УДК 519.862.6 DOI: 10.26102/2310-6018/2020.31.4.026

Оценивание линейно-неэлементарных регрессионных моделей с помощью метода наименьших квадратов

М.П. Базилевский
Федеральное государственное бюджетное образовательное учреждение высшего образования «Иркутский государственный университет путей сообщения», Иркутск, Российская Федерация

Резюме: При построении регрессионных моделей проблема выбора структурной спецификации является первостепенной. На сегодняшний день таких спецификаций разработано великое множество. В данной работе приводится краткое описание следующих форм связи между переменными: линейная регрессия, линейно-элементарная регрессия, линейно-мультипликативная регрессия, производственная функция Леонтьева и индексная регрессия. Благодаря смешению линейной и кусочно-линейной регрессий сформулирована новая спецификация – линейно-неэлементарная регрессия, регрессорами в которой являются как входные переменные, так и бинарные операции всех возможных комбинаций их пар. Показано, что присваивание определенным параметрам таких моделей конкретных значений делает их квазилинейными, что позволяет оценивать их с помощью метода наименьших квадратов. Установлены области определения этих параметров. Разработан алгоритм приближенного оценивания линейно-неэлементарных регрессий с помощью метода наименьших квадратов. Работа алгоритма продемонстрирована на примере моделирования потребления электроэнергии в Иркутской области. Качество построенной линейно-неэлементарной регрессии по коэффициенту детерминации оказалась выше, чем у полученных ранее моделей. Показано, что в линейно-неэлементарных регрессиях с течением времени меняется характер влияния входных переменных на выходную.

Ключевые слова: регрессионная модель, производственная функция Леонтьева, линейно-неэлементарная регрессия, метод наименьших квадратов, потребление электроэнергии.

Для цитирования: Базилевский М.П. Оценивание линейно-неэлементарных регрессионных моделей с помощью метода наименьших квадратов. *Моделирование, оптимизация и информационные технологии.* 2020;8(4). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=872 DOI: 10.26102/2310-6018/2020.31.4.026

Estimation linear non-elementary regression models using ordinary least squares

M.P. Bazilevskiy
Federal State Budgetary Educational Institution of Higher Education "Irkutsk State Transport University", Irkutsk, Russian Federation

Abstract: When constructing regression models, the problem of choosing a structural specification is paramount. To date, a great variety of such specifications have been developed. This paper provides a brief description of the following forms of relationship between variables: linear regression, linear elementary regression, linear multiplicative regression, Leontief production function, and index regression. Due to the mixing of linear and piecewise linear regressions, a new specification has been formulated - linear non-elementary regression, in which regressors are both input variables and binary operations of all possible combinations of their pairs. It is shown that the assignment of specific values to certain parameters of such models makes them quasilinear, which makes it possible to estimate them

using the ordinary least squares. Areas of definition of these parameters are established. An algorithm for approximate estimation of linear non-elementary regressions using the ordinary least squares is developed. The operation of the algorithm is demonstrated by the example of modeling electricity consumption in the Irkutsk region. The quality of the constructed linear non-elementary regression by the coefficient of determination turned out to be higher than that of the previously obtained models. It is shown that in linear non-elementary regressions, the nature of the influence of input variables on the output changes over time.

Keywords: regression model, Leontief production function, linear non-elementary regression, ordinary least squares, electricity consumption.

For citation: Bazilevskiy M.P. Estimation linear non-elementary regression models using ordinary least squares. *Modeling, Optimization and Information Technology*. https://moitvivt.ru/ru/journal/pdf?id=872 DOI: 10.26102/2310-6018/2020.31.4.026 (In Russ).

Введение

При построении регрессионной модели ключевой проблемой является выбор её структурной спецификации, т.е. состава переменных и математической формы связи между ними. Описание большинства таких спецификаций можно найти в работах [1-10].

Чаще всего регрессионный анализ начинается с оценивания самой простой зависимости – модели множественной линейной регрессии:

$$y_i = \alpha_0 + \sum_{j=1}^{m} \alpha_j x_{ij} + \varepsilon_i, \qquad i = \overline{1,n}, \tag{1}$$

где y_i, $i = \overline{1,n}$ – наблюдаемые значения объясняемой (выходной) переменной y; x_{ij}, $i = \overline{1,n}$, $j = \overline{1,m}$ – наблюдаемые значения объясняющих (входных) переменных x_1, x_2, …, x_m; ε_i, $i = \overline{1,n}$ – ошибки аппроксимации; α_0, α_1, α_2, …, α_m – неизвестные параметры.

В [3] рассмотрены вопросы построения линейно-элементарных регрессий (ЛЭР) вида

$$y_i = \alpha_0 + \sum_{k=1}^{l} \sum_{j=1}^{m} \alpha_{kj} f_k\left(x_{ij}\right) + \varepsilon_i, \quad i = \overline{1,n}, \tag{2}$$

в которых в качестве преобразований f_k, $k = \overline{1,l}$ выступают элементарные функции.

В [5,6] приводится описание линейно-мультипликативных регрессий (ЛМР) вида

$$y_i = \alpha_0 + \sum_{j=1}^{2^m-1} \alpha_j \prod_{k=1}^{m} x_{ik}^{\lambda_{jk}} + \varepsilon_i, \qquad i = \overline{1,n}, \tag{3}$$

где λ_{jk} – элементы бинарной матрицы $\Lambda_{\left(2^m-1\right)\times m}$, строками которой являются размещения с повторениями из m элементов по 2 (без нулевой строки).

Модели (1) – (3) являются линейными по параметрам, поэтому легко оцениваются с помощью метода наименьших квадратов (МНК).

В [3,7,8] исследованы кусочно-линейные регрессионные модели (производственные функции Леонтьева [4]) вида

$$y_i = \min\left\{\alpha_1 x_{i1}, \alpha_2 x_{i2}, ..., \alpha_m x_{im}\right\} + \varepsilon_i, \qquad i = \overline{1,n}, \tag{4}$$

а в [9] предложено их обобщение – индексные регрессии:

$$y_i = \alpha_0 + \mathrm{ind}_G\left\{\alpha_1 x_{i1}, \alpha_2 x_{i2}, ..., \alpha_m x_{im}\right\} + \varepsilon_i, \qquad i = \overline{1,n}, \tag{5}$$

где G – заданный индексный вектор.

Модели (4) и (5) принято оценивать с помощью метода наименьших модулей (МНМ). Успешная попытка оценивания кусочно-линейных регрессий была предпринята в работе [10], в которой рассматривалась двухфакторная спецификация вида

$$y_i = \alpha_0 + \min\{\alpha_1 x_{i1}, \alpha_2 x_{i2}\} + \varepsilon_i, \qquad i = \overline{1,n}. \tag{6}$$

Считая, что переменные x_1 и x_2 строго положительны, модель (6) можно представить в виде

$$y_i = \alpha_0 + \alpha_1 \min\{x_{i1}, \lambda x_{i2}\} + \varepsilon_i, \qquad i = \overline{1,n}, \tag{7}$$

где $\lambda = \alpha_2 / \alpha_1$.

Тогда, придавая параметру λ конкретные значения из некоторого интервала, можно легко определить МНК-оценки параметров α_0 и α_1 регрессии (7). В работе [10] установлено, что спецификация модели (7) зависит от величины параметра λ следующим образом:

$$y = \begin{cases} \alpha_0 + \alpha_1(\lambda x_2) + \varepsilon, & \text{при } \lambda \in (0, \lambda_{\min}], \\ \alpha_0 + \alpha_1 \min\{x_1, \lambda x_2\} + \varepsilon, & \text{при } \lambda \in (\lambda_{\min}, \lambda_{\max}), \\ \alpha_0 + \alpha_1 x_1 + \varepsilon, & \text{при } \lambda \in [\lambda_{\max}, \infty), \end{cases}$$

где $\lambda_{\min} = \min\left\{\frac{x_{11}}{x_{12}}, \frac{x_{21}}{x_{22}}, ..., \frac{x_{n1}}{x_{n2}}\right\}$, $\lambda_{\max} = \max\left\{\frac{x_{11}}{x_{12}}, \frac{x_{21}}{x_{22}}, ..., \frac{x_{n1}}{x_{n2}}\right\}$.

Из этого следует, что для приближенного МНК-оценивания регрессии (7) достаточно перебрать значения параметра λ из интервала $(\lambda_{\min}, \lambda_{\max})$.

Целью данной работы является разработка на основе смешения конструкций (1) и (6) новой спецификации регрессионных моделей и алгоритма приближенного оценивания её неизвестных параметров с помощью МНК.

Линейно-неэлементарные регрессионные модели

Рассмотрим бинарную операцию (т.е. над двумя числами) минимум $\min\{a,b\}$, которая возвращает наименьшее из двух чисел a и b. Сформируем модель, в которой помимо регрессоров x_1, x_2, ..., x_m будем использовать бинарные операции всех возможные комбинаций пар этих переменных:

$$y_i = \alpha_0 + \sum_{j=1}^{m} \alpha_j x_{ij} + \sum_{j=1}^{C_m^2} \alpha_{j+m} \min\left\{x_{i,\mu_{j1}}, \lambda_j x_{i,\mu_{j2}}\right\} + \varepsilon_i, \qquad i = \overline{1,n}, \tag{8}$$

где μ_{j1} и μ_{j2}, $j = \overline{1, C_m^2}$ – элементы первого и второго столбца матрицы $\mathrm{M}_{C_m^2 \times 2}$ пар индексов переменных.

Например, если $m = 3$, то регрессионная модель (8) имеет вид:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \alpha_3 x_{i3} + \\ + \alpha_4 \min\{x_{i1}, \lambda_1 x_{i2}\} + \alpha_5 \min\{x_{i1}, \lambda_2 x_{i3}\} + \alpha_6 \min\{x_{i2}, \lambda_3 x_{i3}\} + \varepsilon_i, \qquad i = \overline{1,n},$$

а матрица $\mathrm{M}_{3\times 2} = \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 2 & 3 \end{pmatrix}$.

Будем считать, что значения всех объясняющих переменных, входящих в регрессионную модель (8), являются положительными, т.е. $x_{ij} > 0$, $i = \overline{1,n}$, $j = \overline{1,m}$.

Как видно, регрессионная модель (8) является расширенной версией, как линейной регрессии (1), так и модели (7). Вообще говоря, если в регрессии (8) параметры λ_j, $j = \overline{1, C_m^2}$ неизвестны, то она является в значительной степени нелинейной и её оценивание становится чрезвычайно сложной задачей. Однако если эти коэффициенты заданы, то модель (8) представляет собой квазилинейную регрессию только с неизвестными параметрами α_0, α_1, ..., $\alpha_{m+C_m^2}$, для оценивания которых можно применить обычный МНК. Тогда с помощью простого перебора значений коэффициентов λ_j, $j = \overline{1, C_m^2}$ можно приближенно получить оценки нелинейной модели (8).

Если в выражении (8) значения параметров λ_j, $j = \overline{1, C_m^2}$, фиксированы, то полученную квазилинейную регрессию будем называть линейно-неэлементарной регрессией (ЛНР), поскольку входящие в неё бинарные операции минимум являются неэлементарными функциями.

Стоит отметить, что, используя в ЛНР (8) вместо бинарных операций минимум бинарные операции максимум, можно записать следующую спецификацию ЛНР:

$$y_i = \alpha_0 + \sum_{j=1}^{m} \alpha_j x_{ij} + \sum_{j=1}^{C_m^2} \alpha_{j+m} \max\left\{x_{i,\mu_{j1}}, \lambda_j x_{i,\mu_{j2}}\right\} + \varepsilon_i, \qquad i = \overline{1,n}. \tag{9}$$

Алгоритм приближенного МНК-оценивания

Для приближенного оценивания моделей (8) и (9) воспользуемся результатами, полученными для регрессий (7). Как отмечено выше, в моделях (7) нет смысла перебирать все возможные значения параметра, а нужно ограничиться лишь промежутком $\lambda \in [\lambda_{\min}, \lambda_{\max}]$, где $\lambda_{\min} = \min\left\{\frac{x_{11}}{x_{12}}, ..., \frac{x_{n1}}{x_{n2}}\right\}$, $\lambda_{\max} = \max\left\{\frac{x_{11}}{x_{12}}, ..., \frac{x_{n1}}{x_{n2}}\right\}$. Данный результат справедлив и для моделей (8), за исключением того, что нельзя использовать точки на концах отрезка $[\lambda_{\min}, \lambda_{\max}]$ из-за возникновения совершенной мультиколлинеарности факторов. Тогда сформируем следующие интервалы изменения параметров λ_j, $j = \overline{1, C_m^2}$:

$$\lambda_j \in \left(\lambda_{\min}^{(j)}, \lambda_{\max}^{(j)}\right), \tag{10}$$

где $\lambda_{\min}^{(j)} = \min\left\{\frac{x_{1,\mu_{j1}}}{x_{1,\mu_{j2}}}, \frac{x_{2,\mu_{j1}}}{x_{2,\mu_{j2}}}, ..., \frac{x_{n,\mu_{j1}}}{x_{n,\mu_{j2}}}\right\}$, $\lambda_{\max}^{(j)} = \max\left\{\frac{x_{1,\mu_{j1}}}{x_{1,\mu_{j2}}}, \frac{x_{2,\mu_{j1}}}{x_{2,\mu_{j2}}}, ..., \frac{x_{n,\mu_{j1}}}{x_{n,\mu_{j2}}}\right\}$.

Пусть для каждого промежутка (10) задано одинаковое число точек его разбиения – k. Тогда для приближенного оценивания регрессионной модели (8) необходимо с помощью МНК идентифицировать $k^{C_m^2}$ штук ЛНР и выбрать из них лучшую на основании одного или нескольких критериев адекватности. Алгоритм такого оценивания представлен на Рисунке 1.

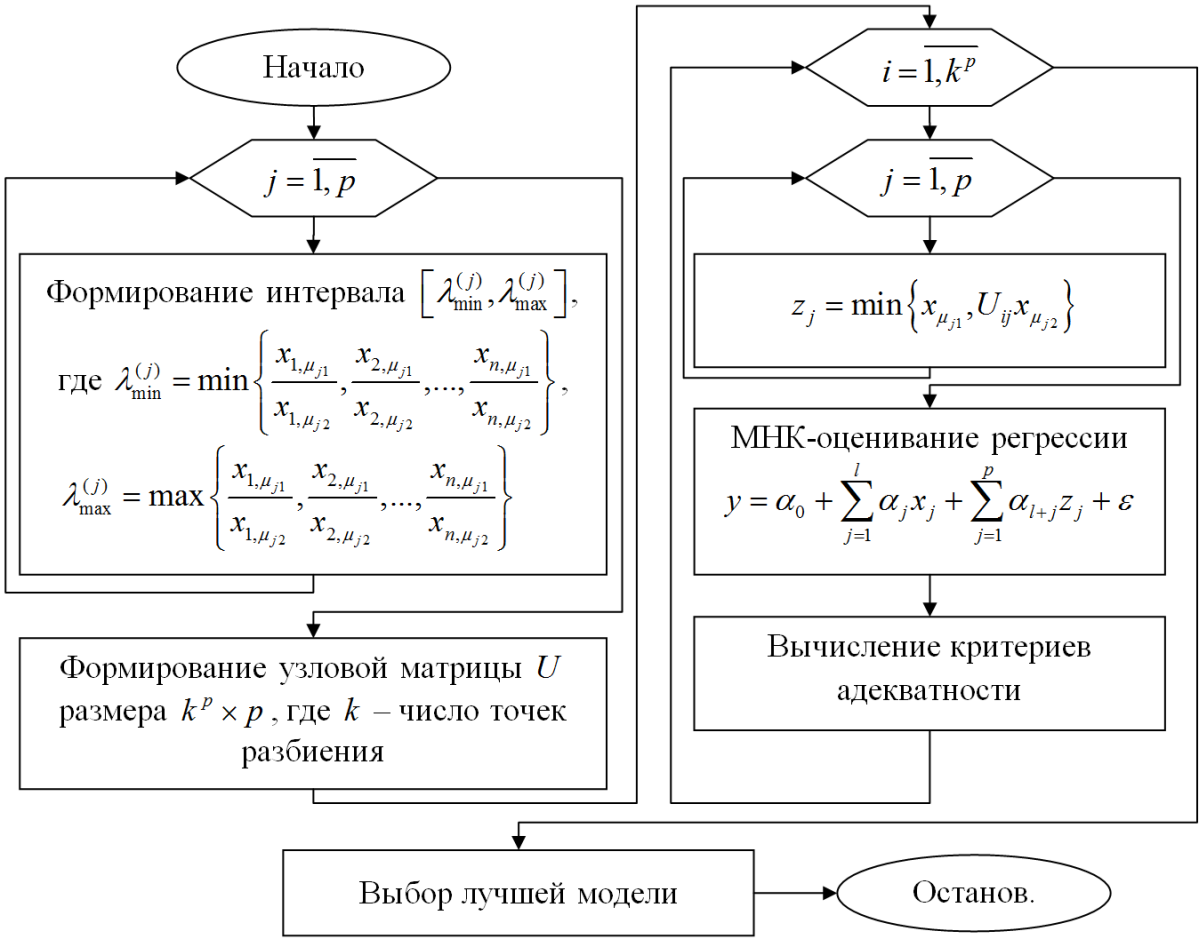

Рисунок 1 – Блок-схема алгоритма приближенного оценивания
Figure 1 – Block diagram of the approximate estimation algorithm

На Рисунке 1 переменная $p=C_m^2$.

Пример построения ЛНР

В работе [10] приведены статистические данные, которые были использованы для моделирования потребления электроэнергии в Иркутской области. Эти данные собраны для следующих переменных:

y – потребление электроэнергии, млрд кВт*ч;

x_1 – валовой региональный продукт, млрд. руб.;

x_2 – строительство жилых домов, тыс. кв. м.

Оцененная по этим данным с помощью МНК модель множественной линейной регрессии имеет вид

$$\tilde{y}=49,793+0,00395x_1+0,00468x_2, \quad (11)$$

а оцененная модель (7):

$$\tilde{y}=49,691+0,0103\min\left\{x_1,0.8473x_2\right\}. \quad (12)$$

Аппроксимационное качество регрессия (12), для которой величина коэффициента детерминации равна 0,7686, оказалось выше, чем у регрессии (11), для которой $R^2=0,732$.

На основе этих данных проводилось МНК-оценивание ЛНР вида

$$y_i = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 \min\{x_1, \lambda x_2\}, \qquad i = \overline{1,n}.$$

Для этого в эконометрическом пакете Gretl был разработан специальный скрипт. Число k задавалось равным 100. Было найдено, что $\lambda_{\min} = 0,7003$, а $\lambda_{\max} = 1,2657$. На основе представленного на Рисунке 1 алгоритма была найдена зависимость суммы квадратов ESS регрессии от величины λ, которая изображена на Рисунке 2.

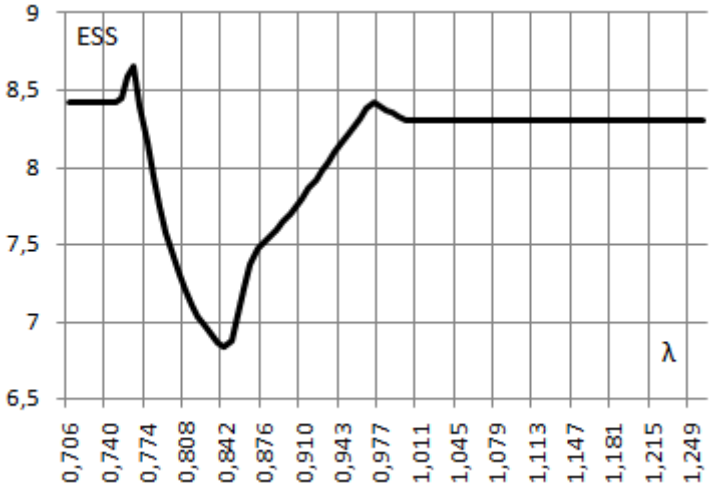

Рисунок 2 – Зависимость ESS от λ

Figure 2 – Dependence between ESS and λ

По этому рисунку видно, что существует единственная точка глобального минимума при $\lambda = 0,842$. В этой точке ESS=6,834, а оцененная ЛНР имеет вид

$$\tilde{y} = 50,2408 + 0,000689 x_1 - 0,00975 x_2 + 0,0204 \min\{x_1, 0.842 x_2\}. \qquad (13)$$

Коэффициент детерминации регрессии (13) составляет 0,801, т.е. её качество ожидаемо выше, чем у моделей (11) и (12).

Заметим, что регрессионную модель (13) можно записать в следующей кусочно-заданной форме:

$$\tilde{y} = \begin{cases} 50,2408 + 0,02116 x_1 - 0,00975 x_2, & \text{при } x_1 / x_2 < 0,842, \\ 50,2408 + 0,000689 x_1 + 0,00748 x_2, & \text{при } x_1 / x_2 \geq 0,842. \end{cases} \qquad (14)$$

Как видно по (14), в результате оценивания ЛНР была определена точка переключения $x_1 / x_2 = 0,842$, при переходе через которую множественная регрессия меняет свое аналитическое выражение (коэффициенты при объясняющих переменных). Таким образом, в ЛНР, в отличие от классических линейных регрессий, с течением времени меняется характер влияния входных переменных на выходную.

Заключение

В данной работе рассмотрены основные структурные спецификации регрессионных моделей. На основе смешения линейных регрессий и производственных функций Леонтьева впервые предложены линейно-неэлементарные модели. Разработан алгоритм приближенного МНК-оценивания линейно-неэлементарных регрессий. Проведено моделирование потребления электроэнергии в Иркутской области. Линейно-неэлементарные регрессионные модели могут успешно применяться как для

прогнозирования значений выходной переменной, так и для содержательной интерпретации функционирования исследуемых объектов или процессов.

ЛИТЕРАТУРА

1. Harrell Jr., Frank E. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer Series in Statistics. 2015.
2. Kuhn M., Johnson K. Applied predictive modeling. Springer. 2018.
3. Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. Иркутск : Облинформпечать. 1996.
4. Клейнер Г.Б. Производственные функции: Теория, методы, применение. М.: Финансы и статистика. 1986.
5. Базилевский М.П. Программный комплекс построения линейно-мультипликативных регрессий. *Прикладная информатика*. 2018;3(75): 110-123.
6. Базилевский М.П., Носков С.И. Формализация задачи построения линейно-мультипликативной регрессии в виде задачи частично-булевого линейного программирования. *Современные технологии. Системный анализ. Моделирование*. 2017;3(55): 101-105.
7. Иванова Н.К., Лебедева С.А., Носков С.И. Идентификация параметров некоторых негладких регрессий. *Информационные технологии и проблемы математического моделирования сложных систем*. 2016;17: 107-110.
8. Носков С.И., Хоняков А.А. Программный комплекс построения некоторых типов кусочно-линейных регрессий. *Информационные технологии и математическое моделирование в управлении сложными системами*. 2019;3(4): 47-55.
9. Базилевский М.П., Носков С.И. Оценивание индексных моделей регрессии с помощью метода наименьших модулей. *Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление*. 2020;1: 17-23.
10. Базилевский М.П. МНК-оценивание параметров специфицированных на основе функций Леонтьева двухфакторных моделей регрессии. *Южно-Сибирский научный вестник*. 2019; 2(26): 66-70.

REFERENCES

1. Harrell Jr., Frank E. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer Series in Statistics. 2015.
2. Kuhn M., Johnson K. Applied predictive modeling. Springer. 2018.
3. Noskov S.I. Tekhnologiya modelirovaniya ob"ektov s nestabil'nym funktsionirovaniem i neopredelennost'yu v dannykh. Irkutsk: RIC GP «Oblinformpechat'» Publ. 1996.
4. Kleyner G.B. Proizvodstvennye funktsii: Teoriya, metody, primenenie. Moscow: Finance and Statistics Publ. 1986.
5. Bazilevskiy M.P. Programmnyj kompleks postroenija linejno-mul'tiplikativnyh regressij. *Prikladnaja informatika*. 2018;3(75):110-123.
6. Bazilevskiy M.P., Noskov S.I. Formalizacija zadachi postroenija linejno-mul'tiplikativnoj regressii v vide zadachi chastichno-bulevogo linejnogo programmirovanija. *Sovremennye tehnologii. Sistemnyj analiz. Modelirovanie*. 2017;3(55):101-105.

7. Ivanova N.K., Lebedeva S.A., Noskov S.I. Identifikacija parametrov nekotoryh negladkih regressij. *Informacionnye tehnologii i problemy matematicheskogo modelirovanija slozhnyh sistem*. 2016;17:107-110.
8. Noskov S.I., Honyakov A.A. Programmnyj kompleks postroenija nekotoryh tipov kusochno-linejnyh regressij. *Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami*. 2019;3(4):47-55.
9. Bazilevskiy M.P., Noskov S.I. Ocenivanie indeksnyh modelej regressii s pomoshh'ju metoda naimen'shih modulej. *Vestnik Rossijskogo novogo universiteta. Serija: Slozhnye sistemy: modeli, analiz i upravlenie*. 2020;1:17-23.
10. Bazilevskiy M.P. MNK-ocenivanie parametrov specificirovannyh na osnove funkcij Leont'eva dvuhfaktornyh modelej regressii. *Juzhno-Sibirskij nauchnyj vestnik*. 2019; 2(26): 66-70.

ИНОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Базилевский Михаил Павлович, канд. техн. наук, доцент, кафедра математики, ФГБОУ ВО "Иркутский государственный университет путей сообщения", Иркутск, Российская Федерация.
e-mail: mik2178@yandex.ru
ORCID: 0000-0002-3253-5697

Mikhail P. Bazilevskiy, PhD (Tech.), Associate Professor, Mathematics Department, Federal State Budgetary Educational Institution of Higher Education "Irkutsk State Transport University", Irkutsk, Russian Federation