

УДК 519.242.5

**АДАПТИВНОЕ ОЦЕНИВАНИЕ ПАРАМЕТРОВ
РЕГРЕССИОННЫХ ЗАВИСИМОСТЕЙ
ПРИ НЕОДНОРОДНОСТИ СЛУЧАЙНЫХ ОШИБОК****В.С. Тимофеев, Е.А. Хайленко***Новосибирский государственный технический университет*

Рассмотрена задача оценивания параметров регрессионных моделей. Предложен алгоритм адаптивного оценивания с использованием полупараметрического подхода к оцениванию функции плотности распределения случайных ошибок с учетом неоднородности распределения ошибок наблюдений на области определения входных факторов. Проведено сравнение точности оценивания параметров регрессионных зависимостей данного метода с результатами, полученными, разработанными авторами ранее, адаптивными методами на основе универсального лямбда-распределения и полупараметрической оценки функции плотности распределения ошибок.

Ключевые слова: регрессионная зависимость, адаптивное оценивание, полупараметрическая оценка, обобщенное лямбда-распределение, метод максимального правдоподобия.

DOI: 10.17212/1727-2769-2014-4-115-123

Введение

На практике исследователям часто приходится сталкиваться с задачей нахождения взаимосвязи между входными факторами, описывающими условия функционирования, и выходными факторами, характеризующими результат этого функционирования. Для решения такой задачи используют методы регрессионного анализа.

Классическим методом нахождения неизвестных параметров регрессионных зависимостей является метод максимального правдоподобия (ММП) [1], однако для его корректного применения необходима априорная информация о виде распределения ошибок наблюдений, которой, как правило, у исследователя нет. Поэтому ранее в работах [2–4] авторами были предложены методы адаптивного оценивания параметров регрессионных зависимостей с использованием универсальных распределений, которые позволяют получать ММП-оценки при неизвестном распределении ошибок наблюдений.

Следует отметить, что данные методы разработаны для случая, когда ошибки являются одинаково распределенными на всей области определения входных факторов, однако на практике распределение ошибок на отдельных участках может отличаться. Поэтому в данной работе предлагается модифицировать уже существующий адаптивный метод оценивания параметров регрессионных зависимостей с использованием полупараметрического восстановления функции плотности распределения на основе обобщенного лямбда-распределения (GL-распределения) на случай неоднородности распределения ошибок наблюдений на области определения входных факторов. В результате появляется возможность более гибко реагировать на изменение условий экспериментов.

Работа выполнена при финансовой поддержке Министерства образования и науки РФ, по государственному заданию № 2014/138, проект № 1689

1. Постановка задачи

Рассмотрим регрессионное уравнение вида

$$y = X\theta + \varepsilon, \quad (1)$$

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 столб-

цовый ранг, т.е. $rg(X) = m$, m – количество регрессоров, n – количество испытаний, $f_1(x), \dots, f_m(x)$ – известные действительные функции, x_{ij} – заданные значения входных факторов в n наблюдениях, $y = (y_1, \dots, y_n)^T$ – вектор отклика, $\theta = (\theta_1, \dots, \theta_m)^T$ – вектор неизвестных параметров, подлежащих оцениванию; $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$ – вектор независимых ошибок наблюдений с унимодальными и, как правило, неизвестными плотностями распределения $g(x)$.

Также имеют место следующие предположения [5]:

$$E(\varepsilon) = 0, \quad E(\varepsilon\varepsilon^T) = \sigma^2 I, \quad \sigma^2 < \infty, \quad rg(X) = m. \quad (2)$$

Кроме того, предполагается, что на отдельных участках области определения входных факторов наблюдается различное распределение случайных ошибок. На рис. 1 представлен пример неоднородного распределения ошибок для одномерной области.

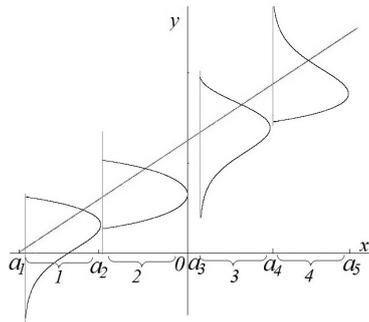


Рис. 1 – Распределение случайной ошибки на области определения входных факторов

Fig. 1 – Distribution of errors on domain of definition the input data

На рис. 1 входной фактор x определен на отрезке $[a_1; a_5]$, на участках $[a_1; a_2)$, $[a_2; a_3)$, $[a_3; a_4)$ и $[a_4; a_5]$ наблюдаются различные распределения случайных ошибок наблюдений, количество участков неоднородности в данном случае равно 4. В общем случае входных факторов может быть несколько, и область может быть поделена на любое количество участков. Обозначим через k количество таких участков, через n_i – количество ошибок наблюдений, соответствующих i -му участку, $i = \overline{1, k}$.

Задача состоит в том, чтобы по имеющимся данным (значениям отклика и входных факторов) восстановить распределение ошибок наблюдений на каждом участке и оценить вектор неизвестных параметров уравнения регрессии (1).

2. Полупараметрическая оценка функции плотности

Для идентификации распределения случайных ошибок на отдельных участках области определения входных факторов предлагается использовать полупараметрическую оценку функции плотности. Ранее авторами в работах [2, 3] предложена такая оценка неизвестной функции плотности с использованием GL-распределения. В этом случае искомая функция плотности имеет вид

$$g(x, \hat{\lambda}, \alpha) = (1 - \alpha)\varphi(x, \hat{\lambda}) + \alpha\hat{\varphi}(x), \quad (3)$$

где $\varphi(x, \hat{\lambda})$ – параметрическая компонента функции плотности, $\hat{\varphi}(x)$ – непараметрическая, α – неизвестная величина ($0 \leq \alpha \leq 1$), которая оценивается с использованием метода максимального правдоподобия [6].

В качестве параметрической компоненты предлагается использовать функцию распределения универсального GL-распределения, которая определяется с точки зрения квантилей распределения следующим образом [7]:

$$Q(u, \lambda_1, \lambda_2, \lambda_3, \lambda_4) = \lambda_1 + \frac{1}{\lambda_2} \left[\frac{u^{\lambda_3}}{\lambda_3} - \frac{(1-u)^{\lambda_4}}{\lambda_4} \right], \quad 0 \leq u \leq 1, \quad (4)$$

$$x = Q(u, \lambda_1, \lambda_2, \lambda_3, \lambda_4).$$

Соответствующая функция плотности имеет вид

$$\varphi(x, \lambda) = \frac{\lambda_2}{u^{\lambda_3-1} + (1-u)^{\lambda_4-1}}, \quad 0 \leq u \leq 1. \quad (5)$$

Как известно из [7, 8], GL-распределение полностью определяется своими первыми четырьмя моментами, поэтому для вычисления оценок параметров распределения можно использовать метод моментов [1].

В качестве непараметрической компоненты предлагается использовать оценку функции плотности, предложенную Розенблатом–Парзенем [9]:

$$\hat{\varphi}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right), \quad (6)$$

где h – ширина ядра (окна сглаживания), $K(r)$ – функция ядра, в качестве которой было взято ядро Гаусса [9],

$$K(r) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}r^2\right\}.$$

Таким образом, искомая функция плотности распределения ошибок наблюдений $g(x, \hat{\lambda}, \alpha)$ находится путем подстановки (5) и (6) в (3).

3. Адаптивный метод оценивания параметров регрессии

Используя идею идентификации распределения остатков регрессионных зависимостей, вычисляемых как $e = y - X\hat{\theta}$, на каждом участке области определения входных факторов на основе полупараметрического способа оценивания функции плотности можно предложить адаптивный метод оценивания параметров регрессионных зависимостей, который учитывает такую неоднородность (рис. 2). Для этого предлагается:

- для каждого i -го участка неоднородности распределения ошибок ($i = \overline{1, k}$) выбрать остатки, соответствующие наблюдениям из данного участка, определить их количество n_i и записать их в вектор $e^{(i)} = (e_1^{(i)}, e_2^{(i)}, \dots, e_{n_i}^{(i)})$;

- найти полупараметрическую оценку функции плотности распределения остатков $e^{(i)}$ по формуле (3);

- подставить полученные оценки плотностей в функцию правдоподобия.

В данном случае функция правдоподобия имеет вид

$$L(e, \theta) = \prod_{i=1}^k \prod_{j=1}^{n_i} g_i(e_j^{(i)}, \hat{\lambda}, \alpha_i),$$

где функция плотности $g_i(e_j^{(i)}, \hat{\lambda}, \alpha_i)$ определяется по соотношению (3), α_i – доля непараметрической компоненты в оценке функции плотности для i -го участка. Поскольку ошибки наблюдений являются независимыми, то логарифм функции правдоподобия имеет вид

$$\ln(L(e, \theta)) = \sum_{i=1}^k \sum_{j=1}^{n_i} \ln(g_i(e_j^{(i)}, \hat{\lambda}, \alpha_i)).$$

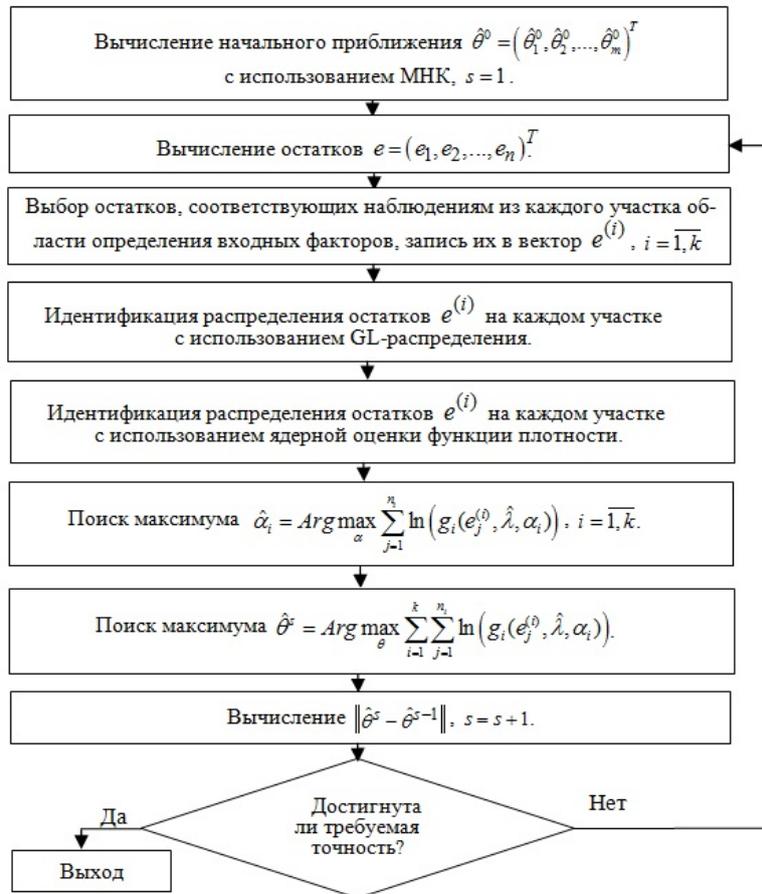


Рис. 2 – Алгоритм адаптивного метода оценивания параметров регрессионных моделей

Fig. 2 – Algorithm of adaptive parameters estimation of regression models

На каждой итерации алгоритма производится идентификация распределения остатков с использованием параметрической и непараметрической оценок функции плотности, вычисляется полупараметрическая оценка плотности, определяются оптимальные значения $\hat{\alpha}_i$, $i = \overline{1, k}$ и $\hat{\theta}$. Для поиска оценок $\hat{\theta}^s$ и $\hat{\alpha}_i$, $i = \overline{1, k}$ на s -й итерации используется симплексный метод Нелдера–Мида или поиск по деформируемому многограннику [10].

4. Результаты исследований

Проведем исследование точности оценивания параметров регрессионных моделей адаптивными методами с параметрическим, основанном только на GL-распределении и полупараметрическими без учета и с учетом неоднородности распределения на области определения входных факторов способами идентификации распределения случайной компоненты.

В качестве исследуемой использовалась следующая модель:

$$y_i = \theta_1 + \theta_2 x_i + \theta_3 x_i^2 + \varepsilon_i, \quad i = 1, \dots, n,$$

где количество регрессоров $m = 3$, количество испытаний $n = 500$, значения входных факторов x_i выбирались из отрезка $[0; 1]$, $\theta_{\text{ист}} = (1; 1,5; 2)^T$, случайные ошибки ε_i , $i = 1, \dots, n$, моделировались независимыми, причем область определения x была разделена на два участка неоднородности распределения ошибок: $x \in [0; 0,5]$ и $x \in [0,5; 1]$.

Распределение на участках области определения x будем моделировать в виде смеси двух нормальных распределений с функцией плотности:

$$F(x) = (1 - \mu)N(0; 0,01) + \mu N(0; 0,5), \quad (7)$$

где $\mu \in [0; 1]$. Представление функции распределения ошибок наблюдений в виде (7) позволяет моделировать ошибки с различной степенью засорений. При $x \in [0; 0,5]$ – доля выбросов $\mu = 0,1$, при $x \in [0,5; 1]$ – доля выбросов $\mu = 0$, т. е. распределение является нормальным и является частным случаем GL-распределения.

В табл. 1 представлены результаты оценивания параметров GL-распределения и доли непараметрической компоненты функции плотности остатков на каждом участке области определения x . По табл. 1 видно, что при появлении в выборке выбросов, на участке $x \in [0; 0,5]$, распределение ошибок существенно отклоняется от нормального закона и при идентификации распределения случайной ошибки требуется использование непараметрической компоненты (доля непараметрического распределения $\hat{\alpha} = 0,470$). При нормальном распределении ошибок наблюдений, на участке $x \in [0,5; 1]$, использование непараметрической компоненты практически не требуется ($\hat{\alpha} = 0,012$).

Таблица 1 / Table 1

Полупараметрическая оценка параметров плотности распределения остатков
Semiparametric density function estimation of residuals distribution

Параметры GL-распределения	λ_1	λ_2	λ_3	λ_4	$\hat{\alpha}$
$x \in [0; 0,5]$	-0,103	6,444	1,201	-0,183	0,470
$x \in [0,5; 1]$	-0,024	13,101	0,333	0,163	0,012

Теперь проведем исследование качества оценивания параметров регрессионных зависимостей от доли выбросов для всех трех адаптивных методов. В качестве показателя точности нахождения оценок использовалось соотношение

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

Проводилось по 100 вычислительных экспериментов, каждый из которых заключался в моделировании выборки исходных данных, а также ошибок наблюдений, и с последующим оцениванием параметров этой модели выбранными методами.

В табл. 2 представлены результаты оценивания. Для удобства обозначим адаптивный метод на основе GL-распределения как метод 1; адаптивный метод на основе полупараметрической оценки функции плотности без учета неоднородности распределения ошибок наблюдений – как метод 2; адаптивный метод на основе полупараметрической оценки функции плотности с учетом неоднородности распределения ошибок наблюдений – как метод 3.

Таблица 2 / Table 2

Точность оценивания параметров регрессии при распределении ошибок, представленных в виде смеси двух нормальных распределений
Estimation accuracy of regression model parameters when errors distribution presented as mixture of two normal distributions

Метод	Доля выбросов (μ)	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	Ψ
1	0,1	0,928	1,759	1,831	4,228E-02
	0,2	1,097	1,175	2,236	7,019E-02
2	0,1	0,980	1,555	1,983	1,824E-03
	0,2	0,991	1,691	1,780	2,838E-02
3	0,1	1,004	1,489	2,008	8,663E-05
	0,2	1,026	1,415	2,087	5,740E-03

По табл. 2 можно сделать следующие выводы. С увеличением доли выбросов в выборке распределение случайной ошибки в большей степени отклоняется от нормального закона и хуже описывается в терминах GL-распределения, таким образом, точность оценивания параметров регрессионных моделей методом 1 ниже по сравнению с методами 2 и 3. Метод 3 показал наиболее точные результаты оценивания, это связано с тем, что непараметрическая компонента в большей степени используется при оценке функции плотности распределения ошибок только на участке, на котором присутствуют выбросы (в данном случае при $x \in [0; 0,5)$).

Рассмотрим случай, когда ошибки имеют различное GL-распределение на отдельных участках (табл. 3). При $x \in [0; 0,5)$ – симметричное GL-распределение с параметрами GLD (0; 10; 0,5; 0,5), и при $x \in [0,5; 1]$ – несимметричное с правой асимметрией распределение GLD (0; 10; 0,5; 0,002).

Таблица 3 / Table 3

Точность оценивания параметров регрессии при GL-распределении ошибок
Estimation accuracy of regression model parameters when errors have Generalized lambda-distribution

Метод	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	Ψ
1	0,843	1,919	1,852	1,082E-01
2	1,072	1,217	2,256	5,719E-02
3	0,988	1,593	1,943	4,833E-03

Как видно из табл. 3, метод 1 показал наименее точные результаты, это связано с тем, что данный метод строит единую оценку плотности распределения ошибок в терминах GL-распределения для всей области определения входных факторов.

Однако при наличии неоднородности распределения на отдельных участках итогового распределения может не быть в классе GL-распределения. Метод 2, за счет подключения непараметрической оценки в процедуру идентификации распределения ошибок на всей области, показал точность выше, чем метод 1. Однако метод 3 показал наиболее точные результаты, поскольку при вычислении функции правдоподобия он учитывает распределение на каждом участке области определения входных факторов.

Заключение

В статье рассмотрена задача оценивания параметров регрессионных зависимостей. Предложен и исследован алгоритм адаптивного оценивания неизвестных параметров регрессионных зависимостей, который учитывает неоднородность распределения ошибок наблюдений на области определения входных факторов в процессе вычисления функции правдоподобия. С помощью вычислительных экспериментов подтверждена работоспособность данного алгоритма. Проведено сравнение результатов работы данного метода с результатами, полученными адаптивными методами на основе GL-распределения и полупараметрической оценки функции плотности. Получено, что при появлении в выборке выбросов предложенный алгоритм показал наиболее точные результаты. При различных GL-распределениях ошибок наблюдений на отдельных участках области определения входных факторов наиболее точные результаты также показал разработанный метод.

ЛИТЕРАТУРА

1. **Боровков А.А.** Математическая статистика. Оценка параметров, проверка гипотез. – М.: Наука, 1984. – 472 с.
2. **Денисов В.И., Тимофеев В.С., Хайленко Е.А.** Полупараметрическое восстановление функции плотности на основе обобщенного лямбда-распределения в задаче идентификации регрессионных моделей // Сибирский журнал индустриальной математики. – 2014. – Т. 17, № 3. – С. 71–77.
3. **Тимофеев В.С.** Адаптивное восстановление регрессионных зависимостей на основе полупараметрической оценки плотности случайной компоненты // Научный вестник НГТУ. – 2013. – № 4 (53). – С. 24–30.
4. **Тимофеев В.С., Хайленко Е.А.** Адаптивное оценивание параметров регрессионных моделей с использованием обобщенного лямбда-распределения // Доклады Академии наук высшей школы Российской Федерации. – 2010. – № 2 (15). – С. 25–36.
5. **Айвазян С.А., Енюков И.С., Мешалкин Л.Д.** Прикладная статистика. – М.: Финансы и статистика, 1985. – 488 с.
6. **Olkin I., Spiegelman C.H.** A semiparametric approach to density estimation // Journal of the American Statistical Association. – Vol. 82, iss. 399. – P. 858–865.
7. **Karian Z.A., Dudewicz E.J.** Fitting statistical distributions: the generalized lambda distribution and generalized bootstrap methods. – New York: CRC Press, 2000. – 435 p.
8. **Lakhany A., Mausser H.** Estimation the parameters of the generalized lambda distribution // ALGO Research Quarterly. – 2000. – Vol. 3, iss. 3. – P. 27–58.
9. **Pagan A., Ullah A.** Nonparametric econometrics. – New York: Cambridge University Press, 1999. – 424 p.
10. **Банди Б.** Методы оптимизации. Вводный курс: пер. с англ. – М.: Радио и связь, 1988. – 128 с.

ADAPTIVE ESTIMATION OF REGRESSION MODEL PARAMETERS WITH ERROR DISTRIBUTION INHOMOGENEITY

Timofeev V.S.¹, Khailenko E.A.²

¹*Novosibirsk State Technical University, Novosibirsk, Russia*

²*Novosibirsk State Technical University, Novosibirsk, Russia*

The problem of estimation of regression model parameters is considered. An adaptive estimation method of regression model parameters using a semi-parametric approach to the estimation of the density function of the distribution of observation errors taking into account their inhomogeneity on the domain of input factor definition is investigated. The comparison of the accuracy of estimating the regression model parameters using this method with the results obtained by the adaptive methods based on Generalized Lambda-Distribution developed by the authors earlier is made.

Keywords: regression models; adaptive estimation; semi-parametric estimation; Generalized Lambda-Distribution; the maximum likelihood method.

REFERENCES

1. Borovkov A.A. *Matematicheskaya statistika. Otsenka parametrov, proverka gipotez* [Mathematical Statistics. Parameters estimating, hypothesis testing]. Moscow, Nauka Publ., 1984. 472 p.
2. Denisov V.I., Timofeev V.S., Khailenko E.A. Poluparametricheskoe vosstanovlenie funktsii plotnosti na osnove obobshchennogo lyambda-raspredeleniya v zadache identifikatsii regressionnykh modelei [Semiparametric reconstruction of the density function based on a generalized lambda-distribution in the problem of identification of regression models]. *Sibirskii zhurnal industrial'noi matematiki – Journal of Applied and Industrial Mathematics*. 2014, vol. 17, no. 3, pp. 71–77. (In Russian)
3. Timofeev V.S. Adaptivnoe vosstanovlenie regressionnykh zavisimostei na osnove poluparametricheskoi otsenki plotnosti sluchainoi komponenty [Adaptive construction of regression models based on semiparametric estimation of disturbance density function]. *Nauchnyi vestnik NGTU – Science Bulletin of Novosibirsk State Technical University*, 2013, no. 4 (53), pp. 24–30.
4. 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, pp. 25–36.
5. Aivazyan S.A. Enyukov I.S. Meshalkin L.D. *Prikladnaya statistika* [Applied Statistics]. Moscow, Finansy i statistika, 1985. 488 p.
6. Olkin I., Spiegelman C.H. A semiparametric approach to density estimation. *Journal of the American Statistical Association*, vol. 82, iss. 399, pp. 858–865.
7. Karian Z.A., Dudewicz E.J. *Fitting statistical distributions: the generalized lambda distribution and generalized bootstrap methods*. New York, CRC Press, 2000. 435 p.
8. Lakhany A., Mausser H. Estimation the parameters of the generalized lambda distribution. *ALGO Research Quarterly*, 2000, vol. 3, iss. 3, pp. 27–58.
9. Pagan A., Ullah A. *Nonparametric econometrics*. New York, Cambridge University Press, 1999. 424 p.
10. Bunday B.D. *Basic optimization methods. Introductory course*. London, Edward Arnold Publ., 1984 (Russ. ed.: Bandi B. *Metody optimizatsii. Vvodnyi kurs*. Moscow, Radio i svyaz', 1988. 128 p.).

СВЕДЕНИЯ ОБ АВТОРАХ



Тимофеев Владимир Семенович – родился в 1972 году, доктор технических наук, доцент, профессор, кафедра программных систем и баз данных, НГТУ. Область научных интересов: разработка и исследование устойчивых методов и алгоритмов анализа многофакторных объектов, в том числе с использованием непараметрической статистики. Опубликовано более 70 научных работ. (Адрес: 630073, Россия, Новосибирск, проспект К. Маркса, 20. Email: v.timofeev@corp.nstu.ru).

Timofeev Vladimir Semenovich (b. 1972) – D. Sc. (Eng.), Associate professor, professor, Department of Software Systems and Databases, NSTU. His research interests are currently focused on developing and investigating robust methods and algorithms for the analysis of multivariate objects using nonparametric statistics. He is author of more than 70 scientific papers. (Address: 20, K. Marx Prospect, Novosibirsk, 630073, Russia. Email: v.timofeev@corp.nstu.ru).



Хайленко Екатерина Алексеевна – родилась в 1985 году, кандидат технических наук, научный сотрудник, кафедра программных систем и баз данных, НГТУ. Область научных интересов: разработка и исследование алгоритмов устойчивого и адаптивного оценивания параметров регрессионных зависимостей и планирование эксперимента. Опубликовано 17 научных работ. (Адрес: 630073, Россия, Новосибирск, проспект К. Маркса, 20. Email: xajlenko@corp.nstu.ru).

Khailenko Ekaterina Alekseevna (b. 1985) – PhD (Eng.), research associate, Department of Software Systems and Databases, NSTU. Her research interests are currently focused on developing and investigating algorithms of robust and adaptive estimation parameters of regression models and design of experiment. He is author of 17 scientific papers. (Address: 20, K. Marx Prospect, Novosibirsk, 630073, Russia. Email: xajlenko@corp.nstu.ru).

*Статья поступила 13 ноября 2014 г.
Received November 13, 2014*

To Reference

Timofeev V.S., Khailenko E.A. Adaptivnoe otsenivanie parametrov regressionnykh zavisimostei pri neodnorodnosti sluchainykh oshibok [Adaptive estimation of regression model parameters with error destrebuton inhomogeneity]. *Doklady Akademii nauk vysshei shkoly Rossiiskoi Federatsii – Proceedings of the Russian higher school Academy of sciences*, 2014, no. 4, pp. 115–123.