.

Define the functionals

$$T_i(F) = \int_a^b x^i dF(x), \quad i \in \{0,1,2\}.$$

For any path of distributions G_ε such that $\|G - G_\varepsilon\|_\infty \to 0$,

$$\frac{1}{\varepsilon}\Big(T_1(F+\varepsilon(G_\varepsilon - F)) - T_1(F)\Big) = \int_a^b x d(G_\varepsilon(x) - F(x)) =$$
$$\int_a^b x d(G(x) - F(x)) + \int_a^b x d(G_\varepsilon(x) - G(x)).$$

Here

$$\Big|\int_a^b x d(G_\varepsilon(x) - G(x))\Big| = \Big| x(G(x) - G_\varepsilon(x))\Big|_a^b - \int_a^b (G_\varepsilon(x) - G(x))dx\Big| \le$$
$$\le 2\big(|b| + |a|\big)\big\|G - G_\varepsilon\big\|_\infty \xrightarrow[\varepsilon\to 0]{} 0,$$

and hence $T_1(F)$ is Hadamard differentiable with derivative

$$\dot{T}_{1F}(G-F) = \int \psi_1 d(G-F),$$

where $\psi_1(y) = y\mathbf{1}_{\{y\in(a,b]\}}$. The corresponding influence function is

$$L_{1F}(x) = \dot{T}_{1F}(\delta_x - F) = \psi_1(x) - T_1(F)$$

Similarly, the functional

$$T_0(F) = F(b) - F(a) = \int_a^b dF(x),$$

is Hadamard differentiable with influence function

$$L_{0F}(x) = \psi_0(x) - T_0(F)$$

where $\psi_0(x) = \mathbf{1}_{\{x\in(a,b]\}}$.

The functional in question is $T(F) = h(T_0(F), T_1(F))$ where $h(t_0,t_1) = t_1/t_0$ with the gradient

$$\nabla h(t) = \left(-\frac{t_1}{t_0^2}, \frac{1}{t_0}\right)$$

which is a smooth function on $\{t \in \mathbb{R}^2 : t_0 \neq 0\}$. Hence for F such that $T_0(F) > 0$, the influence function of $T(F)$ is

$$L_F(x) = -\frac{T_1(F)}{T_0(F)^2} L_{0F}(x) + \frac{1}{T_0(F)} L_{1F}(x) = T(F)\Big(\frac{\psi_1(x)}{T_1(F)} - \frac{\psi_0(x)}{T_0(F)}\Big).$$

If F is such that $T_0(F) = 0$, the influence function is identically zero. The limit variance is therefore given by

$$\begin{aligned} V_F =& T(F)^2 \int \Big(\frac{\psi_1(x)}{T_1(F)} - \frac{\psi_0(x)}{T_0(F)}\Big)^2 dF(x) = \\ & T(F)^2 \int_a^b \Big(\frac{x}{T_1(F)} - \frac{1}{T_0(F)}\Big)^2 dF(x) = \\ & T(F)^2 \Big(\frac{T_2(F)}{T_1(F)^2} - \frac{1}{T_0(F)}\Big) \end{aligned}$$

if $T_0(F) > 0$ and $V_F = 0$ otherwise. As expected, this reduces to the empirical variance if $T_0(F) = 1$.

(4) Argue that the plug-in estimator of V_F is consistent and specify the asymptotic $1-\alpha$ confidence interval for $T(F)$, obtained by means of nonparametric Delta-method.

For F with $F(b) = F(a)$ the claim is trivial. Otherwise, by the LLN, $T_i(\widehat{F}_n) \to T_i(F)$, $\mathbb{P}_F$-a.s. and consistency follows by continuity. The confidence interval in question has the endpoints

$$T(\widehat{F}_n) \pm \frac{1}{\sqrt{n}} \sqrt{V_{\widehat{F}_n}} z_{1-\alpha/2}.$$

(5) Show that the limit minimax risk of $T(\widehat{F}_n)$ on the set of all distributions $\mathcal{F}$ is infinite,

$$\lim_{n\to\infty} \sup_{F\in\mathcal{F}} \mathbb{E}_F \big|\sqrt{n}\big(T(\widehat{F}_n) - T(F)\big)\big|^p = \infty,$$

for any $p > 0$.

Hint: show that there exists a positive constant a_p such that

$$\sup_{c\in\mathbb{R}} \mathbb{E}_{F_c} \big|T(\widehat{F}_n) - T(F_c)\big|^p \geq a_p, \quad \forall n,$$

where F_c is the Laplace distribution with the density

$$f_c(x) = \frac{1}{2} e^{-|x-c|}, \quad x \in \mathbb{R}, \quad c \in \mathbb{R}.$$

Obviously, the hinted lower bound implies the claim. For the Laplace distribution

$$T(F_c) = \frac{\int_a^b x e^{-|x-c|} dx}{\int_a^b e^{-|x-c|} dx} \overset{\dagger}{=} \frac{\int_a^b x e^{-(c-x)} dx}{\int_a^b e^{-(c-x)} dx} = \frac{\int_a^b x e^{x} dx}{\int_a^b e^{x} dx} =: R,$$

where † holds for all c large enough, i.e., $c > b$. Also, for all such c,

$$\mathbb{P}_{F_c}(T(\widehat{F}_n) = 0) \geq \mathbb{P}_{F_c}\big(T_0(\widehat{F}_n) = 0\big) =$$
$$\mathbb{P}_{F_c}(X_1 \notin (a,b])^n = \Big(1 - \tfrac{1}{2}e^{-c}\int_a^b e^x dx\Big)^n.$$

It follows that for all sufficiently large c,

$$\mathbb{E}_{F_c}\big|T(\widehat{F}_n) - T(F_c)\big|^p = \mathbb{E}_{F_c}\big|T(\widehat{F}_n) - R\big|^p \geq |R|^p \mathbb{P}_{F_c}(T(\widehat{F}_n) = 0),$$

and hence, for any n,

$$\sup_{F\in\mathcal{F}} \mathbb{E}_F\big|T(\widehat{F}_n) - T(F)\big|^p \geq \varliminf_{c\to\infty} \mathbb{E}_{F_c}\big|T(\widehat{F}_n) - T(F_c)\big|^p \geq |R|^p.$$

This proves the hinted bound, if $R \neq 0$. If $R = 0$, the same argument can be applied with $c \to -\infty$, since in this case [2]

$$T(F_c) = \frac{\int_a^b xe^{-x}dx}{\int_a^b e^{-x}dx} \neq 0.$$

Hi Amit,

Nothing to do with Markov. Here are the details

$$\mathbb{E}_{F_c}|T(\widehat{F}_n) - R|^p \geq \mathbb{E}_{F_c}|T(\widehat{F}_n) - R|^p 1_{\{T(\widehat{F}_n)=0\}} = |R|^p \mathbb{P}_{F_c}(T(\widehat{F}_n) = 0)$$

where the inequality holds since the random variable under expectation is nonnegative. Agree ?

P.

[2] Note that $\int_a^b x(e^x - e^{-x})dx > 0$ for any $a < b$, since the integrand, being a product of two odd functions, is positive on $\mathbb{R}\setminus\{0\}$.

DEPARTMENT OF STATISTICS, THE HEBREW UNIVERSITY, MOUNT SCOPUS, JERUSALEM 91905, ISRAEL
E-mail address: `Pavel.Chigansky@mail.huji.ac.il`