Моделирование и анализ данных
2023. Том 13. № 2. С. 85–98
DOI: https://doi.org/10.17759/mda.2023130205
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2023. Vol. 13, no. 2, pp. 85–98
DOI: https://doi.org/10.17759/mda.2023130205
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

◇◇◇◇◇◇◇ МЕТОДЫ ОПТИМИЗАЦИИ ◇◇◇◇◇◇◇

УДК 519.862.6

Обобщение неэлементарных линейных регрессий

Базилевский М.П.*
Иркутский государственный университет путей сообщения (ФГБОУ ВО ИрГУПС), г. Иркутск, Российская Федерация
ORCID: https://orcid.org/0000-0002-3253-5697
e-mail: mik2178@yandex.ru

Ранее автором была разработана неэлементарная линейная регрессия, состоящая из линейной части и всех возможных комбинаций бинарных операций min и max. Данная статья посвящена её обобщению. Впервые введена неэлементарная линейная регрессия с линейной частью и всеми возможными комбинациями бинарных, тернарных, ..., *l*-арных операций min и max. Предложенная модель обобщает как линейную регрессию, так и функцию Леонтьева, и может эффективно применяться как для прогнозирования, так и для интерпретации функционирования объекта исследования. Разработан алгоритм оценивания с помощью метода наименьших квадратов неэлементарных линейных регрессий без линейной части и с *l*-арной операцией min (max), т.е. регрессий со спецификацией в виде функции Леонтьева. Суть алгоритма состоит в формировании множества возможных значений угловых коэффициентов, из которого выбирается точка с минимальной величиной суммы квадратов остатков. Идентифицирована система линейных неравенств, позволяющая формировать такое множество. С помощью алгоритма построена модель валового регионального продукта Иркутской области и дана её интерпретация.

Ключевые слова: машинное обучение, регрессионная модель, неэлементарная линейная регрессия, метод наименьших квадратов, функция Леонтьева, мультиколлинеарность.

Для цитаты:
Базилевский М.П. Обобщение неэлементарных линейных регрессий // Моделирование и анализ данных. 2023. Том 13. № 2. С. 85–98. DOI: https://doi.org/10.17759/mda.2023130205

***Базилевский Михаил Павлович,** кандидат технических наук, доцент кафедры математики, Иркутский государственный университет путей сообщения (ФГБОУ ВО ИрГУПС), г. Иркутск, Российская Федерация, ORCID: https://orcid.org/0000-0002-3253-5697, e-mail: mik2178@yandex.ru

1. ВВЕДЕНИЕ

В настоящее время машинное обучение [1,2], вероятно, самая перспективная область искусственного интеллекта. Предназначение машинного обучения в том, чтобы запрограммировать искусственный интеллект действовать как человек, или даже лучше него, при решении различных прикладных задач. Обучение осуществляется на основе больших массивов статистических данных. Как отмечено в работе [1], «машинное обучение быстро превращается в двигатель современной экономики, управляемой данными». Совсем недавно начало выделяться новое направление – интерпретируемое машинное обучение [3, 4].

С помощью машинного обучения решаются различные типы задач: классификация, кластеризация, регрессия, понижение размерности данных, обнаружение аномалий и т.д. Данная статья посвящена задаче регрессии [5–7], состоящей, как правило, в прогнозировании одной или нескольких характеристик по имеющимся статистическим данным – выборке. На сегодняшний день известно множество математических форм связи между переменными в регрессионных моделях: линейные [5–7], полиномиальные [8, 9], степенные [10, 11], степенно-показательные [12], линейно-логарифмические [12], логистические [13, 14], функции с фиксированными пропорциями факторов (функции Леонтьева) [15] и т.д. Тем не менее, процесс поиска новых спецификаций регрессионных моделей, позволяющих извлекать новые знания о функционировании изучаемого процесса или явления, продолжается.

В работе [16] впервые было проведено смешение линейных регрессий с двухфакторными функциями Леонтьева. Полученный синтез был назван неэлементарной линейной регрессией (НЛР). В той же работе был предложен алгоритм численного оценивания НЛР с помощью метода наименьших квадратов (МНК). В [17] были предложены алгоритмы выбора оптимальной структуры НЛР. А в [18] впервые были введены НЛР с бинарными операциями min и max:

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

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; $p=\overline{C_l^2}$ – 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.

Цель данной работы состоит в обобщении НЛР (1), в разработке алгоритма её оценивания с помощью МНК и в решении задачи моделирования валового регионального продукта (ВРП) Иркутской области.

2. ОБОБЩЕНИЕ НЛР

Введем в рассмотрение НЛР с бинарными, тернарными, кватернарными, ..., l -арными операциями min и max:

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \sum_{j=1}^{p_1} \alpha_j^{\min,2} \min\{x_{i,\mu_{j1}^{(1)}}, k_j^{\min,2} x_{i,\mu_{j2}^{(1)}}\} + \sum_{j=1}^{p_1} \alpha_j^{\max,2} \max\{x_{i,\mu_{j1}^{(1)}}, k_j^{\max,2} x_{i,\mu_{j2}^{(1)}}\} +$$

$$+ \sum_{j=1}^{p_2} \alpha_j^{\min,3} \min\{x_{i,\mu_{j1}^{(2)}}, k_{j1}^{\min,3} x_{i,\mu_{j2}^{(2)}}, k_{j2}^{\min,3} x_{i,\mu_{j3}^{(2)}}\} + \sum_{j=1}^{p_2} \alpha_j^{\max,3} \max\{x_{i,\mu_{j1}^{(2)}}, k_{j1}^{\max,3} x_{i,\mu_{j2}^{(2)}}, k_{j2}^{\max,3} x_{i,\mu_{j3}^{(2)}}\} + \ldots +$$

$$+ \sum_{j=1}^{p_{l-2}} \alpha_j^{\min,l-1} \min\{x_{i,\mu_{j1}^{(l-2)}}, k_{j1}^{\min,l-1} x_{i,\mu_{j2}^{(l-2)}}, \ldots, k_{j,l-2}^{\min,l-1} x_{i,\mu_{j,l-1}^{(l-2)}}\} +$$

$$+ \sum_{j=1}^{p_{l-2}} \alpha_j^{\max,l-1} \max\{x_{i,\mu_{j1}^{(l-2)}}, k_{j1}^{\max,l-1} x_{i,\mu_{j2}^{(l-2)}}, \ldots, k_{j,l-2}^{\max,l-1} x_{i,\mu_{j,l-1}^{(l-2)}}\} +$$

$$+ \alpha_1^{\min,l} \min\{x_{i1}, k_1^{\min,l} x_{i2}, \ldots, k_{l-1}^{\min,l} x_{il}\} +,$$

$$\alpha_1^{\max,l} \max\{x_{i1}, k_1^{\max,l} x_{i2}, \ldots, k_{l-1}^{\max,l} x_{il}\} + \varepsilon_i$$

$$i = \overline{1,n}, \qquad (2)$$

где min (max) – бинарные, тернарные, ..., l -арные операции, возвращающие минимум (максимум) двух, трех, ..., l чисел; $\mu_{jh}^{(s-1)}$ ($s = \overline{2,l}$, $j = \overline{1,p_{s-1}}$, $h = \overline{1,s}$) – элемент j -й строки и h – столбца матрицы M_{s-1} размера $p_{s-1} \times s$, где $p_{s-1} = C_l^s$, содержащей по строкам в лексикографическом порядке индексы всех возможных сочетаний из l входных переменных по s ; α_j , $j = \overline{0,l}$ – неизвестные параметры линейной части; $\alpha_j^{\min,s}$, $\alpha_j^{\max,s}$ ($s = \overline{2,l}$, $j = \overline{1,p_{s-1}}$) – неизвестные параметры при s -арных операциях min и max, включающих j -ю комбинацию входных переменных; $k_{jh}^{\min,s}$, $k_{jh}^{\max,s}$ ($s = \overline{2,l}$, $j = \overline{1,p_{s-1}}$, $h = \overline{1,s-1}$) – h -е неизвестные угловые коэффициенты в s -арных операциях min и max, включающих j -ю комбинацию входных переменных.

Рассмотрим частные случаи НЛР (2):

- если $\alpha_j^{\min,s} = 0$, $\alpha_j^{\max,s} = 0$ ($s = \overline{2,l}$, $j = \overline{1,p_{s-1}}$), то имеем линейную регрессию;
- если $\alpha_j^{\min,s} = 0$, $\alpha_j^{\max,s} = 0$ ($s = \overline{3,l}$, $j = \overline{1,p_{s-1}}$), то имеем НЛР (1) только с бинарными операциями min и max;
- если $\alpha_j = 0$ ($j = \overline{0,l}$), $\alpha_j^{\min,s} = 0$, $\alpha_j^{\max,s} = 0$ ($s = \overline{2,l-1}$, $j = \overline{1,p_{s-1}}$), $\alpha_1^{\max,l} = 0$, то имеем функцию Леонтьева;
- если $\alpha_j = 0$ ($j = \overline{0,l}$), $\alpha_j^{\min,s} = 0$, $\alpha_j^{\max,s} = 0$ ($s = \overline{2,l-1}$, $j = \overline{1,p_{s-1}}$), то имеем кусочно-линейную регрессию [19].

НЛР (2) содержит один свободный член α_0, l параметров при входных переменных, $2\cdot\left(\tilde{N}_l^2 + C_l^3 + ... + C_l^{l-1} + 1\right)$ параметров при операциях min и max, $2\cdot\left(\tilde{N}_l^2 + 2C_l^3 + ... + (l-2)C_l^{l-1} + l - 1\right)$ угловых коэффициентов внутри операций min и max. Итого в НЛР (2) содержится $\left(1 + l + 2\sum_{j=2}^{l} j\cdot C_l^j\right)$ неизвестных параметров. Их количество значительно увеличится, если в операциях min и max использовать свободные члены [20]. Таким образом, НЛР (2) можно отнести к очень гибкому инструменту регрессионного моделирования.

3. НЛР С *l*-АРНОЙ ОПЕРАЦИЕЙ MIN (MAX) БЕЗ ЛИНЕЙНОЙ ЧАСТИ

Рассмотрим частный случай модели (2) – НЛР с l-арной операцией min:

$$y_i = \alpha_0 + \alpha_1 \min\left\{x_{i1}, k_1 x_{i2}, k_2 x_{i3}, ..., k_{l-1} x_{il}\right\} + \varepsilon_i,\ i = \overline{1,n}. \qquad (3)$$

Заметим, что оценивание модели (3) равносильно оцениванию l-факторной функции Леонтьева со свободным членом.

Будем считать, что все значения входных переменных в (3) положительны.

Изначально регрессия (3) нелинейна по оцениваемым параметрам. Но если придать коэффициентам k_1, k_2, ..., k_{l-1} определенные значения, то НЛР (3) становится линейной по параметрам α_0 и α_1, оценки которых можно найти с помощью МНК. Возникает вопрос – в какой области D лежат оптимальные МНК-оценки параметров k_1, k_2, ..., k_{l-1}? Тот факт, что такую область можно выделить, не вызывает сомнения. Например, нет смысла использовать в (3) значения k_1, k_2, ..., k_{l-1} существенно превосходящие значения входных переменных, поскольку очевидно, что при $k_1 \to \infty$, $k_2 \to \infty$, ..., $k_{l-1} \to \infty$ в (3) всегда будет срабатывать только переменная x_1. Иными слова, параметры k_1, k_2, ..., k_{l-1} должны быть такими, чтобы каждая переменная на выборке срабатывала хотя бы один раз.

Будем формировать область D возможных значений параметров $k_1, k_2, ..., k_{l-1}$ следующим образом. Исключим из универсального множества U область $\overline{D}$, в которой для всех наблюдений не срабатывает хотя бы одна переменная. Эта область представляет собой совокупность линейных неравенств:

$$\begin{cases} x_{i1} \ge k_1 x_{i2},\ x_{i1} \ge k_2 x_{i3},\ ...,\ x_{i1} \ge k_{l-1} x_{il}, \\ k_1 x_{i2} \ge x_{i1},\ k_1 x_{i2} \ge k_2 x_{i3},\ ...,\ k_1 x_{i2} \ge k_{l-1} x_{il}, \\ ... \\ k_{l-1} x_{il} \ge x_{i1},\ k_{l-1} x_{il} \ge k_1 x_{i2},\ ...,\ k_{l-1} x_{il} \ge k_{l-2} x_{i,l-1}, \end{cases} \quad i = \overline{1,n}. \qquad (4)$$

Решения каждого неравенства совокупности (4) представлены в таблице 1. В ней $\lambda_{ij}^{\min} = \min\left\{\frac{x_{1i}}{x_{1j}}, \frac{x_{2i}}{x_{2j}}, ..., \frac{x_{ni}}{x_{nj}}\right\}$, $\lambda_{ij}^{\max} = \max\left\{\frac{x_{1i}}{x_{1j}}, \frac{x_{2i}}{x_{2j}}, ..., \frac{x_{ni}}{x_{nj}}\right\}$, $i = \overline{1,l-1}$, $j = \overline{i+1,l}$.

Таблица 1

Решения неравенств (4)

–	$k_1 \le \lambda_{12}^{\min}$	$k_2 \le \lambda_{13}^{\min}$	…	$k_{l-2} \le \lambda_{1,l-1}^{\min}$	$k_{l-1} \le \lambda_{1,l}^{\min}$
$k_1 \ge \lambda_{12}^{\max}$	–	$k_2 \le \lambda_{23}^{\min} k_1$	…	$k_{l-2} \le \lambda_{2,l-1}^{\min} k_1$	$k_{l-1} \le \lambda_{2,l}^{\min} k_1$
$k_2 \ge \lambda_{13}^{\max}$	$k_2 \ge \lambda_{23}^{\max} k_1$	–	…	$k_{l-2} \le \lambda_{3,l-1}^{\min} k_2$	$k_{l-1} \le \lambda_{3,l}^{\min} k_2$
…	…	…	…	…	…
$k_{l-2} \ge \lambda_{1,l-1}^{\max}$	$k_{l-2} \ge \lambda_{2,l-1}^{\max} k_1$	$k_{l-2} \ge \lambda_{3,l-1}^{\max} k_2$	…	–	$k_{l-1} \le \lambda_{l-1,l}^{\min} k_{l-2}$
$k_{l-1} \ge \lambda_{1,l}^{\max}$	$k_{l-1} \ge \lambda_{2,l}^{\max} k_1$	$k_{l-1} \ge \lambda_{3,l}^{\max} k_2$	…	$k_{l-1} \ge \lambda_{l-1,l}^{\max} k_{l-2}$	–

Исключение из области U области $\overline{D}$ означает, что необходимо заменить знаки всех неравенств в таблице 1 на противоположные и внести полученные неравенства в систему. Таким образом, область D представляет собой решение системы линейных неравенств, приведенных в таблице 2.

Таблица 2

Неравенства из системы, идентифицирующей область D

–	$k_1 > \lambda_{12}^{\min}$	$k_2 > \lambda_{13}^{\min}$	…	$k_{l-2} > \lambda_{1,l-1}^{\min}$	$k_{l-1} > \lambda_{1,l}^{\min}$
$k_1 < \lambda_{12}^{\max}$	–	$k_2 > \lambda_{23}^{\min} k_1$	…	$k_{l-2} > \lambda_{2,l-1}^{\min} k_1$	$k_{l-1} > \lambda_{2,l}^{\min} k_1$
$k_2 < \lambda_{13}^{\max}$	$k_2 < \lambda_{23}^{\max} k_1$	–	…	$k_{l-2} > \lambda_{3,l-1}^{\min} k_2$	$k_{l-1} > \lambda_{3,l}^{\min} k_2$
…	…	…	…	…	…
$k_{l-2} < \lambda_{1,l-1}^{\max}$	$k_{l-2} < \lambda_{2,l-1}^{\max} k_1$	$k_{l-2} < \lambda_{3,l-1}^{\max} k_2$	…	–	$k_{l-1} > \lambda_{l-1,l}^{\min} k_{l-2}$
$k_{l-1} < \lambda_{1,l}^{\max}$	$k_{l-1} < \lambda_{2,l}^{\max} k_1$	$k_{l-1} < \lambda_{3,l}^{\max} k_2$	…	$k_{l-1} < \lambda_{l-1,l}^{\max} k_{l-2}$	–

Заметим, что для НЛР с l -арной операцией max область D будет точно такой же.

Решение системы линейных неравенств, перечисленных в таблице 2, представляет собой открытый выпуклый многогранник (симплекс) в $(l-1)$ -мерном пространстве. Поскольку в НЛР (3) отсутствует линейная часть, то все неравенства в таблице 2 можно взять нестрогими. Тогда решение будет представлять собой замкнутый выпуклый многогранник.

Таким образом, для численного оценивания с помощью МНК НЛР (3) необходимо выбрать в области D некоторое множество точек, в каждой из них найти МНК-оценки параметров α_0 и α_1, и выбрать оценки, при которых сумма квадратов остатков регрессии минимальна.

Стоит отметить, что схожий алгоритм оценивания функций Леонтьева можно найти в монографии [21]. Однако в ней область D предложено формировать как l-мерный параллелепипед. Тем самым, представленный в настоящей работе алгоритм, очевидно, эффективнее с вычислительной точки зрения.

4. ПРИМЕР

Для демонстрации предложенного математического аппарата решалась задача моделирования ВРП Иркутской области. Для этого на сайте Федеральной службы государственной статистики (https://rosstat.gov.ru/) были собраны ежегодные статистические данные за период с 2000 по 2020 гг. (таблица 3) по следующим переменным:

y – ВРП (млн руб.);

x_1 – продукция сельского хозяйства (млн руб.);

x_2 – инвестиции в основной капитал (млн руб.);

x_3 – объем работ, выполненных по виду экономической деятельности «Строительство» (млн руб.).

Таблица 3

Статистические данные

Год	y	x_1	x_2	x_3	Год	y	x_1	x_2	x_3
2000	103013,8	10006,09	10814	6511,9	2011	634561,4	40990,2	145537	63825,4
2001	120240	14543,88	15233,84	8400,4	2012	737971,6	44079,1	177641	89331,9
2002	140195,9	14894,12	17313,01	7577,2	2013	805197,5	46630	200063	94617
2003	167927,1	15568,4	22122,58	10193,8	2014	916317,5	51765,4	214422	89312,6
2004	213244,2	17824,91	26013,87	14917,2	2015	1001718	53600,8	206075	98839,4
2005	258095,5	19670,4	36675	20544,4	2016	1139207	58721,7	247954,2	131836
2006	330834,3	21925,7	70671,53	28107,2	2017	1268312	61900,4	270018,6	130347,8
2007	402654,7	25942,8	121877,8	45445,2	2018	1460512	63549	318786,9	113826,4
2008	438852,4	29996,7	129951	50022,9	2019	1540238	62154	366723,7	158311,4
2009	458774,9	33196,1	106550	47795,8	2020	1505151	67043	389990,1	164413,5
2010	546141	35119,9	119395	55017,7					

Все перечисленные переменные тесно коррелируют между собой. Так, коэффициент корреляции между переменными y и x_1 равен 0,9798, между y и x_2–0,9874, между y и x_3–0,9784, между x_1 и x_2–0,9689, между x_1 и x_3–0,976, между x_2 и x_3–0,9828. Таким образом, заранее можно предположить, что при построении модели множественной линейной регрессии будет иметь место частичная мультиколлинеарность, которая, возможно, исказит знаки коэффициентов уравнения.

Действительно, построенная по данным из таблицы 3 линейная регрессия имеет вид:

$$\tilde{y} = -80519,7 + 9,995x_1 + 2,657x_2 - 0,383x_3. \quad (5)$$

Как видно, в уравнении (5) из-за мультиколлинеарности знак коэффициента при переменной x_3 противоречит содержательному смыслу задачи. Тем самым модель (5)

теряет способность быть интерпретируемой. Однако для прогнозирования её использовать можно, поскольку её коэффициент детерминации R^2 достаточно высок и составляет 0,983742.

Затем с помощью МНК оценивалась НЛР с тернарной операцией min и без линейной части. Для этого предварительно были вычислены следующие характеристики:

$$\lambda_{12}^{\min} = \min\left\{\frac{x_{11}}{x_{12}}, \frac{x_{21}}{x_{22}}, ..., \frac{x_{n1}}{x_{n2}}\right\} = 0,169485\,,\ \lambda_{12}^{\max} = \max\left\{\frac{x_{11}}{x_{12}}, \frac{x_{21}}{x_{22}}, ..., \frac{x_{n1}}{x_{n2}}\right\} = 0,954709\,,$$

$$\lambda_{13}^{\min} = \min\left\{\frac{x_{11}}{x_{13}}, \frac{x_{21}}{x_{23}}, ..., \frac{x_{n1}}{x_{n3}}\right\} = 0,392606\,,\ \lambda_{13}^{\max} = \max\left\{\frac{x_{11}}{x_{13}}, \frac{x_{21}}{x_{23}}, ..., \frac{x_{n1}}{x_{n3}}\right\} = 1,965649\,,$$

$$\lambda_{23}^{\min} = \min\left\{\frac{x_{12}}{x_{13}}, \frac{x_{22}}{x_{23}}, ..., \frac{x_{n2}}{x_{n3}}\right\} = 1,660652\,,\ \lambda_{23}^{\max} = \max\left\{\frac{x_{12}}{x_{13}}, \frac{x_{22}}{x_{23}}, ..., \frac{x_{n2}}{x_{n3}}\right\} = 2,800641\,.$$

Для идентификации области D с помощью таблицы 2 была составлена следующая система линейных неравенств:

$$\begin{cases} k_1 \geq 0,169485, \\ k_2 \geq 0,392606, \\ k_1 \leq 0,954709, \\ k_2 \geq 1,660652k_1, \\ k_2 \leq 1,965649, \\ k_2 \leq 2,800641k_1. \end{cases} \tag{6}$$

Область D решений системы линейных неравенств (6) представлена на рис. 1.

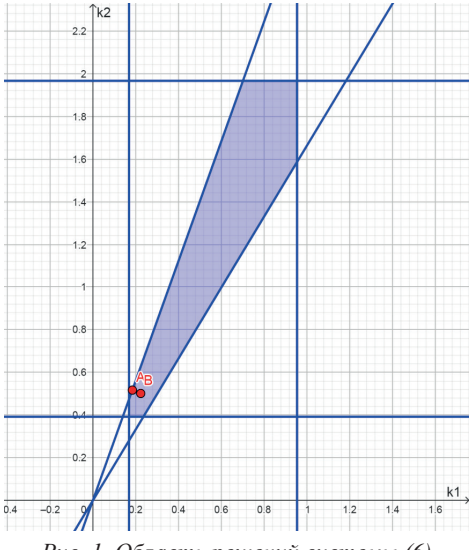

Рис. 1. Область решений системы (6)

Как следует из рис. 1, область D на плоскости представляет собой замкнутый выпуклый шестиугольник. Подчеркнём, что в системе (6) во всех неравенствах были взяты нестрогие знаки, поскольку оцениваемая НЛР представляет собой функцию Леонтьева и не содержит линейной части.

Для численного оценивания НЛР был разработан специальный скрипт на языке программирования hansl пакета Gretl. Скрипт работает по следующему алгоритму. Прямоугольник $k_1 \geq 0,169485 \wedge k_1 \leq 0,954709 \wedge k_2 \geq 0,392606 \wedge k_2 \leq 1,965649$ равномерно разбивается десятью тысячами точек. В каждой такой точке и на границе области D проверяется выпол-

нение условий $k_2 \geq 1,660652k_1 \wedge k_2 \leq 2,800641k_1$. Если условия выполнены, то точка принадлежит области D, поэтому для неё находятся МНК-оценки НЛР. Всего в область D попало 3422 точки. Лучшая регрессия по величине суммы квадратов остатков была зафиксирована в точке $A(0.185034, 0.517203)$ вблизи границы области D (см. рис. 1). Полученной точке A соответствует следующая НЛР с тернарной операцией min:

$$\tilde{y} = \underset{(29,12)}{35627,5 + 23,026} \min\{x_1, 0.185034x_2, 0.517203x_3\}. \quad (7)$$

В уравнении (7) в скобках под коэффициентом 23,026 приведено значение t-критерия Стьюдента, подтверждающее значимость регрессора min $\{x_1, 0.185034x_2, 0.517203x_3\}$. Для модели (7) $R^2 = 0,978088$. Мультиколлинеарности в регрессии (7) по определению нет, поэтому все коэффициенты при переменных можно интерпретировать. Недостаток НЛР (7) в том, что практически всегда в тернарной операции min срабатывает только переменная x_2. Так происходит в 18 наблюдениях из 21 (2000–2017 гг.). Переменная x_1 срабатывает всего 2 раза (2019 и 2020 год), а переменная x_3 – 1 раз (2018 год). Всё это сигнализирует о том, что вместо модели (7) можно было обойтись простой парной линейной регрессией y от x_2. Действительно, такая регрессия имеет вид

$$\tilde{y} = \underset{(27,21)}{54617,3 + 4,058}\, x_2, \quad (8)$$

для которой $R^2 = 0,974977$. Как видно, все коэффициенты и аппроксимационные характеристики моделей (7) и (8) практически не отличаются. В такой ситуации предпочтение следует отдать более простой модели (8). Заметим, что так произошло потому, что точка A (см. рис. 1) оказалась практически на границе области D, на которой нет гарантии срабатывания каждой переменной на выборке хотя бы 1 раз.

После чего с помощью МНК оценивалась НЛР с тернарной операцией max и без линейной части. Для такой модели область D имеет точно такую же конфигурацию, что и на рис. 1. С помощью того же скрипта была найдена лучшая регрессия по величине суммы квадратов остатков. Она была зафиксирована в точке $B(0.223906, 0.501629)$ внутри области D (см. рис. 1). Точке B соответствует следующая НЛР с тернарной операцией max:

$$\tilde{y} = \underset{(35,98)}{-152442 + 20,4354} \max\{x_1, 0.223906x_2, 0.501629x_3\}. \quad (9)$$

Коэффициент детерминации НЛР (9) равен 0,985536, что больше чем у любой из приведенных выше моделей (5), (7), (8). Коэффициент при регрессоре $\max\{x_1, 0.223906x_2, 0.501629x_3\}$ значим по t-критерию Стьюдента. Переменная x_1 срабатывает 13 раз (2000–2006, 2008–2011, 2014 и 2015 годы), переменная x_2 – 4 раза (2007, 2018–2020 годы), а переменная x_3 – 4 раза (2012, 2013, 2016, 2017 годы).

Мультиколлинеарности в регрессии (9) нет, поэтому все коэффициенты при переменных можно интерпретировать. Для этого представим НЛР (9) в кусочно-заданной форме:

$$\tilde{y} = \begin{cases} -152442 + 20,435x_1, \text{ если } \dfrac{x_1}{x_2} \geq 0,223906, \ \dfrac{x_1}{x_3} \geq 0,501629, \\ -152442 + 4,576x_2, \text{ если } \dfrac{x_1}{x_2} < 0,223906, \ \dfrac{x_2}{x_3} \geq 2,240355, \\ -152442 + 10,251x_3, \text{ если } \dfrac{x_1}{x_3} < 0,501629, \ \dfrac{x_2}{x_3} < 2,240355. \end{cases}$$

Тогда модель (9) можно интерпретировать следующим образом.

1. Если отношение продукции сельского хозяйства x_1 к инвестициям в основной капитал x_2 не меньше 0,223906 и отношение x_1 к объемам строительных работ x_3 не меньше 0,501629, то на ВРП оказывает влияние только продукция сельского хозяйства x_1. Причем, с увеличением x_1 на 1 млн руб. ВРП y увеличивается в среднем на 20,435 млн руб.
2. Если отношение x_1 к x_2 меньше 0,223906 и отношение x_2 к x_3 не меньше 2,240355, то на ВРП оказывают влияние только инвестиции в основной капитал x_2. Причем, с увеличением x_2 на 1 млн руб. ВРП y увеличивается в среднем на 4,576 млн руб.
3. Если отношение x_1 к x_3 меньше 0,501629 и отношение x_2 к x_3 меньше 2,240355, то на ВРП оказывают влияние только объемы строительных работ x_3. Причем, с увеличением x_3 на 1 млн руб. ВРП y увеличивается в среднем на 10,251 млн руб.

5. ЗАКЛЮЧЕНИЕ

В статье впервые введены неэлементарные линейные регрессии с линейной частью и со всеми возможными комбинациями бинарных, тернарных, ..., l -арных операций min и max. Такие модели обобщают многие известные на сегодняшний день регрессионные модели, в частности, специфицированные на основе функций Леонтьева регрессии. Предложен алгоритм оценивания с помощью МНК НЛР с l -арной операцией min (max). На первом шаге алгоритма строится область возможных значений угловых коэффициентов, представляющая собой решение системы линейных неравенств (таблица 2). На втором шаге из этой области выбирается точка, в которой сумма квадратов остатков НЛР минимальна. С помощью предложенного алгоритма решена задача моделирования ВРП Иркутской области. В результате была построена НЛР с тернарной операцией max, качество аппроксимации которой оказалось выше, чем у линейной регрессии. Дана интерпретация построенной модели. Таким образом, обобщенные НЛР (2) представляют собой довольно гибкий инструмент математического моделирования, просто интерпретируются и могут эффективно применяться в прогнозировании.

Очевидно, что для оценивания с помощью МНК обобщенной НЛР (2) требуется для каждой входящей в неё операции min или max формировать свою область возможных значений угловых коэффициентов, а затем в каждой из этих областей выбирать по одной точке так, чтобы минимизировать сумму квадратов остатков. Такая

задача довольно сложна с вычислительной точки зрения, поэтому требует в будущем разработки специализированного программного продукта.

Литература

1. *Хенрик Б., Джозеф Р., Марк Ф.* Машинное обучение. СПб.: Питер, 2017. 336 с.
2. *Флах П.* Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / пер. с англ. А.А. Слинкина. М.: ДМК Пресс, 2015. 400 с.
3. *Molnar C.* Interpretable machine learning. Lulu. com, 2020.
4. *Doshi-Velez F., Kim B.* Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. 2017.
5. *Montgomery D.C., Peck E.A., Vining G.G.* Introduction to linear regression analysis. John Wiley & Sons, 2021.
6. *Keith T.Z.* Multiple regression and beyond: An introduction to multiple regression and structural equation modeling. Routledge, 2019.
7. *Gelman A., Hill J., Vehtari A.* Regression and other stories. Cambridge University Press, 2020.
8. *Брачунова У.В.* Численное моделирование зарядного баланса легкового автомобиля // Известия Тульского государственного университета. Технические науки. 2022. № 9. С. 453–458.
9. *Ярымбаш Д.С., Коцур М.И., Ярымбаш С.Т., Килимник И.М.* Моделирование электромагнитных процессов при работе силовых трансформаторов под нагрузкой и в режиме холостого хода // Проблемы региональной энергетики. 2020. № 1 (45). С. 1–13.
10. *Балгарина Л., Джумабаев С., Шокаманов Ю.* Производственная функция Кобба–Дугласа: опыт применения в Северо-Казахстанской области // Экономическая серия вестника Евразийского национального университета имени Л.Н. Гумилева. 2022. Т. 141. № 4.
11. *Чесноков Е.А.* Сравнение регрессионных моделей экономического развития России // Московский экономический журнал. 2021. № 7. С. 96–105.
12. *Базилевский М.П.* Построение степенно-показательных и линейно-логарифмических регрессионных моделей // Проблемы управления. 2021. № 3. С. 25–32.
13. *Рева С.А., Арнаутов А.В., Клиценко О.А., Петров С.Б.* Прогностическая значимость логистической регрессионной модели для оценки риска рецидива у больных раком предстательной железы после радикальной простатэктомии // Research'n Practical Medicine Journal. 2022. Т. 9. № 4. С. 96–105.
14. *Кокоулина М.В., Епифанова А., Пелиновский Е.Н., Куркина О.Е., Куркин А.А.* Анализ динамики распространения коронавируса с помощью обобщенной логистической модели // Труды НГТУ им. Р.Е. Алексеева. 2020. № 3 (130). С. 28–41.
15. *Клейнер Г.Б.* Производственные функции: Теория, методы, применение. М.: Финансы и статистика, 1986. 239 с.
16. *Базилевский М.П.* Оценивание линейно-неэлементарных регрессионных моделей с помощью метода наименьших квадратов // Моделирование, оптимизация и информационные технологии. 2020. Т. 8. № 4 (31).
17. *Базилевский М.П.* Отбор информативных операций при построении линейно-неэлементарных регрессионных моделей // International Journal of Open Information Technologies. 2021. Т. 9. № 5. С. 30–35.
18. *Базилевский М.П.* Метод построения неэлементарных линейных регрессий на основе аппарата математического программирования // Проблемы управления. 2022. № 4. С. 3–14.
19. *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). С. 47–55.

20. *Базилевский М.П.* Оценка методом наименьших квадратов простейших неэлементарных линейных регрессий с линейным аргументом в бинарной операции // Вестник кибернетики. 2022. № 4 (48). С. 69–76.
21. *Носков С.И.* Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. Иркутск, РИЦ ГП «Облинформпечать», 1996. 320 с.

Generalization of Non-elementary Linear Regressions

Mikhail P. Bazilevskiy*
Irkutsk State Transport University (ISTU), Irkutsk, Russia
ORCID: https://orcid.org/0000-0002-3253-5697
e-mail: mik2178@yandex.ru

Earlier, the author developed a non-elementary linear regression consisting of a linear part and all possible combinations of min and max binary operations. This article is devoted to its generalization. For the first time a non-elementary linear regression with a linear part and all possible combinations of binary, ternary, ..., *l*-ary operations min and max has been introduced. The proposed model generalizes both linear regression and the Leontief function, and can be effectively used both for predicting and for interpreting the study object functioning. An estimation algorithm was developed using the method of least squares for non-elementary linear regressions without a linear part and with an *l*-ary operation min (max), i.e. regressions with specification in the form of a Leontief function. The essence of the algorithm is to form a set of possible values of slope coefficients, from which a point is selected with the minimum value of the residual sum of squares. A system of linear inequalities is identified that makes it possible to form such a set. Using the algorithm, a model of the gross regional product of the Irkutsk region was construct and its interpretation was given.

Keywords: machine learning, regression model, non-elementary linear regression, ordinary least squares method, Leontief function, multicollinearity.

For citation:
Bazilevskiy M.P. Generalization of Non-elementary Linear Regressions. *Modelirovanie i analiz dannykh = Modelling and Data Analysis*, 2023. Vol. 13, no. 2, pp. 85–98. DOI: 10.17759/mda.2023130205 (In Russ., abstr. in Engl.).

References

1. Khenrik B., Dzhozef R., Mark F. *Mashinnoe obuchenie* [Machine Learning]. Saint Petersburg, Piter, 2017. 336 p.
2. Flakh P. *Mashinnoe obuchenie. Nauka i iskusstvo postroeniya algoritmov, kotorye izvlekayut znaniya iz dannykh* [Machine Learning. The Art and Science of Algorithms that Make Sense of Data]. Moscow, DMK Press, 2015. 400 p.
3. Molnar C. *Interpretable machine learning*. Lulu. com, 2020.
4. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.
5. Montgomery D.C., Peck E.A., Vining G.G. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.

Mikhail P. Bazilevskiy, PhD (Engineering), Associate Professor, Department of Mathematics, Irkutsk State Transport University (ISTU), Irkutsk, Russia, ORCID: https://orcid.org/0000-0002-3253-5697, e-mail: mik2178@yandex.ru

6. Keith T.Z. *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling*. Routledge, 2019.
7. Gelman A., Hill J., Vehtari A. *Regression and other stories.* Cambridge University Press, 2020.
8. Brachunova U.V. Chislennoe modelirovanie zaryadnogo balansa legkovogo avtomobilya [Numerical simulation of the charging balance of a passenger car], *Proceedings of the TSU. Technical Sciences*, 2022, no. 9, pp. 453–458.
9. Yarymbash D.S., Kotsur M.I., Yarymbash S.T., Kilimnik I.M. Modelirovanie elektromagnitnykh protsessov pri rabote silovykh transformatorov pod nagruzkoy i v rezhime kholostogo khoda [Electromagnetic Processes Simulation of Power Transformers in Operation and in No-load Mods], *Problemele Energeticii Regionale*, 2020, no. 1 (45), pp. 1–13.
10. Balgarina L., Dzhumabaev S., Shokamanov Yu. Proizvodstvennaya funktsiya Kobba–Duglasa: opyt primeneniya v Severo-Kazakhstanskoy oblasti [Cobb – Douglas Production Function: application experience in the North Kazakhstan region], *Economic Series of the Bulletin of the L.N. Gumilyov ENU*, 2022, vol. 141, no. 4.
11. Chesnokov E.A. Sravnenie regressionnykh modeley ekonomicheskogo razvitiya Rossii [Comparison of regression models of economic development in Russia], *Moscow economic journal*, 2021, no. 7, pp. 96–105.
12. Bazilevskiy M.P. Postroenie stepenno-pokazatel'nykh i lineyno-logarifmicheskikh regressionnykh modeley [Constructing power-exponential and linear-logarithmic regression models], *Control Sciences*, 2021, no. 3, pp. 25–32.
13. Reva S.A., Arnautov A.V., Klitsenko O.A., Petrov S.B. Prognosticheskaya znachimost' logisticheskoy regressionnoy modeli dlya otsenki riska retsidiva u bol'nykh rakom predstatel'noy zhelezy posle radikal'noy prostatektomii [Prognostic significance of the logistic regression model for assessing the risk of recurrence in patients with prostate cancer after radical prostatectomy], *Research'n Practical Medicine Journal*, 2022, vol. 9, no. 4, pp. 96–105.
14. Kokoulina M.V., Epifanova A., Pelinovskiy E.N., Kurkina O.E., Kurkin A.A. Analiz dinamiki rasprostraneniya koronavirusa s pomoshch'yu obobshchennoy logisticheskoy modeli [Analysis of coronavirus dynamics using the generalized logistic model], *Proceedings of NSTU n.a. R.E. Alekseev*, 2020, no. 3 (130), pp. 28–41.
15. Kleyner G.B. *Proizvodstvennye funktsii: Teoriya, metody, primenenie* [Production functions: Theory, methods, application]. Moscow: Finance and Statistics, 1986. 239 p.
16. Bazilevskiy M.P. Otsenivanie lineyno-neelementarnykh regressionnykh modeley s pomoshch'yu metoda naimen'shikh kvadratov [Estimation linear non-elementary regression models using ordinary least squares], *Modeling, optimization and information technology*, 2020, vol. 8, no. 4 (31).
17. Bazilevskiy M.P. 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*, 2021, vol. 9, no. 5, pp. 30–35.
18. Bazilevskiy M.P. Metod postroeniya neelementarnykh lineynykh regressiy na osnove apparata matematicheskogo programmirovaniya [A method for constructing nonelementary linear regressions based on mathematical programming], *Control Sciences*, 2022, no. 4, pp. 3–14.
19. Noskov S.I., Khonyakov A.A. Programmnyy kompleks postroeniya nekotorykh tipov kusochno-lineynykh regressiy [Software complex for building some types pieces of linear regressions], *Information technology and mathematical modeling in the management of complex systems*, 2019, no. 3 (4), pp. 47–55.
20. Bazilevskiy M.P. 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], *Proceedings in Cybernetics*, 2022, no. 4 (48), pp. 69–76.

21. Noskov S.I. *Tekhnologiya modelirovaniya ob"ektov s nestabil'nym funktsionirovaniem i neopredelennost'yu v dannykh* [Technology for modeling objects with unstable operation and uncertainty in data]. Irkutsk, RITs GP «Oblinformpechat'», 1996. 320 p.

Получена 24.04.2023 Received 24.04.2023
Принята в печать 19.05.2023 Accepted 19.05.2023