Идентификация неизвестных параметров неэлементарных регрессионных моделей с целочисленными функциями и с бинарными логическими операциями

М. П. Базилевский

***Аннотация*—Статья посвящена проблеме построения регрессионных моделей по выборке, содержащей булеву выходную переменную и непрерывные входные переменные. Исследованы такие известные методы построения моделей с булевыми переменными, как логистическая, логическая и псевдобулева регрессии. Рассмотрена предложенная ранее автором регрессионная модель с целочисленными функциями «пол» и «потолок», идентификация которой с помощью метода наименьших модулей сводится к решению задачи частично целочисленного линейного программирования. Показано, что эта модель может быть использована и для анализа данных по выборке, содержащей булеву выходную переменную. В последнем случае идентификация сводится к задаче частично булевого линейного программирования. Её решение приводит к бинаризации линейной комбинации объясняющих переменных. На основе объединения с помощью бинарных логических операций двух бинаризованных предложенным способом линейных комбинаций объясняющих переменных введено 8 новых спецификаций регрессионных моделей. Для этого использованы операции конъюнкция, дизъюнкция, исключающее «или», эквиваленция, импликация, обратная импликация, штрих Шеффера и стрелка Пирса. Задачи идентификации параметров каждой из предложенных моделей сведены к задачам частично булевого линейного программирования. На примере задачи исследования вероятности невозврата кредита предприятия торговли доказана корректность разработанного математического аппарата. Причём, при использовании только одной целочисленной функции «пол» точность предсказания полученной модели оказалась 80%, что выше, чем у логистической регрессии с точностью 72%. При использовании двух целочисленных функций «пол» сразу 6 моделей с логическими операциями дизъюнкция, исключающее «или», эквиваленция, импликация, обратная импликация и штрих Шеффера показали точность 96%.**

***Ключевые слова*—неэлементарная регрессионная модель, целочисленная функция «пол», функция «потолок», метод наименьших модулей, задача частично булевого линейного программирования, бинарная логическая операция, конъюнкция, дизъюнкция.**

Статья получена 4 марта 2025.

Базилевский Михаил Павлович, Иркутский государственный университет путей сообщения, Иркутск, Российская Федерация (e-mail: mik2178@yandex.ru).

I. Введение

Развитие вычислительных технологий в современном мире влечёт за собой появление новых научных проблем, связанных, в частности, с обработкой статистических данных, и математических методов их решения. Так, довольно быстрыми темпами сегодня эволюционирует область машинного обучения [1,2]. Регрессионный анализ [3,4], предназначенный для построения моделей, связывающих математически выходную переменную с одной или несколькими входными предикторами, считается неотъемлемой компонентой машинного обучения, поэтому его методы также быстро совершенствуются.

Существуют различные подходы к построению регрессионных моделей с булевыми переменными, принимающими либо значение «0», либо «1».

1. Логистическая регрессия [5,6], в которой выходная переменная булева, а входные предикторы – непрерывны, т.е. принимают любые значения из некоторого диапазона. Оценки параметров такой модели определяются методом максимального правдоподобия. В логистической регрессии для преобразования прогнозных значений отклика в вероятности применяются логистические функции.

2. Логическая регрессия, предложенная И. Ручински, Ч. Купербергом и М. Лебланом в [7,8], в которой входные предикторы булевы, а выходная переменная может быть любого типа. Такая регрессия имеет вид

$$g(Y) = a_0 + \sum_{j=1}^{t} a_j L_j ,$$

где t – число логических деревьев модели; Y – выходная переменная; L_j – булева переменная (логическое дерево), полученная в результате преобразования с помощью логических операций конъюнкция, дизъюнкция и инверсия входных булевых переменных X_1, X_2, …, X_l; a_0, a_1, …, a_t – неизвестные параметры. Для непрерывной выходной переменной преобразование $g(Y) = Y$, а для булевой $g(Y) = \log\left(\frac{Y}{1-Y}\right)$. В последнем случае логическая регрессия трансформируется в логистическую регрессию с бинарными входными переменными. Оценки параметров логической регрессии находятся с помощью алгоритма имитации отжига. В 2005 г. Ч.

Куперберг и И. Ручински в [9] предложили симбиоз логической регрессии и марковской цепи Монте-Карло, а в 2008 г. Х. Швендер и К. Икштадт в [10] разработали метод, основанный на использовании логической регрессии на выборках бутстрэпа. В настоящее время логическая регрессия находит широкое применение при решении реальных исследовательских задач. Так, например, в [11] она применена для выявления факторов, влияющих на иммунитет человека, в [12] – для отслеживания микробных источников кишечных инфекций, в [13] – для изучения воздействий экстремальных тепловых явлений на здоровье людей, в [14] – для раннего выявления рака поджелудочной железы, в [15] – для выявления факторов, связанных со смертью людей на месте дорожно-транспортных происшествий, в [16] – для оценки важности взаимодействия между компонентами в сети.

3. Псевдобулева регрессия [17], в которой входные предикторы булевы, а выходная переменная непрерывна. Однако, как отмечено в [17], с помощью алгоритмов бинаризации любую непрерывную переменную можно преобразовать в булеву. Оценки параметров такой модели определяются методом наименьших модулей (МНМ), для чего требуется решить специальным образом сформированную оптимизационную задачу. В [17] подчеркивается сложность поиска логических деревьев в логической регрессии и эвристичность метода оценки её параметров, в отличие от псевдобулевой регрессии.

В настоящей статье предлагаются новые структурные спецификации регрессионных моделей, идентифицируемые с помощью МНМ по выборке, содержащей булеву выходную переменную и непрерывные входные переменные. Предпосылками к их разработке послужили работы автора [18,19], в которых были изобретены модульные линейные регрессии, относящиеся к классу неэлементарных. Затем в [20] была исследована «глубокая» (многослойная) модульная регрессия, а в [21] – «широкая» модульная регрессия. Параллельно в [22,23] были введены неэлементарные регрессии с целочисленными функциями «пол» и «потолок». Статьи [18-23] в совокупности с известными методами логистической, логической и псевдобулевой регрессии привели к идее использовать в неэлементарных моделях таких логических операций, как конъюнкция, дизъюнкция и прочих. Цель настоящей статьи состоит в разработке математического аппарата для идентификации параметров нескольких новых неэлементарных регрессионных моделей с целочисленными функциями «пол» и «потолок» и с бинарными логическими операциями.

II. Регрессионная модель с целочисленной функцией «пол»

Пусть имеется выборка объема n, содержащая значения y_i, x_{i1}, …, x_{il}, $i=\overline{1,n}$, для зависимой (объясняемой) переменной y и для l независимых (объясняющих) переменных x_1, …, x_l. Рассмотрим введенную в [22] регрессионную модель с целочисленной функцией «пол»:

$$y_i = \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor + \varepsilon_i, \qquad i=\overline{1,n}, \tag{1}$$

где α_0, α_1, …, α_l – неизвестные параметры; ε_i, $i=\overline{1,n}$ – ошибки регрессии; скобками $\lfloor\ \rfloor$ обозначена целочисленная функция «пол», которая, как отмечено в [24], возвращает округленное до ближайшего целого в меньшую сторону число. Например, $\lfloor 4{,}8 \rfloor = 4$. Аналогичным образом определяется целочисленная функция «потолок», округляющая число до ближайшего целого, но в большую сторону. Соответствующую ей регрессионную модель также можно найти в [22].

В [22] показано, что МНМ-оценки неизвестных параметров регрессии (1) могут быть найдены в результате решения следующей задачи частично целочисленного линейного программирования (ЧЦЛП):

$$\sum_{i=1}^{n} (u_i + v_i) \to \min, \tag{2}$$

$$y_i = \theta_i + u_i - v_i, \quad i=\overline{1,n}, \tag{3}$$

$$\theta_i \in Z, \qquad i=\overline{1,n}, \tag{4}$$

$$\theta_i \le \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \le \theta_i + 1 - \Delta, \quad i=\overline{1,n}, \tag{5}$$

$$u_i \ge 0,\ v_i \ge 0, \quad i=\overline{1,n}, \tag{6}$$

где Δ – близкое к нулю положительное число; θ_i, $i=\overline{1,n}$ – целочисленные переменные; u_i, v_i, $i=\overline{1,n}$ – неотрицательные переменные, удовлетворяющие условиям:

$$u_i = \begin{cases} y_i - \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor, \text{ если } y_i - \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor > 0, \\ 0, \text{ в противном случае,} \end{cases}$$

$$v_i = \begin{cases} -y_i + \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor, \text{ если } y_i - \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor < 0, \\ 0, \text{ в противном случае.} \end{cases}$$

Пусть построенная в результате решения задачи ЧЦЛП (2) – (6) регрессионная модель имеет вид

$$\tilde{y} = \left\lfloor \tilde{\alpha}_0 + \sum_{j=1}^{l} \tilde{\alpha}_j x_{ij} \right\rfloor, \tag{7}$$

где $\tilde{\alpha}_0$, …, $\tilde{\alpha}_l$ – МНМ-оценки неизвестных параметров; $\tilde{y}$ – прогнозные значения зависимой переменной y. Очевидно, что для любых значений объясняющих переменных x_1, …, x_l прогнозное значение $\tilde{y}$, судя по (7), всегда целое. В этой связи регрессионную модель (1) целесообразно применять тогда, когда зависимая переменная может принимать по смыслу только целочисленные значения. Например, когда y это число вагонов, количество дней, людей и пр.

Предположим, что зависимая переменная y по смыслу булева, т.е. может принимать либо значение «0», либо «1». Тогда для МНМ-оценки параметров регрессионной модели (1) с булевой выходной переменной y достаточно заменить в задаче (2) – (6)

ограничения (4) целочисленности переменных на ограничения булевости переменных вида

$$\theta_i \in \{0,1\}, \qquad i = \overline{1,n}. \tag{8}$$

Решение задачи частично булевого линейного программирования (ЧБЛП) (2), (3), (5), (6), (8) также приводит к уравнению (7). Заметим, что для любого наблюдения с выборки прогнозное значение $\tilde{y}$ зависимой переменной всегда равно либо «0», либо «1», что следует из ограничений (5) задачи ЧБЛП. Но если значения объясняющих переменных x_1, …, x_l взять не с выборки, то $\tilde{y}$ может оказаться любым целым, например, «-1», «2» и т.д. Никаких преград в таком случае для применения оцененной с помощью МНМ регрессии (1) нет. Достаточно просто считать, что $\tilde{y} = \begin{cases} 0, \text{ при } \tilde{y} \le 0, \\ 1, \text{ при } \tilde{y} \ge 1. \end{cases}$ Во избежание таких ситуаций, как известно из математической статистики, выборка данных должна обладать свойством репрезентативности.

III. Регрессионные модели с бинарными логическими операциями

Рассмотренное в предыдущем разделе МНМ-оценивание регрессионной модели (1) с булевой зависимой переменной фактически представляет собой способ бинаризации линейной комбинации объясняющих переменных. Возникает идея в «склеивании» двух бинаризованных предложенным способом линейных комбинаций переменных. Для «склеивания» подойдут известные бинарные логические операции конъюнкция, дизъюнкция и пр. Подобный принцип стыковки линейных комбинаций переменных использован в статье [21] при конструировании «широких» модульных регрессионных моделей.

В табл. 1 представлена таблица истинности для всех восьми использованных в данной работе бинарных логических операций.

Таблица 1 – Таблица истинности бинарных логических операций

Логические переменные		Бинарные логические операции							
		Конъюнкция	Дизъюнкция	Исключающее «или»	Эквиваленция	Импликация	Обратная импликация	Штрих Шеффера	Стрелка Пирса
A	B	A∧B	A∨B	A⊕B	A↔B	A→B	A←B	A↑B	A↓B
0	0	0	0	0	1	1	1	1	1
0	1	0	1	1	0	1	0	1	0
1	0	0	1	1	0	0	1	1	0
1	1	1	1	0	1	1	1	0	0

Далее последовательно будем вводить новые спецификации регрессионных моделей, сопровождая их соответствующим математическим аппаратом, позволяющим идентифицировать с помощью МНМ неизвестные оценки параметров.

1. Регрессионная модель с логической операцией конъюнкция (∧):

$$y_i = \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor \wedge \left\lfloor \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} \right\rfloor + \varepsilon_i, \; i = \overline{1,n}, \tag{9}$$

где β_0, …, β_l – неизвестные параметры.

МНМ-оценивание регрессии (9) сводится к задаче ЧБЛП с целевой функцией (2), ограничениями (6) и

$$y_i = z_i + u_i - v_i, \qquad i = \overline{1,n}, \tag{10}$$

$$\theta_i^{(1)}, \theta_i^{(2)} \in \{0,1\}, \qquad i = \overline{1,n}, \tag{11}$$

$$\theta_i^{(1)} \le \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \le \theta_i^{(1)} + 1 - \Delta, \quad i = \overline{1,n}, \tag{12}$$

$$\theta_i^{(2)} \le \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} \le \theta_i^{(2)} + 1 - \Delta, \quad i = \overline{1,n}, \tag{13}$$

$$z_i \le \theta_i^{(1)}, \qquad i = \overline{1,n}, \tag{14}$$

$$z_i \le \theta_i^{(2)}, \qquad i = \overline{1,n}, \tag{15}$$

$$z_i \ge \theta_i^{(1)} + \theta_i^{(2)} - 1, \qquad i = \overline{1,n}, \tag{16}$$

$$0 \le z_i \le 1, \qquad i = \overline{1,n}, \tag{17}$$

где $\theta_i^{(1)} = \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor$, $\theta_i^{(2)} = \left\lfloor \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} \right\rfloor$, $i = \overline{1,n}$ – булевы переменные; $z_i = \theta_i^{(1)} \wedge \theta_i^{(2)}$, $i = \overline{1,n}$ – значения переменной z, найденные с помощью логической операции конъюнкция.

В этой задаче ограничения (10) – (13) следуют из (3), (5), (8), а ограничения (14) – (17) предназначены для воздействия конъюнкции на булевы переменные. Так, если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 0$, то $z_i = 0$; если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 1$, то $z_i = 0$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 0$, то $z_i = 0$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 1$, то $z_i = 1$. Соответствие такого преобразования конъюнкции подтверждает табл. 1. Заметим, что на переменную z не нужно ставить ограничение целочисленности, поскольку для любых $\theta_i^{(1)}$ и $\theta_i^{(2)}$ она всегда принимает либо значение «0», либо «1».

2. Регрессионная модель с логической операцией дизъюнкция (∨):

$$y_i = \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor \vee \left\lfloor \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} \right\rfloor + \varepsilon_i,\ i = \overline{1,n}. \qquad (18)$$

МНМ-оценивание регрессии (18) сводится к задаче ЧБЛП с целевой функцией (2), ограничениями (6), (10) – (13), (17) и

$$z_i \ge \theta_i^{(1)}, \qquad i = \overline{1,n}, \qquad (19)$$

$$z_i \ge \theta_i^{(2)}, \qquad i = \overline{1,n}, \qquad (20)$$

$$z_i \le \theta_i^{(1)} + \theta_i^{(2)}, \qquad i = \overline{1,n}. \qquad (21)$$

Из (17), (19) – (21) следует, что если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 0$, то $z_i = 0$; если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 1$, то $z_i = 1$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 0$, то $z_i = 1$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 1$, то $z_i = 1$. По табл. 1 видно, что такое преобразование логических переменных соответствует дизъюнкции.

3. Регрессионная модель с логической операцией исключающее «или» ($\oplus$):

$$y_i = \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor \oplus \left\lfloor \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} \right\rfloor + \varepsilon_i,\ i = \overline{1,n}. \qquad (22)$$

Известно, что $a \oplus b = \left(\bar{a} \wedge b\right) \vee \left(a \wedge \bar{b}\right)$, т.е. сначала определяются две конъюнкции, а потом берётся дизъюнкция. Поэтому МНМ-оценивание регрессии (22) сводится к задаче ЧБЛП с целевой функцией (2), ограничениями (6), (10) – (13), (17) и

$$z_i^{(1)} \le 1 - \theta_i^{(1)}, \qquad i = \overline{1,n}, \qquad (23)$$

$$z_i^{(1)} \le \theta_i^{(2)}, \qquad i = \overline{1,n}, \qquad (24)$$

$$z_i^{(1)} \ge -\theta_i^{(1)} + \theta_i^{(2)}, \qquad i = \overline{1,n}, \qquad (25)$$

$$0 \le z_i^{(1)} \le 1, \qquad i = \overline{1,n}, \qquad (26)$$

$$z_i^{(2)} \le \theta_i^{(1)}, \qquad i = \overline{1,n}, \qquad (27)$$

$$z_i^{(2)} \le 1 - \theta_i^{(2)}, \qquad i = \overline{1,n}, \qquad (28)$$

$$z_i^{(2)} \ge \theta_i^{(1)} - \theta_i^{(2)}, \qquad i = \overline{1,n}, \qquad (29)$$

$$0 \le z_i^{(2)} \le 1, \qquad i = \overline{1,n}, \qquad (30)$$

$$z_i = z_i^{(1)} + z_i^{(2)}, \qquad i = \overline{1,n}, \qquad (31)$$

где $z^{(1)}$ – переменная, равная конъюнкции инверсии булевой переменной $\theta^{(1)}$ и переменной $\theta^{(2)}$; $z^{(2)}$ – переменная, равная конъюнкции булевой переменной $\theta^{(1)}$ и инверсии переменной $\theta^{(2)}$; z – переменная, равная дизъюнкции $z^{(1)}$ и $z^{(2)}$.

Из (17), (23) – (31) следует, что если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 0$, то $z_i^{(1)} = 0$, $z_i^{(1)} = 0$, поэтому $z_i = 0$; если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 1$, то $z_i^{(1)} = 1$, $z_i^{(1)} = 0$, поэтому $z_i = 1$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 0$, то $z_i^{(1)} = 0$, $z_i^{(1)} = 1$, поэтому $z_i = 1$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 1$, то $z_i^{(1)} = 0$, $z_i^{(1)} = 0$, поэтому $z_i = 0$. По табл. 1 видно, что такое преобразование логических переменных соответствует исключающему «или».

4. Регрессионная модель с логической операцией эквиваленция ($\leftrightarrow$):

$$y_i = \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor \leftrightarrow \left\lfloor \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} \right\rfloor + \varepsilon_i,\ i = \overline{1,n}. \qquad (32)$$

Поскольку $a \leftrightarrow b = \overline{a \oplus b}$, то МНМ-оценивание регрессии (32) сводится к задаче ЧБЛП с целевой функцией (2), ограничениями (6), (10) – (13), (17), (23) – (30) и

$$z_i = 1 - z_i^{(1)} - z_i^{(2)}, \qquad i = \overline{1,n}. \qquad (33)$$

Из (17), (23) – (30), (33) следует, что если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 0$, то $z_i^{(1)} = 0$, $z_i^{(1)} = 0$, поэтому $z_i = 1$; если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 1$, то $z_i^{(1)} = 1$, $z_i^{(1)} = 0$, поэтому $z_i = 0$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 0$, то $z_i^{(1)} = 0$, $z_i^{(1)} = 1$, поэтому $z_i = 0$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 1$, то $z_i^{(1)} = 0$, $z_i^{(1)} = 0$, поэтому $z_i = 1$. По табл. 1 видно, что такое преобразование логических переменных соответствует эквиваленции.

5. Регрессионная модель с логической операцией импликация ($\rightarrow$):

$$y_i = \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor \rightarrow \left\lfloor \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} \right\rfloor + \varepsilon_i,\ i = \overline{1,n}. \qquad (34)$$

МНМ-оценивание регрессии (34) сводится к задаче ЧБЛП с целевой функцией (2), ограничениями (6), (10) – (13), (17), (20) и

$$z_i \ge 1 - \theta_i^{(1)}, \qquad i = \overline{1,n}, \qquad (35)$$

$$z_i \le 1 - \left(\theta_i^{(1)} - \theta_i^{(2)}\right), \qquad i = \overline{1,n}. \qquad (36)$$

Из (17), (20), (35), (36) следует, что если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 0$, то $z_i = 1$; если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 1$, то $z_i = 1$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 0$, то $z_i = 0$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 1$, то $z_i = 1$. По табл. 1 видно, что такое преобразование логических переменных соответствует импликации.

6. Регрессионная модель с логической операцией обратная импликация ($\leftarrow$):

$$y_i = \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor \leftarrow \left\lfloor \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} \right\rfloor + \varepsilon_i,\ i = \overline{1,n}. \qquad (37)$$

МНМ-оценивание регрессии (37) сводится к задаче ЧБЛП с целевой функцией (2), ограничениями (6), (10) – (13), (17), (19) и

$$z_i \ge 1 - \theta_i^{(2)}, \qquad i = \overline{1,n}, \qquad (38)$$

$$z_i \le 1 - \left(\theta_i^{(2)} - \theta_i^{(1)}\right), \qquad i = \overline{1,n}. \qquad (39)$$

Из (17), (19), (38), (39) следует, что если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 0$, то $z_i = 1$; если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 1$, то $z_i = 0$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 0$, то $z_i = 1$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 1$, то $z_i = 1$. По табл. 1 видно, что такое преобразование логических переменных соответствует обратной импликации.

7. Регрессионная модель с логической операцией штрих Шеффера ($\uparrow$):

$$y_i = \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor \uparrow \left\lfloor \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} \right\rfloor + \varepsilon_i,\ i = \overline{1,n}. \qquad (40)$$

МНМ-оценивание регрессии (40) сводится к задаче ЧБЛП с целевой функцией (2), ограничениями (6), (10) – (13), (17), (35), (38) и

$$z_i \le 2 - \left(\theta_i^{(1)} + \theta_i^{(2)}\right), \qquad i = \overline{1,n}. \qquad (41)$$

Из (17), (35), (38), (41) следует, что если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 0$, то $z_i = 1$; если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 1$, то $z_i = 1$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 0$, то $z_i = 1$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 1$, то

$z_i = 0$. По табл. 1 видно, что такое преобразование логических переменных соответствует штриху Шеффера.

8. Регрессионная модель с логической операцией стрелка Пирса (↓):

$$y_i = \left\lfloor \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} \right\rfloor \downarrow \left\lfloor \beta_0 + \sum_{j=1}^{l} \beta_j x_{ij} \right\rfloor + \varepsilon_i,\ i = \overline{1,n}. \quad (42)$$

МНМ-оценивание регрессии (42) сводится к задаче ЧБЛП с целевой функцией (2), ограничениями (6), (10) – (13), (17) и

$$z_i \le 1 - \theta_i^{(1)}, \qquad i = \overline{1,n}, \quad (43)$$

$$z_i \le 1 - \theta_i^{(2)}, \qquad i = \overline{1,n}, \quad (44)$$

$$z_i \ge 1 - \left(\theta_i^{(1)} + \theta_i^{(2)}\right), \qquad i = \overline{1,n}. \quad (45)$$

Из (17), (43) – (45) следует, что если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 0$, то $z_i = 1$; если $\theta_i^{(1)} = 0$, $\theta_i^{(2)} = 1$, то $z_i = 0$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 0$, то $z_i = 0$; если $\theta_i^{(1)} = 1$, $\theta_i^{(2)} = 1$, то $z_i = 0$. По табл. 1 видно, что такое преобразование логических переменных соответствует стрелке Пирса.

IV. Пример

Корректность предложенного математического аппарата было решено проверить по выборке объема *n*=50 из книги [25]. Банк исследует вероятность невозврата кредита предприятиями торговли. Булева переменная *y* принимает значение «1», если заемщик кредит возвращает, «0» – в противном случае. В качестве объясняющих переменных выбраны x_1 – коэффициент рентабельности, x_2 – коэффициент оборачиваемости оборотных активов.

Сначала по этим данным с помощью пакета Gretl была оценена логистическая регрессия, уравнение которой имеет вид

$$\tilde{y} = \left(1 + \exp\left(1,327 - 10,8098x_1 - 0,2181x_2\right)\right)^{-1}. \quad (46)$$

Для (46) количество корректно предсказанных случаев составило 36 из 50, т.е. 72%.

Затем с помощью решателя LPSolve при $\Delta = 0,0001$ было найдено уравнение регрессионной модели (1) с целочисленной функцией «пол»:

$$\tilde{y} = \left\lfloor 0,7905 + 2,2003x_1 + 0,04074x_2 \right\rfloor. \quad (47)$$

Для (47) количество корректно предсказанных случаев составило 40 из 50 (80%). Как видно, предсказательные качества модели (47) выше, чем у логистической регрессии (46) на 8%.

После чего с помощью LPSolve при $\Delta = 0,0001$ оценивались все предложенные в данной статье неэлементарные регрессионные модели с бинарными логическими операциями конъюнкция, дизъюнкция, исключающее «или», эквиваленция, импликация, обратная импликация, штрих Шеффера, стрелка Пирса. Были получены следующие уравнения:

$$\tilde{y} = \left\lfloor 1 \right\rfloor \wedge \left\lfloor 0,7905 + 2,2003x_1 + 0,04074x_2 \right\rfloor, \quad (48)$$

$$\tilde{y} = \left\lfloor 1,5656 - 6,764x_1 + 0,0147x_2 \right\rfloor \vee \vee \left\lfloor 7,4078x_1 - 1,73 \cdot 10^{-5} x_2 \right\rfloor, \quad (49)$$

$$\tilde{y} = \left\lfloor 2,1088 - 8,3953x_1 + 0,00754x_2 \right\rfloor \oplus \oplus \left\lfloor 0,4342 + 6,764x_1 - 0,0147x_2 \right\rfloor, \quad (50)$$

$$\tilde{y} = \left\lfloor 2,1088 - 8,3953x_1 + 0,00754x_2 \right\rfloor \leftrightarrow \leftrightarrow \left\lfloor 1,5656 - 6,764x_1 + 0,0147x_2 \right\rfloor, \quad (51)$$

$$\tilde{y} = \left\lfloor 2,1088 - 8,3953x_1 + 0,00754x_2 \right\rfloor \to \to \left\lfloor 1,5656 - 6,764x_1 + 0,0147x_2 \right\rfloor, \quad (52)$$

$$\tilde{y} = \left\lfloor 1,0017 - 0,02x_1 + 2,9585 \cdot 10^{-5} x_2 \right\rfloor \leftarrow \leftarrow \left\lfloor 2,1088 - 8,3953x_1 + 0,0075x_2 \right\rfloor, \quad (53)$$

$$\tilde{y} = \left\lfloor 2,1088 - 8,3953x_1 + 0,00754x_2 \right\rfloor \uparrow \uparrow \left\lfloor 0,4342 + 6,764x_1 - 0,0147x_2 \right\rfloor, \quad (54)$$

$$\tilde{y} = \left\lfloor 1,508 - 2,41x_1 - 0,0566x_2 \right\rfloor \downarrow \downarrow \left\lfloor 1,508 - 2,41x_1 - 0,0566x_2 \right\rfloor. \quad (55)$$

Для (48), (55) количество корректно предсказанных случаев составило 40 из 50 (80%), для (49) – (54) – 48 из 50 (96%). Таким образом, использование в моделях логических операций конъюнкция и стрелка Пирса не привело к повышению предсказательных качеств регрессии (47). Применение же логических операций дизъюнкция, исключающее «или», эквиваленция, импликация, обратная импликация и штрих Шеффера привело к увеличению предсказательных качеств на 16%.

V. Заключение

В работе за счёт использования бинарных логических операций конъюнкция, дизъюнкция, исключающее «или», эквиваленция, импликация, обратная импликация, штрих Шеффера и стрелка Пирса предложено 8 структурных спецификаций неэлементарных регрессионных моделей с целочисленными функциями «пол». Для идентификации неизвестных параметров каждой из предложенных регрессий с помощью МНМ сформулирована своя задача ЧБЛП. Проведённые эксперименты доказали корректность разработанного математического аппарата. Заметим, что аналогичным образом можно оценивать параметры моделей с целочисленными функциями «потолок».

Библиография

[1] Telikani A., Tahmassebi A., Banzhaf W., Gandomi A.H. Evolutionary machine learning: A survey // ACM Computing Surveys (CSUR). 2021. Vol. 54. No. 8. P. 1-35.

[2] Cohen S. The evolution of machine learning: Past, present, and future // Artificial Intelligence in Pathology. 2025. P. 3-14.

[3] Montgomery D.C., Peck E.A., Vining G.G. Introduction to linear regression analysis. John Wiley & Sons, 2021.

[4] Chatterjee S., Hadi A.S. Regression analysis by example. John Wiley & Sons, 2015.

[5] Bailly A., Blanc C., Francis É., Guillotin T., Jamal F., Wakim B., Roy P. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models // Computer Methods and Programs in Biomedicine. 2022. Vol. 213. P. 106504.

[6] Zaidi A., Al Luhayb A.S.M. Two statistical approaches to justify the use of the logistic function in binary logistic regression // Mathematical Problems in Engineering. 2023. No. 1. P. 5525675.

[7] Ruczinski I., Kooperberg C., LeBlanc M. Logic regression // Journal of Computational and graphical Statistics. 2003. Vol. 12. No. 3. P. 475-511.

[8] Ruczinski I., Kooperberg C., LeBlanc M.L. Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications // Journal of Multivariate Analysis. 2004. Vol. 90. No. 1. P. 178-195.
[9] Kooperberg C., Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression // Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2005. Vol. 28. No. 2. P. 157-170.
[10] Schwender H., Ickstadt K. Identification of SNP interactions using logic regression // Biostatistics. 2008. Vol. 9. No. 1. P. 187-198.
[11] Ayoub A. Discovering interactions that affect immune recognization using logic regression. MS thesis, 2023.
[12] Yu D., Andersson-Li M., Maes S., Andersson-Li L., Neumann N.F., Odlare M., Jonsson A. Development of a logic regression-based approach for the discovery of host-and niche-informative biomarkers in Escherichia coli and their application for microbial source tracking // Applied and Environmental Microbiology. 2024. Vol. 90. No. 7. P. e00227-24.
[13] Jiang S., Warren J.L., Scovronick N., Moss S.E., Darrow L.A., Strickland M.J., Newman A.J., Chen Y., Ebelt S.T., Chang H.H. Using logic regression to characterize extreme heat exposures and their health associations: a time-series study of emergency department visits in Atlanta // BMC Medical Research Methodology. 2021. Vol. 21. No. 87.
[14] Huang Y., Dasgupta S. Biomarker Panel Development Using Logic Regression in the Presence of Missing Data // The New England Journal of Statistics in Data Science. 2024. Vol. 2. No. 1. P. 3.
[15] Jamali-Dolatabad M., Sadeghi-Bazargani H., Salemi S., Sarbakhsh P. Identifying interactions among factors related to death occurred at the scene of traffic accidents: Application of «logic regression» method // Heliyon, 2024. Vol. 10. No. 11.
[16] Rocco C.M., Hernandez-Perdomo E., Mun J. Application of logic regression to assess the importance of interactions between components in a network // Reliability Engineering & System Safety. 2021. Vol. 205. P. 107235.
[17] Bonates T.O. Optimization in logical analysis of data. Rutgers The State University of New Jersey, School of Graduate Studies, 2007.
[18] Базилевский М.П., Ойдопова А.Б. Оценивание модульных линейных регрессионных моделей с помощью метода наименьших модулей // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. 2023. № 45. С. 130-146.
[19] Базилевский М.П. Программное обеспечение для оценивания модульных линейных регрессий // Информационные и математические технологии в науке и управлении. 2023. № 3 (31). С. 136-146.
[20] Базилевский М.П. Оценивание неизвестных параметров многослойной модульной регрессии методом наименьших модулей // Моделирование, оптимизация и информационные технологии. 2024. Т. 12. № 2 (45).
[21] Базилевский М.П. Оценивание регрессионных моделей с регрессорами в виде модулей линейных комбинаций объясняющих переменных // System Analysis and Mathematical Modeling. 2024. Т. 6. № 3. С. 269-281.
[22] Базилевский М.П. Оценивание с помощью метода наименьших модулей регрессионных моделей с целочисленными функциями пол и потолок // International Journal of Open Information Technologies. 2024. Т. 12. № 10. С. 56-61.
[23] Базилевский М.П. Оценивание методом наименьших квадратов параметров неэлементарных линейных регрессий с равномерно квантованными объясняющими переменными // Информационные и математические технологии в науке и управлении. 2024. № 4 (36). С. 46-56.
[24] Грэхем Р., Кнут Д., Паташник О. Конкретная математика. Основание информатики: Пер. с англ. М. : Мир, 1998. 703 с.
[25] Исмагилов И.И., Кадочникова Е.И. Специальные модели эконометрики в среде Gretl. Казань: Казан. ун-т, 2018. 91 с.

Базилевский Михаил Павлович, к.т.н., доцент кафедры математики Иркутского государственного университета путей сообщения, Иркутск, Россия; ORCID 0000-0002-3253-5697 (e-mail: mik2178@yandex.ru)

Identification of Unknown Parameters of Non-Elementary Regression Models with Integer Functions and Binary Logical Operations

M. P. Bazilevskiy

***Abstract*—The article is devoted to the problem of constructing regression models based on a sample containing a Boolean output variable and continuous input variables. Well-known methods for constructing such models, such as logistic, logical, and pseudo-Boolean regression, are investigated. A regression model with integer functions floor and ceiling, proposed by the author earlier, is considered. The identification of this model using the least absolute deviations method is reduced to solving the problem of mixed integer linear programming. It has been shown that this model can also be used for analyzing data based on samples containing Boolean output variables. In this case, identification reduces to a mixed 0-1 integer linear programming, whose solution leads to binarization of the linear combination of explanatory variables. Based on the combination of two linear combinations of explanatory variables, binarized by the proposed method using binary logical operations, eight new specifications of regression models were introduced. For this purpose, operations such as conjunction, disjunction, exclusive OR, equivalence, implication, inverse implication, Sheffer stroke, and Pierce arrow were used. The parameters of each of these models were identified by solving mixed 0-1 linear programming problems. The correctness of the mathematical apparatus developed was proven using an example problem: studying the probability of non-repayment of a loan by a trading company. Moreover, using only one integer function floor, the accuracy of prediction of the model was 80%, higher than that of logistic regression, which had an accuracy of 72%. When using two integer functions floor, at once, six models with logical operations disjunction, exclusive OR, equivalence, implication, inverse implication, and Sheffer stroke, showed an accuracy of 96%.**

***Keywords*—non-elementary regression model, integer function floor, function ceiling, least absolute deviations, mixed 0-1 integer linear programming, binary logical operation, conjunction, disjunction.**

REFERENCES

[1] Telikani A., Tahmassebi A., Banzhaf W., Gandomi A.H. Evolutionary machine learning: A survey // ACM Computing Surveys (CSUR). 2021. Vol. 54. No. 8. P. 1-35.

[2] Cohen S. The evolution of machine learning: Past, present, and future // Artificial Intelligence in Pathology. 2025. P. 3-14.

[3] Montgomery D.C., Peck E.A., Vining G.G. Introduction to linear regression analysis. John Wiley & Sons, 2021.

[4] Chatterjee S., Hadi A.S. Regression analysis by example. John Wiley & Sons, 2015.

[5] Bailly A., Blanc C., Francis É., Guillotin T., Jamal F., Wakim B., Roy P. Effects of dataset size and interactions on the prediction performance of logistic regression and deep learning models // Computer Methods and Programs in Biomedicine. 2022. Vol. 213. P. 106504.

[6] Zaidi A., Al Luhayb A.S.M. Two statistical approaches to justify the use of the logistic function in binary logistic regression // Mathematical Problems in Engineering. 2023. No. 1. P. 5525675.

[7] Ruczinski I., Kooperberg C., LeBlanc M. Logic regression // Journal of Computational and graphical Statistics. 2003. Vol. 12. No. 3. P. 475-511.

[8] Ruczinski I., Kooperberg C., LeBlanc M.L. Exploring interactions in high-dimensional genomic data: an overview of logic regression, with applications // Journal of Multivariate Analysis. 2004. Vol. 90. No. 1. P. 178-195.

[9] Kooperberg C., Ruczinski I. Identifying interacting SNPs using Monte Carlo logic regression // Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2005. Vol. 28. No. 2. P. 157-170.

[10] Schwender H., Ickstadt K. Identification of SNP interactions using logic regression // Biostatistics. 2008. Vol. 9. No. 1. P. 187-198.

[11] Ayoub A. Discovering interactions that affect immune recognization using logic regression. MS thesis, 2023.

[12] Yu D., Andersson-Li M., Maes S., Andersson-Li L., Neumann N.F., Odlare M., Jonsson A. Development of a logic regression-based approach for the discovery of host-and niche-informative biomarkers in Escherichia coli and their application for microbial source tracking // Applied and Environmental Microbiology. 2024. Vol. 90. No. 7. P. e00227-24.

[13] Jiang S., Warren J.L., Scovronick N., Moss S.E., Darrow L.A., Strickland M.J., Newman A.J., Chen Y., Ebelt S.T., Chang H.H. Using logic regression to characterize extreme heat exposures and their health associations: a time-series study of emergency department visits in Atlanta // BMC Medical Research Methodology. 2021. Vol. 21. No. 87.

[14] Huang Y., Dasgupta S. Biomarker Panel Development Using Logic Regression in the Presence of Missing Data // The New England Journal of Statistics in Data Science. 2024. Vol. 2. No. 1. P. 3.

[15] Jamali-Dolatabad M., Sadeghi-Bazargani H., Salemi S., Sarbakhsh P. Identifying interactions among factors related to death occurred at the scene of traffic accidents: Application of «logic regression» method // Heliyon, 2024. Vol. 10. No. 11.

[16] Rocco C.M., Hernandez-Perdomo E., Mun J. Application of logic regression to assess the importance of interactions between components in a network // Reliability Engineering & System Safety. 2021. Vol. 205. P. 107235.

[17] Bonates T.O. Optimization in logical analysis of data. Rutgers The State University of New Jersey, School of Graduate Studies, 2007.

[18] Bazilevskiy M.P., Oydopova A.B. Otsenivanie modul'nykh lineynykh regressionnykh modeley s pomoshch'yu metoda naimen'shikh moduley // Vestnik Permskogo natsional'nogo issledovatel'skogo politekhnicheskogo universiteta. Elektrotekhnika, informatsionnye tekhnologii, sistemy upravleniya. 2023. No. 45. P. 130-146.

[19] Bazilevskiy M.P. Programmnoe obespechenie dlya otsenivaniya modul'nykh lineynykh regressiy // Informatsionnye i matematicheskie tekhnologii v nauke i upravlenii. 2023. No. 3 (31). P. 136-146.

[20] Bazilevskiy M.P. Otsenivanie neizvestnykh parametrov mnogosloynoy modul'noy regressii metodom naimen'shikh moduley // Modelirovanie, optimizatsiya i informatsionnye tekhnologii. 2024. Vol. 12. No. 2 (45).

[21] Bazilevskiy M.P. Otsenivanie regressionnykh modeley s regressorami v vide moduley lineynykh kombinatsiy ob"yasnyayushchikh peremennykh // System Analysis and Mathematical Modeling. 2024. Vol. 6. No. 3. P. 269-281.

[22] Bazilevskiy M.P. Otsenivanie s pomoshch'yu metoda naimen'shikh moduley regressionnykh modeley s tselochislennymi funktsiyami pol i potolok // International Journal of Open Information Technologies. 2024. Vol. 12. No. 10. P. 56-61.

[23] Bazilevskiy M.P. Otsenivanie metodom naimen'shikh kvadratov parametrov neelementarnykh lineynykh regressiy s ravnomerno kvantovannymi ob"yasnyayushchimi peremennymi // Informatsionnye i matematicheskie tekhnologii v nauke i upravlenii. 2024. No. 4 (36). P. 46-56.
[24] Grekhem R., Knut D., Patashnik O. Konkretnaya matematika. Osnovanie informatiki: Per. s angl. Moscow : Mir, 1998. 703 p.
[25] Ismagilov I.I., Kadochnikova E.I. Spetsial'nye modeli ekonometriki v srede Gretl. Kazan': Kazan. un-t, 2018. 91 p.

Bazilevskiy Mikhail Pavlovich, Ph.D., Associate Professor of the Department of Mathematics, Irkutsk State Transport University, Irkutsk, Russia; ORCID 0000-0002-3253-5697 (e-mail: mik2178@yandex.ru)