

*М. П. Базилевский*¹

1 Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация

ИНТЕРПРЕТАЦИЯ НЕЭЛЕМЕНТАРНЫХ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Аннотация. Статья посвящена исследованию интерпретационных характеристик неэлементарных линейных регрессионных моделей. Рассмотрен алгоритм приближенного оценивания таких моделей с помощью метода наименьших квадратов и одна из возможных стратегий их построения. С использованием встроенных в эконометрический пакет Gretl статистических данных была построена неэлементарная линейная регрессия с семью объясняющими переменными. Для этого впервые использовалось варьирование бинарных операций минимум и максимум. Полученная модель характеризуется низкой степенью мультиколлинеарности и существенно превзошла линейную регрессию по величине коэффициента детерминации. Интерпретация полученной модели позволила выявить новые закономерности функционирования зависимой переменной, недоступные при использовании классической линейной регрессии.

Ключевые слова: неэлементарная линейная регрессия, структурная спецификация, интерпретируемость, метод наименьших квадратов, мультиколлинеарность, бинарные операции.

*M.P. Bazilevskiy*¹

¹ Irkutsk State Transport University, Irkutsk, Russia

INTERPRETATION OF NON-ELEMENTARY LINEAR REGRESSION MODELS

Abstract. The article is devoted to the study of interpretive characteristics of non-elementary linear regression models. An algorithm for the approximate estimation of such models using the ordinary least squares and one of the possible strategies for their construction are considered. Using the statistical data built-in Gretl econometric package, a non-elementary linear regression with seven explanatory variables was construct. For this, for the first time, the variation of the binary operations minimum and maximum was used. The resulting model is characterized by a low degree of multicollinearity and significantly outperformed linear regression in terms of the coefficient of determination. The interpretation of the resulting model made it possible to identify new patterns in the functioning of the dependent variable, which are not available when using classical linear regression.

Keywords: non-elementary linear regression, structural specification, interpretability, ordinary least squares, multicollinearity, binary operations.

Введение.

Приступая к регрессионному моделированию [1], исследователь в первую очередь должен определиться со структурной спецификацией оцениваемого регрессионного уравнения. К настоящему времени в рамках регрессионного анализа разработано значительное количество хорошо известных математических форм связи между переменными. К ним относится традиционная модель множественной линейной регрессии, полиномиальная, степенная и логарифмическая регрессии, логистическая регрессия и т.д. В [2–5] приводится обзор часто применяемых в экономике так называемых производственных функций: Кобба-Дугласа, Леонтьева, Солоу, Джири, CES (Consistent Elasticity of Substitution), Реванкара, Сато, Лиу-Хильдебрандта, Лу-Флетчера, Кадияла, Бруно, Солоу-Мукерджи, Сато-Гофмана. Вместе с тем продолжается процесс создания новых структурных спецификаций регрессионных моделей. Так, в [6] разработан алгоритм построения линейно-мультипликативных регрессий, а в [7] приводится описание реализующего его программного комплекса. В [8] предложены индексные модели регрессии и рассмотрены вопросы их оценивания. В [9] предложено два обобщения производственной функции Кобба-Дугласа: регрессия со степенями в виде линейных комбинаций объясняющих переменных и в виде линейных комбинаций их натуральных логарифмов. В [10] обсуждаются вопросы

построения степенно-показательных и линейно-логарифмических регрессионных моделей. В монографии [11] рассмотрено совместное определение точек переключения и параметров кусочно-линейной зависимости. В [12] исследуется возможность построения кусочно-линейных регрессий с нелинейными границами переключения. В [13] рассмотрена регрессия, противоположная по смыслу производственной функции Леонтьева, а в [14] она получила название «функция риска». В [15] предложен симбиоз производственной функции Леонтьева и функции риска. Заметим, что в работах [11–15] для нахождения неизвестных оценок параметров моделей применяется метод наименьших модулей. В [16] рассмотрены вопросы оценивания параметров специфицированных на основе функций Леонтьева двухфакторных моделей регрессии с помощью метода наименьших квадратов (МНК). На основе этой работы в [17] предложены неэлементарные линейные регрессии (НЛР) и предложен алгоритм их приближенного МНК-оценивания. В [18] сформулирована задача отбора информативных операций в НЛР и разработаны две стратегии её решения.

В настоящее время актуальным научным направлением является построение интерпретируемых моделей. К сожалению, пока еще никем не сформулирован четкий ответ на вопрос, что же такое интерпретируемость модели. Так, в монографии [19], посвященной интерпретируемому машинному обучению, автор придерживается следующих двух определений, ссылаясь на работы [20] и [21]. Первое из них дословно формулируется так: «интерпретируемость – это степень, с которой человек может понять причину решения». Второе гласит: «интерпретируемость – степень, с которой человек может предварительно предсказывать результат модели». Популярными сегодня нейронные сети лишены такой способности, поскольку относятся к моделям типа «черного ящика». Как отмечено в [22], они являются очень хорошим инструментом для решения конкретных задач машинного обучения, но не дают никакой информации аналитикам о том, как работает модель. Целью данной работы является демонстрация на примере решения конкретной задачи того, что предложенные в работах [16–18] НЛР обладают хорошими интерпретационными свойствами.

1. Неэлементарная линейная регрессия

Модель НЛР [17, 18] имеет вид:

$$y_i = \alpha_0 + \sum_{j=1}^l \alpha_j x_{ij} + \sum_{j=1}^{C_l^2} \alpha_{j+l} \min \{ x_{i, \mu_{j1}}, \lambda_j x_{i, \mu_{j2}} \} + \varepsilon_i, \quad i = \overline{1, n}, \quad (1)$$

где n – объем выборки; l – количество объясняющих переменных; y_i , $i = \overline{1, n}$ – значения объясняемой переменной y ; $x_{ij} > 0$, $i = \overline{1, n}$, $j = \overline{1, l}$ – значения объясняющих переменных $x_1, x_2, \dots, x_l$; μ_{j1} и μ_{j2} , $j = \overline{1, C_l^2}$ – элементы первого и второго столбца индексной матрицы $M_{C_l^2 \times 2}$, содержащей по строкам всевозможные комбинации пар индексов переменных; α_j , $j = \overline{0, l + C_l^2}$, λ_j , $j = \overline{1, C_l^2}$ – неизвестные параметры; ε_i , $i = \overline{1, n}$ – ошибки аппроксимации.

К сожалению, на сегодняшний день не существует точного метода оценивания модели (1). Приближенный метод предложен в [17]. Его суть состоит в том, что при фиксировании параметров λ_j , $j = \overline{1, C_l^2}$ НЛР становится линейной, поэтому для нахождения оценок её параметров α_j , $j = \overline{0, l + C_l^2}$ можно применять МНК. Тогда обычным перебором значений параметров λ_j , $j = \overline{1, C_l^2}$ из некоторых интервалов можно получить близкие к оптимальным по выбранному критерию оценки всех параметров НЛР.

В [16] установлено, что для перебора параметров λ_j , $j = \overline{1, C_l^2}$ достаточно ограничиться

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

$\lambda_{\max}^{(j)} = \max \left\{ \frac{x_{1, \mu_{j1}}}{x_{1, \mu_{j2}}}, \frac{x_{2, \mu_{j1}}}{x_{2, \mu_{j2}}}, \dots, \frac{x_{n, \mu_{j1}}}{x_{n, \mu_{j2}}} \right\}$. Точки $\lambda_j = \lambda_{\min}^{(j)}$ и $\lambda_j = \lambda_{\max}^{(j)}$ нельзя использовать из-за

возникновения совершенной мультиколлинеарности переменных.

Блок-схема алгоритма приближенного оценивания НЛР (1) изображена на рис. 1.

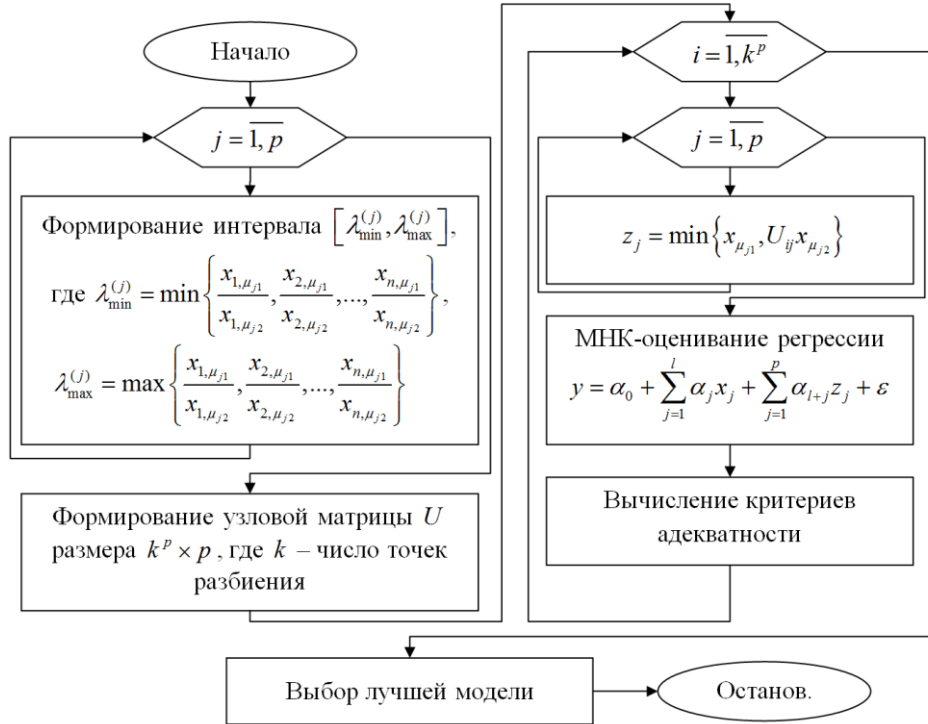


Рис. 1. Блок-схема алгоритма приближенного оценивания НЛР

На рис. 1 число $p = C_l^2$, а k – заданное число точек разбиения промежутков $\lambda_j \in (\lambda_{\min}^{(j)}, \lambda_{\max}^{(j)})$, $j = \overline{1, C_l^2}$. Тогда для получения приближенных оценок параметров НЛР требуется оценить $k^{C_l^2}$ линейных регрессий.

В НЛР (1) с ростом числа объясняющих переменных резко возрастает число бинарных операций минимум. Поэтому в [18] для НЛР была сформулирована следующая задача отбора информативных операций: из общего числа l переменных и C_l^2 бинарных операций требуется выбрать m штук на основе некоторого критерия качества. Для решения этой задачи в той же работе было предложено две стратегии. Рассмотрим вторую из них.

Вторая стратегия построения НЛР формулируется следующим образом: НЛР имеет наибольшее количество бинарных операций и включает в себя s объясняющих переменных ровно по одному разу. В этом случае количество бинарных операций bin определяется по формуле $\text{bin} = \left\lfloor \frac{s}{2} \right\rfloor$, где $\lfloor \cdot \rfloor$ – целая часть числа. Общее число линейных моделей r , которое

необходимо оценить для построения НЛР по второй стратегии, находится по формуле

$$r = k^{\text{bin}} C_l^s \frac{\prod_{q=0}^{\text{bin}-1} C_{s-2q}^2}{\text{bin}!}. \quad (2)$$

s переменных из l , затем для каждой из них – $\frac{\prod_{q=0}^{\text{bin}-1} C_{s-2q}^2}{\text{bin}!}$ комбинаций спецификаций НЛР, после чего каждая из них оценивается по алгоритму, изображенному на рис. 1. Заметим, что в составе полученной в результате оценивания НЛР будет содержаться s переменных, но число её степеней свободы $df = n - 1 - s + \text{bin}$, что больше, чем для линейной регрессии, для которой $df = n - 1 - s$, поэтому реализация второй стратегии – это хороший инструмент для борьбы с мультиколлинеарностью.

2. Моделирование

Моделирование осуществлялось с использованием встроенных в пакет Gretl статистических данных (data7-12) о ценах и характеристиках двухдверных седанов и хэтчбеков. Объем выборки – 82. Зависимой переменной выступает *price* – цена автомобиля (в тысячах долларов). Фиктивные переменные *hatch* и *trans*, а также целочисленная переменная *cyl* были исключены из рассмотрения. В результате был сформирован следующий список из семи объясняющих переменных:

wbase – колесная база (расстояние между осями передних и задних колес) (в дюймах);

length – длина автомобиля (в дюймах);

width – ширина автомобиля (в дюймах);

height – высота автомобиля (в дюймах);

weight – вес автомобиля (в сотнях фунтов);

liters – объем двигателя (в литрах);

gasmpg – расход топлива (миль на галлон).

Для удобства обозначим $y = price$, $x_1 = wbase$, $x_2 = length$, $x_3 = width$, $x_4 = height$, $x_5 = weight$, $x_6 = liters$, $x_7 = gasmpg$.

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

$$\hat{y} = 58,238 + 0,0582 x_1 - 0,409 x_2 - 0,097 x_3 - 0,723 x_4 + 2,219 x_5 + 1,366 x_6 + 0,2 x_7. \quad (3)$$

(0,209) (-2,877) (-0,398) (-1,598) (4,729) (1,336) (0,907)

В уравнении (3) в скобках под коэффициентами указаны значения t-критерия Стьюдента. Как видно, значимыми оказались только переменные x_2 и x_5 . Коэффициент детерминации линейной регрессии (3) $R^2 = 0,6$, а число её степеней свободы $df = 74$.

Корреляционная матрица объясняющих переменных изображена на рис. 2.



Рис. 2. Корреляционная матрица

По матрице видно, что переменные x_2 и x_5 , x_2 и x_7 , x_5 и x_6 , x_5 и x_7 довольно тесно коррелируют друг с другом, что сигнализирует о наличии в модели (3) мультиколлинеарности. О том же самом можно судить по величине определителя этой матрицы $\Delta = 0,002189$.

Линейная регрессия (3) легко интерпретируется. Например, коэффициент при переменной x_1 , равный 0,0582, показывает, что с увеличением колесной базы x_1 на 1 дюйм, стоимость автомобиля увеличивается в среднем на 58,2 доллара.

Для приближенного МНК-оценивания модели НЛР по второй стратегии на языке программирования `hansl` пакета `Gretl` был разработан специальный скрипт. Ставилось условие, что в НЛР каждая из семи объясняющих переменных должна входить только один раз, т.е. $s = 7$. В таком случае число входящих в модель бинарных операций $\text{bin} = 3$. Число точек разбиения промежутков $\lambda_j \in (\lambda_{\min}^{(j)}, \lambda_{\max}^{(j)})$, $j = \overline{1, C_l^2}$, было выбрано $k = 9$. Тогда, согласно формуле (2), для построения НЛР требовалось оценить

$$r = k^{\text{bin}} C_l^s \frac{\prod_{q=0}^{\text{bin}-1} C_{s-2q}^2}{\text{bin}!} = 9^3 C_7^7 \frac{C_7^2 \cdot C_5^2 \cdot C_3^2}{3!} = 76545 \text{ линейных моделей с четырьмя переменными.}$$

Выбранное в результате перебора вариантов лучшее по величине коэффициента детерминации R^2 регрессионное уравнение имеет вид:

$$\begin{aligned} \tilde{y} = & 67,9984 - 0,069 \min_{(-0,922)} \{x_1, 32,122x_6\} - 0,506 \min_{(-5,111)} \{x_2, 9,1881x_7\} - \\ & - 0,497 \min_{(-2,409)} \{x_3, 1,3541x_4\} + 2,827 x_5, \end{aligned} \quad (4)$$

для которого $R^2 = 0,64497$, $\Delta = 0,0294$.

НЛР (4) имеет в своем составе все 7 объясняющих переменных, но её число степеней свободы $df = 77$ больше, чем у линейной регрессии (3). Как видно, по коэффициенту детерминации НЛР (4) также превосходит линейную регрессию (3). В уравнении (4) незначимым оказался только коэффициент при бинарной операции $\min\{x_1, 32,122x_6\}$. За счет увеличения числа степеней свободы по сравнению с (3) снизился эффект мультиколлинеарности.

НЛР (4) содержит в себе три одинаковых бинарных операции минимум. Но существует также бинарная операция максимум. Поэтому было принято решение расширить множество альтернативных вариантов НЛР за счет варьирования в них бинарных операций $\min$ и $\max$. Для этого с помощью МНК было оценено еще 7 НЛР:

$$\begin{aligned} \tilde{y} = & 8,0536 + 0,258 \min_{(1,195)} \{x_1, 1,8581x_3\} - 0,561 \min_{(-5,564)} \{x_2, 9,1881x_7\} + \\ & + 0,455 \max_{(3,459)} \{x_4, 12,781x_6\} + 2,109 x_5, \end{aligned} \quad (5)$$

$$\begin{aligned} \tilde{y} = & 72,178 - 0,47 \min_{(-4,2)} \{x_2, 9,1881x_7\} + 0,694 \max_{(8,996)} \{x_3, 3,3064x_5\} - \\ & - 0,947 \min_{(-2,085)} \{x_4, 50,042x_6\} + 0,1456 x_1, \end{aligned} \quad (6)$$

$$\begin{aligned} \tilde{y} = & 3,731 - 0,558 \min_{(-5,424)} \{x_2, 9,1881x_7\} + 0,636 \max_{(8,271)} \{x_3, 3,3064x_5\} + \\ & + 0,459 \max_{(3,436)} \{x_4, 12,781x_6\} + 0,293 x_1, \end{aligned} \quad (7)$$

$$\begin{aligned} \tilde{y} = & 49,068 + 1,392 \max_{(9,255)} \{x_1, 3,2212x_5\} - 0,505 \min_{(-5,745)} \{x_2, 137,59x_6\} - \\ & - 0,687 \min_{(-6,608)} \{x_3, 1,5374x_7\} - 1,057 x_4, \end{aligned} \quad (8)$$

$$\begin{aligned} \tilde{y} = & -7,18 + 1,361 \max\{x_1, 3,2212x_5\} - 0,657 \min\{x_3, 1,5374x_7\} + \\ & + 0,316 \max\{x_4, 12,781x_6\} - 0,584 x_2, \end{aligned} \quad (9)$$

$$\begin{aligned} \tilde{y} = & 40,339 + 1,429 \max\{x_1, 3,2212x_5\} - 0,26 \max\{x_3, 3,4445x_7\} - \\ & - 1,172 \min\{x_4, 45,385x_6\} - 0,468 x_2, \end{aligned} \quad (10)$$

$$\begin{aligned} \tilde{y} = & -34,01 + 1,471 \max\{x_1, 3,2212x_5\} - 0,276 \max\{x_3, 2,9678x_7\} + \\ & + 0,468 \max\{x_4, 12,781x_6\} - 0,57 x_2. \end{aligned} \quad (11)$$

Для НЛР (5) коэффициент детерминации $R^2 = 0,66699$, а определитель матрицы коэффициентов корреляции объясняющих переменных $\Delta = 0,0932$; для (6) – $R^2 = 0,63238$, $\Delta = 0,0843$; для (7) – $R^2 = 0,66327$, $\Delta = 0,0895$; для (8) – $R^2 = 0,71378$, $\Delta = 0,1337$; для (9) – $R^2 = 0,71051$, $\Delta = 0,1131$; для (10) – $R^2 = 0,6956$, $\Delta = 0,1026$; для (11) – $R^2 = 0,6946$, $\Delta = 0,1384$.

Стоит отметить, что задача построения НЛР с варьированием бинарных операций автором решается впервые. Никогда прежде ему не приходилось встречать регрессионных моделей вида (5) – (11). Понятно, что варьирование бинарных операций увеличивает общее число линейных моделей r , определяемое по формуле (2), в 2^{bin} раз.

Лучшей по величине коэффициента детерминации R^2 моделью из НЛР (5) – (11) оказалась регрессия (8). По величине коэффициента детерминации она превосходит линейную регрессию (3) на 0,113, что существенно для выборки объема $n = 82$. Графики фактических и расчетных по НЛР (8) значений зависимой переменной y изображены на рис. 3.

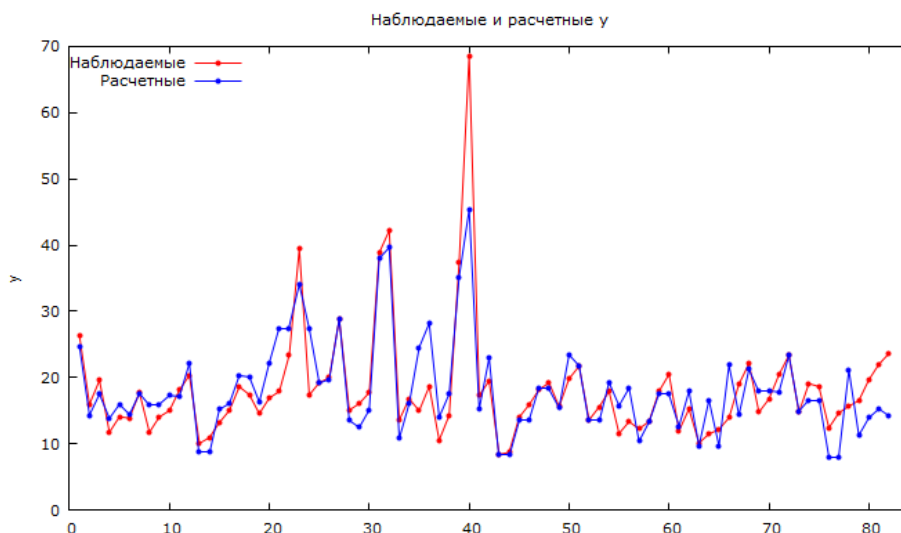


Рис. 3. Графики наблюдаемых и расчетных по НЛР (8) значений переменной y

В НЛР (8) все коэффициенты значимы по t-критерию Стьюдента. Судя по величине Δ , эффект мультиколлинеарности в НЛР (8) гораздо ниже, чем в НЛР (4).

Перейдем теперь к интерпретации НЛР (8). Для этого представим её в кусочно-заданном виде:

$$\tilde{y} = \begin{cases} 49,068 + 1,392x_1 - 1,057x_4 - 69,483x_6 - 1,056x_7, & \text{при } \frac{x_1}{x_5} \geq 3,2212, \frac{x_2}{x_6} \geq 137,59, \frac{x_3}{x_7} \geq 1,5374; \\ 49,068 + 1,392x_1 - 0,687x_3 - 1,057x_4 - 69,483x_6, & \text{при } \frac{x_1}{x_5} \geq 3,2212, \frac{x_2}{x_6} \geq 137,59, \frac{x_3}{x_7} < 1,5374; \\ 49,068 + 1,392x_1 - 0,505x_2 - 1,057x_4 - 1,056x_7, & \text{при } \frac{x_1}{x_5} \geq 3,2212, \frac{x_2}{x_6} < 137,59, \frac{x_3}{x_7} \geq 1,5374; \\ 49,068 + 1,392x_1 - 0,505x_2 - 0,687x_3 - 1,057x_4, & \text{при } \frac{x_1}{x_5} \geq 3,2212, \frac{x_2}{x_6} < 137,59, \frac{x_3}{x_7} < 1,5374; \\ 49,068 - 1,057x_4 + 4,484x_5 - 69,483x_6 - 1,056x_7, & \text{при } \frac{x_1}{x_5} < 3,2212, \frac{x_2}{x_6} \geq 137,59, \frac{x_3}{x_7} \geq 1,5374; \\ 49,068 - 0,687x_3 - 1,057x_4 + 4,484x_5 - 69,483x_6, & \text{при } \frac{x_1}{x_5} < 3,2212, \frac{x_2}{x_6} \geq 137,59, \frac{x_3}{x_7} < 1,5374; \\ 49,068 - 0,505x_2 - 1,057x_4 + 4,484x_5 - 1,056x_7, & \text{при } \frac{x_1}{x_5} < 3,2212, \frac{x_2}{x_6} < 137,59, \frac{x_3}{x_7} \geq 1,5374; \\ 49,068 - 0,505x_2 - 0,687x_3 - 1,057x_4 + 4,484x_5, & \text{при } \frac{x_1}{x_5} < 3,2212, \frac{x_2}{x_6} < 137,59, \frac{x_3}{x_7} < 1,5374. \end{cases} \quad (12)$$

Как видно по (12), в зависимости от выполненных условий меняется состав влияющих на переменную y объясняющих переменных, а идентифицированные в процессе оценивания НЛР (8) параметры $\lambda_1 = 3,2212$, $\lambda_2 = 137,59$ и $\lambda_3 = 1,5374$ играют роль точек переключения для следующих трех автоматически сформированных показателей:

- 1) отношение колесной базы (x_1) к весу автомобиля (x_5);
- 2) отношение длины автомобиля (x_2) к объему двигателя (x_6);
- 3) отношение ширины автомобиля (x_3) к расходу топлива (x_7).

Тогда справедлива следующая интерпретация.

1. Если показатель x_1/x_5 не меньше, чем $\lambda_1 = 3,2212$, то на цену автомобиля влияет колесная база x_1 , а вес x_5 не влияет. При этом с увеличением колесной базы x_1 на 1 дюйм (при неизменных значениях остальных переменных) цена автомобиля y возрастает в среднем на 1392 доллара. А если показатель x_1/x_5 меньше, чем $\lambda_1 = 3,2212$, то на цену автомобиля влияет вес x_5 , а колесная база x_1 не влияет. При этом с увеличением веса x_5 на 1 сотню фунтов (при неизменных значениях остальных переменных) цена автомобиля y возрастает в среднем на 4484 доллара.

2. Если показатель x_2/x_6 не меньше, чем $\lambda_2 = 137,59$, то на цену автомобиля влияет объем двигателя x_6 , а длина x_2 не влияет. При этом с увеличением объема двигателя x_6 на 1 литр (при неизменных значениях остальных переменных) цена автомобиля y падает в среднем на 69483 доллара. А если показатель x_2/x_6 меньше, чем $\lambda_2 = 137,59$, то на цену автомобиля влияет длина x_2 , а объем двигателя x_6 не влияет. При этом с увеличением длины x_2 на 1 дюйм (при неизменных значениях остальных переменных) цена автомобиля y падает в среднем на 505 долларов.

3. Если показатель x_3/x_7 не меньше, чем $\lambda_3 = 1,5374$, то на цену автомобиля влияет расход топлива x_7 , а ширина x_3 не влияет. При этом с увеличением расхода топлива x_7 на 1 милю на галлон (при неизменных значениях остальных переменных) цена автомобиля y падает в среднем на 1056 долларов. А если показатель x_3/x_7 меньше, чем $\lambda_3 = 1,5374$, то на цену автомобиля влияет ширина x_3 , а расход топлива x_7 не влияет. При этом с увеличением ширины x_3 на 1 дюйм (при неизменных значениях остальных переменных) цена автомобиля y падает в среднем на 687 долларов.

4. Переменная x_4 (высота) всегда влияет на цену автомобиля вне зависимости от соотношений других объясняющих переменных. При этом с ростом высоты x_4 на 1 дюйм

(при неизменных значениях остальных переменных) цена автомобиля у падает в среднем на 1057 долларов.

Отметим, что в результате построения НЛР (12) произошло разделение всех содержащихся в выборке 82 автомобилей на 8 кластеров.

Самым большим кластером, содержащим 60 автомобилей, оказался тот, который соответствует третьему уравнению выражения (12). Это означает, что для большинства автомобилей в выборке отношение колесной базы (x_1) к весу автомобиля (x_5) превосходит значение 3,2212, отношение длины автомобиля (x_2) к объему двигателя (x_6) меньше, чем 137,59, отношение ширины автомобиля (x_3) к расходу топлива (x_7) превосходит величину 1,5374, а на цену таких автомобилей у влияют переменные x_1, x_2, x_4, x_7 .

Вторым по объему кластером, содержащим 18 автомобилей, оказался тот, который соответствует седьмому уравнению выражения (12). Для таких автомобилей справедливы условия $\frac{x_1}{x_5} < 3,2212$, $\frac{x_2}{x_6} < 137,59$, $\frac{x_3}{x_7} \geq 1,5374$, а на их цену влияют переменные x_2, x_4, x_5, x_7 .

Третьим и четвертым по объему кластерами, содержащими по 2 автомобиля каждый, оказались те, которые соответствуют второму и четвертому уравнениям выражения (12). Оставшиеся 4 кластера оказались пустыми.

Заключение. В данной работе на примере моделирования цен на автомобили в зависимости от их характеристик с помощью МНК была построена НЛР с семью объясняющими переменными. Впервые для выбора её спецификации использовалось варьирование бинарных операций минимум и максимум. Полученная НЛР характеризуется низкой степенью мультиколлинеарности и существенно превзошла линейную регрессию по величине коэффициента детерминации. Интерпретационные характеристики построенной НЛР оказались богаче и разнообразнее тех же характеристик традиционной линейной регрессии.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Pardoe I. Applied regression modeling. – Wiley, 2020. – 336 p.
2. Клейнер Г.Б. Производственные функции: теория, методы, применение. – М.: Финансы и статистика, 1986. – 239 с.
3. Клейнер Г.Б. Экономика. Моделирование. Математика. Избранные труды. – М.: ЦЭМИ РАН, 2016. – 856 с.
4. Горбунов В.К. Производственные функции: теория и построение. – Ульяновск: УлГУ, 2013. – 84 с.
5. Хацкевич Г.А., Проневич А.Ф., Чайковский М.В. Двухфакторные производственные функции с заданной предельной нормой замещения // Экономическая наука сегодня. – 2019. – № 10. – С. 169-181.
6. Базилевский М.П., Носков С.И. Алгоритм построения линейно-мультипликативной регрессии // Современные технологии. Системный анализ. Моделирование. – 2011. – № 1 (29). – С. 88-92.
7. Базилевский М.П. Программный комплекс построения линейно-мультипликативных регрессий // Прикладная информатика. – 2018. – Т. 13, № 3 (75). – С. 110-123.
8. 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. – С. 17-23.
9. Базилевский М.П. Построение степенно-показательных регрессионных моделей и их интерпретация // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2020. – № 4. – С. 19-28.
10. Базилевский М.П. Построение степенно-показательных и линейно-логарифмических регрессионных моделей // Проблемы управления. – 2021. – № 3. – С. 25-32.

11. Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. – Иркутск: Облформпечать, 1996. – 321 с.
12. Базилевский М.П. Исследование возможности построения кусочно-линейных регрессий с нелинейными границами переключения // *System Analysis and Mathematical Modeling*. – 2021. – Т. 3, № 2. – С. 99-112.
13. Иванова Н.К., Лебедева С.А., Носков С.И. Идентификация параметров некоторых негладких регрессий // *Информационные технологии и проблемы математического моделирования сложных систем*. – 2016. – № 17. – С. 107-110.
14. Носков С.И., Хоняков А.А. Применение функции риска для моделирования экономических систем // *Южно-Сибирский научный вестник*. – 2020. – № 5 (33). – С. 85-92.
15. 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.
16. Базилевский М.П. МНК-оценивание параметров специфицированной на основе функций Леонтьева двухфакторной модели регрессии // *Южно-Сибирский научный вестник*. – 2019. – № 2 (26). – С. 66-70.
17. Базилевский М.П. Оценивание линейно-неэлементарных регрессионных моделей с помощью метода наименьших квадратов // *Моделирование, оптимизация и информационные технологии*. – 2020. – Т. 8. № 4 (31). – С. 26-27.
18. 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.
19. Molnar C. Interpretable machine learning. – Lulu. com, 2020.
20. Miller T. Explanation in artificial intelligence: Insights from the social sciences // *Artificial intelligence*. – 2019. – V. 267. – P. 1-38.
21. Kim B., Khanna R., Koyejo O. O. Examples are not enough, learn to criticize! criticism for interpretability // *Advances in neural information processing systems*. – 2016. – V. 29.
22. Коротеев М. В. Обзор некоторых современных тенденций в технологии машинного обучения // *E-Management*. – 2018. – Т. 1. – №. 1. – С. 26-35.

REFERENCES

1. Pardoe I. *Applied regression modeling*. Wiley, 2020. 336 p.
2. Kleyner G.B. *Proizvodstvennye funktsii: teoriya, metody, primeneniye* [Production functions: theory, methods, application]. Moscow, Finance and Statistics, 1986. 239 p.
3. Kleyner G.B. *Ekonomika. Modelirovanie. Matematika. Izbrannye trudy* [Economy. Modeling. Maths. Selected writings]. Moscow, TsEMI RAN, 2016. 856 p.
4. Gorbunov V.K. *Proizvodstvennye funktsii: teoriya i postroenie* [Production functions: theory and construction]. Ul'yanovsk, UIGU, 2013. 84 p.
5. Khatskevich G.A., Pronevich A.F., Chaykovskiy M.V. *Dvukhfaktornye proizvodstvennye funktsii s zadannoy predel'noy normoy zameshcheniya* [Two-factor production functions with given marginal rate of substitutions]. *Ekonomicheskaya nauka segodnya* [Economics today]. 2019, no. 10, pp. 169-181.
6. Bazilevskiy M.P., Noskov S.I. *Algoritm postroeniya lineynno-mul'tiplikativnoy regressii* [Algorithm for constructing a linear-multiplicative regression]. *Sovremennye tekhnologii. Sistemnyy analiz. Modelirovanie* [Modern technologies. System analysis. Modeling]. 2011, no. 1, vol. 29, pp. 88-92.
7. Bazilevskiy M.P. *Programmnyy kompleks postroeniya lineynno-mul'tiplikativnykh regressiy* [Software package for constructing linear-multiplicative regressions]. *Prikladnaya informatika* [Applied Informatics]. 2018, no. 3, vol. 75, pp. 110-123.
8. Bazilevskiy M.P., Noskov S.I. *Otsenivanie indeksnykh modeley regressii s pomoshch'yu metoda naimen'shikh moduley* [Estimation of index regression models using the method of least

modules]. *Vestnik Rossiyskogo novogo universiteta. Seriya: Slozhnye sistemy: modeli, analiz i upravlenie* [Bulletin of the Russian New University. Series: Complex systems: models, analysis and control]. 2020, no. 1, pp. 17-23.

9. Bazilevskiy M.P. *Postroenie stepenno-pokazatel'nykh regressionnykh modeley i ikh interpretatsiya* [Construction of exponential regression models and their interpretation]. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: Sistemnyy analiz i informatsionnye tekhnologii* [Bulletin of the Voronezh State University. Series: System Analysis and Information Technology]. 2020, no. 4, pp. 19-28.

10. Bazilevskiy M.P. *Postroenie stepenno-pokazatel'nykh i lineynno-logarifmicheskikh regressionnykh modeley* [Construction of exponential and linear-logarithmic regression models]. *Problemy upravleniya* [Control Sciences]. 2021, no. 3, pp. 25-32.

11. Noskov S.I. *Tekhnologiya modelirovaniya ob'ektov s nestabil'nym funkcionirovaniem i neopredelennost'yu v dannykh* [Modeling technology for objects with unstable operation and data uncertainty]. Irkutsk, RIC GP «Oblinformpechat» Publ., 1996. 321 p.

12. Bazilevskiy M.P. *Issledovanie vozmozhnosti postroeniya kusochno-lineynykh regressiy s nelineynymi granitsami pereklyucheniya* [Study of the possibility of constructing piecewise linear regressions with non-linear switching boundaries]. *System Analysis and Mathematical Modeling*. 2021, no. 2, vol. 3, pp. 99-112.

13. Ivanova N.K., Lebedeva S.A., Noskov S.I. *Identifikatsiya parametrov nekotorykh negladkikh regressiy* [Identification of the parameters of some non-smooth regressions]. *Informatsionnye tekhnologii i problemy matematicheskogo modelirovaniya slozhnykh sistem* [Information technologies and problems of mathematical modeling of complex systems]. 2016, no. 17, pp. 107-110.

14. Noskov S.I., Khonyakov A.A. *Primenenie funktsii riska dlya modelirovaniya ekonomicheskikh sistem* [Application of the risk function for modeling economic systems]. *Yuzhno-Sibirskiy nauchnyy vestnik* [South Siberian Scientific Bulletin]. 2020, no. 5, vol. 33, pp. 85-92.

15. Noskov S.I., Khonyakov A.A. *Programmnyy kompleks postroeniya nekotorykh tipov kusochno-lineynykh regressiy* [A software package for constructing some types of piecewise linear regressions]. *Informatsionnye tekhnologii i matematicheskoe modelirovanie v upravlenii slozhnyimi sistemami* [Information technology and mathematical modeling in the management of complex systems]. 2019, no. 3, vol. 4, pp. 47-55.

16. Bazilevskiy M.P. *MNK-otsenivanie parametrov spetsifitsirovannykh na osnove funktsiy Leont'eva dvukhfaktornykh modeley regressii* [OLS estimation of parameters of two-factor regression models specified on the basis of Leontiev functions]. *Yuzhno-Sibirskiy nauchnyy vestnik* [South Siberian Scientific Bulletin]. 2019, no. 2, vol. 26, pp. 66-70.

17. Bazilevskiy M.P. *Otsenivanie lineynno-neelementarnykh regressionnykh modeley s pomoshch'yu metoda naimen'shikh kvadratov* [Estimating Linear Non-Elementary Regression Models Using Least Squares]. *Modelirovanie, optimizatsiya i informatsionnye tekhnologii* [Modeling, optimization and information technology]. 2020, no. 4, vol. 31, pp. 26-27.

18. Bazilevskiy M.P. *Otbor informativnykh operatsiy pri postroenii lineynno-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.

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

20. Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*. 2019, vol. 267, pp. 1-38.

21. Kim B., Khanna R., Koyejo O. O. Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*. 2016, vol. 29.

22. Koroteev M. V. *Obzor nekotorykh sovremennykh tendentsiy v tekhnologii mashinnogo obucheniya* [Overview of some current trends in machine learning technology]. *E-Management*. 2018, no. 1, vol. 1, pp. 26-35.

Информация об авторах

Михаил Павлович Базилевский – к. т. н., доцент, доцент кафедры «Математика», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: mik2178@yandex.ru

Authors

Mikhail Pavlovich Bazilevskiy – Ph. D. in Engineering Science, Associate Professor, the Subdepartment of Mathematics, Irkutsk State Transport University, Irkutsk, e-mail: mik2178@yandex.ru

Для цитирования

Базилевский М.П. Интерпретация неэлементарных линейных регрессионных моделей // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2022. – №1(13). – С. 5-15 – DOI: 10.26731/2658-3704.2022.1(13).5-15 – Режим доступа: <http://ismm-irgups.ru/toma/113-2022>, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 26.04.2022)

For citations

Bazilevskiy M.P. Interpretation of non-elementary linear regression models // *Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal* [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal], 2022. No. 1(13). P. 5-15. DOI: 10.26731/2658-3704.2022.1(13).5-15. [Accessed 26/04/22]