

УДК 519.862.6

DOI 10.17513/snt.40385

ОПТИМАЛЬНАЯ МАРКИРОВКА ЗНАЧЕНИЙ КАТЕГОРИАЛЬНОЙ ЗАВИСИМОЙ ПЕРЕМЕННОЙ ПРИ ПОСТРОЕНИИ РЕГРЕССИОННЫХ МОДЕЛЕЙ С ЦЕЛОЧИСЛЕННОЙ ФУНКЦИЕЙ ПОЛ

Базилевский М.П.

*ФГБОУ ВО «Иркутский государственный университет путей сообщения»,
Иркутск, e-mail: mik2178@yandex.ru*

Статья посвящена проблеме оценки с помощью метода наименьших модулей неизвестных параметров регрессионных моделей с целочисленной функцией пол. Исследуется ситуация, когда зависимая переменная носит категориальный характер. Для включения в регрессионную модель категориальной переменной предварительно проводится ее маркировка (кодирование), состоящая в присвоении каждой категории уникального целого значения. Цель данной работы состоит в формализации в терминах аппарата частично целочисленного линейного программирования задачи идентификации не только неизвестных параметров регрессионной модели с целочисленной функцией пол, но и оптимальных маркеров категорий зависимой переменной. На примере решения задачи классификации семи видов животных доказана корректность разработанного математического аппарата. При этом сначала продемонстрировано, что при случайной маркировке категориальной переменной качество регрессии существенно меняется. Затем предложенным способом найдена оптимальная маркировка категорий зависимой переменной. Полученная модель, по которой было неверно классифицировано только 12 животных из 101, превзошла по качеству все построенные автором регрессии. Предложенный способ оценки параметров регрессионных моделей с автоматической идентификацией маркеров зависимой категориальной переменной может успешно применяться при решении задач классификации.

Ключевые слова: регрессионный анализ, категориальная переменная, метод наименьших модулей, оптимизация, задача частично целочисленного линейного программирования, целочисленная функция пол

OPTIMAL MARKING OF CATEGORICAL DEPENDENT VARIABLE VALUES WHEN CONSTRUCTING REGRESSION MODELS USING INTEGER FLOOR FUNCTION

Bazilevskiy M.P.

Irkutsk State Transport University, Irkutsk, e-mail: mik2178@yandex.ru

The article is devoted to the problem of estimating using the least absolute deviations method of unknown parameters in regression models with integer floor function. The situation under investigation is when the dependent variable is categorical, and to include a categorical variable in a regression model, it must be preliminarily coded, which consists of assigning a unique integer value to each category. The purpose of this study is to formalize, in terms of mixed integer linear programming, the problem of finding not only the unknown parameters of a regression model with integer floor function, but also the optimal markers for the categories of the dependent variables. Using the example of solving a problem of classifying seven types of animals, we prove the correctness of our mathematical approach. At the same time, we demonstrate that with random coding of a categorical variables, the quality of the regression changes significantly. Then, using the proposed method, optimal marking of categories of the dependent variable was found. The resulting model, according to which only 12 out of 101 animals were incorrectly classified, surpassed all regressions constructed by the author in quality. The proposed method for estimating parameters of regression models with automatic identification of markers for a dependent categorical variable can successfully be applied in solving classification problems.

Keywords: regression analysis, categorical variable, least absolute deviations method, optimization, mixed integer linear programming problem, integer floor function

Введение

При обработке статистических данных на практике довольно часто возникает ситуация, когда исследуемая переменная носит категориальный или качественный характер [1; 2]. Например, к категориальной переменной [3] относится «пол человека», «группа крови», «тип заболевания», «страна», «политическая партия» и т.д. Для включения таких категориальных переменных в регрессионную модель их предварительно маркируют (кодируют), то есть

присваивают каждой категории уникальное целое значение. Если порядок категорий важен, то переменную называют порядковой, а если нет – номинальной. Существует множество методов кодирования категориальных переменных (например, [4–6]). Так, в работе Д.В. Новиковой сравнивается эффективность следующих методов: порядковое, горячее, бинарное, восьмеричное, частотное и целевое кодирование, а также кодирование Стейна, Гельмерта и метод обратной разницы [6].

Аппарат математического программирования уже много десятилетий используется в рамках регрессионного анализа. Основой для этого послужила работа Х.М. Вагнера, в которой задача оценки неизвестных параметров линейной регрессии с помощью метода наименьших модулей (МНМ) была сведена к задаче линейного программирования [7]. К настоящему времени в регрессионном анализе появилось несколько новых разновидностей регрессионных моделей, оценка параметров которых с помощью МНМ предполагает решение задач частично булевого линейного программирования (ЧБЛП). Сегодня эффективность решателей таких задач, как отмечено в работе Т. Коха, Т. Бертхолда, Я. Педерсена и Ч. Ванарета, заметно повысилась [8]. Так, в работе С.И. Носкова задача оценки параметров производственной функции Леонтьева с помощью МНМ сведена к задаче ЧБЛП [9, с. 120–121]. В статье [10] введены модульные регрессии, которые послужили фундаментом для разработки их многослойного аналога [11] и «широкой» модульной регрессии [12]. В работе [13] автором описана задача оценки параметров регрессионных моделей с целочисленными функциями пол и потолок с применением аппарата частично целочисленного линейного программирования (ЧЦЛП). Такие модели целесообразно строить, когда зависимая переменная принимает целые значения. Следовательно, такие регрессии отлично подходят для ситуаций, когда зависимая переменная носит категориальный характер. Проблема состоит в том, каким образом следует маркировать зависимую переменную, чтобы качество регрессии с выбранной целочисленной функцией было наилучшим.

Цель исследования состоит в формализации задачи ЧЦЛП для идентификации с помощью МНМ параметров регрессионной модели с целочисленной функцией пол с одновременной маркировкой значений зависимой переменной.

Материалы и методы исследования

Для оценки параметров регрессионных моделей применяется МНМ, реализуемый в виде задачи линейного программирования. Интеграция функции пол в эту задачу производится с помощью целочисленных переменных.

Результаты исследования и их обсуждение

Пусть исследуется влияние l объясняющих (независимых) переменных $x_1, x_2, \dots, x_l$ на объясняемую (зависимую) переменную y .

Предположим, что объясняющие переменные имеют значения x_{ij} , $i = \overline{1, n}$, $j = \overline{1, l}$, где n – объем выборки.

Допустим, что зависимая переменная y имеет категориальный характер.

Будем считать, что число категорий равно p . Тогда значения переменной y можно промаркировать, например, следующим образом:

$$y_i = \begin{cases} 1, & \text{если } i \in I_1, \\ 2, & \text{если } i \in I_2, \\ \dots & \\ p, & \text{если } i \in I_p, \end{cases} \quad (1)$$

где I_k – множество номеров наблюдений, относящихся к k -й категории. Очевидно, что справедливы следующие соотношения между этими множествами:

$$I_k \cap I_q = \emptyset, \quad k, q = \overline{1, p}, \quad k \neq q,$$

$$I_1 \cup I_2 \cup \dots \cup I_p = \{1, 2, 3, \dots, n\}.$$

Выражение (1) означает, что маркером для первой категории переменной y выступает целое число 1, маркером для второй категории – целое число 2, и т.д. Тогда для исследования зависимости y от $x_1, x_2, \dots, x_l$ целесообразно использовать предложенную в [13] регрессионную модель с целочисленной функцией пол:

$$y_i = \left\lfloor \alpha_0 + \sum_{j=1}^l \alpha_j x_{ij} \right\rfloor + \varepsilon_i, \quad i = \overline{1, n}, \quad (2)$$

где $\alpha_0, \alpha_1, \dots, \alpha_l$ – неизвестные параметры; ε_i , $i = \overline{1, n}$ – ошибки регрессии; скобками $\lfloor \cdot \rfloor$ обозначена рассмотренная в работе Р. Грэхема, Д. Кнута и О. Паташника целочисленная функция пол, которая округляет заключенное в них значение до ближайшего целого числа в меньшую сторону [14, с. 88]. Например, $\lfloor 5,91 \rfloor = 5$.

Модель (2) относится к классу нелинейных регрессий. Для оценки ее неизвестных параметров с помощью МНМ требуется решить задачу:

$$\sum_{i=1}^n \left| y_i - \left\lfloor \alpha_0 + \sum_{j=1}^l \alpha_j x_{ij} \right\rfloor \right| \rightarrow \min. \quad (3)$$

В работе автора [13] показано, что решение задачи (3) сводится к решению задачи ЧЦЛП. Учитывая, что зависимая переменная по правилу (1) принимает ограниченное множество значений, такая задача будет иметь следующий вид:

$$\sum_{i=1}^n (u_i + v_i) \rightarrow \min, \quad (4)$$

$$y_i = \theta_i + u_i - v_i, \quad i = \overline{1, n}, \quad (5)$$

$$\theta_i \leq \alpha_0 + \sum_{j=1}^l \alpha_j x_{ij} \leq \theta_i + 1 - \Delta, \quad i = \overline{1, n}, \quad (6)$$

$$1 \leq \theta_i \leq p, \quad i = \overline{1, n} \quad (7)$$

$$\theta_i \in Z, \quad i = \overline{1, n}, \quad (8)$$

$$u_i \geq 0, v_i \geq 0, \quad i = \overline{1, n}, \quad (9)$$

где Z – множество целых чисел; если ошибка регрессии в i -м наблюдении положительна, то ее модуль численно равен значению переменной u_i , а если отрицательна, то значению переменной v_i , причем, $u_i \cdot v_i = 0$ для любого i ; θ_i – предсказанный по модели маркер (категория) зависимой переменной в i -м наблюдении; Δ – выбранное исследователем близкое к нулю положительное число.

Задача ЧЦЛП (4)–(9) подходит для моделирования порядковой категориальной зависимой переменной y , маркировка которой производится по правилу (1). Например, как отмечено в работе С. Баладрам, такой порядковой переменной может служить «температура», принимающая значения «очень низкая», «низкая», «высокая» и «очень высокая» [3]. Отметим, что в таком случае маркировать значения зависимой переменной y можно не только целыми числами в диапазоне от 1 до p , но и в любом диапазоне от $(1+k)$ до $(p+k)$, где k – целое число. Действительно, пусть задача (4)–(9), в которой переменная y закодирована по правилу (1), имеет оптимальное решение $\theta_i = \theta_i^*$, $u_i = u_i^*$, $v_i = v_i^*$, $i = \overline{1, n}$, $\alpha_j = \alpha_j^*$, $j = \overline{0, l}$. Промаркируем теперь зависимую переменную y по правилу

$$y_i = \begin{cases} 1+k, & \text{если } i \in I_1, \\ 2+k, & \text{если } i \in I_2, \\ \dots & \\ p+k, & \text{если } i \in I_p. \end{cases} \quad (10)$$

И рассмотрим задачу (4)–(9), в которой линейные ограничения (7) заменены на двойные неравенства $1+k \leq \theta_i \leq p+k$, $i = \overline{1, n}$. Очевидно, что эта задача будет иметь оптимальное решение $\theta_i = \theta_i^*$, $u_i = u_i^*$, $v_i = v_i^*$, $i = \overline{1, n}$, $\alpha_j = \alpha_j^*$, $j = \overline{1, l}$, $\alpha_0 = \alpha_0^* + k$. Таким образом, при маркировке зависимой переменной y по правилу (10) для любого k оптимальные значения всех параметров, кроме α_0 , сохраняются.

Рассмотрим теперь ситуацию, когда категориальная зависимая переменная y носит номинальный характер. Такой переменной может служить, например, «прогноз погоды», категории которой «солнечно», «пасмурно» и «дождливо» [3]. Порядок маркировки этих категорий не важен. Иными словами, категории «солнечно» можно присвоить число 1, «пасмурно» – 2, «дождливо» – 3. Либо категории «солнечно» можно поставить в соответствие число 2, «пасмурно» – 3, «дождливо» – 1. Естественно, что выбранное правило маркировки значений зависимой переменной y будет влиять на решение задачи (4)–(9). Это доказывается на примере в следующем разделе статьи.

Возникает задача выбора оптимальной маркировки значений зависимой переменной y наряду с идентификацией неизвестных параметров регрессионной модели (2) с помощью МНМ. Введем следующее правило маркировки значений категориальной переменной y :

$$y_i = \begin{cases} z_1, & \text{если } i \in I_1, \\ z_2, & \text{если } i \in I_2, \\ \dots & \\ z_p, & \text{если } i \in I_p, \end{cases} \quad (11)$$

где $z_1, z_2, \dots, z_p$ – маркеры каждой категории переменной y , принимающие целые и отличные друг от друга значения из диапазона от 1 до p .

Введем булевы переменные σ_{qk} , $q = \overline{1, p}$, $k = \overline{1, p}$ по следующему правилу:

$$\sigma_{qk} = \begin{cases} 1, & \text{если для } q\text{-й категории маркер } z_q = k, \\ 0, & \text{в противном случае.} \end{cases}$$

В задачу математического программирования ограничения на булевы переменные интегрируются в виде

$$\sigma_{qk} \in \{0, 1\}, \quad q, k = \overline{1, p}. \quad (12)$$

Тогда выбор маркера для первой категории реализуется через следующее линейное ограничение:

$$1\sigma_{11} + 2\sigma_{12} + 3\sigma_{13} + \dots + p\sigma_{1p} = z_1.$$

Если в этом ограничении, например, $\sigma_{13} = 1$, а все оставшиеся булевы переменные равны 0, то маркер z_1 принимает значение 3. В общем виде все маркеры z_p , $q = \overline{1, p}$ связаны с булевыми переменными равенствами

$$\sum_{k=1}^p k \cdot \sigma_{qk} = z_q, \quad q = \overline{1, p}. \quad (13)$$

Для обеспечения выбора только одного маркера для каждой категории требуется ввести ограничения

$$\sum_{k=1}^p \sigma_{qk} = 1, \quad q = \overline{1, p}, \quad (14)$$

а для обеспечения различия маркеров из разных категорий – ограничения

$$\sum_{q=1}^p \sigma_{qk} = 1, \quad k = \overline{1, p}. \quad (15)$$

Вместо фиксированных значений маркеров y_i в (5) с учетом правила (11) используем переменные z_q :

$$z_q = \theta_i + u_i - v_i, \quad q = \overline{1, p}, \quad i \in I_q. \quad (16)$$

Тогда решение задачи ЧЦЛП с целевой функцией (4) и с линейными ограничениями (6)–(9), (12)–(16) дает оптимальные по МНМ значения параметров $\theta_i, u_i, v_i, i = \overline{1, n}, \alpha_j, j = \overline{0, l}, \sigma_{qk}, q, k = \overline{1, p}, z_q, q = \overline{1, p}$.

Для проведения вычислительных экспериментов была использована известная выборка данных, содержащая информацию о биологических видах животных [15]. Объем выборки составляет 101 наблюдение. Зависимая переменная y является категориальной и выражает категорию (класс, тип) животного. Всего представлено 7 категорий. К первой относится 41 вид (медведь, слон, жираф, заяц, волк и пр.). Ко второй – 20 видов (курица, утка, страус и пр.). К третьей – 5 видов (морская змея, черепаха и пр.). К четвертой – 13 видов (окунь, карп, щука, пиранья и пр.). К пятой – 3 вида (лягушка, тритон и жаба). К шестой – 8 видов (комар, пчела, муха и пр.). К седьмой – 10 видов (краб, рак, осьминог, скорпион и пр.). По умолчанию каждой категории был присвоен соответствующий ей номер, то есть $z_q = q, q = \overline{1, 7}$. Категория животного предсказывается 15 бинарными переменными: x_1 – волосы; x_2 – перья; x_3 – яйца; x_4 – молоко; x_5 – воздушное; x_6 – водное; x_7 – хищник; x_8 – зубастое; x_9 – позвоночник; x_{10} – дышащее; x_{11} – ядовитое; x_{12} – плавники; x_{14} – хвост; x_{15} – домашнее; x_{16} – кошачий. И одна объясняющая переменная x_{13} – число лап, которая принимает целые значения 0, 2, 4, 5, 6 и 8.

Предварительно общий список из 16 объясняющих переменных был сокращен до 3 наиболее сильно влияющих на y переменных – x_4, x_9 и x_{10} . Для их выбора применен метод последовательного исключения, алгоритм которого встроен в пакет Gretl, с использованием двухстороннего r -значения 0,01. Построенная по этим пере-

менным с помощью МНМ линейная регрессия имеет уравнение

$$\tilde{y} = 8 - x_4 - 4x_9 - 2x_{10}. \quad (17)$$

Для модели (17) сумма модулей ошибок (SAE) составила 27.

Построенная модель (2) с целочисленной функцией пол имеет вид

$$\tilde{y} = \lfloor 7,999 - x_4 - 3,001x_9 - 1,999x_{10} \rfloor. \quad (18)$$

Для регрессии (18) SAE = 20, то есть ее качество выше, чем у модели множественной линейной регрессии (17).

Сначала маркировка категориальной переменной y проводилась вручную. В первый раз категориям случайно были присвоены маркеры $z_1 = 3, z_2 = 6, z_3 = 1, z_4 = 7, z_5 = 2, z_6 = 5, z_7 = 4$. Решение задачи ЧЦЛП (4)–(9) позволило идентифицировать модель

$$\tilde{y} = \lfloor 4,999 - 3x_4 + 2,999x_9 - 0,999x_{10} \rfloor, \quad (19)$$

для которой SAE = 50.

Во второй раз категориям случайно были присвоены значения $z_1 = 1, z_2 = 5, z_3 = 4, z_4 = 7, z_5 = 6, z_6 = 2, z_7 = 3$. В результате была построена модель

$$\tilde{y} = \lfloor 3,999 - 4x_4 + 3,999x_9 - 1,999x_{10} \rfloor, \quad (20)$$

для которой SAE = 14.

Как видно, при разной маркировке категориальной переменной качество регрессионной модели (2) меняется. Так, регрессия (19) оказалась хуже модели (18) с кодами категорий по умолчанию, а регрессия (20) лучше, чем (18).

После чего решалась задача ЧЦЛП (4), (6)–(9), (12)–(16). В ограничениях (6) число $\Delta = 0,0001$. В результате была построена модель

$$\tilde{y} = \lfloor 2,999 + 3x_4 + 3,999x_9 - 1,999x_{10} \rfloor, \quad (21)$$

для которой SAE = 12. При этом оптимальные маркеры категорий переменной y оказались следующими: $z_1 = 7, z_2 = 4, z_3 = 5, z_4 = 6, z_5 = 3, z_6 = 1, z_7 = 2$. Как видно, модель (21) превосходит по качеству все построенные регрессии (17)–(20). Значение SAE = 12 означает, что только в 12 наблюдениях из 101 тип животного был классифицирован неверно. Это наблюдения под номерами 26, 27, 53, 63, 73, 77, 81, 82, 90, 91, 92 и 100.

Заключение

В результате проведенных исследований задача одновременной идентификации с помощью МНМ и маркеров категориальной переменной y , и параметров регрессионной модели с целочисленной функцией пол све-

дена к задаче ЧЦП. На примере решения задачи классификации видов животных доказана корректность разработанного математического аппарата. Построенная предложенным способом регрессионная модель превзошла по качеству регрессии, оцененные по известным алгоритмам. В дальнейшем планируется использовать полученные результаты для модернизации метода оценки списочных регрессионных моделей.

Список литературы

1. Лучинин А.С., Лянгузов А.В. Модель логистической регрессии для прогнозирования летальности в отделении интенсивной терапии: проблемы и решения // *Качественная клиническая практика*. 2022. № 2. С. 13–20. DOI: 10.37489/2588-0519-2022-2-13-20.
2. Коржавых В.В. Сравнение эффективности алгоритмов машинного обучения на примере прогнозирования среднемесячного потребления электроэнергии интервальных приборов учета потребителей // *Научный результат. Информационные технологии*. 2024. Т. 9. № 1. С. 58–72. DOI: 10.18413/2518-1092-2024-9-1-0-7.
3. Baladram S. Encoding Categorical Data, Explained: A Visual Guide with Code Example for Beginners [Электронный ресурс]. URL: <https://aiappnews.org/encoding-categorical-data-explained-a-visual-guide-with-code-example-for-beginners-by-samy-baladram-sep-2024/> (дата обращения: 05.03.2025).
4. Alkharusi H. Categorical variables in regression analysis: A comparison of dummy and effect coding // *International Journal of Education*. 2012. Vol. 4, Is. 2. P. 202. DOI: 10.5296/ije.v4i2.1962.
5. Venkataramana M., Subbarayudu M., Rajani M., Sreenivasulu K.N. Regression analysis with categorical variables // *International Journal of Statistics and Systems*. 2016. Vol. 11, Is. 2. P. 135–143. URL: https://www.researchgate.net/profile/Rajani-Meejuru/publication/348421623_Regression_Analysis_with_Categorical_Variables/links/5ffe5a2f45851553a03d60a8/Regression-Analysis-with-Categorical-Variables.pdf (дата обращения: 05.03.2025).
6. Новикова Д.В. Сравнительный анализ эффективности методов кодирования категориальных переменных в задаче прогнозирования безопасности терапии рассеянного склероза // *Заметки по информатике и математике*. 2022. С. 61–69. URL: <https://elibrary.ru/item.asp?id=49215267&pff=1> (дата обращения: 15.03.2025).
7. Wagner H.M. Linear programming techniques for regression analysis // *Journal of the American Statistical Association*. 1959. Vol. 54, Is. 285. P. 206–212. DOI: 10.1080/01621459.1959.10501506.
8. Koch T., Berthold T., Pedersen J., Vanaret C. Progress in mathematical programming solvers from 2001 to 2020 // *EURO Journal on Computational Optimization*. 2022. Vol. 10. P. 100031. DOI: 10.1016/j.ejco.2022.100031.
9. Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. Иркутск: РИЦ ГП «Облформпечать», 1996. 321 с.
10. 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. DOI: 10.15593/2224-9397/2023.1.06.
11. Базилевский М.П. Оценивание неизвестных параметров многослойной модульной регрессии методом наименьших модулей // *Моделирование, оптимизация и информационные технологии*. 2024. Т. 12. № 2 (45). URL: <https://moitvvt.ru/ru/journal/pdf?id=1581> (дата обращения: 05.04.2025). DOI: 10.26102/2310-6018/2024.45.2.039.
12. 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. DOI: 10.17150/2713-1734.2024.6(3).269-281.
13. 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. URL: <http://injoit.org/index.php/j1/article/view/1926/1782> (дата обращения: 15.03.2025).
14. Грэхем Р., Кнут Д., Паташник О. Конкретная математика. Основание информатики: Пер. с англ. М.: Мир, 1998. 703 с.
15. UCI Machine Learning Repository [Электронный ресурс]. URL: <https://archive.ics.uci.edu/dataset/111/zoo> (дата обращения: 05.03.2025).