Т. С. Бородина, М. С. Тихов (Нижний Новгород, ННГУ). **Оценки максимального правдоподобия в моделях пропорциональных и непропорциональных рисков при случайном цензурировании.**

Пусть случайные величины $X_1, X_2, \ldots, X_n$ независимы и одинаково распределены с непрерывной функцией распределения $F = F(x) = \mathbf{P}\{X \leqslant x\}$, $x \in \mathbf{R}$, с функцией времени жизни $S_X = 1 - F$, а последовательность независимых случайных величин $Y_1, Y_2, \ldots, Y_n$ с непрерывной функцией распределения G и $S_Y = 1 - G$ независима от X и цензурирует их справа: на каждом шаге мы наблюдаем минимум из двух величин $Z_j = X_j \wedge Y_j$ и бинарную случайную величину $\delta_j = I\{X_j \leqslant Y_j\}$, $j = 1, 2, \ldots, n$, где $I(A)$ обозначает индикатор события A. Если H есть функция распределения величины Z, где $(Z, \delta) = (Z_j, \delta_j)$, то $S_Z = 1 - H = S_X\, S_Y$. Упорядочим пары (Z_j, δ_j) по первой компоненте $(Z_n^{(i)}, \delta_n^{[i]})$ т. е. пусть $Z_n^{(1)} < Z_n^{(2)} < \cdots < Z_n^{(n)}$ есть вариационный ряд, тогда $\delta_n^{[i]}$ есть i-я индуцированная порядковая статистика: если $Z_n^{(i)} = Z_j$, то $\delta_n^{[i]} = \delta_j$. Для функции распределения $F(x)$ рассмотрим множительную оценку Каплана–Мейера (КМ-оценку):

$$\widehat{F}_n(x) = 1 - \prod_{j=1}^{n} \left(1 - \frac{\delta_n^{[j]}}{n - j + 1}\right)^{I\{Z_n^{(j)} \leqslant x\}} = 1 - \prod_{j=1}^{n} \left(1 - \frac{\delta_n^{[j]} I\{Z_n^{(j)} \leqslant x\}}{n - j + 1}\right),$$

асимптотическая дисперсия которой равна

$$d(x) = \int_{-\infty}^{x} \frac{d\widetilde{H}(u)}{S_Z^2(u)} = \int_{-\infty}^{x} \frac{dF(u)}{S_X^2(u) S_Y(u)},$$

где $\widetilde{H}(x) = \mathbf{P}\{Z \leqslant x, \delta = 1\} = \int_{-\infty}^{x} S_Y(u)\, S_X(u)\, du$ и $H_n(x) = n^{-1} \sum_{j=1}^{n} I\{Z_j \leqslant x\}$ есть эмпирическая функция распределения величины Z.

Если $S_Y(x) = S_X^{\beta}(x)$ для некоторого $0 < \beta < \infty$ и любого $x > 0$, то соответствующая схема называется *моделью пропорциональных рисков* (ПР). В модели пропорциональных рисков $\mathbf{P}\{I = 1\} = \mathbf{P}\{X \leqslant Y\} = 1/(1+\beta)$ и $S_Z(x) = S_X^{1+\beta}(x)$. Здесь для оценивания функции $S_X(x)$ обычно используют оценку Абдушукурова–Ченга–Лина (ACL-оценку) $\widetilde{S}_X(x) = (\widehat{S}_Z(x))^{\overline{\delta}}$, где $\widehat{S}_Z(x) = n^{-1} \sum_{j=1}^{n} I\{Z_j > x\}$ есть эмпирическая функция времени жизни величины Z и $\overline{\delta} = n^{-1} \sum_{j=1}^{n} \delta_j$.

В работе [1] в модели ПР была подсчитана относительная асимптотическая эффективность $\varepsilon(x; \beta) = (1+\beta)^{-1} + \beta(1+\beta) x^{1+\beta} \ln^2 x / (1 - x^{1+\beta})$, $0 < x < 1$, КМ-оценки по сравнению с ACL-оценкой.

Таблица. Значение относительной эффективности КМ-оценки и ACL-оценки для значений параметра $\beta = 0,2,\ 2,\ 10,\ 20$

	Значение переменной x								
β	0,1	0,3	0,5	0,7	0,8	0,85	0,9	0,95	0,99
0,2	0,919	0,941	0,922	0,890	0,872	0,863	0,853	0,843	0,835
2	0,365	0,575	0,745	0,732	0,647	0,586	0,513	0,428	0,353
10	0,091	0,091	0,117	0,373	0,606	0,675	0,649	0,473	0,186
20	0,048	0,048	0,048	0,077	0,242	0,426	0,620	0,618	0,228

Из таблицы видно, что если вероятность $\gamma = \mathbf{P}\{X \leqslant Y\}$ мала, то КМ-оценка малоэффективна по сравнению с ACL-оценкой.

В [2] для модели пропорциональных рисков была предложена оценка максимального правдоподобия (HLS-оценка), где в качестве правдоподобия рассматривалась следующая функция:

$$\begin{aligned}\mathcal{L} &= \prod_{j=1}^{n} \mathbf{P}\{Z_i = z_i, \delta_i = d_i\} \\ &= \prod_{j=1}^{n} \Big\{[S_X(z_i^-) - S_X(z_i)]^{d_i}\Big[S_X^{\beta}(z_i^-)\Big]^{d_i}\Big[S_X^{\beta}(z_i^-) - S_X^{\beta}(z_i)\Big]^{1-d_i}[S_X(z_i)]^{1-d_i}\Big\}. \quad (1)\end{aligned}$$

В [2] был рассмотрен также следующий пример, показывающий, что ACL-оценка не является МП-оценкой на множестве всех непрерывных и дискретных распределений. Именно, для выборки (4,0), (6,1), (8,0), (9,1) функция правдоподобия равна

$$\begin{aligned}\mathcal{L} = {} & [S_X^{\beta}(4^-) - S_X^{\beta}(4)][S_X(4)]\,[S_X(6^-) - S_X(6)][S_X^{\beta}(6^-)] \\ & \times [S_X^{\beta}(8^-) - S_X^{\beta}(8)][S_X(8)]\,[S_X(9^-) - S_X(9)][S_X^{\beta}(9^-)]. \quad (2)\end{aligned}$$

Заметим, что здесь $\overline{\delta} = 0,5$ и ACL-оценка $\widetilde{S}_X(x)$ равна 1, $\sqrt{3}/2 \approx 0,866$, $1/\sqrt{2} \approx 0,707$, 0,5 и 0, соответственно на интервалах $[0,4)$, $[4,6)$, $[6,8)$, $[8,9)$ и $[9,\infty)$. Подставляя $\widetilde{S}_X(x)$ вместо $S_X(x)$ и $\widehat{\beta} = (1-\overline{\delta})/\overline{\delta} = 1$ вместо β в формулу (2), мы получаем $\mathcal{L} = [1-0,866][0,866-0,707][0,707-0,5]\,(0,866)^2\,(0,5)^3 = 0,00041$.

Рассмотрим теперь функцию времени жизни S_X (которая является МП-оценкой), равную 1, 0,924, 0,0808, 0,677, 0, соответственно на указанных интервалах. Тогда для выбранной так функции S_X и $\beta = 2,5$ получаем следующее значение правдоподобия:

$$\mathcal{L} = [1-(0,924)^{2,5}][0,924-0,808][(0,808)^{2,5}-(0,677)^{2,5}](0,924)^{3,5}(0,677)^{4,5} = 0,00057,$$

что в 1,39 раза больше правдоподобия для ACL-оценки (см. [2]).

Однако, если взять функции S_X и S_Y, которые равны на указанных интервалах значениям 1, 0,960, 0,846, 0,712, 0 и 1, $0,669 = 0,96^{9,832}$, $0,669 = 0,846^{2,4}$, $0,337 = 0,712^{3,2}$, 0, соответственно, то

$$\mathcal{L} = [1-(0,96)^{9,832}]\,0,96[0,96-0,846]\,0,96^{9,832}[(0,846)^{2,4}-(0,712)^{3,2}](0,712)^{5,2} = 0,00138,$$

что в 2,4 раза больше предыдущего значения. Последняя оценка получена максимизацией следующей функции правдоподобия (ФП):

$$L = \prod_{j=1}^{n}[S_X(z_{i-1}) - S_X(z_i)]^{d_i}[S_X^{\beta_{i-1}}(z_{i-1})]^{d_i}[S_X^{\beta_{i-1}}(z_{i-1}) - S_X^{\beta_i}(z_i)]^{1-d_i}[S_X(z_i)]^{1-d_i}$$

при ограничениях $1 \geqslant S_X^{\beta_{i-1}}(z_{i-1}) \geqslant S_X^{\beta}(z_i) \geqslant 0$, $i = 1, 2, \ldots, n$.

Мы предлагаем применять данную ФП для нахождения оценок функции S_X вместо КМ-оценок при малых значениях вероятности γ. Кроме того, примеры показывают, что при умеренных значениях вероятности γ и достаточно больших объемах выборки полученные значения оценок в модели ПР близки к ACL-оценкам.

СПИСОК ЛИТЕРАТУРЫ

1. *Hollander M.*, *Proschan F.*, *Sconing J.* Efficiency loss with the Kaplan–Meier estimator. FSU Statistics Report M707, 1985. http://stat.fsu.edu/techreports/M707a.
2. *Hollander M.*, *Laird G.*, *Song K. S.* Maximum likelihood estimation in the proportional Hazars model of random censorship. — Statistics, 2001, v. 35, p. 245–258.