КОМПЬЮТЕРНОЕ МОДЕЛИРОВАНИЕ
COMPUTER SIMULATION HISTORY

УДК 519.862.6
DOI 10.52575/2687-0932-2023-50-2-367-379

Оценивание параметров неэлементарных линейных регрессий методом наименьших квадратов

Базилевский М.П.
Иркутский государственный университет путей сообщения,
Россия, 664074, Иркутская область, г. Иркутск, ул. Чернышевского, д. 15
E-mail: mik2178@yandex.ru

Аннотация. Статья посвящена проблеме оценивания неэлементарных линейных регрессий методом наименьших квадратов. Предложено их обобщение – неэлементарные линейные регрессии с линейным аргументом в бинарной операции. Рассмотрено три вида неэлементарных линейных регрессий: с произвольными угловыми коэффициентами в бинарных операциях, с единичными угловыми коэффициентами и произвольными свободными членами, с произвольными угловыми коэффициентами и свободными членами. Для каждого вида предложен алгоритм численного оценивания. На основе алгоритмов разработана программа численного оценивания неэлементарных линейных регрессий методом наименьших квадратов (НЕЭЛИН). Рассмотрен алгоритм работы программы, в котором параметры моделей хранятся в виде трехмерного массива данных. Показано, как нужно задавать структуру модели в НЕЭЛИН. С помощью программы НЕЭЛИН проведено моделирование железнодорожных грузовых перевозок Республики Башкортостан. Полученная в результате неэлементарная линейная регрессия оказалась лучше по качеству, чем модели, полученные традиционными инструментами.

Ключевые слова: неэлементарная линейная регрессия, бинарная операция, метод наименьших квадратов, численное оценивание, программное обеспечение, мультиколлинеарность, интерпретация, железнодорожные грузоперевозки

Для цитирования: Базилевский М.П. 2023. Оценивание параметров неэлементарных линейных регрессий методом наименьших квадратов. Экономика. Информатика, 50(2): 367–379. DOI: 10.52575/2687-0932-2023-50-2-367-379

Estimation Non-Elementary Linear Regressions Parameters Using Ordinary Least Squares Method

Mikhail P. Bazilevskiy
Irkutsk State Transport University
15 Chernyshevskogo St, Irkutsk, 664074, Russia
E-mail: mik2178@yandex.ru

Abstract. This article is devoted to the problem of non-elementary linear regressions estimation by the ordinary least squares method. Their generalization is proposed - non-elementary linear regressions with a linear argument in a binary operation. Three types of non-elementary linear regressions are considered - with arbitrary slopes in binary operations, with unit slopes and arbitrary free terms, and with arbitrary slopes and free terms. For each type, a numerical estimation algorithm is proposed. Based on the algorithms, a program for the numerical estimation of non-elementary linear regressions by the ordinary least squares method (NEELIN) was developed. The algorithm of the program operation, in which the parameters of the models are stored in the form of a three-dimensional data array, is considered. It is shown how to set the model structure

in NEELIN. With the help of the NEELIN program, modeling of rail freight traffic in the Republic of Bashkortostan was carried out. The resulting non-elementary linear regression proved to be better in quality than the models generated by traditional tools.

Keywords: non-elementary linear regression, binary operation, ordinary least squares, numerical estimation, software, multicollinearity, interpretation, rail freight

For citation: Bazilevskiy M.P. 2023. Estimation Non-Elementary Linear Regressions Parameters Using Ordinary Least Squares Method. Economics. Information technologies, 50(2): 367–379 (in Russian). DOI: 10.52575/2687-0932-2023-50-2-367-379

Введение

Методы регрессионного анализа [Keith, 2019; Gelman et al., 2020] активно развиваются в настоящее время и находят широкое применение в различных отраслях науки. Так, в работе [Karakurt, Aydin, 2023] предложены регрессионные модели для прогнозирования выбросов углекислого газа в странах БРИКС, в [Wang et al., 2022] исследуются характеристики различных регрессионных моделей для диагностики рака молочной железы по данным РНК-секвенирования, в [Luo et al., 2022] предложены модели прогнозирования электрической нагрузки в энергетической отрасли, в [Tian et al., 2022] проведен анализ распространения риска в финансовом секторе Китая на основе новой квантильной копулярной GARCH регрессии, в [Gao et al., 2022] предлагается географически взвешенная регрессионная модель зависимости особенностей застроенной среды от пассажиропотока метро в городе Шэньчжэнь.

Как известно, процесс построения регрессионной модели состоит из нескольких этапов. Возможно, ключевым из них является этап выбора структурной спецификации регрессии – состава переменных и математической формы связи между ними. Спецификаций регрессионных моделей существует много (см., например, [Клейнер, 1986; Базилевский, Носков, 2017; Keith, 2019; Носков, Хоняков, 2019; Gelman et al., 2020]). Универсального метода выбора лучшей из них нет. К тому же новой тенденцией в машинном обучении становится построение интерпретируемых моделей [Molnar, 2020; Du et al., 2019]. Так, в [Letzgus et al., 2022] рассмотрены методы объяснимого искусственного интеллекта для регрессионных моделей.

Проблема поиска новых спецификаций регрессионных моделей с интересными свойствами актуальна и по сей день. В работах [Базилевский, 2020, 2021] автором предложены и исследованы неэлементарные линейные регрессии (НЛР) с бинарными операциями min, а в [Базилевский, 2022а] введена НЛР с бинарными операциями min и max:

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \sum_{j=1}^{p} \alpha_j^{\min} \min\{x_{i,\mu_{j1}}, k_j^{\min} x_{i,\mu_{j2}}\} + \\ + \sum_{j=1}^{p} \alpha_j^{\max} \max\{x_{i,\mu_{j1}}, k_j^{\max} x_{i,\mu_{j2}}\} + \varepsilon_i, \quad i = \overline{1,n}, \tag{1}$$

где n – объем выборки; l – число входных переменных; y_i – i-е значение выходной переменной; x_{ij} – i-е значение j-й входной переменной; min (max) – бинарные операции, возвращающие минимум (максимум) двух чисел; $p = C_l^2$ – число всех возможных комбинаций пар входных переменных; μ_{j1}, μ_{j2}, $j = \overline{1,p}$ – элементы первого и второго столбца матрицы M размера $p \times 2$, содержащей по строкам в лексикографическом порядке индексы всех возможных комбинаций пар входных переменных; α_j, $j = \overline{0,l}$, $\alpha_j^{\min}$, $\alpha_j^{\max}$, $k_j^{\min}$, $k_j^{\max}$, $j = \overline{1,p}$ – неизвестные параметры; ε_i – i-я ошибка аппроксимации.

Стоит сделать следующие замечания касательно НЛР (1).

1. НЛР (1) содержит слишком много неизвестных параметров – $l+1+4C_l^2$ штук. Если $l=3$, то их число уже равно 16. Поэтому для оценки модели (1) потребуется выборка довольного крупного объема. В этой связи в [Базилевский, 2021] предложены алгоритмы отбора в уравнении (1) только информативных регрессоров.

2. НЛР (1) является нелинейной по параметрам. Но если придать параметрам $k_j^{\min}$, $k_j^{\max}$, $j=\overline{1,p}$ определенные значения, то она становится линейной по параметрам α_j, $j=\overline{0,l}$, $\alpha_j^{\min}$, $\alpha_j^{\max}$. В таком случае для оценивания можно воспользоваться традиционным методом наименьших квадратов (МНК).

3. Если в НЛР (1) назначить параметрам $k_j^{\min}$, $k_j^{\max}$, $j=\overline{1,p}$ слишком большие или слишком малые значения, то в некоторых или во всех бинарных операциях min и max сработает только одна переменная, что в итоге приведет к эффекту совершенной мультиколлинеарности, а, следовательно, к невозможности оценки параметров α_j, $j=\overline{0,l}$, $\alpha_j^{\min}$, $\alpha_j^{\max}$. Правильный выбор параметров $k_j^{\min}$, $k_j^{\max}$, $j=\overline{1,p}$ обсуждается в работах [Базилевский, 2020, 2021, 2022а].

В [Базилевский, 2022в] рассмотрены вопросы построения вполне интерпретируемых НЛР (1).

Работы [Базилевский, 2020, 2021, 2022а, в] объединяет то, что предложенное в их рамках программное обеспечение в первую очередь предназначено для выбора оптимальной структуры НЛР (1). Программного продукта, позволяющего оценивать с помощью МНК заданную пользователем структуру НЛР (1) со многими переменными, до сегодняшнего дня разработано не было.

Обобщение НЛР и алгоритмы их численного оценивания

В работе [Базилевский, 2022б] хорошо зарекомендовали себя простейшие НЛР как с угловым коэффициентом, так и со свободным членом в бинарной операции. На основе этого введем НЛР с линейным аргументом в бинарных операциях:

$$y_i=\alpha_0+\sum_{j=1}^{l}\alpha_j x_{ij}+\sum_{j=1}^{p}\alpha_j^{\min}\min\{x_{i,\mu_{j1}},k_j^{\min}x_{i,\mu_{j2}}+b_j^{\min}\}+$$
$$+\sum_{j=1}^{p}\alpha_j^{\max}\max\{x_{i,\mu_{j1}},k_j^{\max}x_{i,\mu_{j2}}+b_j^{\max}\}+\varepsilon_i,\quad i=\overline{1,n},\qquad(2)$$

где $b_j^{\min}$, $b_j^{\max}$, $j=\overline{1,p}$ – неизвестные параметры. Будем называть их свободными членами бинарных операций, а параметры $k_j^{\min}$, $k_j^{\max}$, $j=\overline{1,p}$ – угловыми коэффициентами бинарных операций. НЛР (2) представляет собой более гибкий инструмент математического моделирования, нежели НЛР (1).

Рассмотрим 3 случая НЛР (2) и алгоритмы их численного оценивания.

1. Параметры $b_j^{\min}$, $b_j^{\max}$, $j=\overline{1,p}$ равны 0. В этом случае НЛР (2) вырождается в НЛР (1) с угловыми коэффициентами в бинарных операциях. Алгоритм численного МНК-оценивания таких моделей хорошо изучен (см., например, [Базилевский, 2020, 2021]). Сначала определяются границы возможных значений угловых коэффициентов $k_j^{\min}$, $k_j^{\max}$, $j=\overline{1,p}$ по формулам:

$$\text{low}_j=\min\left\{\frac{x_{1,\mu_{j1}}}{x_{1,\mu_{j2}}},\frac{x_{2,\mu_{j1}}}{x_{2,\mu_{j2}}},...,\frac{x_{n,\mu_{j1}}}{x_{n,\mu_{j2}}}\right\},\quad \text{up}_j=\max\left\{\frac{x_{1,\mu_{j1}}}{x_{1,\mu_{j2}}},\frac{x_{2,\mu_{j1}}}{x_{2,\mu_{j2}}},...,\frac{x_{n,\mu_{j1}}}{x_{n,\mu_{j2}}}\right\},\ j=\overline{1,p}.\qquad(3)$$

Затем каждый полученный интервал $\left[\text{low}_j, \text{up}_j\right]$ разбивается равномерно r точками. После чего, используя все возможные комбинации этих точек вместо угловых коэффициентов НЛР (1), с помощью МНК оценивается r^{2p} линейных регрессий и выбирается лучшая из них по величине коэффициента детерминации R^2. Стоит отметить, что если в структуру НЛР (1) каждая объясняющая переменная будет входить ровно 1 раз, то концы интервалов $\left[\text{low}_j, \text{up}_j\right]$ тоже можно использовать при формировании линейных регрессий, поскольку совершенной мультиколлинеарности не будет.

2. Параметры $k_j^{\min}$, $k_j^{\max}$, $j=\overline{1,p}$ равны 1. В этом случае НЛР (2) вырождается в НЛР с единичными угловыми коэффициентами и со свободными членами в бинарных операциях:

$$y_i = \alpha_0 + \sum_{j=1}^{l}\alpha_j x_{ij} + \sum_{j=1}^{p}\alpha_j^{\min}\min\{x_{i,\mu_{j1}}, x_{i,\mu_{j2}} + b_j^{\min}\} +$$
$$+\sum_{j=1}^{p}\alpha_j^{\max}\max\{x_{i,\mu_{j1}}, x_{i,\mu_{j2}} + b_j^{\max}\} + \varepsilon_i, \qquad i=\overline{1,n}. \tag{4}$$

Используя результаты работы [Базилевский, 2022б], не трудно установить, что алгоритм численного МНК-оценивания НЛР (4) будет отличаться от алгоритма идентификации НЛР (1) только формулами для определения границ возможных значений свободных членов $b_j^{\min}$, $b_j^{\max}$, $j=\overline{1,p}$:

$$\text{low}_j = \min\left\{x_{1,\mu_{j1}} - x_{1,\mu_{j2}}, x_{2,\mu_{j1}} - x_{2,\mu_{j2}}, ..., x_{n,\mu_{j1}} - x_{n,\mu_{j2}}\right\},$$
$$\text{up}_j = \max\left\{x_{1,\mu_{j1}} - x_{1,\mu_{j2}}, x_{2,\mu_{j1}} - x_{2,\mu_{j2}}, ..., x_{n,\mu_{j1}} - x_{n,\mu_{j2}}\right\}, \qquad j=\overline{1,p}. \tag{5}$$

3. Параметры $k_j^{\min}$, $k_j^{\max}$, $b_j^{\min}$, $b_j^{\max}$, $j=\overline{1,p}$ отличны от нуля. Это самый интересный, но проблематичный случай. Точно определить области возможных значений изменения угловых коэффициентов и свободных членов затруднительно. Поэтому предлагается действовать следующим образом. Сначала выбрать произвольные границы (например, по формулам (3)) возможных значений угловых коэффициентов $k_j^{\min}$, $k_j^{\max}$, $j=\overline{1,p}$ и разбить полученные интервалы r точками. А затем для каждой такой точки $k_j^{(s)}$, $j=\overline{1,p}$, $s=\overline{1,r}$ определить границы возможных значений свободных членов $b_j^{\min}$, $b_j^{\max}$, $j=\overline{1,p}$:

$$\text{low}_{js} = \min\left\{x_{1,\mu_{j1}} - k_j^{(s)}x_{1,\mu_{j2}}, x_{2,\mu_{j1}} - k_j^{(s)}x_{2,\mu_{j2}}, ..., x_{n,\mu_{j1}} - k_j^{(s)}x_{n,\mu_{j2}}\right\},$$
$$\text{up}_{js} = \max\left\{x_{1,\mu_{j1}} - k_j^{(s)}x_{1,\mu_{j2}}, x_{2,\mu_{j1}} - k_j^{(s)}x_{2,\mu_{j2}}, ..., x_{n,\mu_{j1}} - k_j^{(s)}x_{n,\mu_{j2}}\right\}, \quad j=\overline{1,p}. \tag{6}$$

После чего снова разбить полученные отрезки точками и осуществить процедуру выбора лучшей регрессии методом полного перебора r^{4p} комбинаций. Как видно, вычислительная сложность такой задачи гораздо выше, чем в предыдущих двух случаях. Также её решение не гарантирует близость оценок полученной НЛР к оптимальным МНК-оценкам, в отличие от предыдущих двух алгоритмов. Тем не менее, высока вероятность того, что качество НЛР (2), оцененной представленным алгоритмом, будет лучше, чем качество НЛР (1) и (4).

Предложенные в этом разделе алгоритмы численного МНК-оценивания НЛР представляют собой довольно сложные вычислительные задачи. Однако они рассмотрены в условиях, когда абсолютно все $2p$ бинарных операций входят в модель. На практике же чаще возникают

задачи построения НЛР произвольной структуры с ограниченным числом бинарных операций. Такие задачи, естественно, будут решаться разработанными алгоритмами значительно быстрее.

Выделим некоторые существующие пока особенности применения НЛР на практике.

1. На сегодняшний день НЛР применяются [Базилевский, 2021, 2022а, в] только для решения задач, в которых все переменные строго положительны. Решать задачи с отрицательными значениями переменных тоже возможно, но пока не приходилось.

2. Чтобы избежать проблем с совершенной мультиколлинеарностью факторов, к спецификации НЛР предъявляется требование, чтобы каждая переменная входила в неё не более одного раза.

3. Угловые коэффициенты НЛР выбираются положительными. Тогда для обеспечения интерпретируемости НЛР [Базилевский, 2022в] вводится ограничение, чтобы знаки коэффициентов корреляции входящих в каждую бинарную операцию переменных с объясняемой переменной y были одинаковы и совпадали со знаком МНК-оценки при этой бинарной операции.

Программа численного оценивания НЛР методом наименьших квадратов

Для оценивания НЛР произвольной структуры с помощью МНК в среде программирования Delphi была разработана специальная программа НЕЭЛИН. Алгоритм работы программы представлен на рис. 1.

Сначала пользователю нужно ввести данные в программу. Делается это нажатием кнопки «Загрузка» на главной форме. В открывшемся окне нужно выбрать текстовый файл формата «*.txt». Данные в нём должны быть структурированы в виде таблицы. Первый столбик должен содержать значения объясняемой переменной, второй – первой объясняющей переменной, третий – второй объясняющей переменной и т.д. Значения в столбцах отделяются друг от друга символом «Tab», а в вещественных числах сепаратором служит «,». Если всё сделано верно, то данные отобразятся в поле «Статистические данные» главной формы.

Затем нужно выбрать структуру НЛР. Делается это в поле «Структура модели» главной формы. Структура задается в виде таблицы. Образец такой таблицы представлен на рис. 2. Её строки соответствуют числу регрессоров НЛР, а их количество регулируется в поле «Число регрессоров». Предусмотрена возможность задать от 1 до 10 регрессоров. В каждую строку таблицы вводится информация о типе регрессора и его параметрах. Столбец «Преобр.» обязателен для заполнения и содержит информацию о типе регрессора. Типы бывают следующие:

- «lin» – линейный регрессор;
- «min» – бинарная операция min;
- «max» – бинарная операция max.

Столбец «Перем». обязателен для заполнения и содержит информацию о номерах переменных, входящих в регрессор. Если выбран тип регрессора «lin», то в этот столбик вводится один номер, а если «min» или «max», то два. В последнем случае номера переменных вводятся в порядке возрастания и отделяются друг от друга пробелом. Важно, чтобы входящие в регрессор переменные коррелировали с зависимой переменной с одинаковым знаком. Иначе оцененная НЛР может получиться не интерпретируемой.

Столбец «Тип» обязателен для заполнения для регрессоров только типов «min» и «max». Для линейных регрессоров соответствующую ячейку таблицы надо оставить пустой. В столбец «Тип» вносится информация о типе бинарной операции. Типы бывают следующие:

- «s» – в бинарную операцию входит только угловой коэффициент (slope);
- «i» – в бинарную операцию входит единичный угловой коэффициент и свободный член (intercept);
- «si» – в бинарную операцию входит линейный аргумент.

Столбец «Slope» заполняется только для бинарных операций типа «si». Он содержит нижнюю и верхнюю границы изменения углового коэффициента. Границы задаются в порядке возрастания и отделяются друг от друга пробелом.

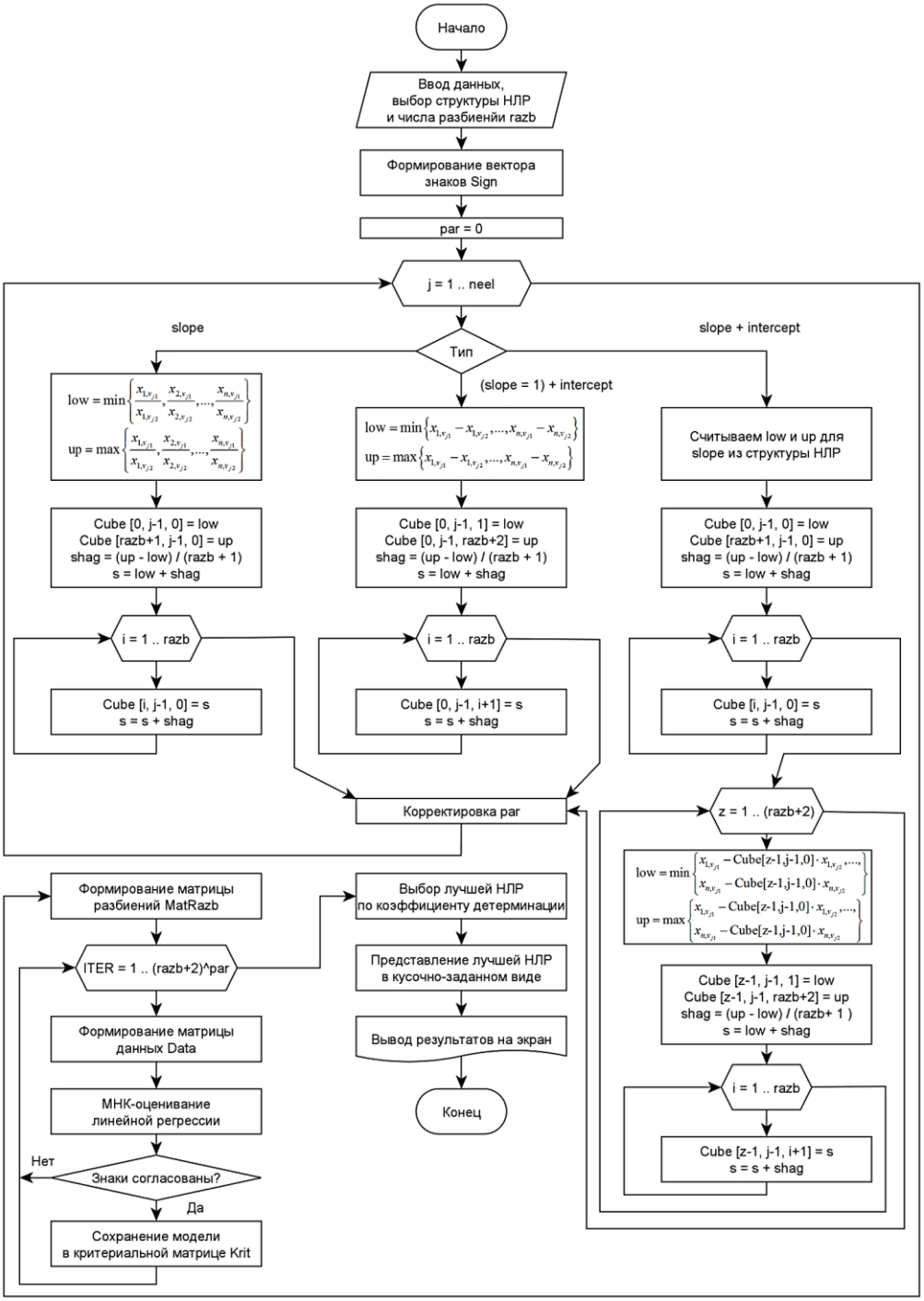

Рис. 1. Алгоритм работы программы НЕЭЛИН
Fig. 1. The algorithm of the program NEELIN

Например, приведенная на рис. 2 таблица в поле «Структура модели» соответствует следующей спецификации:

$$y_i = \alpha_0 + \alpha_1 \min\{x_{i1}, k_1 x_{i3} + b\} + \alpha_2 \max\{x_{i2}, k_2 x_{i5}\} + \alpha_3 x_{i4} + \varepsilon_i,\ i = \overline{1,n}.$$

Если пользователь затрудняется с выбором границ углового коэффициента бинарной операции типа «si», то может нажать на кнопку «Оценить slope» и посмотреть границы для всех угловых коэффициентов, найденных по формулам (3). Это поможет ему сделать выбор.

Перед оцениванием НЛР также нужно выбрать число разбиений (razb) интервалов изменения угловых коэффициентов и свободных членов в поле «Число разбиений» и установить

переключатель «Согласованность знаков», отвечающий за то, чтобы знаки МНК-оценок модели были или не были согласованы со знаками входящих в регрессоры переменных в нужное положение.

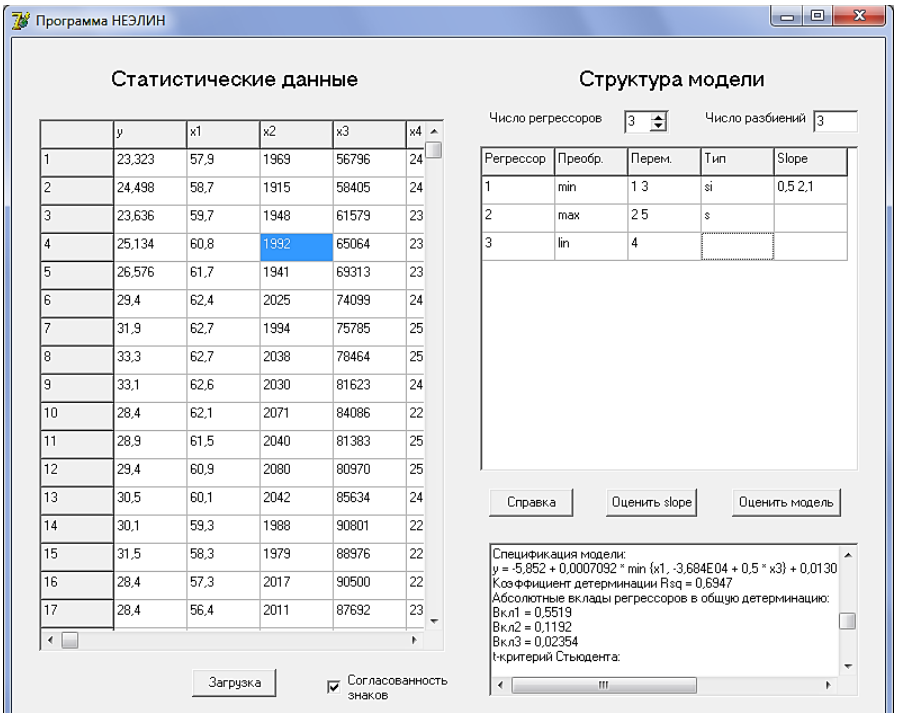

Рис. 2. Главная форма программы НЕЭЛИН
Fig. 2. The main form of the NEELIN program

Далее нужно запустить процесс оценивания НЛР, нажав кнопку «Оценить модель». Перед его началом проверяется корректность введенной пользователем структуры модели. Если все проверки пройдены, то сначала формируется вектор знаков корреляции регрессоров Sign, компонента которого равна «+1», если коэффициент корреляции произвольной переменной регрессора с зависимой переменной положителен, и «–1», если коэффициент корреляции отрицателен.

Затем начинает формироваться трехмерный массив Cube (рис. 1) размера (razb+2)*neel*(razb+3), где neel – число неэлементарных функций, содержащий для выбранной структуры НЛР все возможные значения угловых коэффициентов и свободных членов в бинарных операциях. Будем называть этот массив кубом. Структура куба представлена на рис. 3. Он состоит из трех измерений – «Угловой коэффициент», «Свободный член», «Бинарная операция».

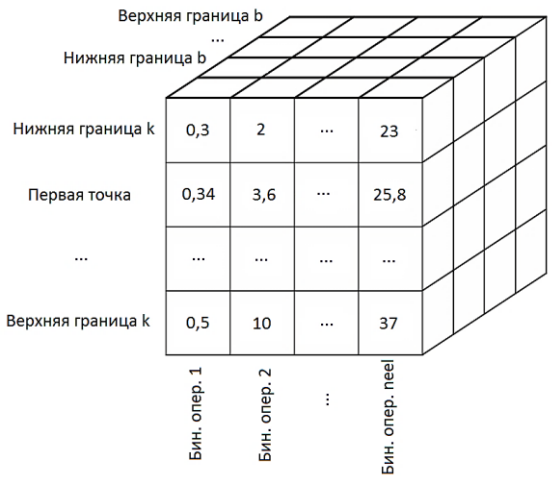

Рис. 3. Структура куба
Fig. 3. The structure of the cube

Постепенно в цикле (рис. 1) заполняются срезы куба по каждой бинарной операции, представляющие собой двумерные массивы двух измерений – «Угловой коэффициент» и «Свободный член». При этом структура каждого такого среза зависит от типа бинарной операции. Если выбран тип «s», то нижняя (low) и верхняя (up) границы углового коэффициента определяются по формулам (3), затем полученный интервал разбивается razb точками на (razb+1) интервалов одинаковой длины shag и все эти точки (вместе с low и up) сохраняются в срез куба в виде одномерного массива измерения «Угловой коэффициент». Если выбран тип «i», то границы low и up свободного члена определяются по формулам (5), затем полученный интервал разбивается точками и все они вместе с low и up сохраняются в срез куба в виде одномерного массива измерения «Свободный член». Если выбран тип «si», то сформированный по заданным пользователем границам углового коэффициента low и up интервал сначала разбивается точками, потом для каждой из этих точек вместе с low и up определяются границы свободного члена по формулам (6), затем полученные интервалы снова разбиваются razb точками и все они сохраняются в срез куба в виде двумерного массива измерений «Угловой коэффициент» и «Свободный член».

После чего автоматически формируется матрица разбиений MatRazb размера [(razb+2)^par]*par, где par – общее количество угловых коэффициентов и свободных членов в бинарных операциях выбранной НЛР. Комбинации разбиений формируются в лексикографическом порядке.

На следующем этапе начинается цикл по строкам матрицы MatRazb. Для каждой строки реализуется следующая последовательность действий:

1) с помощью куба формируется матрица данных Data, содержащая в первом столбце значения зависимой переменной, а в остальных столбцах значения переменных, преобразованных в соответствии с заданной структурой НЛР;

2) по матрице Data находятся МНК-оценки линейной регрессии;

3) если пользователем активирован переключатель «Согласованность знаков», то знаки всех МНК-оценок проверяются на соответствие знакам вектора Sign. Если знаки абсолютно всех оценок согласованы, то модель и её характеристики сохраняются в критериальной матрице Krit, а если нет, то сохранение не производится.

И, наконец, по критериальной матрице Krit выбирается лучшая с точки зрения коэффициента детерминации НЛР и сразу в окне результатов на главной форме программы выводится следующая информация: общее количество оцененных вариантов НЛР, число НЛР с согласованными по знакам МНК-оценками, значения преобразованных переменных для лучшей НЛР, а также её спецификация, коэффициент детерминации, абсолютные вклады регрессоров в общую детерминацию [Базилевский, 2022а], t-критерий Стьюдента и её представление в кусочно-заданной форме.

Моделирование грузовых перевозок железнодорожного транспорта в Республике Башкортостан

Республика Башкортостан входит в Приволжский федеральный округ Российской Федерации. Вопросы моделирования экономики Республики Башкортостан рассмотрены в работах [Дегтярев, 2020; Муратов, Дегтярев, 2021]. Транспорт – одна из важнейших отраслей экономики Башкортостана. По объемам отправления грузов железнодорожным транспортом Республика Башкортостан в 2020 году заняла третье место с объемом 26,4 млн тонн среди четырнадцати субъектов Приволжского федерального округа. Поэтому задача моделирования грузовых перевозок в Башкортостане весьма актуальна в настоящее время.

В качестве объясняемой переменной y выступает отправление грузов железнодорожным транспортом общего пользования (млн тонн) в Республике Башкортостан. На сайте Росстата (https://rosstat.gov.ru/) для этой переменной были собраны ежегодные статистические данные за период с 2000 по 2020 гг. Также предварительно был сформирован список из 30 объясняющих переменных, предположительно влияющих на y, для каждой из которых были собраны данные с того же сайта. С помощью корреляционного анализа из этого списка были исключены те переменные, которые либо совсем слабо коррелируют с y, либо у которых знаки коэффициентов корреляции противоречат содержательному смыслу решаемой задачи. В результате исключения осталось 7 переменных:

x_1 – население в трудоспособном возрасте (%);

x_2 – численность рабочей силы (тыс. человек);

x_3 – число предприятий и организаций;

x_4 – производство электроэнергии (млрд киловатт-часов);

x_5 – продукция сельского хозяйства (млн руб.);

x_6 – объем работ, выполненных по виду экономической деятельности «Строительство» (млн руб.);

x_7 – валовой региональный продукт (млн руб.).

Заметим, что все эти переменные коррелируют с y со знаком «+».

Сначала с помощью МНК в пакете Gretl (https://gretl.sourceforge.net/ru.html) была построена линейная регрессия со всеми семью объясняющими переменными:

$$\tilde{y} = \underset{}{-25,446} + \underset{(1,195)}{0,629}\,x_1 - \underset{(-1,026)}{0,0109}\,x_2 + \underset{(1,431)}{0,000263}\,x_3 + \underset{(1,318)}{0,769}\,x_4 - \\ - \underset{(-0,733)}{2,732\cdot 10^{-5}}\,x_5 + \underset{(1,075)}{3,77\cdot 10^{-5}}\,x_6 - \underset{(-0,420)}{2,118\cdot 10^{-6}}\,x_7 . \qquad (7)$$

Коэффициент детерминации R^2 регрессии (7) равен 0,855918, что говорит об её хорошем качестве. Но абсолютно все коэффициенты, судя по указанным в скобках под ними значениям t-критерия Стьюдента, незначимы даже для уровня значимости 0,1. Это следствие такого негативного эффекта, как частичная мультиколлинеарность. Действительно, пять из семи коэффициентов вздутия дисперсии VIF (для переменных x_1, x_3, x_5, x_6 и x_7) превосходят пороговое значение 10, что свидетельствует о наличии мультиколлинеарности. Из-за неё знаки коэффициентов при переменных x_2, x_5 и x_7 противоречат их содержательному смыслу, поэтому модель (7) не годится для интерпретации.

Поэтому было принято решение перестроить модель (7), воспользовавшись в Gretl методом автоматического исключения незначимых факторов. Для этого уровень значимости был задан 0,05. В результате исключения получилась модель:

$$y = \underset{}{-37,23} + \underset{(4,127)}{0,448}\,x_1 + \underset{(8,161)}{0,000244}\,x_3 + \underset{(3,412)}{0,834}\,x_4 . \qquad (8)$$

Для регрессии (8) $R^2 = 0,822860$, поэтому она хуже по качеству аппроксимации, чем модель (7), но зато все её коэффициенты значимы по t-критерию Стьюдента, а также она лишена эффекта мультиколлинеарности и может быть интерпретирована. И всё же хотелось бы иметь модель более высокого качества, чем (8).

Для построения НЛР сначала нужно было определиться с её структурой. Для этого была использована программа ВИнтер-2 [Базилевский, 2022в], в которой по заданным настройкам формируется и решается задача частично-булевого линейного программирования, выбирающая оптимальную по коэффициенту детерминации структуру НЛР (1). К таким настройкам относится число точек, разбивающих отрезки (3), и минимальное пороговое значение для коэффициентов корреляции регрессоров с y. Первая настройка была выбрана равной 4, вторая – 0,2. В результате работы программы за 81,9 сек. была выбрана следующая структура НЛР:

$$\tilde{y} = \underset{}{-43,79} + \overset{(0,0626)}{\underset{(3,152)}{1,033}}\,x_4 + \overset{(0,4356)}{\underset{(2,286)}{0,00277}}\min\{x_2, 0.0317x_6\} + \\ + \overset{(0,3251)}{\underset{(3,026)}{0,695}}\max\{x_1, 0.00071x_3\} + \overset{(0,0313)}{\underset{(0,4136)}{4,769\cdot 10^{-6}}}\max\{x_5, 0.117x_7\} . \qquad (9)$$

Для регрессии (9) $R^2 = 0,854706$, т.е. качество её аппроксимации примерно такое же, что и у линейной регрессии со всеми переменными (7). Но НЛР (9), в отличие от модели (9), лишена эффекта мультиколлинеарности и все её коэффициенты удовлетворяют смыслу задачи.

Но всё же по t-критерию Стьюдента коэффициент НЛР (9) при регрессоре $\max\{x_5, 0.117x_7\}$ оказался не значим для уровня значимости 0,05. О его слабом влиянии на y также можно судить по величине абсолютного вклада этого регрессора в общую детерминацию (0,0313) [Базилевский, 2022а]. Эти вклады указаны в скобках над коэффициентами модели (9). Поэтому было принято решение исключить регрессор $\max\{x_5, 0.117x_7\}$. Переоцененная с помощью МНК новая структура НЛР получилась следующей:

$$\tilde{y} = -37,21 + \underset{(3,809)}{0,948}\, x_4 + \underset{(6,491)}{0,00322}\min\{x_2, 0.0317x_6\} + \underset{(4,849)}{0,616}\max\{x_1, 0.00071x_3\}. \quad (10)$$

Для регрессии (10) $R^2 = 0,8532$, т.е. её качество практически не отличается от качества более сложной модели (9). При этом отсутствует эффект мультиколлинеарности, все коэффициенты интерпретируемы и значимы по t-критерию Стьюдента даже для уровня 0,01.

В принципе, НЛР (10) уже может применяться на практике. Однако она получена при условии, когда число точек разбиения отрезков (3) выбиралось равным 4. Для уточнения МНК-оценок модели (10) была применена программа НЕЭЛИН. Число разбиений отрезков (3) в ней было выбрано равным 100. В результате была получена уточненная НЛР с угловыми коэффициентами в бинарных операциях:

$$\tilde{y} = -38,73 + \underset{(3,864)}{1,003}\, x_4 + \underset{(5,803)}{0,00323}\min\{x_2, 0.0361x_6\} + \underset{(4,559)}{0,613}\max\{x_1, 0.000721x_3\}. \quad (11)$$

Для регрессии (11) $R^2 = 0,8608$, т.е. она лучше по качеству любой из построенных выше моделей (7) – (10). Никаких проблем с мультиколлинеарностью, интерпретацией и значимостью её коэффициентов по-прежнему нет. Заметим, что уравнение (11) было выбрано из 10404 вариантов моделей, из которых только в 9031 регрессиях все МНК-оценки оказались интерпретируемыми. Таким образом, НЕЭЛИН позволяет уточнять МНК-оценки НЛР, структура которой выбрана в программе ВИнтер-2.

Затем в НЕЭЛИН оценивалась НЛР (4) с единичными угловыми коэффициентами и произвольными свободными членами в бинарных операциях. Число разбиений по-прежнему было выбрано равным 100. В результате из 10404 вариантов моделей только в 6031 регрессиях все МНК-оценки оказались интерпретируемыми. Лучшая из них модель имеет вид:

$$\tilde{y} = -36,6 + \underset{(2,117)}{0,4514}\, x_4 + \underset{(9,194)}{5,703\cdot 10^{-5}}\min\{x_2, -1,45\cdot 10^5 + x_6\} + \underset{(8,127)}{0,9625}\max\{x_1, -9,074\cdot 10^4 + x_3\}. \quad (12)$$

Для регрессии (12) $R^2 = 0,8541$, т.е. она примерно такая же по качеству, что и НЛР (11). В (12) все коэффициенты значимы для уровня 0,05, интерпретируемы и мультиколлинеарность отсутствует. При этом содержательная интерпретация НЛР (12) будет принципиально отличаться от интерпретации НЛР (11) [Базилевский, 2022б].

И, наконец, с помощью НЕЭЛИН оценивалась НЛР с произвольными угловыми коэффициентами и свободными членами в бинарных операциях. Интервалы возможных значений угловых коэффициентов НЛР были выбраны [0,01;0,05] для регрессора с бинарной операцией

min, и [0,0001;0,001] для регрессора с бинарной операцией max. Число разбиений было выбрано равным 25. В результате из 1048576 моделей, из которых в 690718 регрессиях все МНК-оценки оказались интерпретируемыми, была выбрана следующая:

$$\tilde{y} = -67,17 + \underset{(4,773)}{0,9127}\, x_4 + \underset{(8,785)}{0,0008464}\min\{x_2, -5066 + 0,05x_6\} + \underset{(9,405)}{1,227}\max\{x_1, 31,12 + 0,0003323x_3\}. \quad (13)$$

Для регрессии (13) $R^2 = 0,8891$, т.е. по качеству она оказалась гораздо лучше, чем НЛР (11) и (12). И в ней как и прежде все коэффициенты значимы, интерпретируемы и отсутствует мультиколлинеарность.

Представим модель (13) в кусочно-заданном виде:

$$\tilde{y} = \begin{cases} -71,46 + 0,9127x_4 + 4,232\cdot 10^{-5}x_6 + 1,227x_1, \text{ при } x_2 \ge -5066 + 0,05x_6,\ x_1 \ge 31,12 + 0,000332x_3, \\ -33,29 + 0,9127x_4 + 4,232\cdot 10^{-5}x_6 + 0,000407x_3, \text{ при } x_2 \ge -5066 + 0,05x_6,\ x_1 < 31,12 + 0,00033x_3, \\ -67,17 + 0,9127x_4 + 0,0008464x_2 + 1,227x_1, \text{ при } x_2 < -5066 + 0,05x_6,\ x_1 \ge 31,12 + 0,000332x_3, \\ -29 + 0,9127x_4 + 0,0008464x_2 + 0,000407x_3, \text{ при } x_2 < -5066 + 0,05x_6,\ x_1 < 31,12 + 0,000332x_3. \end{cases}$$

Тогда НЛР (13) можно интерпретировать следующим образом.

1. Если производство электроэнергии x_4 увеличится на 1 млрд киловатт-часов (при неизменных значениях остальных переменных), то ж/д грузоперевозки y в Республике Башкортостан вырастут в среднем на 0,9127 млн тонн.

2. Если $x_2 \ge -5066 + 0,05x_6$, то на y влияют объемы строительных работ x_6, а численность рабочей силы x_2 не влияет. При этом с увеличением x_6 на 100000 млн руб (при неизменных значениях остальных переменных) y возрастает в среднем на 4,232 млн тонн. Если $x_2 < -5066 + 0,05x_6$, то на y, наоборот, влияет x_2, а не x_6. При этом с увеличением x_2 на 1 тыс. человек (при неизменных значениях остальных переменных) y возрастает в среднем на 846,4 тонны.

3. Если $x_1 \ge 31,12 + 0,000332x_3$, то на y влияет численность населения в трудоспособном возрасте x_1, а число предприятий и организаций x_3 не влияет. При этом с увеличением x_1 на 1% (при неизменных значениях остальных переменных) y возрастает в среднем на 1,227 млн тонн. Если $x_1 < 31,12 + 0,000332x_3$, то на y, наоборот, влияет x_3, а не x_1. При этом с увеличением x_3 на 10000 штук (при неизменных значениях остальных переменных) y возрастает в среднем на 4,07 млн тонн.

Заключение

В работе сформулировано обобщение НЛР с угловыми коэффициентами в бинарных операциях – НЛР с линейными аргументами в бинарных операциях. Рассмотрено три их вида, для каждого из которых предложен численный алгоритм МНК-оценивания. На основе этих алгоритмов была разработана программа НЕЭЛИН, позволяющая численно оценивать с помощью МНК НЛР произвольной структуры. Рассмотрен алгоритм работы НЕЭЛИН. Показано, что в этом алгоритме возможные значения угловых коэффициентов и свободных членов хранятся в виде трехмерного массива. Продемонстрировано, как в НЕЭЛИН нужно задавать структуру НЛР. С помощью НЕЭЛИН решена задача моделирования железнодорожных грузовых перевозок Республики Башкортостан. Разработанная программа НЕЭЛИН универсальна и может применяться для решения различных задач анализа данных.

Список литературы

Базилевский М.П. 2020. Оценивание линейно-неэлементарных регрессионных моделей с помощью метода наименьших квадратов. Моделирование, оптимизация и информационные технологии, 8, 4 (31). https://doi.org/10.26102/2310-6018/2020.31.4.026

Базилевский М.П. 2021. Отбор информативных операций при построении линейно-неэлементарных регрессионных моделей. International Journal of Open Information Technologies, 9 (5): 30–35.

Базилевский М.П. 2022. Метод построения неэлементарных линейных регрессий на основе аппарата математического программирования. Проблемы управления, 4: 3–14. https://doi.org/10.25728/pu.2022.4.1

Базилевский М.П. 2022. Оценка методом наименьших квадратов простейших неэлементарных линейных регрессий с линейным аргументом в бинарной операции. Вестник кибернетики, 4 (48): 69–76.

Базилевский М.П. 2022. Построение вполне интерпретируемых неэлементарных линейных регрессионных моделей. Вестник Югорского государственного университета, 4 (67): 105–114.

Базилевский М.П., Носков С.И. 2017. Формализация задачи построения линейно-мультипликативной регрессии в виде задачи частично-булевого линейного программирования. Современные технологии. Системный анализ. Моделирование, 3 (55): 101–105.

Дегтярев А.Н. 2020. Анализ современного состояния развития промышленности и нефтедобычи в Республике Башкортостан. Научные труды Вольного экономического общества России, 223 (3): 432–444.

Клейнер Г.Б. 1986. Производственные функции: Теория, методы, применение. Москва, Финансы и статистика, 239 с.

Муратов Р.Х., Дегтярев А.Н. 2021. Условия экономического роста в модели инновационного развития региона (на материалах Республики Башкортостан). Уфимский гуманитарный научный форум, 2 (6): 10–18.

Носков С.И., Хоняков А.А. 2019. Программный комплекс построения некоторых типов кусочно-линейных регрессий. Информационные технологии и математическое моделирование в управлении сложными системами, 3 (4): 47–55.

Du M., Liu N., Hu X. 2019. Techniques for interpretable machine learning. Communications of the ACM, 63 (1): 68-77. https://doi.org/10.1145/3359786

Gao F., Yang L., Han C., Tang J., Li Z. 2022. A network-distance-based geographically weighted regression model to examine spatiotemporal effects of station-level built environments on metro ridership. Journal of Transport Geography, 105: 103472. https://doi.org/10.1016/j.jtrangeo.2022.103472

Gelman A., Hill J., Vehtari A. 2020. Regression and other stories. Cambridge University Press.

Karakurt I., Aydin G. 2023. Development of regression models to forecast the CO2 emissions from fossil fuels in the BRICS and MINT countries. Energy, 263: 125650. https://doi.org/10.1016/j.energy.2022.125650

Keith T.Z. 2019. Multiple regression and beyond: An introduction to multiple regression and structural equation modeling. Routledge.

Letzgus S., Wagner P., Lederer J., Samek W., Müller K.R., Montavon G. 2022. Toward explainable artificial intelligence for regression models: A methodological perspective. IEEE Signal Processing Magazine, 39 (4): 40-58. https://doi.org/10.1109/MSP.2022.3153277

Luo J., Hong T., Gao Z., Fang S.C. 2022. A robust support vector regression model for electric load forecasting. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2022.04.001

Molnar C. 2020. Interpretable machine learning. Lulu. com.

Tian M., Guo F., Niu R. 2022. Risk spillover analysis of China’s financial sectors based on a new GARCH copula quantile regression model. The North American Journal of Economics and Finance, 63: 101817. https://doi.org/10.1016/j.najef.2022.101817

Wang P., Chen S., Yang S. 2022. Recent advances on penalized regression models for biological data. Mathematics, 10 (19): 3695. https://doi.org/10.3390/math10193695

References

Bazilevskiy M.P. 2020. Otsenivanie lineyno-neelementarnykh regressionnykh modeley s pomoshch'yu metoda naimen'shikh kvadratov [Estimation linear non-elementary regression models using ordinary least square]. Modelirovanie, optimizatsiya i informatsionnye tekhnologii, 8, 4 (31). https://doi.org/10.26102/2310-6018/2020.31.4.026

Bazilevskiy M.P. 2021. Otbor informativnykh operatsiy pri postroenii lineyno-neelementarnykh regressionnykh modeley [Selection of informative operations in the construction of linear non-elementary regression models]. International Journal of Open Information Technologies, 9 (5): 30–35.

Bazilevskiy M.P. 2022. Metod postroeniya neelementarnykh lineynykh regressiy na osnove apparata matematicheskogo programmirovaniya [A method for constructing non-elementary linear regressions based on mathematical programming]. Problemy upravleniya, 4: 3–14. https://doi.org/10.25728/pu.2022.4.1

Bazilevskiy M.P. 2022. Otsenka metodom naimen'shikh kvadratov prosteyshikh neelementarnykh lineynykh regressiy s lineynym argumentom v binarnoy operatsii [Ordinary least squares estimation of simple non-

elementary linear regressions with a linear argument in a binary operation]. Vestnik kibernetiki, 4 (48): 69–76.
Bazilevskiy M.P. 2022. Postroenie vpolne interpretiruemykh neelementarnykh lineynykh regressionnykh modeley [Construction of quite interpretable non-elementary linear regression models]. Vestnik Yugorskogo gosudarstvennogo universiteta, 4 (67): 105–114.
Bazilevskiy M.P., Noskov S.I. 2017. Formalizatsiya zadachi postroeniya lineyno-mul'tiplikativnoy regressii v vide zadachi chastichno-bulevogo lineynogo programmirovaniya. Sovremennye tekhnologii [Formalization of the problem of construction of linear multiplicative regressions in the form of a partial-Boolean linear programming problem]. Sistemnyy analiz. Modelirovanie, 3 (55): 101–105.
Degtyarev A.N. 2020. Analiz sovremennogo sostoyaniya razvitiya promyshlennosti i neftedobychi v Respublike Bashkortostan [Analysis of the modern state of the industry and oil production in the Republic of Bashkortostan]. Nauchnye trudy Vol'nogo ekonomicheskogo obshchestva Rossii, 223 (3): 432–444.
Kleyner G.B. 1986. Proizvodstvennye funktsii: Teoriya, metody, primenenie [Production functions: Theory, methods, application]. Moskva, Finansy i statistika, 239 p.
Muratov R.Kh., Degtyarev A.N. 2021. Usloviya ekonomicheskogo rosta v modeli innovatsionnogo razvitiya regiona (na materialakh Respubliki Bashkortostan) [The economic growth in the model of innovate development of the region (based on the materials of the Republic of Bashkortostan)]. Ufimskiy gumanitarnyy nauchnyy forum, 2 (6): 10–18.
Noskov S.I., Khonyakov A.A. 2019. Programmnyy kompleks postroeniya nekotorykh tipov kusochno-lineynykh regressiy [Software complex for building some types pieces of linear regressions]. Informatsionnye tekhnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami, 3 (4): 47–55.
Du M., Liu N., Hu X. 2019. Techniques for interpretable machine learning. Communications of the ACM, 63 (1): 68-77. https://doi.org/10.1145/3359786
Gao F., Yang L., Han C., Tang J., Li Z. 2022. A network-distance-based geographically weighted regression model to examine spatiotemporal effects of station-level built environments on metro ridership. Journal of Transport Geography, 105: 103472. https://doi.org/10.1016/j.jtrangeo.2022.103472
Gelman A., Hill J., Vehtari A. 2020. Regression and other stories. Cambridge University Press.
Karakurt I., Aydin G. 2023. Development of regression models to forecast the CO2 emissions from fossil fuels in the BRICS and MINT countries. Energy, 263: 125650. https://doi.org/10.1016/j.energy.2022.125650
Keith T.Z. 2019. Multiple regression and beyond: An introduction to multiple regression and structural equation modeling. Routledge.
Letzgus S., Wagner P., Lederer J., Samek W., Müller K.R., Montavon G. 2022. Toward explainable artificial intelligence for regression models: A methodological perspective. IEEE Signal Processing Magazine, 39 (4): 40-58. https://doi.org/10.1109/MSP.2022.3153277
Luo J., Hong T., Gao Z., Fang S.C. 2022. A robust support vector regression model for electric load forecasting. International Journal of Forecasting. https://doi.org/10.1016/j.ijforecast.2022.04.001
Molnar C. 2020. Interpretable machine learning. Lulu. com.
Tian M., Guo F., Niu R. 2022. Risk spillover analysis of China's financial sectors based on a new GARCH copula quantile regression model. The North American Journal of Economics and Finance, 63: 101817. https://doi.org/10.1016/j.najef.2022.101817
Wang P., Chen S., Yang S. 2022. Recent advances on penalized regression models for biological data. Mathematics, 10 (19): 3695. https://doi.org/10.3390/math10193695

Конфликт интересов: о потенциальном конфликте интересов не сообщалось.
Conflict of interest: no potential conflict of interest related to this article was reported.

ИНФОРМАЦИЯ ОБ АВТОРЕ

Базилевский Михаил Павлович, кандидат технических наук, доцент кафедры математики, Иркутский государственный университет путей сообщения, г. Иркутск, Россия

INFORMATION ABOUT THE AUTHOR

Mikhail P. Bazilevskiy, Candidate of Technical Sciences, Associate Professor of the Department of Mathematics, Irkutsk State Transport University, Irkutsk, Russia