Носков С. И., Чекалова А. Р.
S. I. Noskov, A. R. Chekalova

МИНИМИЗАЦИЯ РАССТОЯНИЯ МОДУЛЕЙ ОШИБОК АППРОКСИМАЦИИ РЕГРЕССИОННОЙ МОДЕЛИ ДО ИХ СРЕДНЕГО ЗНАЧЕНИЯ

MINIMIZATION OF THE DISTANCE OF REGRESSION MODEL APPROXIMATION ERROR MODULES TO THEIR MEAN VALUE

Носков Сергей Иванович – доктор технических наук, профессор, профессор кафедры «Информационные системы и защита информации» Иркутского государственного университета путей сообщения (Россия, Иркутск); Россия, 664074, ул. Чернышевского, д. 15; тел. 8(914)902-24-94. E-mail: sergey.noskov.57@mail.ru.
Sergey I. Noskov – Doctor of Technical Sciences, Professor, Professor of Information Systems and Information Protection Department, Irkutsk State Transport University (Russia, Irkutsk); Russia, 664074, st. Chernyshevsky, 15; tel. 8(914)902-24-94. E-mail: sergey.noskov.57@mail.ru.

Чекалова Александра Романовна – магистрант Иркутского государственного университета путей сообщения (Россия, Иркутск); Россия, 664074, ул. Чернышевского, д. 15; тел. 8(924)531-25-18. E-mail: chekalova49@gmail.com.
Aleksandra R. Chekalova – Master's Degree Student, Irkutsk State Transport University (Russia, Irkutsk); Russia, 664074, st. Chernyshevsky, 15; tel. 8(924)531-25-18. E-mail: chekalova49@gmail.com.

Аннотация. В работе описан алгоритмический способ уточнения оценок параметров линейного регрессионного уравнения, идентифицированных с помощью метода наименьших модулей, основанный на выравнивании модулей ошибок аппроксимации по отношению к их среднему значению. Сформулированная задача сводится к задаче линейного программирования приемлемой для практических ситуаций размерности. Разработана модель линейного тренда для описания динамики числа пользователей сети Интернет в мире методами наименьших квадратов и модулей, а также с использованием данного способа. Все три варианта трендовой модели обладают высоким качеством, на что указывают значения используемых критериев адекватности: множественной детерминации, Фишера, суммы модулей ошибок аппроксимации и их среднего значения.

Summary. The paper describes an algorithmic way to refine estimates of the parameters of a linear regression equation identified using the method of least modules, based on the alignment of the modules of approximation errors with respect to their average value. The formulated problem is reduced to a linear programming problem of a dimension acceptable for practical situations. A linear trend model has been developed to describe the dynamics of the number of Internet users in the world using least squares and modules methods, as well as using this method. All three variants of the trend model are of high quality, as indicated by the values of the adequacy criteria used: multiple determination, Fisher, the sum of the modules of approximation errors and their average value.

Ключевые слова: регрессионная модель, методы наименьших квадратов и модулей, ошибки аппроксимации, задача линейного программирования, тренд, число пользователей сети Интернет.

Key words: regression model, least squares and modulus methods, approximation errors, linear programming problem, trend, number of Internet users.

УДК 330.4

Введение. Рассмотрим проблему оценивания неизвестных параметров линейного регрессионного уравнения (модели) [1]:

$$y_k = \sum_{i=1}^{m} \alpha_i x_{ki} + \varepsilon_k, \quad k = \overline{1,n}, \tag{1}$$

где y, x_i – соответственно зависимая и i-я независимая переменные; α_i − i-й определяемый параметр; ε_k − ошибки аппроксимации; k − номер наблюдения; n − число наблюдений. Будем считать все переменные модели (1) детерминированными.

Уравнение (1) можно представить в векторной форме:

$$y = X\alpha + \varepsilon = \hat{y} + \varepsilon,$$

где $y = (y_1, \dots, y_n)^T$; $\alpha = (\alpha_1, \dots, \alpha_m)^T$; $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^T$; $X - (n \times m)$ – матрица с элементами x_{ki}; $\hat{y} = X\alpha$ – вектор расчётных значений зависимой переменной.

В регрессионном анализе при идентификации параметров модели (1) широко применяется метод наименьших модулей (МНМ), состоящий в минимизации городского (манхэттенского) расстояния $\rho(y, \hat{y})$ между фактическими (заданными) и расчётными значениями зависимой переменной, сводящийся к решению задачи

$$J(\alpha) = \sum_{k=1}^{n} |\varepsilon_k| \to \min. \quad (2)$$

Так, в работе [2] предлагаются средства построения модели нечёткой регрессии и исследуется её эффективность по отношению к определённой мере ошибки. Имитационные исследования и примеры показывают, что оценивание параметров модели с помощью МНМ даёт меньшую ошибку, чем модель нечёткой регрессии, изученная многими авторами, которые используют метод наименьших квадратов, когда данные содержат нечёткие выбросы. В [3] рассматривается задача формирования физических прототипов на основе трёхмерных моделей проектирования с использованием аддитивного процесса со слоями. При этом анализируется эффект лестницы между двумя последовательными слоями, а затем выводится формула прямого расчёта отклонения объёма всей модели. Вводится термин «взвешенная нормаль по площади», чтобы выразить значительное влияние площади фасета на объёмную ошибку, а задача определения оптимальной ориентации преобразуется в задачу построения линейной регрессии на основе применения МНМ. Статья [4] посвящена новому общему подходу к построению модели нечёткой регрессии, когда выходная переменная и параметры модели представляют собой нечёткие числа. При этом вводится новое определение целевой функции, основанное на различных функциях потерь. Применение предложенного подхода изучается с использованием смоделированного набора данных и некоторых реальных наборов при наличии различных типов выбросов, для обработки которых особенно эффективен МНМ. В работе [5] МНМ используется при исследовании вопроса формирования режима подзарядки группы автономных рабочих роботов в их рабочей среде с помощью регрессионных моделей. В [6] разрабатывается единый метод дисперсионного анализа на основе городского расстояния для проверки линейных гипотез. Как и классический дисперсионный анализ, этот метод является бескоординатным в том смысле, что он инвариантен при любом линейном преобразовании ковариат или параметров регрессии. Более того, он допускает использование единственных матриц проектирования и неоднородных ошибок. Предлагается простая аппроксимация с использованием стохастических возмущений для получения пороговых значений результирующей статистики испытаний. В статье [7] изучаются неточные данные с точки зрения не вполне определённых переменных и предлагается новый надёжный подход в соответствии с принципом наименьших модулей для оценки неизвестных параметров в неопределённых регрессионных моделях. Исследование [8] посвящено способам подавления влияния выбросов на параметры модели. Предлагается робастная регрессия опорного вектора на основе применения МНМ. Кроме того, для решения задачи оптимизации представлен эффективный алгоритм, основанный на методе разделения Брегмана. В [9] изучаются асимптотические свойства МНМ-оценок для моделей нелинейной регрессии. Даны простые достаточные условия сильной состоятельности и асимптотической нормальности оценок. Подтверждено, что распространение этих свойств на широкий класс функций регрессии можно установить, наложив некоторое условие на входные значения. Предлагается доверительная область, основанная на МНМ-оценках, и обсуждаются некоторые желательные

асимптотические свойства, включая асимптотическую относительную эффективность для различных распределений ошибок. В работе [10] рассматривается единая МНМ-оценка для стационарных и нестационарных дробно-интегрированных моделей авторегрессии скользящего среднего с условной гетероскедастичностью. Эксперименты подтверждают сделанные выводы, а результаты абсолютной доходности дневной цены закрытия промышленного индекса Доу-Джонса демонстрируют их полезность при моделировании временных рядов, учитывающих особенности долгой памяти, условной гетероскедастичности и тяжёлых хвостов.

Приближение модулей ошибок аппроксимации регрессионной модели к их среднему значению. Задача (2) сводится к следующей задаче линейного программирования (ЛП) (см., например, [11–13]):

$$\sum_{i=1}^{m} \alpha_i x_{ki} + u_k - v_k = y_k, \qquad k = \overline{1,n}, \tag{3}$$

$$u_k \geq 0,\ v_k \geq 0,\ k = \overline{1,n}, \tag{4}$$

$$J(\alpha) = \sum_{k=1}^{n} (u_k + v_k) \to \min, \tag{5}$$

где $u_k - v_k = \varepsilon_k$; $u_k + v_k = |\varepsilon_k|$. При этом после решения задачи ЛП (3) – (5) $u_k v_k = 0$, $k = \overline{1,n}$.

Одной из характеристик адекватности модели (1) является средняя абсолютная ошибка аппроксимации E:

$$E = \sum_{k=1}^{n} |\varepsilon_k| / n.$$

Важным свойством МНМ является равенство нулю m ошибок аппроксимации (см., например, [14]). Вызывает интерес задача выравнивания их абсолютных значений, «подтягивания» их к среднему значению E без существенного увеличения оптимального значения целевой функции (5). Сформулируем эту задачу формально.

Пусть J^* – оптимальное значение целевой функции в задаче ЛП (3) – (5). Обозначим через $\Delta J > 0$ величину, на которую исследователь может допустить некоторое увеличение значения J^* с тем, чтобы несколько приблизить модули ошибок аппроксимации к их среднему значению.

Сформируем множество $D(\alpha)$:

$$D(\alpha) = \{\alpha \in R^m \mid J(\alpha) \leq J^* + \Delta J\}.$$

Тогда указанная задача примет следующий формальный вид:

$$\min_{\alpha \in D(\alpha)} \sum_{k=1}^{n} \left| |\varepsilon_k| - \sum_{j=1}^{n} |\varepsilon_j| / n \right|. \tag{6}$$

Задача (6), так же как и (2), может быть сведена к задаче ЛП. Действительно, дополним ограничения (3), (4) следующими:

$$\sum_{k=1}^{n} (u_k + v_k) \leq J^* + \Delta J, \tag{7}$$

$$u_k + v_k + c_k - d_k = \sum_{j=1}^{n} \frac{(u_j + v_j)}{n}, \qquad k = \overline{1,n}, \tag{8}$$

$$c_k \geq 0, \qquad d_k \geq 0, \qquad k = \overline{1,n}. \tag{9}$$

Целевая функция примет вид

$$h\sum_{k=1}^{n}(u_k + v_k) + \sum_{k=1}^{n}(c_k + d_k) \to \min, \qquad (10)$$

где h – малая положительная константа.

Применим данный способ выравнивания ошибок к линейному тренду, описывающему динамику числа пользователей сети Интернет в мире в миллионах человек (зависимая переменная *y*). В табл. 1 представлена статистика по этому показателю за 2003-2022 гг. [15; 16].

Таблица 1

Исходные данные

Год	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
y	668	784	917	1040	1166	1383	1578	1763	1990	2177
Год	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022
y	2431	2692	2916	3282	3640	3950	4212	4418	4758	5385

Таким образом, будем строить линейную регрессионную модель:

$$y_t = \alpha_0 + \alpha_1\, t + \varepsilon_t, \qquad t = \overline{1,20}. \qquad (11)$$

Вначале оценим параметры тренда (11) с помощью метода наименьших квадратов (МНК):

$$y_t = 16 + 242\, t + \varepsilon_t, \qquad t = \overline{1,20}, \qquad (12)$$

$R = 0{,}98$, $F = 732{,}9$, $E = 178{,}3$, $J = 3566{,}48$,

где *R* – критерий множественной детерминации, *F* – критерий Фишера. Значения критериев адекватности указывают на высокое качество модели (12).

Теперь оценим параметры модели (11) с помощью МНМ:

$$y_t = -71 + 247{,}4\, t + \varepsilon_t, \qquad t = \overline{1,20}, \qquad (13)$$

где $E = 172{,}3$, $J = 3446$.

Значения критериев *R* и *F* здесь не приведены, т. к. их использование для МНМ не является корректным.

Сумма ошибок аппроксимации *J* для модели (13) меньше на 120,48 (т. е. на 3,4 %), чем для модели (12). Поэтому назначим:

$$\Delta J = 120{,}48.$$

Наконец, оценим параметры модели (11) описанным выше способом путём решения задачи ЛП (3), (4), (7) – (10):

$$y_t = -4 + 241{,}4\, t + \varepsilon_t, \qquad t = \overline{1,20}, \qquad (14)$$

$E = 172{,}3$, $J = 3566{,}48$.

Вполне ожидаемо, значения *E* и *J* моделей (12) и (14) совпадают. Вместе с тем значение свободного члена α_0 в модели (14) меньше, чем для модели (12), но больше, чем для модели (13). А вот значение углового коэффициента α_1 в модели (14) наименьшее из всех трёх моделей.

Окончательное решение по поводу того, какую именно модель из трёх построенных использовать, должен принять исследователь в зависимости от характера решаемой прогнозной и/или аналитической задачи, а также от своих индивидуальных опыта и предпочтений.

Заключение. В работе предложен способ уточнения значений параметров линейной регрессионной модели, вычисленных с помощью метода наименьших модулей, основанный на приближении модулей ошибок аппроксимации к их среднему значению. Соответствующая задача сводится к задаче линейного программирования. Построена модель линейного тренда для описа-

ния динамики числа пользователей сети Интернет в мире методами наименьших квадратов и модулей, а также с использованием данного способа.

ЛИТЕРАТУРА

1. Носков, С. И. Применение многокритериального метода наименьших модулей для моделирования количества дорожно-транспортных происшествий / С. И. Носков // Учёные записки Комсомольского-на-Амуре государственного технического университета. Науки о природе и технике. – 2023. – № V (69). – С. 30-35.
2. Seung Hoe Choi, Buckley J. J. Fuzzy regression using least absolute deviation estimators // Soft Computing. – 2008. – V. 12. – P. 257-263.
3. Nan Luo, Quan Wang. Fast slicing orientation determining and optimizing algorithm for least volumetric error in rapid prototyping // The International Journal of Advanced Manufacturing Technology. – 2016. – V. 83. – P. 1297-1313.
4. Khammar A. H., Arefi M., Akbari M. G. A general approach to fuzzy regression models based on different loss functions // Soft Computing. – 2021. – V. 25. – P. 835-849.
5. Keshmiri S., Payandeh S. Regression Analysis of Multi-Rendezvous Recharging Route in Multi-Robot Environment // International Journal of Social Robotics. – 2012. – V. 4. – P. 15-27.
6. Kani Chen, Zhiliang Ying, Hong Zhang, Lincheng Zhao. Analysis of least absolute deviation // Biometrika. – 2008. – V. 95. – P. 107-122.
7. Zhe Liu, Ying Yang. Least absolute deviations estimation for uncertain regression with imprecise observations // Fuzzy Optimization and Decision Making. – 2020. – V. 19. – P. 33-52.
8. Chen Chuanfa, Li Yanyan, Yan Changqing, Guo Jinyun, Liu Guolin. Least absolute deviation-based robust support vector regression // Knowledge-Based Systems. – 2017. – V. 131. – P. 183-194.
9. Kim Hae Kyung, Park Seung. Hoe Asymptotic Properties of Nonlinear Least Absolute Deviation Estimators // Journal of the Korean Statistical Society. – 1995. – V. 24. – P. 127-139.
10. Guodong Li, Wai Keung Li. Least absolute deviation estimation for fractionally integrated autoregressive moving average time series models with conditional heteroscedasticity // Biometrika. – 2008. – V. 95 – P. 399-414.
11. Носков, С. И. Метод антиробастного оценивания параметров линейной регрессии: число максимальных по модулю ошибок аппроксимации / С. И. Носков // Южно-Сибирский научный вестник. – 2020. – № 1 (29). – С. 51-54.
12. Носков, С. И. L-множество в многокритериальной задаче оценивания параметров регрессионных уравнений / С. И. Носков // Информационные технологии и проблемы математического моделирования сложных систем. – 2004. – № 1. – С. 164-171.
13. Носков, С. И. Обобщённый критерий согласованности поведения в регрессионном анализе / С. И. Носков // Информационные технологии и математическое моделирование в управлении сложными системами. – 2018. – № 1 (1). – С. 14-20.
14. Носков, С. И. О кластеризации данных на основе свойств методов идентификации параметров линейной регрессии / С. И. Носков // Информационные технологии и математическое моделирование в управлении сложными системами. – 2022. – № 4 (16). – С. 82-85.
15. Статистика пользователей интернета в мире в 2022 // Seo блог Алексея Файнгора, 2024. – URL: https://fayngor.ru/blog/statistika-polzovatelej-interneta-v-mire-v-2022/ (дата обращения: 04.11.2023). – Текст: электронный.
16. Мировые пользователи Интернета и статистика населения в 2023 году // Internet World Stats: мировая интернет-статистика, сайт. – URL: https://www.internetworldstats.com/stats.htm (дата обращения: 04.11.2023). – Текст: электронный.