

Научная статья
УДК 519.862.6
EDN MLCDNR
DOI 10.17150/2713-1734.2024.6(3).269-281



М.П. Базилевский

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

Оценивание регрессионных моделей с регрессорами в виде модулей линейных комбинаций объясняющих переменных

Аннотация. Статья посвящена исследованию процесса построения новой структурной спецификации регрессионных моделей, содержащей знаки модулей. Ранее подобные более простые формы связи между переменными уже вводились автором и были названы модульными регрессиями. Предложенная в данной статье регрессия с регрессорами в виде модулей линейных комбинаций объясняющих переменных обобщает ранее рассмотренные формы. Для оценивания предложенных регрессий используется метод наименьших модулей. Сформулировано две задачи частично-булевого линейного программирования для оценивания модульных регрессий. В первой из них знаки перед модулями требуется задавать вручную, а во второй знаки перед модулями определяются автоматически. С использованием реальных данных об уровне безработицы в Российской Федерации проведены вычислительные эксперименты. В качестве решателя оптимизационных задач выбран пакет LPSolve. Установлено, что при одном возможном способе объявления неограниченных переменных в этом пакете полученное решение может оказываться неоптимальным. Показано, что верным будет известный способ объявления неограниченных параметров в виде разности двух неотрицательных переменных. В ходе экспериментов были построены модульные регрессии уровня безработицы с пятью переменными, содержащие один и два модуля, а также с двумя переменными, содержащие один, два и три модуля. Полученные регрессии по сумме модулей остатков оказались лучше линейных моделей. В одном случае была идентифицирована регрессия со всеми нулевыми остатками. Эксперименты показали, что с ростом числа модулей возрастает время оценивания модульных регрессий.

Ключевые слова. Регрессионный анализ, модульная регрессия, метод наименьших модулей, задача частично-булевого линейного программирования, уровень безработицы.

Информация о статье. Дата поступления: 7 августа, 2024; дата принятия к публикации: 1 октября 2024 г.; дата онлайн-размещения: 17 октября 2024 г.

Original article

М.Р. Bazilevskiy

*Irkutsk State Transport University,
Irkutsk, Russian Federation*

Estimation of Regression Models with Regressors in the Explanatory Variables Linear Combinations Modules Form

Abstract. This article is devoted to the study of the constructing process a new structural specification of regression models containing module signs. Earlier, similar

simpler forms of relationships between variables were introduced by the author and were called modular regressions. The regression with regressors in the explanatory variables linear combinations modules form proposed in this article generalizes the previously considered forms. The least absolute deviations method is used to estimate the proposed regressions. Two problems of mixed 0-1 integer linear programming are formulated for estimating modular regressions. In the first of them, the signs in front of the modules must be specified manually, and in the second, the signs in front of the modules are determined automatically. Computational experiments were conducted using real data on the unemployment rate in the Russian Federation. The LPSolve package was chosen as a solver for optimization problems. It was found that with one possible way of declaring unlimited variables in this package, the resulting solution may be suboptimal. It is shown that the known method of declaring unlimited parameters as the difference of two non-negative variables will be correct. During the experiments, modular regressions of the unemployment rate with five variables, containing one and two modules, and with two variables, containing one, two and three modules, were constructed. The resulting regressions by the residuals modules sum were better than linear models. In one case, a regression with all zero residuals was identified. The experiments showed that with an increase in the number of modules, the time for estimating modular regressions increases.

Keywords. Regression analysis, modular regression, least absolute deviations method, mixed 0-1 integer linear programming problem, unemployment rate.

Article info. Received 7 August, 2024; Accepted 1 October, 2024; Available online 17 October, 2024.

Введение

Задача выявления скрытых математических зависимостей между исследуемыми переменными по имеющимся статистическим данным, безусловно, актуальна на сегодняшний день. Если такая зависимость найдена и подтверждено ее высокое качество, то она может быть использована для анализа и прогнозирования поведения исследуемой системы или процесса. Все это позволяет исследователю принимать более обоснованные управленческие решения. Очень часто для построения математической зависимости применяется нейросетевой подход [1; 2], при котором по статистическим данным осуществляется обучение так называемой нейронной сети. С помощью такого подхода решается довольно много прикладных задач (см., например, [3–5]). Однако обученная нейронная сеть обычно представляет собой весьма сложную математическую конструкцию, интерпретация которой затруднительна. Не менее эффективным подходом к анализу данных можно считать регрессионный анализ [6; 7].

Высокие темпы развития вычислительной техники за последние годы привели к появлению новых нелинейных структурных спецификаций регрессионных моделей [8; 9], время оценивания которых ничуть не меньше, чем время обучения нейронных сетей. Скорее всего, грань между нейронными сетями и регрессионным анализом со временем вовсе будет стерта. Постепенно будет формироваться единый подход к построению моделей машинного обучения. Пока же регрессионному анализу предстоит пройти большой путь. Данная статья по-

священа развитию его аппарата, а именно, разработке метода оценивания новой структурной спецификации, содержащей модули.

Регрессии, содержащие в своем уравнении модули, впервые были исследованы в [9] и были названы модульными. Для выборки из n наблюдений для объясняемой переменной y и для объясняющих переменных x_1, x_2, \dots, x_l модульная регрессия с неизвестными параметрами $\alpha_0, \beta_{0j}, \beta_{1j}, j=1, l$, имеет следующий вид:

$$y_i = \alpha_0 + \sum_{j=1}^l (-1)^{\Delta_j} |\beta_{0j} + \beta_{1j} x_{ij}| + \varepsilon_i, i = \overline{1, n}, \quad (1)$$

где $\Delta_j, j = \overline{1, l}$ — заданная исследователем бинарная переменная, с помощью которой контролируются знаки перед модулями; $\varepsilon_i, i = \overline{1, n}$ — ошибки аппроксимации. В той же работе задача оценивания модульной регрессии (1) с помощью метода наименьших модулей (МНМ) была сведена к задаче частично-булевого линейного программирования (ЧБЛП). Как отмечено в [10], в среднем с 2001 по 2020 гг. для задач частично-целочисленного линейного программирования компьютерное оборудование стало примерно в 20 раз быстрее, а алгоритмы улучшились примерно в 50 раз, что дает общее ускорение в 1000 раз. В [11] для оценивания регрессий (1) предложено специальное программное обеспечение. А в [12] сформулирована задача МНМ-оценивания модульных регрессий при неизвестных бинарных переменных $\Delta_j, j = \overline{1, l}$.

В [13] была введена модульная регрессия с одним регрессором в виде модуля линейной комбинации объясняющих переменных:

$$y_i = \alpha_0 + (-1)^{\Delta} \cdot \left| \beta_0 + \sum_{j=1}^l \beta_j x_{ij} \right| + \varepsilon_i, i = \overline{1, n}, \quad (2)$$

где $\beta_1, \beta_2, \dots, \beta_l$ — неизвестные параметры; Δ — бинарная переменная для контроля знака перед модулем. Задача МНМ-оценивания регрессии (2) также формулируется в терминах задачи ЧБЛП. В [14] модель (2) была расширена линейной частью, что позволило ввести в рассмотрение многослойную модульную регрессию, выстроенную по принципу «модуль в модуле». Ее также можно называть «глубокой» модульной регрессией.

Цель данной работы состоит в разработке и тестировании на реальных данных обобщенной для спецификаций (1) и (2) модульной регрессии.

1. Постановка задачи

Введем модульную регрессию с линейной частью и с p регрессорами в виде модулей линейных комбинаций объясняющих переменных:

$$y_i = \alpha_0 + \sum_{j=1}^l \alpha_j x_{ij} + \sum_{k=1}^p (-1)^{\Delta_k} \cdot \left| \beta_{0k} + \sum_{j=1}^l \beta_{jk} x_{ij} \right| + \varepsilon_i, \quad i = \overline{1, n}, \quad (3)$$

где $\alpha_0, \alpha_j, j = \overline{1, l}, \beta_{0k}, \beta_{1k}, \dots, \beta_{lk}, k = \overline{1, p}$ — неизвестные параметры; $\Delta_k, k = \overline{1, p}$ — бинарные переменные для контроля знаков перед модулями.

Как видно, если $p = l, \alpha_j = 0, j = \overline{1, l}, \beta_{jk} = 0, j \neq k$, то модульная регрессия (3) вырождается в (1). Если $p = 1, \alpha_j = 0, j = \overline{1, l}$, то она трансформируется в (2). А если $\beta_{0k}, \beta_{1k}, \dots, \beta_{lk} = 0, k = \overline{1, p}$, то в линейную регрессию.

В отличие от «глубокой» модульной регрессии, введенной в [14], регрессию (3) можно назвать «широкой».

Для оценивания модульной регрессии (3) будем использовать МНМ, который, как известно, предполагает решение следующей задачи оптимизации:

$$J(\alpha, \beta) = \sum_{i=1}^n |e_i| \rightarrow \min, \quad (4)$$

где $e_i, i = \overline{1, n}$ — остатки регрессии, т.е. разность между наблюдаемыми и прогнозными значениями переменной y ; e_i — сумма модулей остатков.

В работе [15] В.Д. Фишер показал, как аппроксимировать линейную функцию с помощью МНМ с использованием линейного программирования. Для этого нужно представить остатки $e_i, i = \overline{1, n}$, в виде разностей двух неотрицательных переменных g_i и $h_i, i = \overline{1, n}$, а целевую функцию $\sum_{i=1}^n |e_i| \rightarrow \min$ заменить на $\sum_{i=1}^n (g_i + h_i) \rightarrow \min$.

С использованием данного приема и результатов работы [9] можно сформулировать равносильную задаче (4) задачу ЧБЛП:

$$\sum_{i=1}^n (g_i + h_i) \rightarrow \min, \quad (5)$$

$$y_i = \alpha_0 + \sum_{j=1}^l \alpha_j x_{ij} + \sum_{k=1}^p (-1)^{\Delta_k} \cdot (u_{ik} + v_{ik}) + g_i - h_i, \quad i = \overline{1, n}, \quad (6)$$

$$\beta_{0k} + \sum_{j=1}^l \beta_{jk} x_{ij} = u_{ik} - v_{ik}, \quad i = \overline{1, n}, \quad k = \overline{1, p}, \quad (7)$$

$$u_{ik} \leq M \cdot \delta_{ik}, \quad i = \overline{1, n}, \quad k = \overline{1, p}, \quad (8)$$

$$v_{ik} \leq M \cdot (1 - \delta_{ik}), \quad i = \overline{1, n}, \quad k = \overline{1, p}, \quad (9)$$

$$u_{ik} \geq 0, \quad v_{ik} \geq 0, \quad i = \overline{1, n}, \quad k = \overline{1, p}, \quad (10)$$

$$\delta_{ik} \in \{0, 1\}, \quad i = \overline{1, n}, \quad k = \overline{1, p}, \quad (11)$$

$$g_i \geq 0, \quad h_i \geq 0, \quad i = \overline{1, n}, \quad (12)$$

где $u_{ik}, v_{ik}, i = \overline{1, n}, k = \overline{1, p}$ — неотрицательные переменные, разности между которыми равны значениям линейных комбинаций в i -м наблюдении при k -м регрессоре с модулем, а суммы — значениям модулей; $M > 0$ — верхняя граница модулей линейных комбинаций объясняющих переменных. Бинарные переменные $\delta_{ik}, i = \overline{1, n}, k = \overline{1, p}$, удовлетворяют правилу:

$$\delta_{ik} = \begin{cases} 1, & \text{если значение линейной комбинации в } i\text{-м наблюдении} \\ & \text{при } k\text{-м регрессоре с модулем неотрицательно,} \\ 0, & \text{в противном случае.} \end{cases}$$

В задаче (5)–(12) требуется самостоятельно контролировать знаки перед регрессорами с модулями, т.е. априори задавать значения бинарных переменных $\Delta_k, k = \overline{1, p}$. Например, если имеется одна объясняющая переменная x_1 , а $p = 3$, то комбинаций знаков перед модулями будет 4:

$$\begin{aligned} y_i &= \alpha_0 + \alpha_1 x_{i1} + |\beta_{01} + \beta_{11} x_{i1}| + |\beta_{02} + \beta_{12} x_{i1}| + |\beta_{03} + \beta_{13} x_{i1}| + \varepsilon_i, & i = \overline{1, n}, \\ y_i &= \alpha_0 + \alpha_1 x_{i1} + |\beta_{01} + \beta_{11} x_{i1}| + |\beta_{02} + \beta_{12} x_{i1}| - |\beta_{03} + \beta_{13} x_{i1}| + \varepsilon_i, & i = \overline{1, n}, \\ y_i &= \alpha_0 + \alpha_1 x_{i1} + |\beta_{01} + \beta_{11} x_{i1}| - |\beta_{02} + \beta_{12} x_{i1}| - |\beta_{03} + \beta_{13} x_{i1}| + \varepsilon_i, & i = \overline{1, n}, \\ y_i &= \alpha_0 + \alpha_1 x_{i1} - |\beta_{01} + \beta_{11} x_{i1}| - |\beta_{02} + \beta_{12} x_{i1}| - |\beta_{03} + \beta_{13} x_{i1}| + \varepsilon_i, & i = \overline{1, n}. \end{aligned}$$

В общем случае, чтобы выбрать оптимальный набор знаков, обеспечивающий минимум функционала (4), требуется решить $(p + 1)$ задачу ЧБЛП (5)–(12).

Сформулируем единую задачу идентификации оптимального набора знаков в модульной регрессии (3) так, как это сделано в работе [12]. Для этого введем матрицу Φ размера $(p + 1) \times p$, элементы φ_{qk} которой заданы по правилу:

$$\varphi_{qk} = \begin{cases} 0, & \text{если } q + k \leq p + 1, \\ 1, & \text{если } q + k > p + 1. \end{cases}$$

Тогда в задаче ЧБЛП (5)–(12) вместо ограничений (6) нужно использовать следующие ограничения:

$$\alpha_0 + \sum_{j=1}^l \alpha_j x_{ij} + \sum_{k=1}^p (-1)^{q_k} \cdot (u_{ik} + v_{ik}) + g_i - h_i \leq y_i + M^* - M^* \cdot \sigma_q, \\ i = \overline{1, n}, \quad q = \overline{1, p+1}, \quad (13)$$

$$\alpha_0 + \sum_{j=1}^l \alpha_j x_{ij} + \sum_{k=1}^p (-1)^{q_k} \cdot (u_{ik} + v_{ik}) + g_i - h_i \geq y_i - M^* + M^* \cdot \sigma_q, \\ i = \overline{1, n}, \quad q = \overline{1, p+1}, \quad (14)$$

$$\sigma_q \in \{0, 1\}, \quad q = \overline{1, p+1}, \quad (15)$$

$$\sum_{q=1}^{p+1} \sigma_q = 1, \quad (16)$$

где M^* — большое положительное число; бинарные переменные σ_q , $q = \overline{1, p+1}$, удовлетворяют правилу:

$$\sigma_q = \begin{cases} 1, & \text{если знаки перед модулями соответствуют } q\text{-й строке матрицы } \Phi, \\ 0, & \text{в противном случае.} \end{cases}$$

Решение задачи ЧБЛП с целевой функцией (5) и с линейными ограничениями (7)–(16) дает оптимальную структуру модульной регрессии (3).

2. Вычислительные эксперименты

Моделированию уровня безработицы в Российской Федерации посвящено много научных работ (см., например, [16–18]). Для этого авторы в основном применяют линейную регрессию. В данной работе моделирование уровня безработицы проводилось с помощью модульных регрессий. Для этого была использована выборка данных объема $n = 16$, приведенная и подробно описанная в [18]. Объясняемой переменной y выступает уровень безработицы (в %) в Российской Федерации, а список объясняющих переменных таков:

- x_1 — прирост населения (тыс. человек);
- x_2 — средняя зарплата (тыс. рублей);
- x_3 — уровень пандемии COVID-19 (%);
- x_4 — средний размер пособия по безработице (тыс. рублей);
- x_5 — минимальный размер оплаты труда (тыс. рублей).

Для проведения вычислительных экспериментов использовался обычный персональный компьютер с процессором AMD Ryzen 3 4300U (2.7 ГГц) и объемом оперативной памяти 16 Гб.

Для решения задач ЧБЛП применялся размещенный в открытом доступе оптимизационный решатель LPSolve. На решение каждой задачи в этом пакете устанавливался лимит времени 20 мин. В результате решения фиксировались МНМ-оценки, сумма модулей остатков J , время решения T в секундах. Числа M и M^* во всех экспериментах равны 1000.

Сначала с помощью МНМ была оценена линейная регрессия со всеми пятью объясняющими переменными:

$$\tilde{y} = 7,551 - 0,000198x_1 - 0,059x_2 + 0,457x_3 - 0,0488x_4 + 0,0412x_5, \quad (17)$$

для которой $J = 5,57383$.

Затем оценивалась модульная регрессия с одним модулем, т.е. решалась задача ЧБЛП (5), (7)–(16) при $p = 1$. Было получено уравнение

$$\tilde{y} = -577,867 - 0,0784x_1 - 0,059x_2 - 51,671x_3 - 0,0488x_4 + 51,614x_5 + |585,418 + 0,0782x_1 + 52,128x_3 - 51,573x_5|, \quad (18)$$

для которого $T = 0,026$ сек., $J = 5,57383$. И тут возникла проблема. Линейная комбинация $585,418 + 0,0782x_1 + 52,128x_3 - 51,573x_5$ неотрицательна для любого наблюдения, поэтому модель (18) это абсолютно та же линейная регрессия (17), записанная в другой форме. Тогда было решено сделать проверку перебором знаков перед модулем, т.е. решить задачу ЧБЛП (5)–(12) при $\Delta_1 = 0$ и при $\Delta_1 = 1$. Лучшая из этих двух моделей обязана совпадать со спецификацией (18). При $\Delta_1 = 0$ была получена еще одна форма линейной регрессии (17) с параметрами $T=0,019$ сек., $J=5,57383$, а при $\Delta_1 = 1$ была идентифицирована новая форма

$$\tilde{y} = 1177,892 + 0,156x_1 + 104,185x_3 - 0,116x_4 - 103,185x_5 - |1169,403 + 0,156x_1 + 0,308x_2 + 105,117x_3 - 0,35x_4 - 104,254x_5|. \quad (19)$$

Для модели (19) $T = 0,017$ сек., $J = 3,36591$. Тем самым было выявлено расхождение в одинаковых с теоретической точки зрения методах оценивания модульной регрессии.

Объяснение этой нестыковки оставалось искать только в реализациях задач в пакете LPSolve. И оно было найдено. Дело в том, что в LPSolve неограниченные переменные $\alpha_0, \alpha_1, \dots, \alpha_5; \beta_{01}, \beta_{11}, \dots, \beta_{51}$ были обозначены следующим образом:

$$\begin{aligned} a_0 &> =-\text{Inf}, a_1 > =-\text{Inf}, a_2 > =-\text{Inf}, a_3 > =-\text{Inf}, a_4 > =-\text{Inf}, a_5 > =-\text{Inf}, \\ b_{0_1} &> =-\text{Inf}, b_{1_1} > =-\text{Inf}, b_{2_1} > =-\text{Inf}, b_{3_1} > =-\text{Inf}, b_{4_1} > =-\text{Inf}, \\ b_{5_1} &> =-\text{Inf}. \end{aligned}$$

А в [15] рекомендуется в задаче математического программирования представлять все неограниченные переменные в виде разности двух неотрицательных переменных. Иными словами, ограничения (7), (13), (14) в задаче (5), (7)–(16) следует переписать в виде

$$\beta_{0k}^* - \beta_{0k}^{**} + \sum_{j=1}^l (\beta_{jk}^* - \beta_{jk}^{**}) x_{ij} = u_{ik} - v_{ik}, \quad i = \overline{1, n}, \quad k = \overline{1, p}, \quad (20)$$

$$\begin{aligned} \alpha_0^* - \alpha_0^{**} + \sum_{j=1}^l (\alpha_j^* - \alpha_j^{**}) x_{ij} + \sum_{k=1}^p (-1)^{\varphi_{qk}} \cdot (u_{ik} + v_{ik}) + \\ + g_i - h_i \leq y_i + M^* - M^* \cdot \sigma_q, \\ i = \overline{1, n}, \quad q = \overline{1, p+1}, \end{aligned} \quad (21)$$

$$\begin{aligned} \alpha_0^* - \alpha_0^{**} + \sum_{j=1}^l (\alpha_j^* - \alpha_j^{**}) x_{ij} + \sum_{k=1}^p (-1)^{\varphi_{qk}} \cdot (u_{ik} + v_{ik}) + \\ + g_i - h_i \geq y_i - M^* + M^* \cdot \sigma_q, \\ i = \overline{1, n}, \quad q = \overline{1, p+1}, \end{aligned} \quad (22)$$

$$\alpha_j^* \geq 0, \quad \alpha_j^{**} \geq 0, \quad j = \overline{0, l}, \quad \beta_{jk}^* \geq 0, \quad \beta_{jk}^{**} \geq 0, \quad j = \overline{0, l}, \quad k = \overline{1, p}. \quad (23)$$

Аналогичным образом следует заменить ограничения (6) и (7) в задаче (5)–(12).

Реализация в LPSolve задачи ЧБЛП с целевой функцией (5) и с линейными ограничениями (8)–(12), (15), (16), (20)–(23) привела к следующей регрессии с одним модулем:

$$\tilde{y} = 4,98 - 0,000295x_1 - 0,0126x_2 - 2,535x_3 + 1,458x_4 - 0,21x_5 - |5,589 + 0,000328x_1 - 0,295x_2 + 2,819x_3 - 1,94x_4 + 1,279x_5|, \quad (24)$$

для которой $T = 0,292$ сек, $J = 1,89861$. Как видно, по качеству аппроксимации модель (24) превзошла все построенные ранее. Перебором знаков модулей при $\Delta_1 = 0$ был получен результат с параметрами $T = 0,101$ сек., $J = 2,1702$, а при $\Delta_1 = 1 - T = 0,098$ сек., $J = 1,89861$. Расхождение не выявлено. Дальнейшие эксперименты проводились только на основе решения задачи (5), (8)–(12), (15), (16), (20)–(23).

Интересно, что при $M = M^* = 1\,000\,000$ была получена иная форма модели (24) с большим числом нулевых оценок:

$$\begin{aligned} \tilde{y} = 0,00139x_1 - 0,182x_2 + 0,284x_3 + 5,536x_4 - 0,149x_5 - \\ - |10,569 - 0,00136x_1 - 0,126x_2 - 6,01896x_4 + 1,218x_5|. \end{aligned}$$

Далее оценивалась модульная регрессия с двумя модулями. В результате была идентифицирована следующая модель:

$$\begin{aligned} \tilde{y} = & 5,357 + 0,0116x_1 - 0,37x_2 + 1,926x_3 + 1,997x_4 + 0,637x_5 + \\ & + |18,265 + 0,00121x_1 - 0,227x_2 + 10,301x_3 - 4,402x_4 + 0,0207x_5| - \\ & - |-18,958 - 0,0117x_1 + 0,189x_2 - 13,66x_3 + 6,469x_4 - 0,784x_5|, \end{aligned} \quad (25)$$

для которой $T=74,873$ сек, $J=0$. Таким образом, найдена «идеальная» регрессия, все остатки которой равны нулю. Однако считать такой результат приемлемым вряд ли возможно. Построенные модульные регрессии (24) и (25) содержат слишком много параметров по сравнению с объемом выборки, поэтому рекомендовать их для прогнозирования не хотелось бы. Так или иначе, проведенные эксперименты позволили выявить нестыковку при реализации задач оценивания модульных регрессий в LPSolve.

В работе [18] было установлено, что только две объясняющие переменные значимо влияют на y — это x_2 и x_5 . Поэтому было принято решение провести дополнительные эксперименты, при этом уменьшить число факторов до двух и увеличить максимальное число модулей в регрессии.

Сначала с помощью МНМ была оценена линейная регрессия

$$\tilde{y} = 8,501 - 0,21x_2 + 0,635x_5, \quad (26)$$

для которой $J = 6,25092$.

Затем была построена модульная регрессия с одним модулем вида

$$\tilde{y} = 8,501 + 45,698x_2 - 177,329x_5 - |45,908x_2 - 177,964x_5|, \quad (27)$$

для которой $T = 1,46$ сек, $J = 4,6681$, и с двумя модулями вида

$$\begin{aligned} \tilde{y} = & 174,596 - 52,577x_2 + 188,656x_5 + \\ & + |-175,855 + 54,053x_2 - 193,978x_5| \\ & - |9,842 - 1,691x_2 + 5,968x_5|, \end{aligned} \quad (28)$$

для которой $T=148,717$ сек, $J=2,37482$.

Задача оценивания модульной регрессии с тремя модулями гораздо сложнее в вычислительном плане. Так, в установленный лимит времени (20 минут) была получена модель, для которой $J = 3,18406$. Понятно, что это не оптимальное решение, поскольку качество регрессии (28) с меньшим числом параметров выше. Поэтому было принято решение построить 4 модели со всеми возможными комбинациями знаков перед модулями. В установленный лимит времени были получены следующие результаты: для модели с комбинацией знаков «+ + +» $J = 3,18406$, «+ + -» $J = 0,65151$, «+ - -» $J = 1,06163$,

«— —» — $J = 3,4319$. Исходя из этого, ни в одном из четырех случаев не была доказана оптимальность решения. Однако модель для случая «+ + —» показала наилучшие результаты по качеству аппроксимации из всех модульных регрессий с тремя модулями. Ее уравнение имеет вид:

$$\begin{aligned} \tilde{y} = & 1,413 + 0,299x_2 - 1,699x_5 + |790,62 + 37,832x_2 - 290,336x_5| + \\ & + |-0,788 + 0,52x_2 - 3,161x_5| - \\ & - |786,155 + 38,0361x_2 - 291,0934x_5|. \end{aligned} \quad (29)$$

Окончательный выбор наилучшей модели из набора (17), (24)–(29) с точки зрения простоты и точности субъективен, поэтому оставим его на усмотрение читателя.

Таким образом, все модульные регрессии превосходили по качеству аппроксимации линейные регрессии. При этом с ростом числа модулей в модульных регрессиях существенно возрастает время их оценивания с помощью МНМ. Так, для пяти факторов при $p = 1$ время составило 0,292 сек, а при $p = 2$ – 74,873 сек. Для двух факторов при $p = 1$ время составило 1,46 сек, при $p = 2$ – 148,717 сек, а при $p = 3$ оптимальное решение не было найдено и за 1200 сек.

Заключение

В работе введена модульная регрессия с линейной частью и с p регрессорами в виде модулей линейных комбинаций объясняющих переменных. Сформулированы две задачи ЧБЛП оценивания таких регрессий с помощью МНМ. В первой из них требуется самостоятельно задавать знаки перед модулями, во второй такого требования нет, поскольку знаки будут идентифицированы автоматически. Проведены вычислительные эксперименты. При реализации задач в пакете LPSolve было выявлено, что обозначение неограниченных переменных в виде «a0 > =-Inf» приводит к некорректному решению. Для корректной работы неограниченные переменные, как рекомендовано в [15], нужно представлять в виде разностей неотрицательных переменных.

Проведенные эксперименты показывают, что использовать модульные регрессии в практике регрессионного анализа целесообразно. За счет регулирования в них числа модулей можно добиваться высококачественных результатов. К тому же модульные регрессии довольно легко интерпретируются, для чего следует представлять их в виде кусочно-заданных функций. Но, к сожалению, чем больше объем выборки и число модулей, тем сложнее становится вычислительная задача оценивания модульной регрессии. Поэтому проблема поиска эффективных алгоритмических и программных средств построения предложенных моделей становится актуальной. К тому же вызывает научный интерес контроль количества ненулевых оценок в модульной регрессии.

Список использованной литературы

1. Taye M.M. Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, Future Directions / M.M. Taye // *Computation*. — 2023. — Vol. 11, no. 3. — P. 52.
2. Krichen M. Convolutional Neural Networks: A Survey / M. Krichen // *Computers*. — 2023. Vol. 12, no. 8. — P. 151.
3. Assessing and Forecasting Water Quality in the Danube River by Using Neural Network Approaches / P.L. Georgescu, S. Moldovanu, C. Iticescu [et al.] // *Science of the Total Environment*. — 2023. — Vol. 879. — P. 162998.
4. Multivariate Energy Forecasting Via Metaheuristic Tuned Long-Short Term Memory and Gated Recurrent Unit Neural Networks / N. Bacanin, L. Jovanovic, M. Zivkovic [et al.] // *Information Sciences*. — 2023. — Vol. 642. — P. 119122.
5. Buaria D. Forecasting Small-Scale Dynamics of Fluid Turbulence Using Deep Neural Networks / D. Buaria, K.R. Sreenivasan // *Proceedings of the National Academy of Sciences*. — 2023. — Vol. 120, no. 30. — P. e2305765120.
6. Айвазян С.А. Прикладная статистика и основы эконометрики / С.А. Айвазян, В.С. Мхитарян. — Москва: Юнити, 1998. — 1005 с.
7. Chatterjee S. Regression Analysis by Example / S. Chatterjee, A.S. Hadi. — New York : John Wiley & Sons, 2015. — 268 p.
8. Носков С.И. Программный комплекс построения некоторых типов кусочно-линейных регрессий / С.И. Носков, А.А. Хоняков. — EDN UTFPOD // *Информационные технологии и математическое моделирование в управлении сложными системами*. — 2019. — № 3 (4). — С. 47–55.
9. Базилевский М.П. Оценивание модульных линейных регрессионных моделей с помощью метода наименьших модулей / М.П. Базилевский, А.Б. Ойдопова. — DOI 10.15593/2224-9397/2023.1.06. — EDN MEKQNE // *Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления*. — 2023. — № 45. — С. 130–146.
10. Progress in Mathematical Programming Solvers from 2001 to 2020 / T. Koch, T. Berthold, J. Pedersen, C. Vanaret // *EURO Journal on Computational Optimization*. — 2022. — Vol. 10. — P. 100031.
11. Базилевский М.П. Программное обеспечение для оценивания модульных линейных регрессий / М.П. Базилевский. — DOI 10.25729/ESI.2023.31.3.013. — EDN PPVLHT // *Информационные и математические технологии в науке и управлении*. — 2023. — № 3 (31). — С. 136–146.
12. Базилевский М.П. Совершенствование алгоритма точного оценивания модульных линейных регрессий с помощью метода наименьших модулей / М.П. Базилевский. — DOI 10.55421/1998-7072_2024_27_4_97. — EDN ADYUDU // *Вестник Технологического университета*. — 2024. — Т. 27, № 4. — С. 97–102.
13. Базилевский М.П. Оценивание регрессионных моделей с мультиарной операцией модуль методом наименьших модулей / М.П. Базилевский. — EDN ICHLWV // *Инженерный вестник Дона*. — 2024. — № 5. — С. 690–697.
14. Базилевский М.П. Оценивание неизвестных параметров многослойной модульной регрессии методом наименьших модулей / М.П. Базилевский. — EDN INJWYZ // *Моделирование, оптимизация и информационные технологии*. — 2024. — Т. 12, № 2 (45). — С. 39.
15. Fisher W.D. A Note on Curve Fitting with Minimum Deviations by Linear Programming / W.D. Fisher // *Journal of the American Statistical Association*. — 1961. — Vol. 56, no. 294. — P. 359–362.

16. Резникова О.С. Экономико-математическое моделирование уровня безработицы молодежи Российской Федерации / О.С. Резникова, Ч. Чжан. — EDN UZCQUF // Геополитика и экогодинамика регионов. — 2024. — Т. 10, № 1. — С. 101–108.

17. Джункеев У.К. Моделирование влияния цифровых технологий на уровень безработицы в России / У.К. Джункеев. — EDN UDFBGD // Вестник Московского университета. Серия 6. Экономика. — 2021. — № 6. — С. 186–201.

18. Антипина Н.В. Построение математической модели уровня безработицы в Российской Федерации / Н.В. Антипина, М.Е. Селиверстова. — DOI 10.17150/2713-1734.2021.3(4).243-249. — EDN APQBJG // System Analysis & Mathematical Modeling. — 2021. — Т. 3, № 4. — С. 243–249.

References

1. Taye M.M. Theoretical Understanding of Convolutional Neural Network: Concepts, Architectures, Applications, future Directions. *Computation*, 2023, vol. 11, no. 3, pp. 52.

2. Krichen M. Convolutional Neural Networks: A Survey. *Computers*, 2023, vol. 12, no. 8, pp. 151.

3. Georgescu P.L., Moldovanu S., Iticescu C., Calmuc M., Calmuc V. Assessing and Forecasting Water Quality in the Danube River by Using Neural Network Approaches. *Science of the Total Environment*, 2023, vol. 879, pp. 162998.

4. Bacanin N., Jovanovic L., Zivkovic M., Kandasamy V., Antonijevic M. 4. Multivariate Energy Forecasting Via Metaheuristic Tuned Long-Short Term Memory and Gated Recurrent Unit Neural Networks. *Information Sciences*, 2023, vol. 642, pp. 119122.

5. Buaria D., Sreenivasan K.R. Forecasting Small-Scale Dynamics of Fluid Turbulence Using Deep Neural Networks. *Proceedings of the National Academy of Sciences*, 2023, vol. 120, no. 30, pp. e2305765120.

6. Aivazyan S.A., Mkhitarany V.S. *Applied Statistics and Basics of Econometrics*. Moscow, Yuniti Publ., 1998. 1005 p.

7. Chatterjee S., Hadi A.S. *Regression Analysis by Example*. New York, John Wiley & Sons, 2015. 268 p.

8. Noskov S.I., Khonyakov A.A. Software Complex for Building Some Types Pieces of 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, pp. 47–55. (In Russian). EDN: UTFPOD.

9. Bazilevskiy M.P., Oydopova A.B. Estimation of Modular Linear Regression Models Using the Least Absolute Deviations. *Vestnik Permskogo natsional'nogo issledovatel'skogo politekhnicheskogo universiteta. Ehlektrotehnika, informatsionnye tekhnologii, sistemy upravleniya = Bulletin of Perm National Research Polytechnic University. Electrical engineering, information technology, control systems*, 2023, no. 45, pp. 130–146. (In Russian). EDN: MEKQHE. DOI: 10.15593/2224-9397/2023.1.06.


10. 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, pp. 100031.

11. Bazilevskiy M.P. Software for Estimating Modular Linear Regressions. *Informatsionnye i matematicheskie tekhnologii v nauke i upravlenii = Information and Mathematical Technologies in Science and Management*, 2023, no. 2, pp. 136–146. (In Russian). EDN: PPVLHT. DOI: 10.25729/ESI.2023.31.3.013.


12. Bazilevskiy M.P. Improving the Algorithm for Exact Estimation of Modular Linear Regressions Using the Least Absolute Deviations. *Vestnik tekhnologicheskogo universiteta = Bulletin of the Technological University*, 2024, no. 4, pp. 97–102. (In Russian). EDN: ADYUDU. DOI: 10.55421/1998-7072_2024_27_4_97.

13. Bazilevskiy M.P. Estimation of Regression Models with Multiary Modulus Operation Using Least Absolute Deviations. *Inzhenernyj vestnik Dona = Engineering journal of Don*, 2024, no. 5, pp. 690–697. (In Russian). EDN: ICHLWV.
14. Bazilevskiy M.P. Unknown Parameters Estimation for Multilayer Modular Regression Using the Least Absolute Deviations Method. *Modelirovanie, optimizaciya i informacionnye tehnologii = Modeling, Optimization and Information Technology*, 2024, vol. 12, no. 2, pp. 39. (In Russian). EDN: INJWYZ.
15. Fisher W.D. A Note on Curve Fitting with Minimum Deviations by Linear Programming. *Journal of the American Statistical Association*, 1961, vol. 56, no. 294, pp. 359–362.
16. Reznikova O.S., Zhang Zh. Economic and Mathematical Modeling of Youth Unemployment in the Russian Federation. *Geopolitika i ehkogeodinamika regionov = Geopolitics and ecogeodynamics of regions*, 2024, vol. 10, no. 1, pp. 101–108. (In Russian). EDN: UZCQUF.
17. Dzhunkeev U.K. Modelling the Impact of Digital Technologies on Unemployment Rate in Russia. *Vestnik Moskovskogo universiteta. Seriya 6, Ekonomika = Moscow University Economics Bulletin*, 2021, no. 6, pp. 186–201. (In Russian). EDN: UDFBGD.
18. Antipina N.V., Seliverstova M.E. Formation a Mathematical Modeling of Unemployment Rate in Russian Federation. *System Analysis & Mathematical Modeling*, 2021, vol. 3, no. 4, pp. 243–249. (In Russian). EDN: APQBJG. DOI: 10.17150/2713-1734.2021.3(4).243-249.

Информация об авторе

Базилевский Михаил Павлович — кандидат технических наук, доцент, кафедра математики, Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация, e-mail: mik2178@yandex.ru,  <https://orcid.org/0000-0002-3253-5697>, SPIN-код: 4347-5028, AuthorID РИНЦ: 679277.

Information about the Author

Mikhail P. Bazilevskiy — PhD in Technical Sciences, Associate Professor, Department of Mathematics, Irkutsk State Transport University, Irkutsk, Russian Federation, e-mail: mik2178@yandex.ru,  <https://orcid.org/0000-0002-3253-5697>, SPIN-Code: 4347-5028, AuthorID RSCI: 679277.

Для цитирования

Базилевский М.П. Оценивание регрессионных моделей с регрессорами в виде модулей линейных комбинаций объясняющих переменных / М.Т. Базилевский. — DOI 10.17150/2713-1734.2024.6(3).269-281 — EDN MLCDNR // System Analysis & Mathematical Modeling. — 2024. — Т. 6, № 3. — С. 269–281.

For Citation

Bazilevskiy M.P. Estimation of Regression Models with Regressors in the Explanatory Variables Linear Combinations Modules Form. *System Analysis & Mathematical Modeling*, 2024, vol. 6, no. 3, pp. 269–281. (In Russian). EDN: MLCDNR. DOI: 10.17150/2713-1734.2024.6(3).269-281.