Оценка параметра степенной модели Лемана-Кокса методом минимизации функционалов типа Колмогорова-Смирнова и Сэвиджа

07, июль 2012

DOI: 10.7463/0712.0410885

Тимонин В. И., Ермолаева М. А.

УДК. 519.22

Россия, МГТУ им. Н.Э. Баумана

timoninmgtu52@mail.ru

ermolaevama@yandex.ru

1. Введение. Для оценки результатов различных экспериментов в науке и технике широко используется непараметрическая степенная модель Кокса-Лемана [1]. Суть модели заключается в том, что функции распределения различных выборок связаны степенной зависимостью, причем показатель степени есть функция от различных факторов эксперимента. Разработаны методы оценки параметров регрессионных зависимостей этого показателя. Вместе с тем отсутствуют методы проверки справедливости самой степенной модели Кокса-Лемана, что часто ставит под сомнение адекватность получаемых статистических выводов. В настоящей работе приводятся два возможных способа проверки адекватности модели Кокса-Лемана. Методом Монте-Карло сравнивается качество оценок, полученных предложенными способами.

Задача формулируется как задача теории надежности, однако, при соответствующем изменении терминологии, результаты могут быть использованы для анализа результатов биологических, социологических и других экспериментов.

2. Постановка задачи. Пусть проводятся испытания двух выборок изделий в различных режимах ε_1 и ε_2 . Обозначим $\bar{\xi} = (\xi_1, \dots, \xi_m)$, $\bar{\eta} = (\eta_1, \dots, \eta_n)$ – моменты отказов в режимах ε_1 и ε_2 соответственно, причем $\xi_i \sim F(x)$, $\eta_j \sim G(x)$, $i = \overline{1, m}$, $j = \overline{1, n}$. Здесь $F(x)$, $G(x)$ – функции распределения наработок до отказа в этих режимах. В дальнейшем предполагается, что для $F(x)$, $G(x)$ выполняется либо соотношение

$$M_1 : F(x) = [G(x)]^k, \quad (1)$$

либо

$$M_2 : 1 - F(x) = [1 - G(x)]^k. \quad (2)$$

Показатель степени $k > 0$ неизвестен и подлежит оценке.

Обе модели относятся к моделям Кокса – Лемана [1]. В работе анализируется случай, когда справедлива первая модель. Вторая модель сводится к первой простым переходом от случайных величин ξ_i, η_j к величинам $\nu_i = -\xi_i, \tau_i = -\eta_j$.

Введем некоторые обозначения. Пусть $\gamma_1 < \gamma_2 < \dots < \gamma_{m+n}$ – объединенный вариационный ряд из элементов $\bar{\xi}, \bar{\eta}$. Определим 2 вектора: $\bar{Z} = (z_1, z_2, \dots, z_{m+n})$ и $\bar{V} = (V_1, V_2, \dots, V_{m+n})$ по правилу:

$$z_i = \begin{cases} 1, & \gamma_i \in \bar{\xi}, \\ 0, & \gamma_i \in \bar{\eta}, \end{cases} \quad i = \overline{1, m+n}, \quad V_i = \sum_{j=1}^i z_j.$$

Очевидно, что вектор $\bar{Z} = (z_1, z_2, \dots, z_{m+n})$ состоит из m единиц и n нулей. Их общее количество равно числу C_{m+n}^m . Для величин V_i справедливо неравенство $\max(0, i - n) \leq V_i \leq \min(i, m), i = \overline{1, m+n}$.

В дальнейшем без ограничения общности будем считать, что $k > 1$.

Лемма 1. (Сэвидж, [2]). Пусть справедливо (1). Тогда вероятность $p(\bar{Z})$ имеет вид

$$p(\bar{Z}) = \frac{k^m m! n!}{\prod_{i=1}^{m+n} (i + V_i (k - 1))}. \quad (3)$$

Обычно параметр k оценивают методом максимального правдоподобия, то есть максимизируют по k вероятность (3). Элементарно можно получить уравнение, которому удовлетворяет оценка $\hat{k}$:

$$\frac{m}{\hat{k}} = \sum_{i=1}^{m+n} \frac{V_i}{i + V_i (\hat{k} - 1)}. \quad (4)$$

Дело, однако, заключается в том, что при небольших объемах выборок с ненулевой вероятностью решение (4) – отрицательно, что невозможно. Другими словами, оценка максимального правдоподобия дает качественно неверные результаты.

3. Описание предлагаемых методов. Рассмотрим две оценки параметра k , основанные на минимизации двух различных функционалов. Эти оценки будут лишены указанного выше недостатка. Кроме того, они позволяют проверить адекватность самой степенной модели Кокса-Лемана.

В работах [3,4] были предложены два критерия проверки статистической гипотезы вида

$$H_0 : F(x) = [G(x)]^{k_0}, \quad (5)$$

где k_0 – известное фиксированное число.

Первый из них является критерием типа Колмогорова-Смирнова со статистикой [3]

$$S(k_0) = \max \frac{\left| \hat{F}_m(t) - (\hat{G}_n(t))^{k_0} \right|}{1 + \mu (\hat{Z}(t))^q (1 - (\hat{Z}(t))^{1-q})}, \quad (6)$$

где $\hat{F}(x), \hat{G}(x)$ – эмпирические функции распределения выборок $\bar{\xi}, \bar{\eta}$;

$\mu = k_0^2 \rho, \rho = \frac{m}{n}, q = \frac{k_0 - 1}{k_0}; \hat{Z}(t) = \frac{1}{m+n} \left(m \hat{F}_m(t) + n (\hat{G}_n(t))^{k_0} \right)$ – объединенная

оценка функции распределения.

Второй критерий является локально наиболее мощным критерием проверки (5) против альтернатив вида $H_1 : F(x) = [G(x)]^k, k \neq k_0$ [4]. Его статистика имеет вид

$$T_{m,n}(k_0) = \sum_{i=1}^{m+n} \frac{V_i}{i + V_i(k_0 - 1)}. \quad (7)$$

Если параметр k неизвестен, то в качестве оценок $\hat{k}$ предлагается брать величины, минимизирующие $S(k)$ и $T_{m,n}^*(k) = \left| \frac{T_{m,n}(k) - ET_{m,n}(k)}{\sqrt{DT_{m,n}(k)}} \right|$. Другими словами,

рассматриваются оценки

$$\widehat{k}_1 = \arg \min_k S(k), \quad \widehat{k}_2 = \arg \min_k T_{m,n}^*(k).$$

Здесь $ET_{m,n}(k)$, $DT_{m,n}(k)$ – соответственно математическое ожидание и дисперсия статистики $T_{m,n}(k)$, которые требуют определения. Дело в том, что в работе [4], где была введена статистика $T_{m,n}(k)$, были разработаны лишь итерационные методы вычисления точных вероятностей $P(T_{m,n}(k)=u)$. Однако они становятся бесполезными уже при $m, n > 15$, так как множество значений $T_{m,n}(k)$ растет пропорционально C_{m+n}^n .

Утверждение. Среднее и дисперсия $T_{m,n}(k)$ имеют вид

$$ET_{m,n} = \frac{m}{k}, \quad DT_{m,n} = \left(\frac{mn}{k} \right) \int_0^1 \int_0^1 (1-u(1-v))^{n-1} (1-u^k(1-v^k))^{m-1} u^k v^k du dv. \quad (8)$$

Кроме того, дисперсия $DT_{m,n}$ может быть вычислена через рекуррентное соотношение с заданными начальными и граничными условиями

$$DT_{uw}(k) = \frac{ku}{ku+w} DT_{u-1,w}(k) + \frac{w}{ku+w} DT_{u,w-1}(k) + \frac{uw}{k(uk+w)^2}, \quad (9)$$

$$DT_{u,0}(k) = 0, \quad DT_{0,w}(k) = 0, \quad u = 1, \dots, m, \quad w = 1, \dots, n.$$

Доказательство. Дифференцируя по k обе части тождества $1 = \frac{\sum_{\bar{z}} k^m m! n!}{\prod_{i=1}^{m+n} (i + V_i(k-1))}$, имеем

$$0 = \sum_{\bar{z}} p(\bar{z}) \cdot \left(\frac{m}{k} - \sum_{i=1}^{m+n} \frac{V_i}{i + V_i(k-1)} \right). \quad \text{Отсюда следует выражение для } ET_{m,n}.$$

Дифференцируя второй раз, аналогичным образом получим выражение

$$0 = \sum_{\bar{z}} p(\bar{z}) \cdot \left(\frac{m}{k} - \sum_{i=1}^{m+n} \frac{V_i}{i + V_i(k-1)} \right)^2 + \sum_{\bar{z}} p(\bar{z}) \cdot \left(-\frac{m}{k^2} + \sum_{\bar{z}} \frac{V_i^2}{(i + V_i(k-1))^2} \right).$$

$$\text{Отсюда следует, что } DT_{m,n} = \frac{m}{k^2} - E \left(\sum_{\bar{z}} \frac{V_i^2}{(i + V_i(k-1))^2} \right).$$

Прямым вычислением математического ожидания суммы в правой части нетрудно получить интегральное представление для $DT_{m,n}$. Рекуррентное соотношение (9) является следствием интегрального представления. <

Адекватность модели проверяется следующим образом. Пусть $\hat{k}_1$ и $\hat{k}_2$ - найденные оценки, полученные минимизацией $S(k)$, $T_{m,n}^*(k)$ соответственно. Тогда можно проверить гипотезу (5) с помощью статистик $S(k)$ - при $k_0 = \hat{k}_1$, и $T_{m,n}^*(k)$ при $k_0 = \hat{k}_2$. Если гипотеза (5) отвергается хотя бы одним из предложенных критериев, то модель Кокса-Лемана - неадекватна.

4. Сравнение оценок методом Монте-Карло. Было проведено обширное статистическое моделирование для сравнения свойств оценок, полученных минимизацией этих двух статистик – $S(k)$, $T_{m,n}^*(k)$. В качестве примера приведем результаты моделирования, когда истинное значение $k=2$, функция распределения $G(x)=1-e^{-x}$, $x>0$.

Моделировались выборки $\bar{\xi}, \bar{\eta}$ одинакового объема, чьи функции распределения удовлетворяют (1) при $k=2$. По этим двум выборкам вычислялись оценки $\hat{k}_1, \hat{k}_2$ параметра k . Процедура повторялась 500 раз. По полученным реализациям строились гистограммы распределения оценок для обоих методов оценки, а также вычислялись средние значения оценок коэффициентов и значения дисперсий оценок.

$$\bar{k}_1 = \frac{1}{500} \sum_{i=1}^{500} k_{1i}, \bar{k}_2 = \frac{1}{500} \sum_{i=1}^{500} k_{2i}, S_1^2 = \frac{1}{499} \sum_{i=1}^{500} (k_{1i} - \bar{k}_1)^2, S_2^2 = \frac{1}{499} \sum_{i=1}^{500} (k_{2i} - \bar{k}_2)^2.$$

На рисунках 1, 2 показаны гистограммы оценок $\hat{k}$ для объемов выборок $m=n=30$. Для них $\bar{k}_1 = 2,50$; $\bar{k}_2 = 2,08$; $S_1^2 = 1,66$; $S_2^2 = 0,62$.

На рисунках 3, 4 показаны гистограммы оценок $\hat{k}$ для объемов выборок $m=n=50$. Для них $\bar{k}_1 = 2,24$; $\bar{k}_2 = 2,07$; $S_1^2 = 1,06$; $S_2^2 = 0,50$.

Анализ результатов моделирования показал безусловное преимущество оценок, полученных минимизацией статистики $T_{m,n}^*(k)$ перед оценками, полученными минимизацией статистики $S(k)$ в том случае, если соотношение (1) справедливо при каком-то значении k . Смещение и разброс оценки k_2 всегда меньше смещения и разброса

оценки k_1 . Вместе с тем при небольших объемах выборок, обе эти оценки имеют значительное систематическое смещение. По этой причине следует пользоваться поправочными коэффициентами, которые также можно рассчитать методом Монте-Карло.

Рис. 1. Гистограмма $\hat{k}_1$ для $k_0=2, m=n=30$.

Рис. 2. Гистограмма $\hat{k}_2$ для $k_0=2, m=n=30$.

Рис.3. Гистограмма $\hat{k}_1$ для $k_0=2, m=n=50$.

Рис.4. Гистограмма $\hat{k}_2$ для $k_0=2, m=n=50$.

Заключение. Предложены два метода оценки параметра степенной модели Кокса-Лемана, позволяющие проверять адекватность этой модели. Показано, что если модель верна, то одна из оценок всегда «лучше» другой. Вместе с тем, авторы считают, что следует всегда использовать оба метода, так как использование статистики $S(k)$ необходимо для проверки адекватности модели, а статистики $T_{m,n}^*(k)$ - для оценки параметра модели.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Кокс Д., Оукс Д. Анализ данных типа времени жизни. М.: Финансы и статистика. 1988. 191с.
2. Savage R. Contributions to the Theory of Rank Order Statistics – the Two – Sample Case // Ann. Math. Stat. 1956. V.27. № 3. P. 590 –615.

3. Тимонин В.И. О предельном распределении статистики одного непараметрического критерия. // Теория вероятностей и ее применение. 1987. Т. 32. №4. С. 790-792.

4. Тимонин В.И. Об одном локально наиболее мощном ранговом критерии// Вероятностные процессы и их приложения. М.: МИЭМ. 1983. С. 74-80.

**Estimation of parameter in Lehmann-Cox power-law model
by minimizing Kolmogorov-Smirnov and Savage functionals**

07, July 2012

DOI: 10.7463/0712.0410885

Timonin V.I., Ermolaeva M.A.

Russia, Bauman Moscow State Technical University

timoninmgtu52@mail.ruermolaevama@yandex.ru

In this paper we consider two estimations of the unknown parameter in the Lehmann-Cox power-law model. These estimates are obtained by minimization of two different functionals: Kolmogorov-Smirnov and Savage ones. The type of these functionals was obtained by using non-parametric statistics proposed by authors in their previous articles. We demonstrated the advantage of one statistics over the other by use of the methods of statistical modeling when the Lehmann-Cox power-law model is valid. It was proposed using a correction coefficient calculated with Monte Carlo method to eliminate bias of estimates for small sample sizes.

Publications with keywords: [Monte Carlo method](#), [statistical estimation](#), [Cox-Lehmann power-law model](#), [Kolmogorov-Smirnov and Savage type statistics](#)

Publications with words: [Monte Carlo method](#), [statistical estimation](#), [Cox-Lehmann power-law model](#), [Kolmogorov-Smirnov and Savage type statistics](#)

References

1. Cox D.R., Oakes D. *Analysis of Survival Data*. London, Chapman and Hall, 1984. 201p. (*Monographs on Statistics and Applied Probability*, Vol. 21). (Russ. ed.: Koks D., Ouks D. *Analiz dannykh tipa vremeni zhizni*. Moscow, Finansy i statistika, 1988. 191 p.).
2. Savage R. Contributions to the theory of rank order statistics – the two-sample case. *Annals of Mathematical Statistics*, 1956, vol. 27, no. 3, pp. 590–615.
3. Timonin V.I. O predel'nom raspredelenii statistiki odnogo neparametricheskogo kriteriia [On the limit distribution of a nonparametric test statistics]. *Teoriia veroiatnostei i ee primeneniia*, 1987, vol. 32, no. 4, pp. 790-792.
4. Timonin V.I. Ob odnom lokal'no naibolee moshchnom rangovom kriterii [On a locally most powerful rank test]. *Veroiatnostnye protsessy i ikh prilozheniia* [Stochastic processes and their applications]. Moscow, MIEM Publ., 1983, pp. 74-80.