УДК 519.2

В.А. Демин, Е.В. Чимитова

ВЫБОР ОПТИМАЛЬНОГО ПАРАМЕТРА СГЛАЖИВАНИЯ ДЛЯ НЕПАРАМЕТРИЧЕСКОЙ ОЦЕНКИ РЕГРЕССИОННОЙ МОДЕЛИ НАДЕЖНОСТИ

Рассматривается один из популярных подходов к непараметрическому оцениванию регрессионной модели надежности, предложенный Бераном. Оценка Берана позволяет оценить функцию надёжности регрессионной модели. Показаны результаты исследования зависимости точности оценки Берана от различных факторов, и предлагается универсальный метод для подбора параметра сглаживания.

Ключевые слова: *функция надёжности*, *регрессионная модель*, *непараметрическая оценка Берана*, *параметр сглаживания.*

В задачах статистического анализа данных типа времени жизни, например, времени безотказной работы технических изделий в теории надежности, времени жизни пациентов в анализе выживаемости, типичной задачей является исследование зависимости функции надежности (выживаемости) наблюдаемой случайной величины от объясняющих переменных. В теории надежности в качестве объясняющих переменных обычно выступают воздействия (нагрузки), оказывающие влияние на продолжительность безотказной работы, такие, как температура, давление, напряжение, механические нагрузки и другие. Для описания зависимости функции надежности от объясняющих переменных, или, как их принято называть в анализе данных типа времени жизни, – ковариат, используют различные параметрические модели, наиболее популярными из которых являются модель ускоренных испытаний и модель пропорциональных интенсивностей. Однако построение любой параметрической модели требует выполнения определенных предположений. На практике же априорные предположения о функциональной зависимости функции надежности от ковариат обычно отсутствуют. В такой ситуации целесообразно применение непараметрических методов, которые позволяют не только оценить функцию надежности при различных значениях ковариаты, но и могут использоваться для построения статистического критерия согласия с некоторой параметрической моделью надежности.

Одним из наиболее популярных подходов к непараметрическому оцениванию регрессионной модели надежности является оценка, предложенная Бераном [1]. Исследования статистических свойств данной оценки для случайного плана эксперимента, когда значение ковариаты не фиксировано, представлены в [2–5]. В [6] исследованы свойства оценки для неслучайного плана, когда значения ковариат определяются заранее.

В литературе, посвященной непараметрическим оценкам, широко представлены различные методы выбора оптимального параметра сглаживания для случая ядерного оценивания функции плотности распределения, например в [7]. В [8] описываются основные подходы к выбору параметра сглаживания при построении непараметрических оценок регрессионных моделей, для которых имеются

значения отклика и факторов, от которых он зависит. К сожалению, в известной авторам литературе проблема выбора оптимального параметра сглаживания для оценки Берана не рассматривается. Тогда как от значения этого параметра существенно зависит качество получаемых оценок. В данной работе предлагается алгоритм выбора оптимального параметра сглаживания при построении непараметрической оценки Берана для регрессионных моделей надежности.

1. Непараметрическая оценка Берана

Обозначим через T_x время безотказной работы исследуемых технических изделий, которое зависит от скалярной ковариаты x. Функция надежности определяется соотношением

$$S(t \mid x) = P(T_x \geq t) = 1 - F(t \mid x), \quad (1)$$

где $F(t \mid x)$ – условная функция распределения случайной величины T_x.

Главной особенностью данных типа времени жизни является наличие цензурированных справа наблюдений, которые можно представить в виде

$$(Y_1, x_1, \delta_1), (Y_2, x_2, \delta_2), ..., (Y_n, x_n, \delta_n),$$

где n – объем выборки, x_i – значение ковариаты для *i*-го объекта, Y_i – время наработки до момента отказа или цензурирования, δ_i – индикатор цензурирования, который принимает значение 1, если наблюдение полное, и 0, если цензурированное.

Оценка Берана имеет следующий вид [1]:

$$\tilde{S}_{h_n}(t \mid x) = \prod_{Y_{(i)} \leq t} \left\{ 1 - \frac{W_n^i(x; h_n)}{1 - \sum_{j=1}^{i-1} W_n^j(x; h_n)} \right\}^{\delta_i}, \quad (2)$$

где x – значение ковариаты, для которой оценивается функция надёжности; $W_n^i(x; h_n), i = 1, \ldots, n$ – веса Надарая – Ватсона, которые можно вычислить по формуле [5]

$$W_n^i(x; h_n) = K\left(\frac{x - x_i}{h_n}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{x - x_j}{h_n}\right), \quad (3)$$

где $K\left(\frac{x - x_i}{h_n}\right)$ – ядерная функция, удовлетворяющая условиям регулярности: $K(y) = K(-y)$, $0 \leq K(y) < \infty$, $\int_{-\infty}^{\infty} K(y)dy = 1$, $h_n > 0$ – параметр сглаживания такой, что $\lim\limits_{n \to \infty} h_n = 0$, $\lim\limits_{n \to \infty} nh_n = \infty$.

Следует отметить, что при значениях весов Надарая – Ватсона $W_n^i(x; h_n) = n^{-1}$ оценка Берана сводится к оценке Каплана – Мейера [5].

С использованием методов компьютерного моделирования и исследования статистических закономерностей нами подтверждены свойства оценки Берана: с увеличением объема выборки точность получаемых оценок растет. В результате

проведенных исследований показано, что точность оценок существенно зависит от значения параметра сглаживания и практически не зависит от вида ядерной функции. При этом выбор параметра сглаживания должен осуществляться, в первую очередь, на основании разницы предполагаемых функций надежности, соответствующих разным значениям ковариаты, тогда как влияние объема выборки и плана эксперимента оказывается несущественным при выборе параметра сглаживания. Проиллюстрируем данный результат на примере.

Рассмотрим следующий план эксперимента: все испытуемые объекты разделены на 10 групп по $n_i = 15$, $i = 1,...,10$, объектов. Каждая группа объектов тестируется при воздействии x равном 0, 0,11, 0,22, 0,33, 0,44, 0,56, 0,67, 0,78, 0,89, 1, соответственно. На основе данного плана эксперимента смоделируем 2 выборки в соответствии с моделью ускоренных испытаний вида

$$S(t \mid x) = \frac{1}{2} - \frac{1}{2\sqrt{\pi}} \cdot \Gamma\left(\frac{1}{2}\ln^2\left(\frac{t}{r(x;\beta)}\right), \frac{1}{2}\right), \qquad (4)$$

где $\Gamma(\cdot, q)$ – неполная гамма-функция, функция от воздействий $r(x,\beta) = e^{\beta x}$. Первая выборка моделировалась при значении регрессионного параметра равном $\beta = 2$, вторая выборка – при $\beta = 5$.

На рис. 1 представлены оценки Берана для функции надежности при $x = 0$ и $x = 0,56$, полученные по первой выборке. Для сравнения также приведены соответствующие истинные функции надежности (4). Оценки Берана, построенные по второй выборке, и соответствующие истинные функции надежности изображены на рис. 2. Параметр сглаживания h_n при построении оценки Берана в обоих случаях взят равным 0,5.

Как видно из рис. 1, оценки Берана достаточно близки к соответствующим функциям надежности, однако, как показано на рис. 2, при таком же плане эксперимента наблюдается существенное отклонение оценок Берана от истинных функций в случае, когда влияние воздействия x более значимо (при большем значении регрессионного параметра).

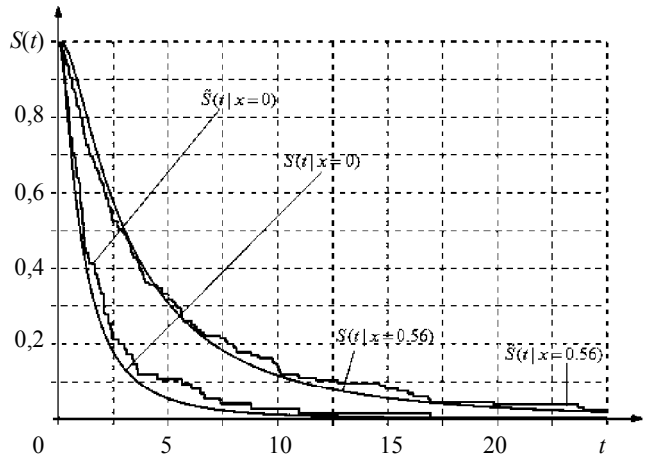

Рис. 1. Функции надёжности и оценки Берана, $h_n = 0,5$, $\beta = 2$

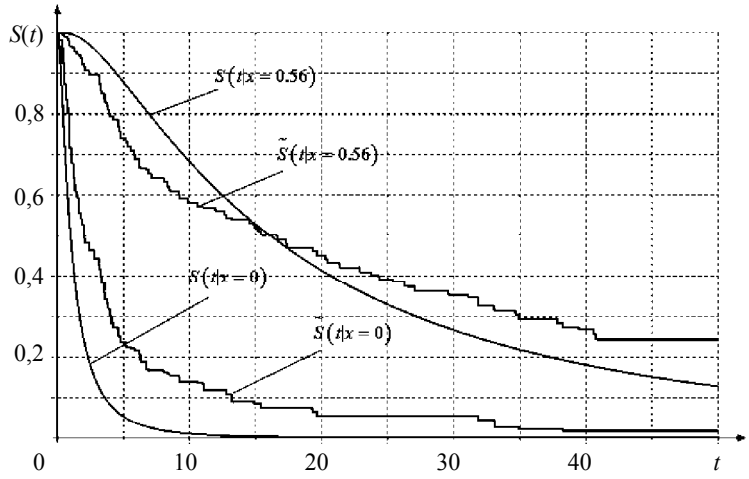

Рис. 2. Функции надёжности и оценки Берана, $h_n = 0{,}5$, $\beta = 5$

2. Выбор оптимального параметра сглаживания

Параметр сглаживания определяет, какие наблюдения будут участвовать в построении оценки Берана, а какие нет: чем больше параметр сглаживания, тем больше наблюдений будет участвовать в построении оценки. Таким образом, меняя параметр сглаживания, мы можем отсеивать «лишние» наблюдения.

В данной работе предлагается алгоритм выбора оптимального параметра сглаживания h_n для оценки Берана, основанный на минимизации среднеквадратического отклонения времен отказов $Y_1, Y_2, ..., Y_n$ от непараметрической оценки обратной функции надежности $S_x^{-1}(p)$. Обозначим обратную функцию надежности через $g(p \mid x)$. Тогда модель (1) можно переписать в виде

$$T_x = g(p \mid x) + \varepsilon , \tag{5}$$

где $p \in (0,1)$, ε – ошибка наблюдения, которая в общем случае может зависеть от p и x .

Ядерная оценка для модели (5) имеет вид

$$\hat{g}(\hat{p}_i \mid x_i) = \frac{1}{n}\sum_{j=1}^{n} W_n^j(\hat{p}_i) \cdot Y_j ,$$

где W_n^j – это уже известные нам веса Надарая – Ватсона, которые в данном случае вычисляются следующим образом:

$$W_n^j(\hat{p}_i) = K\left(\frac{\hat{p}_i - \hat{p}_j}{b_n}\right) \Big/ \sum\nolimits_{k=1}^{n} K\left(\frac{\hat{p}_i - \hat{p}_k}{b_n}\right).$$

Вероятности $\hat{p}_i$ вычисляются с использованием оценки Берана по формуле (2):

$$\hat{p}_i = \tilde{S}_{h_n}(Y_i \mid x_i) ,$$

параметр сглаживания b_n можно рассчитать, например, по формуле [7]

$$b_n = 1{,}059 \cdot \hat{\sigma} \cdot n^{-1/5},\ \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}_i - \frac{1}{n}\sum_{j=1}^{n}\hat{p}_j\right)^2}\ .$$

Таким образом, получить оптимальный параметр сглаживания можно в результате минимизации:

$$h_n^{\text{opt}} = \arg\min_{h_n} \frac{1}{n}\sum_{i=1}^{n}\delta_i \cdot \left(\hat{g}\left(\hat{p}_i \mid x_i\right) - Y_i\right)^2 .$$

Исследуем точность получаемых оценок с использованием предложенного алгоритма выбора оптимального параметра сглаживания. В качестве оценки точности получаемых оценок будем рассчитывать среднее отклонение вида

$$D_{avg} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{n}\left|\frac{\delta_i}{n}\cdot\left(\tilde{S}_{h_h}\left(Y_j \mid x_j\right) - S\left(Y_j \mid x_j\right)\right)\right|, \qquad (6)$$

где N – число моделируемых выборок, соответствующих модели $S(t \mid x)$.

В таблице приведены средние отклонения (6) в случае моделирования $N = 2000$ выборок в соответствии с моделью ускоренных испытаний (4) при различных значениях регрессионного параметра и объемах выборок n. Значения ковариаты в моделируемых выборках генерировались из равномерного на отрезке [0, 1] распределения.

Зависимость точности оценки Берана от параметра сглаживания

β	Объём выборки(n)	$h_n = 0{,}1$	$h_n = 0{,}5$	$h_n = 0{,}9$	h_n^{opt}
4,5	50	0,083	0,066	0,087	0,063
	75	0,067	0,063	0,085	0,053
	100	0,058	0,061	0,083	0,048
7	50	0,061	0,096	0,131	0,057
	75	0,048	0,095	0,130	0,047
	100	0,043	0,094	0,129	0,041

Из таблицы видно, что применение алгоритма выбора оптимального параметра сглаживания позволяет получать более точные оценки Берана: значение отклонения (6) в случае заданных значений параметра сглаживания h_n больше, чем в случае использования оптимального параметра h_n^{opt} при всех рассмотренных объемах выборок и значениях регрессионного параметра.

Вернемся к рассмотренному выше примеру построения оценки Берана для двух выборок, смоделированных в соответствии с моделью ускоренных испытаний (4). Построим по ним оценки Берана с использованием оптимального параметра сглаживания h_n^{opt}. На рис. 3 и 4 представлены теоретические функции надежности и оценки Берана с использованием оптимального параметра сглаживания, полученные по тем же выборкам, для которых на рис. 1 и 2 соответственно представлены оценки Берана с заданным значением параметра сглаживания.

Как видно из рис. 3, оценки Берана достаточно близки к соответствующим функциям надежности, впрочем, как и на рис. 1. Однако на рис. 4 оценки Берана значительно ближе к соответствующим теоретическим функциям надежности,

чем на рис. 2, что свидетельствует о том, что применение алгоритма выбора оптимального параметра сглаживания позволяет существенно повысить точность оценок Берана по сравнению с подходами к выбору параметра сглаживания, основанными на объеме выборки и особенностях плана эксперимента.

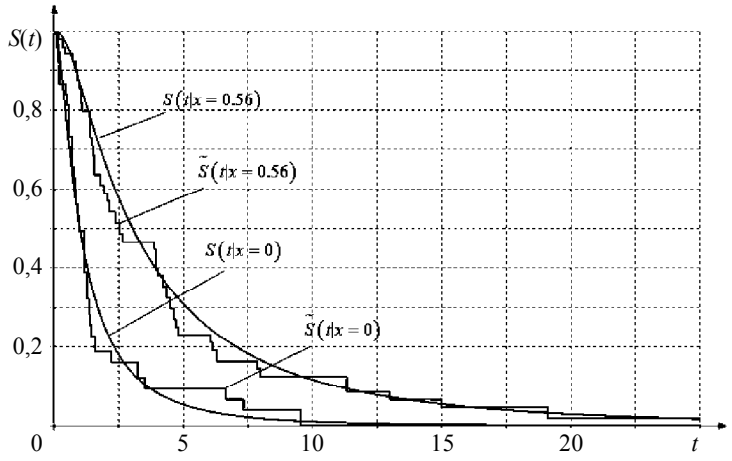

Рис. 3. Функция надёжности и оценки Берана с параметром h_n^{opt}, $\beta = 2$

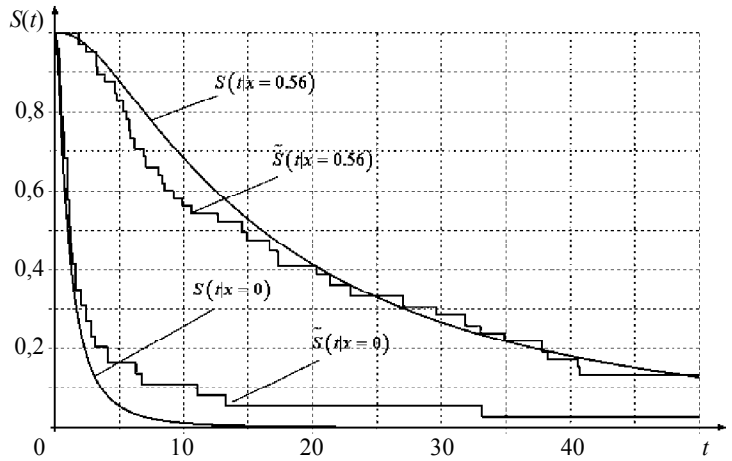

Рис. 4. Функция надёжности и оценки Берана с параметром h_n^{opt}, $\beta = 5$

Заключение

В работе рассматриваются вопросы построения непараметрической оценки Берана для регрессионной модели надежности. Основным фактором, влияющим на точность получаемых оценок, является выбор параметра сглаживания. На примере выборок, смоделированных в соответствии с параметрической моделью ускоренных испытаний, показано, что выбор параметра сглаживания должен осуществляться, в первую очередь, на основании разницы предполагаемых функций надежности, соответствующих разным значениям ковариаты, тогда как влияние

объема выборки и плана эксперимента оказывается несущественным при выборе параметра сглаживания.

В работе предложен алгоритм выбора оптимального параметра сглаживания для построения непараметрической оценки Берана регрессионной модели надежности. Алгоритм основан на минимизации среднеквадратического отклонения времен отказов от непараметрической оценки обратной функции надежности. Оценки Берана, построенные с использованием оптимального параметра сглаживания, оказываются точнее, чем при использовании фиксированного параметра сглаживания, во всех рассмотренных случаях.

ЛИТЕРАТУРА

1. *Beran R.* Nonparametic Regression with Randomly Censored Survival Data. Technical report. Department of Statistics. University of California. Berkeley, 1981.
2. *Dabrowska D.M.* Nonparametric quantile regression with censored data // Sankhya Ser. A. 54. 1992. P. 252–259.
3. *Gonzalez M.W.*, *Cadarso S.C.* Asymptotic properties of a generalized Kaplan-Meier estimator with some application // J. Nonparametric Statistics. 1994. No. 4. P. 65–78.
4. *McKeague I.W.*, *Utikal K.J.* Inference for a nonlinear counting process regression model // Ann. Statist. 1990. V. 18. P. 1172–1187.
5. *Van Keilegom I.*, *Akritas M.G.*, *Veraverbeke N.* Estimation of the conditional distribution in regression with censored data: a comparative study // Computational Statistics & Data Analysis. 2001. V. 35. P. 487–500.
6. *Akritas M.G.* Nearest neighbor estimation of a bivariate distribution under random censoring // Ann. Statist. . 1994V. 22. P. 1299–1327.
7. *Расин Д.* Непараметрическая эконометрика: вводный курс // Квантиль. 2008. № 4. С. 7–26.
8. *Хардле В.* Прикладная непараметрическая регрессия. М.: Мир, 1993. С. 6–45.

Демин Виктор Андреевич
Чимитова Екатерина Владимировна
Новосибирский государственный технический университет
E-mail: vicdemin@gmail.com; ekaterina.chimitova@gmail.com Поступила в редакцию 28 апреля 2012 г.

Demin Victor A., *Chimitova Ekaterina V.* (Novosibirsk State Technical University). **Choice of optimal smoothing parameter for nonparametric estimation of regression reliability model.**

Keywords: reliability function, regression model, nonparametric Beran estimator, smoothing parameter.

The problem of nonparametric estimation of regression reliability model is considered. We consider nonparametric estimates, suggested by Beran. The main factor influencing the quality of estimates is the choice of smoothing parameter. On the example of samples, simulated from the accelerated failure time model it has been shown that the choice of smoothing parameter should be based on the difference between reliability functions corresponding to different values of the covariate, whereas the influence of the sample size and plan of experiment is not significant in the choice of smoothing parameter.

In this paper we propose the algorithm of the choice of optimal smoothing parameter for nonparametric Beran estimate of regression reliability model. The algorithm is based on the minimization of standard deviation of lifetimes from nonparametric estimate of the inverse reliability function. In all considered examples the Beran estimates, obtained with the optimal smoothing parameter, turn out to be more accurate than in the case of using fixed parameter.