ISSN 2713-1734

Научная статья
УДК 519.862.6
EDN ZELKZO
DOI 10.17150/2713-1734.2024.6(2).146-158

М.П. Базилевский

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

Контроль автокорреляции остатков с помощью коэффициента Фехнера в задаче математического программирования для отбора информативных регрессоров в линейной регрессии

Аннотация. Статья посвящена проблеме отбора наиболее информативных регрессоров в линейной регрессии, оцениваемой с помощью метода наименьших квадратов. Ранее эта задача была формализована в виде задачи частично-булевого линейного программирования. Целевой функцией в ней выступает значение коэффициента детерминации, а линейные ограничения позволяют контролировать такие характеристики, как абсолютные вклады переменных в общую детерминацию, критерий Стьюдента, коэффициенты вздутия дисперсии, коэффициенты интеркорреляций. Цель данной статьи состоит в расширении задачи частично-булевого программирования линейными ограничениями, позволяющими контролировать в процессе построения по данным временных рядов степень автокорреляции остатков регрессии. Показано, что для обнаружения автокорреляции первого порядка достаточно вычислить коэффициент корреляции между остатками в текущий и предыдущий момент времени. Использовать коэффициент корреляции Пирсона для интеграции в задачу в виде линейных ограничений не представляется возможным. Поэтому был использован коэффициент Фехнера, зависящий от количества совпадений и несовпадений знаков отклонений двух переменных от их средних величин. Этот коэффициент, как и коэффициент Пирсона, принимает значения от –1 до +1. Чем ближе его абсолютное значение к единице, тем сильнее коррелируют переменные. Использование коэффициента Фехнера при вычислении автокорреляции остатков первого порядка позволило интегрировать его в задачу частично-булевого линейного программирования в виде линейных ограничений. Корректность сформулированной задачи подтверждена решением конкретного примера по реальным статистическим данным. При этом была построена модель с полным отсутствием автокорреляции остатков, уравнение которой совпало с уравнением полученной ранее при других ограничениях регрессии, что снова подтверждает ее адекватность.

Ключевые слова. Линейная регрессия, временные ряды, метод наименьших квадратов, автокорреляция остатков, коэффициент Фехнера, отбор информативных регрессоров, задача частично-булевого линейного программирования.

Информация о статье. Дата поступления: 16 апреля 2024 г.; дата принятия к публикации: 23 мая 2024 г.; дата онлайн-размещения: 19 июня 2024 г.

Original article

M.P. Bazilevskiy
Irkutsk State Transport University,
Irkutsk, Russian Federation

Control of Autocorrelation of Residuals Using the Fechner Coefficient in a Mathematical Programming Problem for Subset Selection in Linear Regression

Abstract. This article is devoted to the problem of subset selection in linear regression estimated using the ordinary least squares method. Previously, this problem was formalized as a mixed 0-1 integer linear programming problem. The target function in it is the value of the coefficient of determination, and linear restrictions allow one to control such characteristics as the absolute contributions of variables to the overall determination, Student's criterion, variance inflation factors, and inter-correlation coefficients. The purpose of this article is to expand the mixed 0-1 integer linear programming problem with linear constraints that make it possible to control the degree of autocorrelation of regression residuals during the construction process based on time series data. It is shown that to detect first-order autocorrelation it is enough to calculate the correlation coefficient between the residuals at the current and previous time. It does not seem possible to use the Pearson correlation coefficient for integration into the problem in the form of linear constraints. Therefore, the Fechner coefficient was used, depending on the number of coincidences and discrepancies in the signs of deviations of two variables from their average values. This coefficient, like the Pearson coefficient, takes values from –1 to +1. The closer its absolute value is to one, the more strongly the variables are correlated. The use of the Fechner coefficient in calculating the autocorrelation of first-order residuals made it possible to integrate it into the problem of mixed 0-1 integer linear programming in the form of linear constraints. The correctness of the formulated problem is confirmed by solving a specific example using real statistical data. At the same time, a model was constructed with a complete absence of autocorrelation of residuals, the equation of which coincided with the equation of the regression obtained earlier under other restrictions, which again confirms its adequacy.

Keywords. Linear regression, time series, ordinary least squares, autocorrelation of residuals, Fechner coefficient, subset selection, mixed 0-1 integer linear programming problem.

Article info. Received 16 April, 2024; Accepted 23 May, 2024; Available online 19 June, 2024.

Введение

Регрессионный анализ [1], относящийся к одному из методов машинного обучения [2], сегодня находит широкое применение при решении конкретных прикладных задач обработки данных. Построение регрессионной модели затруднят то, что сразу не понятно, какие именно объясняющие переменные следует включать в ее уравнение. Эта проблема известна в отечественной литературе как задача отбора информативных регрессоров (ОИР) [3], а в зарубежной как «subset selection» или «feature selection». Существует два способа ее точного решения. Первый из них — метод «всех возможных регрессий» [4], который состоит в полном переборе

всех возможных комбинаций входящих в модель факторов, что представляет собой довольно трудоемкую вычислительную задачу. Второй способ связан с решением специальным образом сформулированной задачи математического программирования (см., например, [5; 6]), что эффективнее, чем метод «всех регрессий».

Усложняет задачу ОИР то, что каждая регрессионная модель характеризуется целой системой различных критериев адекватности, отвечающих за измерение самых разных ее качественных сторон. Так, при использовании метода наименьших квадратов (МНК) оценкой качества аппроксимации регрессии служит коэффициент детерминации, критерий Стьюдента выступает мерой значимости регрессоров, по коэффициентам вздутия дисперсии судят о мультиколлинеарности и т.д. При обработке временных рядов в регрессионной модели часто возникает проблема автокорреляции остатков, из-за которой МНК-оценки становятся не самыми эффективными. Для выявления автокорреляции остатков первого порядка применяется известный в эконометрике критерий Дарбина — Уотсона.

Ранее автору удалось свести задачу ОИР в линейной регрессии, оцениваемой с помощью МНК, к задаче частично-булевого линейного программирования (ЧБЛП) (см, например, [7]). Постепенно эта задача дополнялась линейными ограничениями [8–10], позволяющими контролировать значения таких характеристик модели, как критерий Стьюдента, коэффициенты вздутия дисперсии, абсолютные вклады переменных в общую детерминацию, коэффициенты интеркорреляций. Линейные ограничения, позволяющие контролировать степень автокорреляции остатков линейной регрессии, никогда еще в эту задачу ЧБЛП не интегрировались.

Цель работы состоит в расширении задачи ЧБЛП для ОИР в линейной регрессии линейными ограничениями, позволяющими контролировать в процессе отбора автокорреляцию остатков.

1. Формализация задачи

Модель множественной линейной регрессии записывается в виде

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \varepsilon_i,\ i = \overline{1, n}, \qquad (1)$$

где n — объем выборки (число наблюдений); l — количество объясняющих (входных, независимых) переменных; y_i, $i = \overline{1, n}$, — известные значения объясняемой (выходной, зависимой) переменной; x_{ij}, $i = \overline{1, n}$, , $j = \overline{1, n}$, — известные значения объясняющих переменных; $\alpha_j, j = \overline{1, n}$, — неизвестные параметры; ε_j, $i = \overline{1, n}$, — ошибки аппроксимации.

Если линейная регрессия (1) оценивается с помощью МНК, то одно из допущений, согласно теореме Гаусса — Маркова, состоит в том, что ошибки разных наблюдений должны быть некоррелированы, т.е. $M(\varepsilon_i, \varepsilon_j) = 0$, $i \neq j$, где M — математическое ожидание. Если это свойство не выполняется, то имеет место автокорреляция остатков [11], из-за чего оценки остаются несмещенными и состоятельными, но перестают быть эффективными.

Пусть в регрессионной модели (1) ошибки специфицированы как

$$\varepsilon_i = \rho\varepsilon_{i-1} + u_i,\ i = \overline{2, n}, \tag{2}$$

где u_i — случайные величины, распределенные как «белый шум»; ρ — коэффициент авторегрессии (2), причем, $|\rho| < 1$.

Ошибки линейной регрессии (1) ε_i, $i = \overline{1, n}$, неизвестны, поэтому оценить коэффициент авторегрессии ρ можно, используя остатки $e_i = y_i - \tilde{\alpha}_0 - \sum_{j=1}^{l} \tilde{\alpha}_j x_{ij}$, по следующей формуле:

$$\tilde{\rho} = \sum_{i=2}^{n} e_i e_{i-1} \Big/ \sum_{i=2}^{n} e_{i-1}^2.$$

Если $\tilde{\rho} \approx 0$, то можно считать, что автокорреляция остатков отсутствует; если $\tilde{\rho} \approx 1$, то полная положительная автокорреляция остатков; если $\tilde{\rho} \approx -1$ то полная отрицательная автокорреляция остатков.

Коэффициент автокорреляции первого порядка r_1 между e_i и e_{i-1} находится как коэффициента корреляции Пирсона:

$$r_1 = \frac{(n-1)\sum_{i=2}^{n} e_{i-1}e_i - \sum_{i=2}^{n} e_{i-1}\sum_{i=2}^{n} e_i}{\sqrt{(n-1)\sum_{i=2}^{n} e_{i-1}^2 - \left(\sum_{i=2}^{n} e_{i-1}\right)^2}\sqrt{(n-1)\sum_{i=2}^{n} e_i^2 - \left(\sum_{i=2}^{n} e_i\right)^2}}. \tag{3}$$

Если n велико, то $\sum_{i=2}^{n} e_i \approx \sum_{i=2}^{n} e_{i-1} \approx 0$, $\sum_{i=2}^{n} e_i^2 \approx \sum_{i=2}^{n} e_{i-1}^2$. Тогда $r_1 = \tilde{\rho}$. Из этого следует, что если n — велико, а именно, как отмечено в [12], $n \geq 15$, то практически нет никакой разницы между оцененным коэффициентом авторегрессии $\tilde{\rho}$ и коэффициентом автокорреляции 1-го порядка r_1. Получается, что вывод относительно автокорреляции остатков при $n \geq 15$ зависит от коэффициента r_1. Заметим, что в такой ситуации также используется критерий Дарбина — Уотсона $DW = 2(1 - r_1)$ [11; 12], имеющий статистический характер.

Известно, что простым показателем степени взаимосвязи между двумя переменными x и y является коэффициент (индекс) Фехнера [13], который находится по формуле

$$\Phi = \frac{\upsilon - \omega}{\upsilon + \omega}, \quad (4)$$

где υ — количество совпадений знаков отклонений $x_i - \overline{x}$ и $y_i - \overline{y}$, а ω — количество несовпадений их знаков. Половину отклонений, равных нулю, относят к υ, половину — к ω. Критерий Фехнера так же, как и критерий Пирсона, принимает значений от —1 до +1. При $\Phi > 0$ имеем положительную корреляцию, при $\Phi < 0$ — отрицательную, при $\Phi = 0$ связь отсутствует. Преимущество данного метода, как отмечают авторы статьи [14], в том, что коэффициент Фехнера просто вычисляется, не зависит от закона распределения изучаемых данных и может применяться в условиях малой выборки. В [15] продемонстрирована методика расчета коэффициента Фехнера и Пирсона.

Будем использовать для вычисления коэффициента автокорреляции первого порядка при обнаружении автокорреляции остатков в линейной регрессии (1) не критерий Пирсона (3), а критерий Фехнера (4):

$$\Phi = \frac{\upsilon - \omega}{n - 1}, \quad (5)$$

где

$$\upsilon = \sum_{i=2}^{n} \frac{1}{2}\left|\operatorname{sgn}\left(e_i - \overline{e_{2..n}}\right) + \operatorname{sgn}\left(e_{i-1} - \overline{e_{1..n-1}}\right)\right|,$$

$$\omega = \sum_{i=2}^{n} \frac{1}{2}\left|\operatorname{sgn}\left(e_i - \overline{e_{2..n}}\right) - \operatorname{sgn}\left(e_{i-1} - \overline{e_{1..n-1}}\right)\right|, \ \overline{e_{2..n}} =$$

$$= \frac{1}{n-1}\sum_{i=2}^{n} e_i, \ \overline{e_{1..n-1}} = \frac{1}{n-1}\sum_{i=2}^{n} e_{i-1}, \ \operatorname{sgn}(x) = \begin{cases} 1, \text{ если } x \geq 0, \\ -1, \text{ если } x < 0. \end{cases}$$

Рассмотрим далее одну из возможных форм задачи ЧБЛП [7–10] для ОИР в линейной регрессии.

Предварительно проведем стандартизацию объясняющих переменных по правилам

$$y_i^{\bullet} = \frac{y_i - \overline{y}}{\sigma_y}, \ x_{i1}^{\bullet} = \frac{x_{i1} - \overline{x_1}}{\sigma_{x_1}}, \ \ldots, x_{il}^{\bullet} = \frac{x_{il} - \overline{x_l}}{\sigma_{x_l}}, \ i = \overline{1, n},$$

где $\sigma_y, \sigma_{x_1}, \ldots, \sigma_{x_l}$ — стандартные отклонения переменных.

Введем стандартизованную модель линейной регрессии вида

$$y_i^{\bullet} = \sum_{j=1}^{l} \beta_j x_{ij}^{\bullet} + \varepsilon_i^{\bullet}, \ i = \overline{1, n}, \quad (6)$$

где β_j, $j = \overline{1, l}$ — неизвестные параметры, $\varepsilon_i^{\bullet}$, $i = \overline{1, n}$ — ошибки аппроксимации.

МНК-оценки неизвестных параметров линейной регрессии (6) находятся в результате решения системы линейных алгебраических уравнений

$$R_{xx} \cdot \beta = R_{yx},$$

где $R_{xx} = \begin{pmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_l} \\ r_{x_1x_2} & 1 & \dots & r_{x_2x_l} \\ \dots & \dots & \dots & \dots \\ r_{x_1x_l} & r_{x_2x_l} & \dots & 1 \end{pmatrix}$ — матрица коэффициентов интеркорреляций; $R_{yx} = r_{yx_1} \quad r_{yx_2} \quad \dots \quad r_{yx_l}$ — вектор-столбец коэффициентов корреляции объясняющих переменных с y; $\beta_{yx} = \beta_1 \quad \beta_2 \quad \dots \quad \beta_l$ — вектор-столбец неизвестных параметров.

Сформулируем следующую задачу ОИР [8]: требуется из общего набора l объясняющих переменных выбрать оптимальное число регрессоров так, чтобы коэффициент детерминации $R^2 = \sum_{j=1}^{l} r_{yx_j} \cdot \beta_j$ линейной регрессии (6) был максимален, а абсолютные вклады назначенных переменных в общую детерминацию $C_{x_j}^{\text{абс}} = r_{yx_j} \cdot \beta_j$ были не меньше выбранного из интервала [0, 1] числа θ. В работах [8–10] эта задача сведена к следующей задаче ЧБЛП:

$$\sum_{j=1}^{l} r_{yx_j} \cdot \beta_j \rightarrow \max, \tag{7}$$

$$-\left(1 - \delta_j\right) \cdot M \le \sum_{k=1}^{l} r_{x_jx_k} \cdot \beta_k - r_{yx_j} \le \left(1 - \delta_j\right) \cdot M, \quad j = \overline{1, l}, \tag{8}$$

$$0 \le \beta_j \le \delta_j \cdot M, j \in J^+, \tag{9}$$

$$-\delta_j \cdot M \le \beta_j \le 0, j \in J^-, \tag{10}$$

$$\delta_j \in \{0, 1\}, \; j = \overline{1, l}, \tag{11}$$

$$r_{yx_j} \cdot \beta_j \ge \theta \cdot \delta_j, \; j = \overline{1, l}, \tag{12}$$

где $\delta_j = \begin{cases} 1, & \text{если j-я объясняющая переменная входит в модель,} \\ 0, & \text{в противном случае,} \end{cases}$

J^+, J^- — индексные подмножества, элементы которых удовлетво-

ряют условиям $r_{yx_j} > 0$ и $r_{yx_j} <$ соответственно; M — большое положительное число, выбор нижних границ которого рассмотрен в работах [8–10].

В [10] экспериментально доказано, что ОИР методом решения задачи ЧБЛП (7) — (12) многократно эффективнее ОИР методом «всех регрессий». Задача (7) — (12) может быть дополнена ограничениями [7–10], позволяющими контролировать самые разные характеристики регрессионных моделей. Однако никогда еще в эту задачу не интегрировались линейные ограничения, позволяющие контролировать автокорреляцию остатков.

Остатки $e_i^{\bullet}$, $i = \overline{1, n}$, стандартизованной регрессии (6) находятся по формулам

$$e_i^{\bullet} = y_i^{\bullet} - \sum_{j=1}^{l} \beta_j x_{ij}^{\bullet}, \; i = \overline{1, n}. \tag{13}$$

Заметим, что знаки остатков линейных регрессий (1) и (6) не отличаются.

Если по остаткам (13) вычислять коэффициент автокорреляции первого порядка как критерий Пирсона (3), то линеаризовать полученное выражение не представляется возможным. Поэтому будем вычислять этот коэффициент как критерий Фехнера (5), в котором

$$\upsilon = \sum_{i=2}^{n} \frac{1}{2} \left| \operatorname{sgn}\left(e_i^{\bullet} - \overline{e_{2..n}^{\bullet}}\right) + \operatorname{sgn}\left(e_{i-1}^{\bullet} - \overline{e_{1..n-1}^{\bullet}}\right) \right|, \tag{14}$$

$$\omega = \sum_{i=2}^{n} \frac{1}{2} \left| \operatorname{sgn}\left(e_i^{\bullet} - \overline{e_{2..n}^{\bullet}}\right) - \operatorname{sgn}\left(e_{i-1}^{\bullet} - \overline{e_{1..n-1}^{\bullet}}\right) \right|, \tag{15}$$

где $\overline{e_{2..n}^{\bullet}} = \frac{1}{n-1} \sum_{i=2}^{n} e_i^{\bullet}$, $\overline{e_{1..n-1}^{\bullet}} = \frac{1}{n-1} \sum_{i=1}^{n-1} e_i^{\bullet}$.

Очевидно, что величины (14) и (15) связаны соотношением $\upsilon + \omega = n - 1$, откуда

$$\upsilon = n - 1 - \omega. \tag{16}$$

Введем бинарные переменные σ_i, χ_{i-1}, $i = \overline{2, n}$, по правилам

$$\sigma_i = \begin{cases} 1, \text{ если } e_i^{\bullet} - \overline{e_{2..n}^{\bullet}} > 0, \\ 0, \text{ если } e_i^{\bullet} - \overline{e_{2..n}^{\bullet}} < 0, \end{cases}$$

$$\chi_{i-1} = \begin{cases} 1, \text{ если } e_{i-1}^{\bullet} - \overline{e_{1..n-1}^{\bullet}} > 0, \\ 0, \text{ если } e_{i-1}^{\bullet} - \overline{e_{1..n-1}^{\bullet}} < 0. \end{cases}$$

На их основе сформируем следующие линейные ограничения:

$$-(1 - \sigma_i) \cdot M \le e_i^{\bullet} - \overline{e_{2..n}^{\bullet}} \le \sigma_i \cdot M,\ i = \overline{2, n}, \tag{17}$$

$$-(1 - \chi_{i-1}) \cdot M \le e_{i-1}^{\bullet} - \overline{e_{1..n-1}^{\bullet}} \le \chi_{i-1} \cdot M,\ i = \overline{2, n}. \tag{18}$$

Учитывая (13), ограничения (17), (18) принимают вид

$$-(1 - \sigma_i) \cdot M \le y_i^{\bullet} - \sum_{j=1}^{l} \beta_j x_{ij}^{\bullet} - Z_{2..n} \le \sigma_i \times \times M,\ i = \overline{2, n}, \tag{19}$$

$$-(1 - \chi_{i-1}) \cdot M \le y_{i-1}^{\bullet} - \sum_{j=1}^{l} \beta_j x_{i-1,j}^{\bullet} - Z_{1..n-1} \le \chi_{i-1} \cdot M,\ i = \overline{2, n}, \tag{20}$$

где

$$Z_{2..n} = \frac{1}{n-1} \sum_{i=2}^{n} \left(y_i^{\bullet} - \sum_{j=1}^{l} \beta_j x_{ij}^{\bullet} \right), \tag{21}$$

$$Z_{1..n-1} = \frac{1}{n-1} \sum_{i=1}^{n-1} \left(y_i^{\bullet} - \sum_{j=1}^{l} \beta_j x_{ij}^{\bullet} \right). \tag{22}$$

Тогда формулу (15) для количества несовпадений знаков отклонений можно записать в виде

$$\omega = \sum_{i=2}^{n} |\sigma_i - \chi_{i-1}|. \tag{23}$$

Для раскрытия модулей в (23) используем прием, примененный в работе [16]. Для этого введем следующие неотрицательные переменные:

$$g_i = \begin{cases} \sigma_i - \chi_{i-1}, \text{ если } \sigma_i - \chi_{i-1} \ge 0, \\ 0, \text{ если } \sigma_i - \chi_{i-1} < 0, \end{cases} i = \overline{2, n},$$

$$h_i = \begin{cases} 0, \text{ если } \sigma_i - \chi_{i-1} \ge 0, \\ \sigma_i - \chi_{i-1}, \text{ если } \sigma_i - \chi_{i-1} < 0, \end{cases} i = \overline{2, n}.$$

Тогда аргументы под знаками модуля в (23) удовлетворяют тождествам

$$\sigma_i - \chi_{i-1} = g_i - h_i,\ i = 2, n, \tag{24}$$

а само выражение (23) будет иметь вид

$$\omega = \sum_{i=2}^{n} (g_i + h_i). \quad (25)$$

Поскольку для переменных g_i и h_i должны выполняться условия $g_i \cdot h_i = 0$, $i = \overline{2, n}$, введем следующие булевы переменные:

$$\Delta_i = \begin{cases} 0, \text{ если } \sigma_i - \chi_{i-1} < 0, \\ 1, \text{ если } \sigma_i - \chi_{i-1} \geq 0, \end{cases} i = \overline{2, n}.$$

Тогда для обеспечения срабатывания либо переменной g_i, либо h_i, введем линейные ограничения

$$g_i \leq \Delta_i, i = \overline{2, n}, \quad (26)$$

$$h_i \leq 1 - \Delta_i, i = \overline{2, n}. \quad (27)$$

Допустим, что исследователь желает осуществить ОИР так, чтобы коэффициент Фехнера линейной регрессии не превосходил по абсолютной величине числа Φ^* из интервала [0, 1]. Используя (5), введем двойное неравенство

$$-\Phi^* \leq \frac{\upsilon - \omega}{n - 1} \leq \Phi^*. \quad (28)$$

Очевидно, чем меньше Φ^*, тем жестче требование на отсутствие в регрессии автокорреляции остатков.

Учитывая (16) и (25), перепишем двойное неравенство (28) в виде

$$-\Phi^* \leq 1 - \frac{2}{n - 1} \sum_{i=2}^{n} (g_i + h_i) \leq \Phi^*. \quad (29)$$

Тогда решение задачи ЧБЛП с целевой функцией (7) и с линейными ограничениями (8)–(12), (19)–(22), (24), (26), (27), (29) приводит к построению линейной регрессии с оптимальным по коэффициенту детерминации R^2 числом регрессоров, в которой $C_{x_j}^{\text{абс}} \geq \theta$, а коэффициент Фехнера по модулю $|\Phi| \leq \Phi^*$.

2. Моделирование

Для демонстрации работоспособности предложенного математического аппарата решалась задача моделирования грузовых железнодорожных перевозок в Иркутской области по статистическим данным из статьи [8] объема $n = 21$. Сначала по ним с помощью программы ВИнтер-1 была решена задача ЧБЛП (7) — (12) при $\theta = 0{,}01$. Большое число M в этой задаче ВИнтер-1 определяет

автоматически так, как описано в [10]. В результате была построена следующая линейная регрессия:

$$\tilde{y} = -67{,}791 + \overset{(0{,}4123)}{1{,}235}x_2 - \overset{(0{,}0705)}{0{,}023}x_8 + \\ + \overset{(0{,}2838)}{0{,}00057}x_{18} + \overset{(0{,}1005)}{0{,}359}x_{22}, \qquad (30)$$

где $\tilde{y}$ — расчетное значение отправления грузов ж/д транспортом (млн тонн), x_2 — процент трудоспособного населения, x_8 — число собственных легковых автомобилей на 1 000 чел. (шт.), x_{18} — число предприятий и организаций, x_{22} — производство электроэнергии (млрд киловатт-часов). В скобках над коэффициентами в уравнении (30) указаны абсолютные вклады переменных в общую детерминацию, которые, как и ожидалось, превосходят величину 0,01.

Для регрессии (30) $R^2 = 0{,}867206$, а коэффициент Фехнера $\Phi = 0{,}1$.

Затем решалась задача ЧБЛП (7)–(12) с дополнительными ограничениями (19)–(22), (24), (26), (27), (29) для контроля автокорреляции остатков. При этом число M в ограничениях (8)–(10) определялось автоматически, число θ было выбрано равным 0,01, число M в ограничениях (19), (20) задавалось равным 1 000, а в (29) $\Phi^* = 0$. В результате была построена модель

$$\tilde{y} = -86{,}106 + \overset{(0{,}532)}{1{,}593}x_2 + \overset{(0{,}251)}{0{,}0005}x_{18} + \overset{(0{,}082)}{0{,}293}x_{22}, \qquad (31)$$

для которой коэффициент детерминации чуть меньше, чем для (30), и составляет 0,864777, но зато коэффициент корреляции Фехнера между ее остатками e_i и e_{i-1} уменьшился с 0,1 до 0.

Заметим, что в работе [8] получена точно такая же модель (31), но сделано это на основе решения совершенно другой задачи ЧБЛП, в которой не учитывались ограничения на степень автокорреляции остатков в регрессии.

Заключение

В статье задача построения линейной регрессии с оптимальным по коэффициенту детерминации числом регрессоров, в которой абсолютные вклады переменных в общую детерминацию не меньше числа θ, а коэффициент корреляции Фехнера между остатками в текущий и предыдущий момент времени по модулю не больше числа Φ^*, сведена к задаче частично-булевого линейного программирования. Проведено тестирование сформулированной задачи на примере моделирования грузовых железнодорожных перевозок в Иркутской области. Построенная модель, спецификация которой совпала с полученным ранее при других ограниче-

ниях уравнением, оказалась полностью лишена автокорреляция в остатках. Предложенные в работе линейные ограничения можно использовать для контроля автокорреляции остатков в сочетании с ограничениями, введенными автором ранее для других критериев адекватности.

Список использованной литературы

1. Montgomery D.C. Introduction to Linear Regression Analysis / D.C. Montgomery, E.A. Peck, G.G. Vining. — John Wiley & Sons, 2021. — 704 p.

2. Alpaydin E. Introduction to Machine Learning / E. Alpaydin. — MIT Press, 2020. — 537 p.

3. Стрижов В.В. Методы выбора регрессионных моделей / В.В. Стрижов. — Москва : Вычислительный Центр им. А.А. Дородницына Российской академии наук, 2010. — 60 с.

4. Айвазян С.А. Прикладная статистика и основы эконометрики / С.А. Айвазян, В.С. Мхитарян. — Москва : ЮНИТИ, 1998. — 1022 с.

5. Chung S. A Mathematical Programming Approach for Integrated Multiple Linear Regression Subset Selection and Validation / S. Chung, Y.W. Park, T. Cheong // Pattern Recognition. — 2020. — Vol. 108. — P. 107565.

6. Bertsimas D. Scalable Holistic Linear Regression / D. Bertsimas, M.L. Li. — DOI 10.1016/j.orl.2020.02.008 // Operations Research Letters. — 2020. — Vol. 48, no. 3. — P. 203– 208.

7. Базилевский М.П. Отбор информативных регрессоров с учётом мультиколлинеарности между ними в регрессионных моделях как задача частично-булевого линейного программирования / М.П. Базилевский // Моделирование, оптимизация и информационные технологии. — 2018. — Т. 6, № 2 (21). — С. 104–118.

8. Базилевский М.П. Построение вполне интерпретируемых линейных регрессионных моделей с помощью метода последовательного повышения абсолютных вкладов переменных в общую детерминацию / М.П. Базилевский. — DOI 10.17308/sait/1995-5499/2022/2/5-16. — EDN CNDSSW // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. — 2022. — № 2. — С. 5–16.

9. Базилевский М.П. Формализация процесса отбора информативных регрессоров в линейной регрессии в виде задачи частично-булевого линейного программирования с ограничениями на коэффициенты интеркорреляций / М.П. Базилевский. — DOI 10.17513/snt.39723. — EDN FCOPEL // Современные наукоемкие технологии. — 2023. — № 8. — С. 10–14.

10. Базилевский М.П. Сравнительный анализ эффективности методов построения вполне интерпретируемых линейных регрессионных моделей / М.П. Базилевский. — DOI 10.17759/mda.2023130404. — EDN VXFGBO // Моделирование и анализ данных. — 2023. — Т. 13, № 4. — С. 59–83.

11. Гефан Г.Д. Эконометрика / Г.Д. Гефан. — EDN VAABCZ. — Иркутск : Иркутский государственный университет путей сообщения, 2005. — 84 с.

12. Кремер Н.Ш. Теория вероятностей и математическая статистика / Н.Ш. Кремер. — Москва : Юнити-Дана, 2004. — 573 с.

13. Фёрстер Э. Методы корреляционного и регрессионного анализа / Э. Фёрстер, Б. Рёнц. — Москва : Финансы и статистика, 1983. — 303 с.

14. Демаков В.И. Модификация метода Фехнера для повышения устойчивости анализа данных / В.И. Демаков, А.В. Демаков. — DOI 10.18101/2304-5728-2022-1-35-44. — EDN IGQBCH // Вестник Бурятского государственного университета. Математика, информатика. — 2022. — № 1. — С. 35–44.

15. Саадалов Т. Методика расчета коэффициента корреляции Фехнера и Пирсона, и их области применения / Т. Саадалов, Р. Мырзаибраимов, Ж.Д. Абдуллаева. — DOI 10.33619/2414-2948/71/31. — EDN GNMYZT // Бюллетень науки и практики. — 2021. — Т. 7, № 10. — С. 270–276.

16. 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 MEKQHE // Вестник Пермского национального исследовательского политехнического университета. Электротехника, информационные технологии, системы управления. — 2023. — № 45. — С. 130–146.

References

1. Montgomery D.C., Peck E.A., Vining G.G. *Introduction to Linear Regression Analysis*. John Wiley & Sons, 2021. 704 p.

2. Alpaydin E. *Introduction to Machine Learning*. MIT Press, 2020. 537 p.

3. Strizhov V.V. *Methods for selecting regression models.* Moscow, Vychislitel'nyi Tsentr im. A.A. Dorodnitsyna Rossiiskoi akademii nauk Publ., 2010. 60 p.

4. Aivazyan S.A., Mkhitaryan V.S. *Applied statistics and fundamentals of econometrics.* Moscow, YUNITI Publ., 1998. 1022 p.

5. Chung S., Park Y.W., Cheong T. A mathematical Programming Approach for Integrated Multiple Linear Regression Subset Selection and Validation. *Pattern Recognition*, 2020, vol. 108, pp. 107565.

6. Bertsimas D., Li M.L. Scalable Holistic Linear Regression. *Operations Research Letters,* 2020, vol. 48, no. 3, pp. 203–208. DOI:10.1016/j.orl.2020.02.008.

7. Bazilevskii M.P. Selection of informative regressors taking into account multicollinearity between them in regression models as a partial-boolean linear programming problem. *Modelirovanie, optimizacija i informacionnye tehnologii = Modeling, Optimization and Information Technology,* 2018, vol. 6, no. 2, pp. 104–118. (In Russian).

8. Bazilevskii M.P. Construction of Quite Interpretable Linear Regression Models Using the Method of Successive Increase the Absolute Contributions of Variables to the General Determination. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: Sistemnyi analiz i informatsionnye tekhnologii = Proceedings of Voronezh State University. Series: Systems analysis and information technologies,* 2022, no. 2, pp. 5–16. (In Russian). EDN: CNDSSW. DOI: 10.17308/sait/1995-5499/2022/2/5-16.

9. Bazilevskii M.P. Formalization the Subset Selection Process in Linear Regression as a Mixed Integer 0-1 Linear Programming Problem with Constraints on Intercorrelation Coefficients. *Sovremennye naukoemkie tekhnologii = Modern high technologies,* 2023, no. 8, pp. 10–14. (In Russian). EDN: FCOPEL. DOI: 10.17513/snt.39723.

10. Bazilevskii M.P. Comparative Analysis of the Effectiveness of Methods for Constructing Quite Interpretable Linear Regression Models. *Modelirovanie i analiz dannykh = Modeling and data analysis*, 2023, vol. 13, no. 4, pp. 59–83. (In Russian). EDN: VXFGBO. DOI: 10.17759/mda.2023130404.

11. Gefan G.D. *Econometrics.* Irkutsk, Irkutsk State Railway Transport Engineering University Publ., 2005. 84 p. EDN: VAABCZ.

12. Kremer N.SH. *Theory of Probability and Mathematical Statistics.* Moscow, Yuniti-Dana Publ., 2004. 573 p.

13. Forster E., Ronz B. *Methoden der Korrelations - und RegressionsanalYse*. Berlin, 1979. 504 p. (Russ. ed.: Ferster E., Rents B. *Methods of correlation and regression analysis.* Moscow, Finansy i statistika Publ., 1983. 303 p.).

14. Demakov V.I., Demakov A.V. Modification of the Fechner Method to Increase the Robustness of Data Analysis. *Vestnik Buryatskogo gosudarstvennogo universiteta. Matematika, informatika = Bulletin of the Buryat State University. Mathematics, Informatics,* 2022, no. 1, pp. 35–44. (In Russian). EDN: IGQBCH. DOI: 10.18101/2304-5728-2022-1-35-44.

15. Saadalov T., Myrzaibraimov R., Abdullaeva Zh.D. Calculating Procedure for the Correlation Coefficient of Fechner and Pearson and Their Application Areas. *Byulleten' nauki i praktiki = Bulletin of Science and Practice,* 2021, vol. 7, no. 10, pp. 270–276. (In Russian). EDN: GNMYZT. DOI: 10.33619/2414-2948/71/31.

16. Bazilevskii M.P., Oidopova A.B. Estimating of Modular Linear Regression Models Using the Least Absolute Deviations. *Vestnik Permskogo natsional'nogo issledovatel'skogo politekhnicheskogo universiteta. Ehlektrotekhnika, informatsionnye tekhnologii, sistemy upravleniya = Bulletin of the 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.

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 RSCI: 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. Контроль автокорреляции остатков с помощью коэффициента Фехнера в задаче математического программирования для отбора информативных регрессоров в линейной регрессии / М.П. Базилевский. — DOI 10.17150/2713-1734.2024.6(2).146-158. — EDN ZELKZO // System Analysis & Mathematical Modeling. — 2024. — Т. 6, № 2. — С. 146–158.

For Citation

Bazilevskiy M.P. Control of Autocorrelation of Residuals Using the Fechner Coefficient in a Mathematical Programming Problem for Subset Selection in Linear Regression. *System Analysis & Mathematical Modeling*, 2024, vol. 6, no. 2, pp. 146–158. (In Russian). EDN: ZELKZO. DOI: 10.17150/2713-1734.2024.6(2).146-158.