Круглов А.И. БО5-202 ST 7

1 $X_1, \dots, X_n$ – выборка с эмпир. ф-цией расп. $\hat F_n$. Найти $\operatorname{cov}(\hat F_n(x), \hat F_n(y))$

Лемма: $F(x) = \int_{-\infty}^{x} f(t)\,dt = \int_{\mathbb R} \mathbb 1\{X_i \le x\} f(t)\,dt = \mathbb E\, \mathbb 1\{X_i \le x\}$; эмпирическая функция распределения: $\hat F_n(x) = \frac1n \sum_{i=1}^n \mathbb 1\{X_i \le x\} \Rightarrow \mathbb E \hat F_n(x) = \frac1n \sum_{i=1}^n \mathbb E \mathbb 1\{X_i \le x\} = \frac1n \cdot n F(x) = F(x)$.

Также посчитаем $\mathbb E\left(\hat F_n(x)\,\hat F_n(y)\right) = \frac{1}{n^2} \mathbb E\left(\sum_{i=1}^n \mathbb 1\{X_i \le x\} \cdot \sum_{j=1}^n \mathbb 1\{X_j \le y\}\right) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \mathbb E\left(\mathbb 1\{X_i \le x\}\,\mathbb 1\{X_j \le y\}\right)$,

где $\mathbb E = \begin{cases} \mathbb E(\mathbb 1\{X_i \le x \text{ и } X_j \le y\}), & i = j \\ \mathbb E\,\mathbb 1\{X_i \le x\} \cdot \mathbb E\,\mathbb 1\{X_j \le y\}, & i \ne j \end{cases} = \begin{cases} F(\min(x,y)), & i = j \\ F(x) \cdot F(y), & i \ne j \end{cases} \Rightarrow \mathbb E\left(\hat F_n(x)\,\hat F_n(y)\right) =$

$= \frac{1}{n^2}\left(n F(\min(x,y)) + n(n-1) F(x) F(y)\right) = \frac1n F(\min(x,y)) + \frac{n-1}{n} F(x) F(y)$. Тогда ковариация

$\operatorname{cov}(\hat F_n(x), \hat F_n(y)) = \mathbb E\left(\hat F_n(x)\,\hat F_n(y)\right) - \mathbb E \hat F_n(x) \cdot \mathbb E \hat F_n(y) = \left(\frac1n F(\min(x,y)) + \frac{n-1}{n} F(x) F(y)\right) - F(x) F(y) =$

$= \frac1n F(\min(x,y)) - \frac1n F(x) f(y)$

2 $X_1, \dots, X_n \sim P$. Найти оценку коэф-та эксцесса методом подстановки. Описать схему метода бутстрепа для оценки дисперсии полученной оценки.

Коэф. эксцесса: $\gamma = \frac{1}{\sigma^4} \mathbb E (X-a)^4 - 3$; оценка коэф-та эксцесса методом подстановки:

$\hat\gamma = \frac{1}{\hat\sigma^4} \int_{\mathbb R} (x - \hat a)^4\, d\hat F_n(x) - 3 = \frac{1}{n\hat\sigma^4} \sum_{i=1}^n (X_i - \hat a)^4 - 3$ (аналогично материалу слайдов).

Схема метода бутстрепа для оценки дисперсии оценки коэф. эксцесса

1) сгенерируем B бутстрепных выборок из эмпир. распределения $\hat P_n$, т.е. $X_b^* = (X_{b1}^*, \dots, X_{bn}^*)$, $b \in \overline{1,B}$, каждая выборка – по n случайных (индексы из $U\{1,\dots,n\}$) элов из $\{X_1, \dots, X_n\}$ с возвращением;

реализация выборки

2) по каждой бутстрепной выборке посчитаем $\hat\gamma$, получим $\hat\gamma_1^* = \hat\gamma(X_1^*), \dots, \hat\gamma_B^* = \hat\gamma(X_B^*)$;

3) полученную выборку $\hat\gamma_b^*$, $b \in \overline{1,B}$ используем для асимптотики оценки дисперсии оценки γ, т.е. $\hat V_{boot} = \hat D\,\hat\gamma = \frac1B \sum_{b=1}^B \left(\hat\gamma_b^*\right)^2 - \left(\frac1B \sum_{b=1}^B \hat\gamma_b^*\right)^2$. $\hat V_{boot}$ – искомый ответ

3 Формулировки

а) оценка параметра – статистика (ф-ция от выборки), применяемая для получения приближенного значения некоторого параметра

эмпир. распределение – $\hat F_n(x) = \frac1n \sum_{i=1}^n \mathbb 1\{X_i \le x\}$, грубое приближение ф-ции распределения исходя из данной выборки $X_1, \dots, X_n$

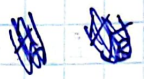
метод подстановки – метод получения оценки, при котором используется эмпир. аналоги $\hat F_n(x)$, как, например, в задаче 2, где $\mathbb E$ посчитано с помощью $d\hat F_n$

метод Монте-Карло – метод численного интегрирования, позволяющий интеграл от всех интегрируемых функций от равном. распределенной СВ: $\int_a^b f(x)\,dx = (b-a) \int f(x) p_{Unif[a,b]}\,dx = (b-a)\,\mathbb E f(X)$, где $X \sim U[a,b]$. Тогда $\int_a^b f(x)\,dx \approx (b-a) \cdot \frac1n \sum_{i=1}^n f(X_i)$. Метод хорош для $\mathbb R^n$, $n \gg 1$.

ОМП – оценка, максимизирующая ф-цию $L_X(\theta) = \prod_{i=1}^n p_\theta(x)$ – ф-ция правдоподобия ↘

3 (продолжение)

а) оценщик метода подстановки: $1/\sqrt{n}$

оценщик метода типа Калло:  $1/n$

Аппроксимация методом МК оценки метода подстановки: $1/\sqrt{B}$

4 $x_1,\dots,x_n$ – реализация выборки, $X_1^*,\dots,X_n^*$ – построенная по ней бутстрепная выборка. С какой вер-тью эл-т x_i попадет в бутстрепную выборку? Посчитать среднее # уник. эл-тов в бутстрепной выборке, если в исх. выборке все наблюдения различны.

1) $P(x_i \in X^*) = 1 - P(x_i \notin X^*) = 1 - P(x_i \neq X_1^*)\cdot\dots\cdot P(x_i \neq X_n^*) = 1 - \left(\frac{n-1}{n}\right)^n = 1 - \left(1-\frac{1}{n}\right)^n$

Интересно, что $\left(1-\frac{1}{n}\right)^n \xrightarrow[n\to\infty]{} e^{-n}$, т.е. для достаточно больших n: $P(x_i \in X^*) \approx 1 - e^{-n} \xrightarrow[n\to\infty]{} 1$

2) # уникальных $= \sum_{i=1}^{n} \mathbb{1}\{x_i \in X^*\} \Rightarrow \mathbb{E}(\#\text{уникальных}) = \sum_{i=1}^{n} P(x_i \in X^*) = n\left(1-\left(1-\frac{1}{n}\right)^n\right) \approx n\left(1-e^{-n}\right) \downarrow_{n\to\infty} n$

(так работает, если в исх. выборке все наблюдения различны) ↑ из линейности математического ожидания

Вывод: при достаточно больших размерах выборки n каждый эл-т почти наверняка попадет в бутстрепную выборку, и почти наверняка эл-ты бутстрепной выборки будут уникальными

5 $X_1,\dots,X_{(n)} \sim U[0,\theta]$, $\hat\theta = X_{(n)}$. Почему бутстреп для оценки распределения $\hat\theta$ сработает плохо?

Обозначим бутстреп. выборку $X_b^* = (X_{b1}^*,\dots,X_{bn}^*)$, $b \in \overline{1,B}$, и пусть $X_b^* := X_{b(n)}^*$. Тогда

$\hat{\mathbb{E}}^{X_{(n)}}(\hat\theta) = \frac{1}{B}\sum_{b=1}^{B} X_{b(n)}^* \leqslant \frac{1}{B}\cdot B X_{(n)} = X_{(n)}$, т.к. максимум выборки может не попасть в бутстреп выборку.

(оценка мах. правд. оценки $X_{(n)}$ для θ методом бутстрепа)

Более того, $\hat{\mathbb{E}}X_{(n)} = X_{(n)}$ лишь если эл-т $X_{(n)}$ попадет в каждую из B бутстреп выборок, что (согласно задаче 4) происходит с вероятностью $\approx 1 - e^{-n}$. То есть $P\left(\hat{\mathbb{E}}X_{(n)} = X_{(n)}\right) = \left(1-e^{-n}\right)^B \sim \left(1-e^{-n}\right)^n < 1$

Т.е. с ненулевой вероятностью оценка мат. ожидания оценки параметра будет занижена, и бутстреп сработает плохо, особенно – для малых выборок

6 $X_1, \dots, X_n$ – выборка, $X_1^*, \dots, X_n^*$ – повт. по ней бутстреп. выборка, $\overline{X^*} := \frac{1}{n}\sum_{i=1}^{n} X_i^*$. Найти $D(\overline{X^*} \mid X_1, \dots, X_n)$ – ? $D\,\overline{X^*}$ – ?

$$D(\overline{X^*} \mid X_1, \dots, X_n) \equiv \hat{D}\,\overline{X^*} = \frac{1}{B}\sum_{b=1}^{B}\left(\overline{X_b^*}\right)^2 - \left(\frac{1}{B}\sum_{b=1}^{B}\overline{X_b^*}\right)^2 \xrightarrow{B=1} \left(\overline{X^*}\right)^2 - \left(\overline{X^*}\right)^2 = 0$$

Действительно, при заданной реализации выборки случайность исчезает и дисперсия однозначно $= 0$.

$$D\,\overline{X^*} = \frac{1}{n^2}\sum_{i=1}^{n} DX_i^* = \frac{1}{n^2}\sum_{i=1}^{n} DX_i = \frac{1}{n^2}\cdot n\,DX_1 = \frac{DX_1}{n} \neq 0$$

В этом случае некоторая случайная погрешность сохраняется и дисперсия $\neq 0$

7 регрессия Y по X, оценка совместной плотности (X,Y) с ядром $q(x)\cdot q(y)$ $E(Y|X)$. Найти оценку и сравнить с оценкой в модели ядерной регрессии

(лекции: $\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} q\left(\frac{x-X_i}{h}\right)$, обобщим: $\hat{f}(x,y) = \frac{1}{nh_x h_y}\sum_{i=1}^{n} q\left(\frac{x-X_i}{h_x}\right) q\left(\frac{y-Y_i}{h_y}\right)$. Пусть $h=1$.

Тогда $E(Y|X) = \int y\, f(y|x)\,dy = \int \frac{y\, f(x,y)}{f(x)}\,dy =$

$$= \int \frac{y \sum q(x-X_i)\, q(y-Y_i)}{\sum q(x-X_i)}\,dy = \frac{\sum\left(q(x-X_i)\int y\, q(y-Y_i)\,dy\right)}{\sum q(x-X_i)} = \frac{\sum q(x-X_i)\, Y_i}{\sum q(x-X_i)},$$

что и есть оценка в модели яд. регрессии