ДОКЛАДЫ АН ВШ РФ
2017 октябрь–декабрь № 4 (37)

ТЕХНИЧЕСКИЕ НАУКИ

УДК 519.242.5

РОБАСТНЫЕ ОЦЕНКИ МОМЕНТОВ ПРИ ИДЕНТИФИКАЦИИ ЛЯМБДА-РАСПРЕДЕЛЕНИЯ В РАМКАХ АДАПТИВНОГО ОЦЕНИВАНИЯ

В.С. Тимофеев, Е.А. Хайленко
Новосибирский государственный технический университет

Рассмотрена задача оценки параметров универсального лямбда-распределения методом моментов. Описаны способы вычисления моментов, такие как классическое среднее, усеченное среднее, винзоризированное среднее, среднее по шорту. С помощью вычислительных экспериментов показана эффективность применения робастных оценок для идентификации лямбда-распределения. Получено, что при отсутствии в выборке аномальных наблюдений наиболее точные результаты оценивания показал метод моментов на основе классических оценок моментов, при появлении выбросов оценки лямбда-распределения полученные методом моментов на основе усеченного среднего и винзоризированного среднего являются наиболее точными, а метод моментов с использованием среднего по шорту некорректно описывает форму распределения. Данная идея идентификации лямбда-распределения была применена в алгоритмах модификаций разработанного авторами ранее метода адаптивного оценивания параметров регрессионных зависимостей. Авторами было проведено исследование данных алгоритмов при различных условиях вычислительных экспериментов. Получено, что при идентификации лямбда-распределения в случае отсутствия в выборке выбросов наиболее точные результаты показал метод моментов на основе классических моментов. При появлении в выборке грубых ошибок наблюдений более точные результаты оценивания дают предложенные модификации метода моментов, что говорит об их устойчивости.

Ключевые слова: регрессионная зависимость, адаптивное оценивание, метод максимального правдоподобия, универсальное лямбда-распределение, идентификация распределения, метод моментов, устойчивые оценки моментов.

DOI: 10.17212/1727-2769-2017-4-101-111

Введение

Проведение статистического анализа данных часто приводит к необходимости построения уравнений, описывающих наблюдаемые взаимосвязи между входными и выходными переменными. Для решения таких задач, как правило, предпочтение отдается методам регрессионного анализа, одной из задач которого является поиск оценок неизвестных параметров регрессионных зависимостей.

Известно, что классические методы оценивания параметров регрессионных моделей, например метод максимального правдоподобия (ММП), требуют наличия априорной информации о виде распределения, которой у исследователей, как правило, нет. Поэтому в работе [1] предложен подход к адаптивному оцениванию параметров регрессионных зависимостей с использованием обобщенного лямбда-распределения (GL-распределение), который позволяет получать оценки при различных распределениях случайных ошибок, таких как нормальное, экспоненциальное, Вейбулла, логнормальное, гамма-распределение и другие [2, 3]. Однако

Работа выполнена при финансовой поддержке Министерства образования и науки РФ по государственному заданию (проект 2.7996.2017/8.9).

для применения этого подхода необходимо на каждой итерации проводить оценивание параметров GL-распределения. Используемый авторами для этого метод моментов сильно зависим от качества оценок моментов, которое, при наличии в выборке даже небольшой доли выбросов, может оказаться весьма низким. Следует также отметить, что чем более высокий порядок моментов используется, тем острее проявляется данная проблема.

В связи с этим задача построения алгоритмов для вычисления робастных оценок моментов является весьма актуальной. Очевидно, что наличие устойчивых оценок моментов позволит не только получить качественные оценки параметров GL-распределения, но и значительно улучшить оценки параметров регрессионных зависимостей.

1. Постановка задачи

Рассмотрим регрессионное уравнение вида

$$y = X\theta + \varepsilon , \tag{1}$$

где $X = \begin{bmatrix} f_1(x_{11}) & \cdots & f_m(x_{1m}) \\ \vdots & \ddots & \vdots \\ f_1(x_{n1}) & \cdots & f_m(x_{nm}) \end{bmatrix}$ – матрица регрессоров, имеющая полный столбцовый ранг, т. е. $rg(X) = m$, m – количество регрессоров, n – количество испытаний, $f_1(x),...,f_m(x)$ – известные действительные функции, x_{ij} – заданные значения входных факторов в n наблюдениях, $y = (y_1,...,y_n)^{\mathrm{T}}$ – вектор отклика, $\theta = (\theta_1,...,\theta_m)^{\mathrm{T}}$ – вектор неизвестных параметров, подлежащих оцениванию; $\varepsilon = (\varepsilon_1,...,\varepsilon_n)^{\mathrm{T}}$ – вектор независимых ошибок наблюдений, имеющих одинаковое GL-распределение. Следовательно, функция распределения зависит от четырех параметров $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ и определяется с точки зрения квантилей распределения следующим образом [2]:

$$Q(u,\lambda_1,\lambda_2,\lambda_3,\lambda_4) = \lambda_1 + \frac{1}{\lambda_2}\left[\frac{u^{\lambda_3}}{\lambda_3} - \frac{(1-u)^{\lambda_4}}{\lambda_4}\right],\ 0 \le u \le 1, \tag{2}$$

$$\varepsilon_i = Q(u, \lambda_1, \lambda_2, \lambda_3, \lambda_4)\ ,\ i = 1, ..., n .$$

В данной работе предполагается унимодальность плотности распределения ошибок, что позволяет применять ММП для оценивания параметров регрессионных моделей. Кроме того, имеют место следующие предположения [3,4]:

$$E(\varepsilon) = 0\ ;\ E(\varepsilon\varepsilon^{\mathrm{T}}) = \sigma^2 I\ ,\ \sigma^2 < \infty\ ,\ rg(X) = m\ . \tag{3}$$

Таким образом, задача состоит в том, чтобы по имеющимся исходным данным (значениям отклика и входных факторов) как можно точнее оценить вектор неизвестных параметров уравнения регрессии (1).

2. Идентификация GL-распределения с использованием робастных оценок моментов

Как уже отмечалось, GL-распределение имеет четыре неизвестных параметра, поэтому для использования метода моментов достаточно иметь информацию о значениях первых четырех моментов. Однако на практике такой информации у исследователя, как правило, нет. Следовательно, разрабатываемые алгоритмы должны извлекать данную информацию из выборки, проводя оценивание значений моментов до требуемого порядка. Именно с этого и начинается алгоритм реализованного в [1] метода моментов [2]:

• для выборки $X_1, X_2, ..., X_N$ вычисляются выборочные моменты [6, 7]:

$$\hat{\alpha}_1 = \bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad \hat{\alpha}_2 = \hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{X})^2,$$
$$\hat{\alpha}_3 = \frac{1}{N\hat{\sigma}^3}\sum_{i=1}^{N}(X_i - \bar{X})^3 \quad \hat{\alpha}_4 = \frac{1}{n\hat{\sigma}^4}\sum_{i=1}^{N}(X_i - \bar{X})^4; \tag{4}$$

• значения данных моментов приравниваются к значениям теоретических для GL-распределения, которые вычисляются по формулам [2]:

$$\alpha_1 = \mu = \lambda_1 + \frac{A}{\lambda_2}, \ \alpha_2 = \sigma^2 = \frac{1}{\lambda_2^2}(B - A^2),$$

$$\alpha_3 = \frac{1}{\lambda_2^3\sigma^3}(C - 3AB + 2A^3), \ \alpha_4 = \frac{1}{\lambda_2^4\sigma^4}(D - 4AC + 6A^2B - 3A^4),$$

где $A = \dfrac{1}{\lambda_3(1+\lambda_3)} + \dfrac{1}{\lambda_4(1+\lambda_4)}$,

$$B = \frac{1}{\lambda_3^2(1+2\lambda_3)} + \frac{1}{\lambda_4^2(1+2\lambda_4)} - \frac{2}{\lambda_3\lambda_4}\beta(1+\lambda_3,\, 1+\lambda_4);$$

$$C = \frac{1}{\lambda_3^3(1+3\lambda_3)} - \frac{1}{\lambda_4^3(1+3\lambda_4)} - \frac{3}{\lambda_3^2\lambda_4}\beta(1+2\lambda_3, 1+\lambda_4) +$$
$$+ \frac{3}{\lambda_4^2\lambda_3}\beta(1+\lambda_3,\, 1+2\lambda_4);$$

$$D = \frac{1}{\lambda_3^4(1+4\lambda_3)} + \frac{1}{\lambda_4^4(1+4\lambda_4)} - \frac{4}{\lambda_3^3\lambda_4}\beta(1+3\lambda_3,\, 1+\lambda_4) +$$
$$+ \frac{6}{\lambda_3^2\lambda_4^2}\beta(1+2\lambda_3,\, 1+2\lambda_4) - \frac{4}{\lambda_3\lambda_4^3}\beta(1+\lambda_3,\, 1+3\lambda_4);$$

• оценки параметров $\lambda_1, ..., \lambda_4$ GL-распределения находятся путем решения системы уравнений

$$\alpha_i = \hat{\alpha}_i, \ i = \overline{1,4}. \tag{5}$$

Нетрудно заметить, что используемые здесь оценки моментов являются классическими оценками, которые очень чувствительны к появлению в выборке грубых ошибок наблюдения. Именно это обстоятельство приводит к резкому снижению качества оценок параметров распределения метода моментов. В этой связи авторами предлагается вместо классических оценок (4) воспользоваться робастными оценками моментов. В качестве таких оценок в данной работе рассматриваются три альтернативных варианта, которые описаны ниже.

Усеченное среднее порядка γ вычисляется следующим образом [8,9]:

$$X_{tr}^{(\gamma)} = \frac{1}{N-2k} \sum_{i=k+1}^{N-k} X_{(i)}, \ \gamma \in [0,1], \tag{6}$$

где $k = \frac{N}{2}(1-\gamma)$. При использовании данной формулы исключаются крайние элементы с двух сторон вариационного ряда.

Винзоризированное среднее порядка γ равно [8, 9]

$$X_{ws}^{(\gamma)} = \frac{1}{N}\left((k+1)\left(X_{(k+1)} + X_{(N-k)} \right) + \sum_{i=k+2}^{N-k-1} X_{(i)} \right), \tag{7}$$

т. е. «крайние» элементы вариационного ряда, исключаемые из выборки, заменяются на $X_{(k+1)}$ и $X_{(N-k)}$ соответственно.

Среднее по шорту представляет собой следующее соотношение [8, 10]:

$$X_{sh}^{(\gamma)} = \frac{1}{\tilde{N}} \sum_{i=1}^{\tilde{N}} \upsilon_{(i)}, \tag{8}$$

где $\tilde{N} = \gamma N$. Новый вариационный ряд $\upsilon_{(1)}, \upsilon_{(2)}, \ldots, \upsilon_{(G)}$ получается из исходного путем поиска части вариационного с наименьшим размахом.

Приведем результаты исследования точности идентификации параметров GL-распределения случайной величины ξ методом моментов с использованием рассмотренных выше способов их оценивания. Для того чтобы провести сравнение точности оценивания параметров $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ GL-распределения при различных засорениях была взята функция распределения случайной величины ξ следующего вида:

$$F(x) = (1-\mu) GLD_1\left(x, \lambda_1^1, \lambda_2^1, \lambda_3^1, \lambda_4^1\right) + \mu GLD_2\left(x, \lambda_1^2, \lambda_2^2, \lambda_3^3, \lambda_4^4\right), \tag{9}$$

где GLD_1 и GLD_2 – функции GL-распределения с параметрами $\left(\lambda_1^1, \lambda_2^1, \lambda_3^1, \lambda_4^1\right)$ и $\left(\lambda_1^2, \lambda_2^2, \lambda_3^2, \lambda_4^2\right)$ соответственно, μ – доля выбросов. При $\mu = 0$ случайная величина ξ будет иметь распределение GLD_1, а при $\mu = 1$ – GLD_2.

В работе были рассмотрены три варианта распределения случайной величины ξ:

1) нормальное распределение $N(0,1)$, которое является частным случаем GL-распределения с параметрами $\lambda^{\text{ист}} = (0,\ 1.418,\ 0.161,\ 0.161)$;

2) симметричное GL-распределение с параметрами $\lambda^{\text{ист}} = (0,\ 1,\ 0.5,\ 0.5)$;

3) асимметричное (с левой асимметрией) GL-распределение с параметрами $\lambda^{\text{ист}} = (0, 1, 0.5, 0.002)$.

В качестве показателя точности нахождения неизвестных параметров $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ было взято следующее соотношение:

$$\psi^{GL} = \sum_{i=1}^{4} \left(\lambda_i - \lambda_i^{\text{ист}} \right)^2 .$$

Все исследования проводились с помощью вычислительных экспериментов, которые заключались в генерации выборок значений случайной величины ξ, имеющей распределения вида 1–3, объемом 1000 наблюдений и оценивании параметров GL-распределения, количество экспериментов 500. Для поиска неизвестных параметров применялся метод моментов с использованием следующих оценок моментов: классические моменты, вычисляемые по (4); усеченное среднее (6); винзоризированное среднее (7); среднее по шорту (8). В качестве итогового было взято усредненное по 500 выборкам значение показателя точности ψ^{GL}.

Приведем результаты исследования точности оценивания параметров GL-распределения при отсутствии выбросов для всех рассмотренных выше вариантов распределений в 1–3. В табл. 1 в столбце, отмеченном 1, приведены результаты идентификации GL-распределения для варианта распределения 1; в столбце, отмеченном 2, представлены результаты для распределения вида 2, и в столбце, отмеченном 3, – результаты для распределения вида 3.

Таблица 1 / Table 1

Точность оценивания параметров GL-распределений ψ^{GL} при отсутствии выбросов

The accuracy of estimating the parameters of GL-distributions ψ^{GL} when there are no outliers

Распределение случайной величины	1	2	3
Усеченное среднее	1,750E-01	1,558E-01	5,126E-01
Винзоризированное среднее	6,376E-02	1,677E-02	4,057E-01
Среднее по шорту	1,571E-01	4,333E-02	6,879E+01
Классическое среднее	3,189E-05	3,054E-05	4,331E-02

Как видно по табл. 1, при отсутствии выбросов в выборке наиболее точные результаты были достигнуты с использованием классических оценок, что можно объяснить использованием всей имеющейся в выборке информации. Кроме того, следует отметить, что наиболее точные результаты оценивания параметров GL-распределения получены всеми методами при симметричном виде распределения (случаи 1 и 2). Также по проведенным исследованиям было замечено, что все методы идентификации сохраняют симметрию либо асимметрию формы распределения случайной величины.

Далее был рассмотрен случай появления выбросов в выборке, т. е. функция распределения ξ представлена в виде (9). В качестве GLD_1 выступают описанные выше распределения для вариантов 1–3. В качестве распределения аномальных наблюдений GLD_2 было рассмотрено то же распределение, но с большей дисперсией. В терминах GL-распределения за величину масштаба отвечает параметр λ_2, связь обратная и для моделирования была взята величина $\lambda_2^2 = 0.35 \cdot \lambda_2^1$, доля выбросов $\mu = 0.1$.

В табл. 2 представлены результаты исследования точности оценивания параметров GL-распределения. По таблице видно, что при появлении в выборке выбросов наиболее точные результаты показали оценки моментов с использованием (6) и (7).

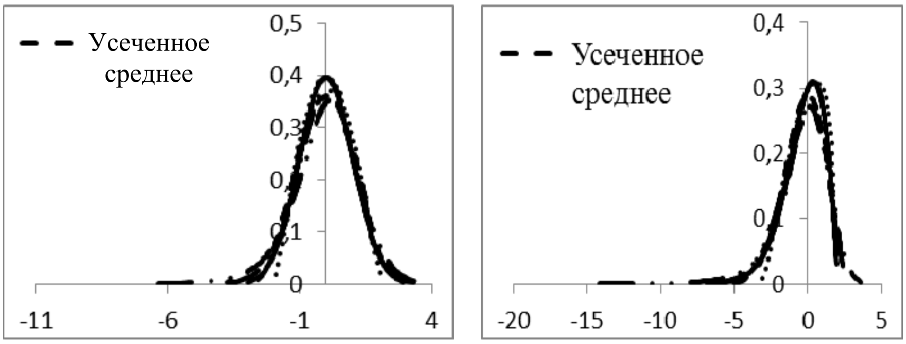
Таблица 2 / Table 2

Точность оценивания параметров GL-распределений при появлении выбросов, $\mu = 0.1$

The accuracy of estimating the parameters of GL-distributions when outliers exist, $\mu = 0.1$

Распределение случайной величины	1	2	3
Усеченное среднее	3,201E-02	6,992E-02	7,647E-02
Винзоризированное среднее	1,781E-02	1,094E-01	5,023E-02
Среднее по шорту	8,784E-02	4,216E-01	1,916E-01
Классическое среднее	3,673E-01	3,660E-01	1,686E-01

Рассмотрим графики (рис. 1) восстановленных функций плотности распределений для засоренного нормального и асимметричного GL-распределений, случаи 1 и 3, на примере одной выборки объемом 1000 наблюдений.



а *б*

Рис. 1 – Графики восстановленной плотности распределения для случаев распределений: *а* – 1; *б* – 2

Fig. 1 – Plots of the reconstructed distribution density for cases of distributions: *a* – 1; *b* – 2

Как видно из рис. 1, восстановленные функции плотности с использованием оценок моментов, полученных на основе соотношений (6)–(7), описывают форму распределения ближе к истинному. Особенно следует отметить, что восстановленная функция плотности, полученная методом моментов с использованием среднего по шорту, наихудшим образом описывает форму истинного распределения, поскольку у нее отсутствуют «хвосты» распределения. Кроме того, представленные в табл. 2 результаты показали, что усредненные оценки параметров GL-распределения, полученные с использованием моментов, вычисленных по (7), также менее точные по сравнению с другими робастными оценками, поэтому в дальнейшем этот способ вычисления значений моментов авторами не рассматривался.

3. Модификации алгоритмов адаптивного оценивания параметров регрессии с использованием робастных оценок моментов

Перейдем к рассмотрению регрессионных зависимостей. Поскольку приведенные выше результаты исследований показали, что идентификация GL-распределения с использованием метода моментов на основе робастных оценок дает более устойчивые результаты к появлению в выборке выбросов, то возникает идея разработки модификаций представленного в работе [1] алгоритма адаптивного метода. В рамках таких модификаций предполагается проводить идентификацию параметров GL-распределения методом моментов с использованием робастных оценок моментов усеченного среднего и винзоризированного среднего.

Будем проводить сравнение точности нахождения оценок $\hat{\theta}$ вектора неизвестных параметров θ с использованием вычислительных экспериментов. Для оценивания параметров регрессии были взяты адаптивные методы с использованием классической и робастных оценок среднего (6) и (7), применяемых в методе моментов для идентификации параметров GL-распределения. Поскольку авторами ранее в работе [1] проводилось сравнение точности оценок, полученных адаптивным методом и методом наименьших квадратов (МНК), были показаны преимущества первого метода, поэтому МНК в данной работе рассматриваться не будет.

В качестве исследуемой использовалась следующая модель:

$$y_i = \theta_1 + \theta_2 x_{1i} + \theta_3 x_{2i} + \varepsilon_i ,\ i = 1,...,n ,$$

где количество регрессоров $m = 3$, количество испытаний $n = 500$, значения входных факторов x_{ij} выбирались из отрезка $[0,1]$, $\theta_{\text{ист}} = (1, 1.5, 2)^{\mathrm{T}}$. Случайные ошибки ε_i, $i = 1,...,n$ моделировались независимыми и одинаково распределенными с функцией распределения вида (9).

В качестве показателя точности нахождения оценок различными методами использовалось следующее соотношение:

$$\psi = \sum_{i=1}^{m} \frac{\left(\theta_i^{\text{ист}} - \hat{\theta}_i\right)^2}{\left(\theta_i^{\text{ист}}\right)^2} .$$

Проведем исследование точности оценивания параметров регрессионных зависимостей при следующих распределениях случайных ошибок.

I. Нормальное распределение $N(0,\ 0.1)$.

II. Асимметричное распределение $GLD(0,\ 4,\ 0.5,\ 0.002)$.

Для различных значений μ проводилось по 500 вычислительных экспериментов. Каждый такой эксперимент заключался в моделировании выборки исходных данных в соответствии с моделью (9) с последующим оцениванием параметров этой модели рассмотренными выше методами: наименьших квадратов и адаптивными методами с использованием классической и робастных оценок среднего. В качестве итогового показателя точности оценивания ψ использовались усредненные по 500 экспериментам значения.

Рассмотрим случаи отсутствия выбросов, т. е. ошибка измерений имеет распределения, описанные для случаев I и II, $\mu = 0$, а также появления в выборке выбросов, т. е. функция распределения ошибки представлена в виде (9). В качестве распределения ошибок аномальных наблюдений GLD_2 будем рассматривать то же распределение, но с большей дисперсией.

В табл. 3 представлены результаты сравнения точности оценивания неизвестных параметров регрессии для данных случаев. Для удобства представления результатов введем следующие обозначения адаптивных методов на основе метода моментов:

- «классический метод» – с использованием классической оценки моментов для идентификации параметров GL-распределения;
- «усеченное среднее» – с использованием оценки усеченного среднего для оценки моментов для идентификации параметров GL-распределения;
- «винзоризированное среднее» – на основе винзоризированного среднего для оценки моментов для идентификации параметров GL-распределения.

Как видно из табл. 3, при отсутствии в выборке выбросов $(\mu = 0)$ при распределении случайной ошибки вида I все исследуемые методы дают близкие результаты оценивания. Наилучшие результаты оценивания показал адаптивный метод на основе классического метода моментов, это связано с тем, что идентификация распределения проводится по всем наблюдениям.

Таблица 3 / Table 3

Точность оценивания параметров регрессионных зависимостей ψ

The accuracy of estimating regression model parameters ψ

	Доля выбросов μ	0	0,05	0,1	0,15
I	«Классический метод»	1.39E-05	1.597E-04	3.503E-01	2.304E-01
	«Усеченное среднее»	4.36E-05	2.700E-04	3.913E-03	1.249E-02
	«Винзоризированное среднее»	1.03E-05	2.048E-04	5.143E-03	1.636E-02
II	«Классический метод»	1.486E-04	3.086E-03	3.503E-01	2.304E-01
	«Усеченное среднее»	4.712E-04	1.026E-03	3.319E-03	1.249E-02
	«Винзоризированное среднее»	7.294E-04	6.367E-04	5.143E-03	1.636E-02

Как видно из табл. 3, при появлении в выборке аномальных наблюдений наиболее точные результаты дают методы адаптивного оценивания на основе робастных оценок (6)–(7). Это свидетельствует о том, что предложенные модификации адаптивного метода обладают свойством устойчивости и их можно рекомендовать для оценивания при наличии в выборке грубых ошибок наблюдений.

Заключение

В статье рассмотрена задача оценивания параметров регрессионных зависимостей. Предложены модификации метода адаптивного оценивания неизвестных параметров регрессионного уравнения, основанные на идентификации GL-распределения с использованием устойчивых методов оценивания моментов. С помощью вычислительных экспериментов подтверждена работоспособность данных модификаций. Проведено сравнение результатов работы данных методов с результатами, полученными разработанным ранее адаптивным методом. Получено, что в случае отсутствия аномальных наблюдений адаптивный метод, основанный на классическом методе моментов для идентификации GL-распределения, дает наиболее точные результаты. При появлении выбросов в выборке наилучшие результаты оценивания показывают адаптивные методы, основанные на робастных методах вычисления моментов, таких как усеченное и винзоризированное средние, для идентификации GL-распределения, что говорит об устойчивости к грубым ошибкам наблюдений данных модификаций.

ЛИТЕРАТУРА

1. **Тимофеев В.С., Хайленко Е.А.** Адаптивное оценивание параметров регрессионных моделей с использованием обобщенного лямбда-распределения // Доклады АН ВШ РФ. – 2010. – № 2 (15). – С. 25–36.
2. **Karian Z.A., Dudewicz E.J.** Fitting statistical distributions: the Generalized Lambda Distribution and Generalized Bootstrap methods. – New York: CRC Press, 2000. – 435 p.
3. **Lakhany A., Mausser H.** Estimation the parameters of the Generalized Lambda Distribution // ALGO Research Quarterly. – 2000. – Vol. 3, N 3. – P. 47–58.
4. **Айвазян С.А. Енюков И.С. Мешалкин Л.Д**. Прикладная статистика. – М.: Финансы и статистика, 1985. – 488 с.
5. **Кендалл М., Стьюарт А**. Теория распределений. – М.: Наука, 1966. – 587 с.
6. **Крамер Г.** Математические методы статистики. – 2-е изд., стер. – М.: Мир, 1975. – 648 с.
7. **Боровков А.А**. Математическая статистика: оценка параметров, проверка гипотез. – М.: Наука, 1984. – 472 с.
8. **Тимофеев В.С., Щеколдин В.Ю.** Об оценивании статистических характеристик при анализе статических многофакторных объектов // Научный вестник НГТУ. – 2006. – № 3 (24). – С. 47–59.
9. **Tukey J.W., McLaughlin D.H.** Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsirization 1 // Sankhyā: The Indian Journal of Statistics, Series A. – 1963. – Vol. 25. – P. 331–352.
10. **Grubel R.** The length of the shorth // The Annals of Statistics. – 1988. – Vol. 16, N 2. – P. 619–628.

ROBUST ESTIMATES OF MOMENTS IN THE IDENTIFICATION OF GENERALIZED LAMBDA-DISTRIBUTION WITHIN THE ADAPTIVE REGRESSION MODEL ESTIMATION

Timofeev V.S., Khailenko E.A.
Novosibirsk State Technical University, Novosibirsk, Russia

The problem of estimation Generalized Lambda-distribution parameters using he method of moments is investigated. The methods for calculating the moments are described, such as the classical mean, truncated mean, vizorized mean, and average by shorts. The effectiveness of applying robust estimates for the identification of Lambda Distribution is shown using computational experiments. It was revealed that the method of moments based on classical moment estimates provided the most accurate results of the estimation lambda-distribution parameters when there were no outliers in the sample. When there were outliers in the sample: estimates of the Generalized Lambda -distribution obtained by the method of moments based on the truncated mean and vinsorized mean, the results were most accurate. The method of moments using the mean by shorts incorrectly describes the shape of the distribution. This idea of identifying the Generalized Lambda -distribution was applied to algorithms of modification developed by the authors before the method of adaptive estimation of regression model parameters. These algorithms were studied under various conditions of computational experiments. The most accurate results were obtained by the adaptive method using the method of moments based on classical estimates in the case when there were no outliers in the sample. When outliers exist in the sample,the most accurate results were obtaied by the adaptive method using the method of moments based on robust estimates. These results show the robustness of these modifications.

Keywords: Regression equation, adaptive estimation, maximum-likelihood method, Generalized Lambda-distribution, identification of distribution, method of moments, robust estimates of moments.

DOI: 10.17212/1727-2769-2017-4-101-111

REFERENCES

1. Timofeev V.S., Khailenko E.A. Adaptivnoe otsenivanie parametrov regressionnykh modelei s ispol'zovaniem obobshchennogo lyambda-raspredeleniya [Adaptive estimation of regression models parameters using Generalized Lambda-Distribution]. *Doklady Akademii nauk vysshei shkoly Rossiiskoi Federatsii – Proceedings of the Russian higher school Academy of sciences*, 2010, no. 2 (15), pp. 25–36.
2. Karian Z.A., Dudewicz E.J. *Fitting statistical distributions: the Generalized Lambda Distribution and Generalized Bootstrap methods*. New York, CRC Press, 2000. 435 p.
3. Lakhany A., Mausser H. Estimation the parameters of the Generalized Lambda Distribution. *ALGO Research Quarterly*, 2000, vol. 3, no. 3, pp. 47–58.
4. Aivazyan S.A. Enyukov I.S. Meshalkin L.D. *Prikladnaya statistika* [Applied statistics: studying dependency]. Moscow, Finansy i statistika Publ., 1985. 488 p.
5. Kendall M., Stuart A. *The advanced theory of statistics*. Vol. 1. *Distribution theory*. London, Griffin, 1963 (Russ. ed.: Kendall M., St'yuart A. *Teoriya raspredelenii*. Moscow, Nauka Publ., 1966. 587 p.).
6. Cramér H. *Mathematical methods of statistics*. Princeton, Princeton university press, 1946 (Russ. ed.: Kramer G. *Matematicheskie metody statistiki*. 2nd ed. Moscow, Mir Publ., 1975. 648 p.).
7. Borovkov A.A. *Matematicheskaya statistika: otsenka parametrov, proverka gipotez* [Mathematical statistics. Estimation of parameters, hypothesis testing]. Moscow, Nauka Publ., 1984. 472 p.
8. Timofeev V.S., Shchekoldin V.Yu. Ob otsenivanii statisticheskikh kharakteristik pri analize staticheskikh mnogofaktornykh ob"ektov [About the estimation of statistical characteristics in the analysis of static multifactor objects]. *Nauchnyi vestnik Novosibirskogo gosudarstvennogo tekhnicheskogo universiteta – Science bulletin of the Novosibirsk state technical university*, 2006, no. 3 (24), pp. 47–59.
9. Tukey J.W., McLaughlin D.H. Less vulnerable confidence and significance procedures for location based on a single sample: Trimming/Winsirization 1. *Sankhyā: The Indian Journal of Statistics, Series A*, 1963, vol. 25, pp. 331–352.
10. Grubel R. The length of the shorth. *The Annals of Statistics*, 1988, vol. 16, no. 2, pp. 619–628.

СВЕДЕНИЯ ОБ АВТОРАХ

Тимофеев Владимир Семенович – родился в 1972 году, д-р техн. наук, профессор, кафедра теоретической и прикладной информатики, Новосибирский государственный технический университет. Основное направление научных исследований – разработка и исследование устойчивых методов и алгоритмов анализа многофакторных объектов, в том числе с использованием непараметрической статистики. Имеет более 100 публикаций, в том числе один учебник. (Адрес: 630073, Россия, г. Новосибирск, пр. Карла Маркса, 20. E-mail: v.timofeev@corp.nstu.ru).

Timofeev Vladimir Semenovich (b. 1972) – Doctor of Sciences (Eng.), professor at the Department of Theoretical and Applied Informatics, Novosibirsk State Technical University. His research interests cover development and research of stable methods and algorithms for the analysis of multifactor objects including nonparametric statistics. He is the author more than 100 scientific papers and 1 textbook. (Address: 20, Karl Marx Av., Novosibirsk, 630073, Russia. E-mail: v.timofeev@corp.nstu.ru).

Хайленко Екатерина Алексеевна – родилась в 1985 году, канд. техн. наук, доцент, кафедра теоретической и прикладной информатики, Новосибирский государственный технический университет. Область научных интересов: разработка и исследование алгоритмов устойчивого и адаптивного оценивания параметров регрессионных зависимостей и планирование эксперимента. Опубликовано 28 научных работ. (Адрес: 630073, Россия, Новосибирск, проспект Карла Маркса, 20. E-mail: xajlenko@corp.nstu.ru).

Khailenko Ekaterina Alekseevna (b. 1985) – Candidate of Sciences (Eng.), associate professor at the Department of Theoretical and Applied Informatics, Novosibirsk State Technical University. Her research interests are currently focused on developing and investigating algorithms of robust and adaptive estimation parameters of regression models and experiment design. She is the author of 28 scientific papers. (Address: 20, Karl Marx Av., Novosibirsk, 630073, Russia. E-mail: xajlenko@corp.nstu.ru).

Статья поступила 11 декабря 2017 г.
Received December 11, 2017

To References:

Timofeev V.S., Khailenko E.A. Robastnye otsenki momentov pri identifikatsii lyambda-raspredeleniya v ramkakh adaptivnogo otsenivaniya [Robust estimates of moments in the identification of generalized lambda-distribution within the adaptive regression model estimation]. *Doklady Akademii nauk vysshei shkoly Rossiiskoi Federatsii – Proceedings of the Russian higher school Academy of sciences*, 2017, no. 4 (37), pp. 101–111. doi: 10.17212/1727-2769-2017-4-101-111