ВЕСТНИК ТОМСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2024 Управление, вычислительная техника и информатика № 69
Tomsk State University Journal of Control and Computer Science

ОБРАБОТКА ИНФОРМАЦИИ

DATA PROCESSING

Научная статья
УДК 519.862.6
doi: 10.17223/19988605/69/4

Интеграция ограничений на коэффициенты интеркорреляций в оптимизационную задачу и условия построения вполне интерпретируемых неэлементарных линейных регрессий

Михаил Павлович Базилевский

Иркутский государственный университет путей сообщения, Иркутск, Россия, mik2178@yandex.ru

Аннотация. Сформулировано строгое определение вполне интерпретируемой неэлементарной линейной регрессии, содержащее девять условий. Для контроля мультиколлинеарности в оптимизационную задачу интегрированы линейные ограничения на интеркорреляции как между внешними регрессорами, так и между внутренними переменными из разных регрессоров. Математический аппарат был реализован в специализированной компьютерной программе. С помощью этой программы по реальным данным успешно построена вполне интерпретируемая неэлементарная линейная регрессия, удовлетворяющая всем девяти сформулированным условиям.

Ключевые слова: регрессионный анализ; неэлементарная линейная регрессия; интерпретация; кусочно-заданная функция; метод наименьших квадратов; задача частично-булевого линейного программирования; мультиколлинеарность; железнодорожные перевозки.

***Для цитирования*:** Базилевский М.П. Интеграция ограничений на коэффициенты интеркорреляций в оптимизационную задачу и условия построения вполне интерпретируемых неэлементарных линейных регрессий // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2024. № 69. С. 31–41. doi: 10.17223/19988605/69/4

Original article
doi: 10.17223/19988605/69/4

Integration of constraints on intercorrelation coefficients into the optimization problem and conditions for constructing quite interpretable non-elementary linear regressions

Mikhail P. Bazilevskiy

Irkutsk State Transport University, Irkutsk, Russian Federation, mik2178@yandex.ru

Abstract. The article formulates a rigorous definition of quite interpretable non-elementary linear regression, containing nine conditions. To control multicollinearity, linear constraints on intercorrelations between both external regressors and internal variables from different regressors are integrated into the optimization problem. The mathematical apparatus was implemented in a specialized computer program. Using this program, a quite interpretable non-elementary linear regression has been successfully constructed using real data, satisfying all nine formulated conditions.

© М.П. Базилевский, 2024

Keywords: regression analysis; non-elementary linear regression; interpretation; piecewise-defined function; ordinary least squares method; mixed 0-1 integer linear programming problem; multicollinearity; rail transportation.

***For citation*:** Bazilevskiy, M.P. (2024) Integration of constraints on intercorrelation coefficients into the optimization problem and conditions for constructing quite interpretable non-elementary linear regressions. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitelnaja tehnika i informatika – Tomsk State University Journal of Control and Computer Science*. 69. pp. 31–41. doi: 10.17223/19988605/69/4

Введение

В настоящее время с помощью моделей машинного обучения решается множество прикладных задач в самых разных областях деятельности человека (см., напр.: [1–3]). К одному из главных свойств моделей, безусловно, относится точность. Любая математическая модель с низкой точностью не имеет никакой практической пользы. При этом высокоточные модели, например в виде многослойных нейронных сетей, зачастую можно охарактеризовать как громоздкие «черные ящики», объяснить механизм функционирования которых не представляется возможным. Если такая «необъяснимая» модель применяется для решения несерьезной задачи, например для прогнозирования числа просмотров видеороликов на некотором интернет-ресурсе, то вряд ли стоит тратить время и пытаться как-то ее объяснить, ведь главное, чтобы она давала точные прогнозы. Но если решается серьезная задача, например с помощью модели машинного обучения ставится медицинский диагноз человеку либо выносится судебное решение, то одной только точности недостаточно. Нужно еще доказать, что модель корректно работает во всех возможных ситуациях и не будет давать сбоев. Иными словами, нужно повысить доверие к модели, т.е. сделать так, чтобы у специалистов не возникло никаких сомнений относительно ее внедрения в производственный процесс. Для этого требуется модель интерпретировать, что означает объяснить механизм ее функционирования в понятных для человека терминах. Интерпретируемое машинное обучение, возникшее относительно недавно [4, 5], является актуальным научным направлением на сегодняшний день [6, 7].

В монографии К. Молнара [5] к интерпретируемым моделям машинного обучения относятся следующие виды регрессионных моделей: линейные [8] и логистические регрессии [9], а также обобщенные линейные [10] и аддитивные модели [11]. При этом универсального строгого математического определения интерпретируемой модели пока не существует. В [12] автором были рассмотрены вопросы построения вполне интерпретируемых линейных регрессий (ВИЛинР), причем идентификация ВИЛинР осуществлялась с помощью метода наименьших квадратов (МНК) [13], а для отбора информативных регрессоров (ОИР) [14, 15] использовался аппарат частично-булевого линейного программирования (ЧБЛП). Ограничения в задаче ЧБЛП ставились на знаки коэффициентов регрессии и абсолютные вклады переменных в общую детерминацию, что позволило эффективно бороться с мультиколлинеарностью [16, 17]. В [18] в задаче построения ВИЛинР появились дополнительные ограничения на коэффициенты интеркорреляций. Строгое математическое определение вполне интерпретируемых квазилинейных регрессий (ВИКЛинР), обобщающих ВИЛинР, появилось в [19].

Процесс поиска новых хорошо интерпретируемых математических форм связи между переменными продолжается. Так, в [20] с использованием бинарных операций min и max введены неэлементарные линейные регрессии (НЛинР). Там же задача ОИР в НЛинР была формализована в виде задачи ЧБЛП. Достоинство таких моделей в том, что они легко интерпретируются в виде кусочно-заданных функций. Поэтому в [21] задача ЧБЛП была расширена ограничениями, с помощью которых можно контролировать абсолютные вклады переменных в общую детерминацию. Полученное в результате ее решения уравнение называется вполне интерпретируемой неэлементарной линейной регрессией (ВИНЛинР). Однако строгого математического определения ВИНЛинР пока еще дано не было. К тому же в задаче ЧБЛП для построения ВИНЛинР не были задействованы ограничения на мультиколлинеарность, что все же не позволяет полностью относить такие модели ко вполне интерпретируемым. Цель данной статьи заключается в решении двух указанных проблем.

1. Вполне интерпретируемые неэлементарные линейные регрессии

Пусть имеется выборочная совокупность объема n, содержащая значения зависимой (объясняемой) переменной y и l независимых (объясняющих) переменных $x_1, x_2, \ldots, x_l$. Рассмотрим введенную в [20] НЛинР с неизвестными параметрами α_0, α_j, $j=\overline{1,l}$, $\alpha_j^{\min}$, $\alpha_j^{\max}$, $k_j^{\min}$, $k_j^{\max}$, $j=\overline{1,C_l^2}$, вида

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{i,j} + \sum_{j=1}^{C_l^2} \alpha_j^{\min} \min\left\{x_{i,\mu_{j,1}}, k_j^{\min} x_{i,\mu_{j,2}}\right\} +$$

$$+\sum_{j=1}^{C_l^2} \alpha_j^{\max} \max\left\{x_{i,\mu_{j,1}}, k_j^{\max} x_{i,\mu_{j,2}}\right\} + \varepsilon_i, \quad i=\overline{1,n}, \tag{1}$$

где C_l^2 – число сочетаний без повторений из l элементов по 2; min, max – бинарные операции, возвращающие минимум и максимум двух чисел соответственно; $\mu_{j,1}$, $\mu_{j,2}$, $j=\overline{1,C_l^2}$ – элементы матрицы M размера $C_l^2 \times 2$, содержащей в строках в лексикографическом порядке все возможные комбинации пар индексов объясняющих переменных; ε_i, $i=\overline{1,n}$, – ошибки аппроксимации; $x_{ij} \neq 0$, $i=\overline{1,n}$, $j=\overline{1,l}$.

Как отмечено в [20], области возможных значений оценок параметров $k_j^{\min}$, $k_j^{\max}$, $j=\overline{1,C_l^2}$, можно записать в виде:

$$k_j^{\min}, k_j^{\max} \in \left(k_j^{\text{нижн}}, k_j^{\text{верхн}}\right), \quad j=\overline{1,C_l^2}, \tag{2}$$

где $k_j^{\text{нижн}} = \min\left\{\frac{x_{1,\mu_{j,1}}}{x_{1,\mu_{j,2}}}, \ldots, \frac{x_{n,\mu_{j,1}}}{x_{n,\mu_{j,2}}}\right\}$, $k_j^{\text{верхн}} = \max\left\{\frac{x_{1,\mu_{j,1}}}{x_{1,\mu_{j,2}}}, \ldots, \frac{x_{n,\mu_{j,1}}}{x_{n,\mu_{j,2}}}\right\}$.

Равномерно разбивая промежутки (2) p точками [20], перейдем от НЛинР (1) к линейной по параметрам регрессии:

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \sum_{j=1}^{C_l^2} \sum_{k=1}^{p} \alpha_{jk}^{-} z_{ijk}^{-} + \sum_{j=1}^{C_l^2} \sum_{k=1}^{p} \alpha_{jk}^{+} z_{ijk}^{+} + \varepsilon_i, \quad i=\overline{1,n}, \tag{3}$$

где $z_{ijk}^{-} = \min\left\{x_{i,\mu_{j,1}}, \lambda_{jk} x_{i,\mu_{j,2}}\right\}$, $z_{ijk}^{+} = \max\left\{x_{i,\mu_{j,1}}, \lambda_{jk} x_{i,\mu_{j,2}}\right\}$, $i=\overline{1,n}$, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$; λ_{jk}, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$ – элементы матрицы Λ размера $C_l^2 \times p$, каждый из которых равен k-й точке разбиения j-го промежутка (2); α_{jk}^{-}, α_{jk}^{+}, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$ – неизвестные параметры.

Используя модель (3), введем строгое определение ВИНЛинР.

Определение. НЛинР (1), оцененная с помощью МНК, называется вполне интерпретируемой, если:

1) каждая объясняющая переменная входит в модель не более одного раза;

2) знаки всех коэффициентов корреляции r_{yx_j}, $j=\overline{1,l}$, $r_{yz_{jk}^{-}}$, $r_{yz_{jk}^{+}}$, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$, удовлетворяют содержательному смыслу решаемой задачи;

3) знаки всех коэффициентов представленной в виде кусочно-заданной функции модели согласуются со знаками соответствующих коэффициентов корреляции r_{yx_j};

4) все абсолютные вклады регрессоров в общую детерминацию R^2 не ниже выбранного числа $\theta \in [0,1)$;

5) все оценки $\tilde{\alpha}_j$, $\tilde{\alpha}_j^{\min}$, $\tilde{\alpha}_j^{\max}$ значимы по t-критерию Стьюдента для выбранного уровня значимости α;

6) все интеркорреляции как между внешними регрессорами, так и между внутренними переменными из разных регрессоров незначимы по t-критерию Стьюдента для выбранного уровня значимости α;

7) все вспомогательные зависимости каждого регрессора от остальных незначимы по *F*-критерию Фишера для выбранного уровня значимости α;

8) модель в целом значима по *F*-критерию Фишера для выбранного уровня значимости α;

9) коэффициент детерминации модели $R^2 \geq 0{,}8$.

Прокомментируем далее некоторые условия сформулированного определения.

Условие 2 можно проверить еще до оценивания НЛинР. Для этого требуется участие экспертов из данной предметной области. Их задача состоит в анализе соответствия содержательному смыслу знаков коэффициентов корреляции между регрессорами и переменной *y*. Если выявлены несоответствия, то использовать такие регрессоры в процессе оценивания модели не следует.

Заметим, что если эксперты согласовали корреляции между объясняющими переменными и *y*, то согласование корреляций между *y* и преобразованными переменными z_{jk}^{-}, z_{jk}^{+}, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$, можно автоматизировать. Например, преобразование $\min\{x_i, k\cdot x_j\}$ не нужно исключать, если для него выполняется одно из четырех условий:

$$1)\ k>0,\ r_{yx_i}>0,\ r_{yx_j}>0,\ r_{y,\min\{x_i,kx_j\}}>0;$$

$$2)\ k>0,\ r_{yx_i}<0,\ r_{yx_j}<0,\ r_{y,\min\{x_i,kx_j\}}<0;$$

$$3)\ k<0,\ r_{yx_i}>0,\ r_{yx_j}<0,\ r_{y,\min\{x_i,kx_j\}}>0;$$

$$4)\ k<0,\ r_{yx_i}<0,\ r_{yx_j}>0,\ r_{y,\min\{x_i,kx_j\}}<0.$$

Условие 3 равносильно выполнению неравенств $\alpha_j \cdot r_{yx_j}>0$, $j=\overline{1,l}$; $\alpha_{jk}^{-}\cdot r_{y,z_{jk}^{-}}>0$, $\alpha_{jk}^{+}\cdot r_{y,z_{jk}^{+}}>0$, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$.

Фигурирующие в условии 4 абсолютные вклады регрессоров в общую детерминацию для модели (3) находятся по формулам $C_{x_j}^{\text{абс}}=r_{yx_j}\cdot\beta_j$, $j=\overline{1,l}$; $C_{z_{jk}^{-}}^{\text{абс}}=r_{yz_{jk}^{-}}\cdot\beta_{jk}^{-}$, $C_{z_{jk}^{+}}^{\text{абс}}=r_{yz_{jk}^{+}}\cdot\beta_{jk}^{+}$, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$, где β_j, $j=\overline{1,l}$; β_{jk}^{-}, β_{jk}^{+}, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$ – неизвестные параметры стандартизованной регрессии вида

$$y_i^{\bullet}=\sum_{j=1}^{l}\beta_j x_{ij}^{\bullet}+\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\beta_{jk}^{-}z_{ijk}^{*}+\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\beta_{jk}^{+}z_{ijk}^{**}+\varepsilon_i^{\bullet},\ \ i=\overline{1,n}, \tag{4}$$

в которой $y_i^{\bullet}=\dfrac{y_i-\overline{y}}{\sigma_y}$, $i=\overline{1,n}$; $x_{ij}^{\bullet}=\dfrac{x_{ij}-\overline{x_j}}{\sigma_{x_j}}$, $i=\overline{1,n}$, $j=\overline{1,l}$; $z_{ijk}^{*}=\dfrac{z_{ijk}^{-}-\overline{z_{jk}^{-}}}{\sigma_{z_{jk}^{-}}}$, $z_{ijk}^{**}=\dfrac{z_{ijk}^{+}-\overline{z_{jk}^{+}}}{\sigma_{z_{jk}^{+}}}$, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$; $\varepsilon_i^{\bullet}$, $i=\overline{1,n}$ – новые ошибки аппроксимации.

При этом справедливо следующее равенство:

$$\sum_{j=1}^{l}C_{x_j}^{\text{абс}}+\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}C_{z_{jk}^{-}}^{\text{абс}}+\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}C_{z_{jk}^{+}}^{\text{абс}}=R^2,$$

где R^2 – коэффициент детерминации, оцененной с помощью МНК регрессии (3) или (4).

Условия 6 и 7 отвечают за контроль эффекта мультиколлинеарности в модели.

Поясним условие 6 на простом примере. Пусть оцененная НЛинР имеет вид:

$$\tilde{y}=2+3x_1+4\min\{x_2,8x_5\}+6\max\{x_3,2x_4\}.$$

Тогда интеркорреляциями между внешними регрессорами являются коэффициенты корреляции между x_1 и $\min\{x_2,8x_5\}$, x_1 и $\max\{x_3,2x_4\}$, $\min\{x_2,8x_5\}$ и $\max\{x_3,2x_4\}$. А интеркорреляциями между внутренними переменными из разных регрессоров являются коэффициенты корреляции между x_1 и x_2, x_1 и x_5, x_1 и x_3, x_1 и x_4, x_2 и x_3, x_2 и x_4, x_5 и x_3, x_5 и x_4.

Если в условии 6 выбран уровень значимости α, то, как отмечено в [21], интеркорреляции будут незначимы по *t*-критерию Стьюдента тогда, когда все они по абсолютной величине будут

меньше числа $r = \dfrac{t_{\text{крит}}(\alpha, n-2)}{\sqrt{n-2+t^2_{\text{крит}}(\alpha, n-2)}}$, где $t_{\text{крит}}(\alpha, n-2)$ – критическая точка распределения Стьюдента.

Как видно, представленное определение ВИНЛинР носит субъективный характер, поскольку окончательный выбор числа θ в условии 4, уровней значимости α в условиях 5, 6, 7, 8 и даже границы R^2 в условии 9 остается за исследователем. Понятно, что от этого выбора будут зависеть результаты моделирования.

Перейдем к формализации задачи ОИР в стандартизованной регрессии (4) с помощью аппарата математического программирования. Решение этой задачи, если оно существует, должно гарантировать выполнение условий вышеприведенного определения ВИНЛинР. Для этого введем бинарные переменные

$$\delta_j \in \{0,1\},\ j=\overline{1,l};\ \delta^-_{jk}, \delta^+_{jk} \in \{0,1\},\ j=\overline{1,C^2_l},\ k=\overline{1,p}, \tag{5}$$

отвечающие за вхождения регрессоров в модель (4). Например, если $\delta^-_{jk}=1$, то регрессор z^-_{jk} входит в регрессию, а если $\delta^-_{jk}=0$, то нет.

Для выполнения условия 1 нужно ввести ограничения

$$\delta_j + \sum_{i\in\Upsilon_j}\sum_{k=1}^{p}\delta^-_{ik} + \sum_{i\in\Upsilon_j}\sum_{k=1}^{p}\delta^+_{ik} \le 1,\quad j=\overline{1,l}, \tag{6}$$

где Υ_j – множество номеров строк матрицы M, содержащих среди своих элементов число *j*.

Условие 2, как уже отмечалось, проверяется до решения задачи.

Выполнение условия 3 гарантируют ограничения

$$0 \le \beta_j \le M\cdot\delta_j,\quad j\in\left\{s \mid r_{yx_s}>0\right\}, \tag{7}$$

$$-M\cdot\delta_j \le \beta_j \le 0,\quad j\in\left\{s \mid r_{yx_s}<0\right\}, \tag{8}$$

$$0 \le \beta^-_{jk} \le M\cdot\delta^-_{jk},\quad (j,k)\in\left\{(s_1,s_2) \mid r_{yz^-_{s_1s_2}}>0\right\}, \tag{9}$$

$$-M\cdot\delta^-_{jk} \le \beta^-_{jk} \le 0,\quad (j,k)\in\left\{(s_1,s_2) \mid r_{yz^-_{s_1s_2}}<0\right\}, \tag{10}$$

$$0 \le \beta^+_{jk} \le M\cdot\delta^+_{jk},\quad (j,k)\in\left\{(s_1,s_2) \mid r_{yz^+_{s_1s_2}}>0\right\}, \tag{11}$$

$$-M\cdot\delta^+_{jk} \le \beta^+_{jk} \le 0,\quad (j,k)\in\left\{(s_1,s_2) \mid r_{yz^+_{s_1s_2}}<0\right\}, \tag{12}$$

где *M* – большое положительное число, процедура выбора которого подробно описана в [20].

Условие 4 контролируется с помощью ограничений

$$r_{yx_j}\cdot\beta_j \ge \theta\cdot\delta_j,\quad j=\overline{1,l}, \tag{13}$$

$$r_{yz^-_{jk}}\cdot\beta^-_{jk} \ge \theta\cdot\delta^-_{jk},\quad j=\overline{1,C^2_l},\quad k=\overline{1,p}, \tag{14}$$

$$r_{yz^+_{jk}}\cdot\beta^+_{jk} \ge \theta\cdot\delta^+_{jk},\quad j=\overline{1,C^2_l},\quad k=\overline{1,p}, \tag{15}$$

в которых число $\theta\in[0,1)$ рекомендуется брать не ниже 0,05.

Для контроля интеркорреляций между внешними регрессорами в условии 6 введем ограничения

$$\left|r_{x_ix_j}\right|\cdot\left(\delta_i+\delta_j-1\right)\le r,\quad i=\overline{1,l-1},\ j=\overline{i+1,l}, \tag{16}$$

$$\left|r_{x_jz^-_{sk}}\right|\cdot\left(\delta_j+\delta^-_{sk}-1\right)\le r,\quad \left|r_{x_jz^+_{sk}}\right|\cdot\left(\delta_j+\delta^+_{sk}-1\right)\le r,\quad j=\overline{1,l},\quad s\in\Upsilon\setminus\Upsilon_j,\quad k=\overline{1,p}, \tag{17}$$

$$\left|r_{z^-_{s_1k_1}z^-_{s_2k_2}}\right|\cdot\left(\delta^-_{s_1k_1}+\delta^-_{s_2k_2}-1\right)\le r,\quad \left|r_{z^+_{s_1k_1}z^+_{s_2k_2}}\right|\cdot\left(\delta^+_{s_1k_1}+\delta^+_{s_2k_2}-1\right)\le r,\quad j_1=\overline{1,l-1},\quad j_2=\overline{j_1+1,l},$$

$$s_1\in\Upsilon_{j_1}\cap\Upsilon_{j_2},\quad k_1=\overline{1,p},\quad s_2\in\Upsilon\setminus\left(\Upsilon_{j_1}\cup\Upsilon_{j_2}\right),\quad s_1<s_2,\ k_2=\overline{1,p}, \tag{18}$$

$$\left|r_{z^-_{s_1k_1}z^+_{s_2k_2}}\right|\cdot\left(\delta^-_{s_1k_1}+\delta^+_{s_2k_2}-1\right)\le r\,,\quad j_1=\overline{1,l-1},\quad j_2=\overline{j_1+1,l}\,,$$

$$s_1\in\Upsilon_{j_1}\cap\Upsilon_{j_2}\,,\quad k_1=\overline{1,p}\,,\quad s_2\in\Upsilon\setminus\left(\Upsilon_{j_1}\cup\Upsilon_{j_2}\right),\quad k_2=\overline{1,p}\,,\tag{19}$$

где Υ – множество номеров строк матрицы M.

Подобный прием для линейной регрессии можно найти в [18]. Ограничения (16) предназначены для контроля интеркорреляций между объясняющими переменными, (17) – между объясняющими переменными и бинарными операциями min и max, (18) – между бинарными операциями min и min, max и max, (19) – между бинарными операциями min и max.

Для контроля интеркорреляций между внутренними переменными из разных регрессоров нужно использовать ограничения

$$\left|r_{x_jx_{\mu_{s1}}}\right|\cdot\left(\delta_j+\delta^-_{sk}-1\right)\le r\,,\quad \left|r_{x_jx_{\mu_{s2}}}\right|\cdot\left(\delta_j+\delta^-_{sk}-1\right)\le r\,,\quad j=\overline{1,l}\,,\quad s\in\Upsilon\setminus\Upsilon_j\,,\quad k=\overline{1,p}\,,\tag{20}$$

$$\left|r_{x_jx_{\mu_{s1}}}\right|\cdot\left(\delta_j+\delta^+_{sk}-1\right)\le r\,,\quad \left|r_{x_jx_{\mu_{s2}}}\right|\cdot\left(\delta_j+\delta^+_{sk}-1\right)\le r\,,\quad j=\overline{1,l}\,,\quad s\in\Upsilon\setminus\Upsilon_j\,,\quad k=\overline{1,p}\,,\tag{21}$$

$$\left|r_{x_{\mu_{s_1,1}}x_{\mu_{s_2,1}}}\right|\cdot\left(\delta^-_{s_1k_1}+\delta^-_{s_2k_2}-1\right)\le r\,,\quad \left|r_{x_{\mu_{s_1,1}}x_{\mu_{s_2,2}}}\right|\cdot\left(\delta^-_{s_1k_1}+\delta^-_{s_2k_2}-1\right)\le r\,,$$

$$\left|r_{x_{\mu_{s_1,2}}x_{\mu_{s_2,1}}}\right|\cdot\left(\delta^-_{s_1k_1}+\delta^-_{s_2k_2}-1\right)\le r\,,\quad \left|r_{x_{\mu_{s_1,2}}x_{\mu_{s_2,2}}}\right|\cdot\left(\delta^-_{s_1k_1}+\delta^-_{s_2k_2}-1\right)\le r\,,$$

$$j_1=\overline{1,l-1},\quad j_2=\overline{j_1+1,l}\,,\quad s_1\in\Upsilon_{j_1}\cap\Upsilon_{j_2}\,,\quad k_1=\overline{1,p}\,,\quad s_2\in\Upsilon\setminus\left(\Upsilon_{j_1}\cup\Upsilon_{j_2}\right),\quad s_1<s_2\,,\quad k_2=\overline{1,p}\,,\tag{22}$$

$$\left|r_{x_{\mu_{s_1,1}}x_{\mu_{s_2,1}}}\right|\cdot\left(\delta^+_{s_1k_1}+\delta^+_{s_2k_2}-1\right)\le r\,,\quad \left|r_{x_{\mu_{s_1,1}}x_{\mu_{s_2,2}}}\right|\cdot\left(\delta^+_{s_1k_1}+\delta^+_{s_2k_2}-1\right)\le r\,,$$

$$\left|r_{x_{\mu_{s_1,2}}x_{\mu_{s_2,1}}}\right|\cdot\left(\delta^+_{s_1k_1}+\delta^+_{s_2k_2}-1\right)\le r\,,\quad \left|r_{x_{\mu_{s_1,2}}x_{\mu_{s_2,2}}}\right|\cdot\left(\delta^+_{s_1k_1}+\delta^+_{s_2k_2}-1\right)\le r\,,$$

$$j_1=\overline{1,l-1},\quad j_2=\overline{j_1+1,l}\,,\quad s_1\in\Upsilon_{j_1}\cap\Upsilon_{j_2}\,,\quad k_1=\overline{1,p}\,,\quad s_2\in\Upsilon\setminus\left(\Upsilon_{j_1}\cup\Upsilon_{j_2}\right),\quad s_1<s_2\,,\quad k_2=\overline{1,p}\,,\tag{23}$$

$$\left|r_{x_{\mu_{s_1,1}}x_{\mu_{s_2,1}}}\right|\cdot\left(\delta^-_{s_1k_1}+\delta^+_{s_2k_2}-1\right)\le r\,,\quad \left|r_{x_{\mu_{s_1,1}}x_{\mu_{s_2,2}}}\right|\cdot\left(\delta^-_{s_1k_1}+\delta^+_{s_2k_2}-1\right)\le r\,,$$

$$\left|r_{x_{\mu_{s_1,2}}x_{\mu_{s_2,1}}}\right|\cdot\left(\delta^-_{s_1k_1}+\delta^+_{s_2k_2}-1\right)\le r\,,\quad \left|r_{x_{\mu_{s_1,2}}x_{\mu_{s_2,2}}}\right|\cdot\left(\delta^-_{s_1k_1}+\delta^+_{s_2k_2}-1\right)\le r\,,$$

$$j_1=\overline{1,l-1},\quad j_2=\overline{j_1+1,l}\,,\quad s_1\in\Upsilon_{j_1}\cap\Upsilon_{j_2}\,,\quad k_1=\overline{1,p}\,,\quad s_2\in\Upsilon\setminus\left(\Upsilon_{j_1}\cup\Upsilon_{j_2}\right),\quad k_2=\overline{1,p}\,.\tag{24}$$

Ограничения (20) предназначены для контроля интеркорреляций между внутренними переменными из регрессоров с одной переменной и регрессоров с операцией min, (21) – из регрессоров с одной переменной и регрессоров с операцией max, (22) – из регрессоров с операциями min, (23) – из регрессоров с операциями max, (24) – из регрессоров с операциями min и max.

В зависимости от бинарных переменных изменение конфигурации системы линейных алгебраических уравнений для нахождения МНК-оценок стандартизованной регрессии можно записать в виде следующих ограничений:

$$-\left(1-\delta_j\right)M\le\sum_{k=1}^{l}r_{x_jx_k}\cdot\beta_k+\sum_{s=1}^{C_l^2}\sum_{k=1}^{p}r_{x_jz^-_{sk}}\cdot\beta^-_{sk}+\sum_{s=1}^{C_l^2}\sum_{k=1}^{p}r_{x_jz^+_{sk}}\cdot\beta^+_{sk}-r_{yx_j}\le(1-\delta_j)M\,,\quad j=\overline{1,l}\,,\tag{25}$$

$$-\left(1-\delta^-_{jk}\right)M\le\sum_{s=1}^{l}r_{x_sz^-_{jk}}\cdot\beta_s+\sum_{s_1=1}^{C_l^2}\sum_{s_2=1}^{p}r_{z^-_{s_1s_2}z^-_{jk}}\cdot\beta^-_{s_1s_2}+$$

$$+\sum_{s_1=1}^{C_l^2}\sum_{s_2=1}^{p}r_{z^+_{s_1s_2}z^-_{jk}}\cdot\beta^+_{s_1s_2}-r_{yz^-_{jk}}\le(1-\delta^-_{jk})M\,,\quad j=\overline{1,C_l^2}\,,\quad k=\overline{1,p}\,,\tag{26}$$

$$-\left(1-\delta^+_{jk}\right)M\le\sum_{s=1}^{l}r_{x_sz^+_{jk}}\cdot\beta_s+\sum_{s_1=1}^{C_l^2}\sum_{s_2=1}^{p}r_{z^-_{s_1s_2}z^+_{jk}}\cdot\beta^-_{s_1s_2}+$$

$$+\sum_{s_1=1}^{C_l^2}\sum_{s_2=1}^{p}r_{z^+_{s_1s_2}z^+_{jk}}\cdot\beta^+_{s_1s_2}-r_{yz^+_{jk}}\le(1-\delta^+_{jk})M\,,\quad j=\overline{1,C_l^2}\,,\quad k=\overline{1,p}\,.\tag{27}$$

Например, если $\delta_{jk}^{-}=1$, что означает, что регрессор z_{jk}^{-} входит в регрессию, то соответствующее двойное неравенство (26) трансформируется в строгое равенство. Иными словами, происходит включение уравнения в систему. В противном случае, когда $\delta_{jk}^{-}=0$, включения не происходит. Подход к выбору большого числа *M* в (25)–(27) также обсуждается в [20].

В качестве критерия оптимизации выступает коэффициент детерминации R^2, который, естественным образом, должен быть максимальным:

$$R^2=\sum_{j=1}^{l} r_{yx_j}\cdot\beta_j+\sum_{j=1}^{C_l^2}\sum_{k=1}^{p} r_{yz_{jk}^{-}}\cdot\beta_{jk}^{-}+\sum_{j=1}^{C_l^2}\sum_{k=1}^{p} r_{yz_{jk}^{+}}\cdot\beta_{jk}^{+}\to\max\,. \qquad (28)$$

Пусть экспертами успешно выполнена проверка условия 2. Тогда решение оптимизационной задачи ЧБЛП с целевой функцией (28) и линейными ограничениями (5)–(27), если оно существует, гарантирует выполнение условий 1, 3, 4 и 6 в построенной НЛинР. Формализация остальных условий в виде линейных ограничений и последующая их интеграция в предложенную задачу возможна. Например, условие 9 с учетом (28) можно записать в виде простого ограничения $R^2\geq 0{,}8$. Однако для обеспечения условий 5, 7 и 8 нужно вводить дополнительные бинарные переменные, что усложнит и без того непростую в вычислительном плане оптимизационную задачу. К тому же условия 5, 7 и 8 по смыслу соотносятся с условиями 4, 6 и 9 соответственно. Так, если абсолютные вклады регрессоров в общую детерминацию высоки, то, скорее всего, и все коэффициенты будут значимы по *t*-критерию Стьюдента. Если интеркорреляции между внешними регрессорами малы, то, вероятно, и зависимости каждого регрессора от множества остальных будут незначимы. А если коэффициент R^2 высок, то модель в большинстве случаев будет значима в целом по *F*-критерию Фишера. Таким образом, очень высоки шансы, что решение задачи (5)–(28) будет гарантировать выполнение всех девяти условий ВИНЛинР.

2. Проверка условий ВИНЛинР на примере логистической задачи

Для тестирования разработанного математического аппарата было решено использовать статистические данные, приведенные в [20], по следующим переменным:

y – отправление грузов ж/д транспортом общего пользования в Иркутской области (млн т);

x_2 – процент трудоспособного населения;

x_3 – численность рабочей силы (тыс. чел.);

x_5 – численность пенсионеров (тыс. чел.);

x_8 – число собственных легковых автомобилей на 1 000 человек населения (шт.);

x_{18} – число предприятий и организаций;

x_{20} – кредиторская задолженность организаций (млн руб.);

x_{22} – производство электроэнергии (млрд кВт·ч);

x_{58} – тарифы на грузовые перевозки ж/д транспортом (усл. ед.).

По этим данным в [20] была построена НЛинР вида

$$\tilde{y}=-24{,}527+\underset{(13{,}98)}{\overset{(0{,}6427)}{1{,}189}}\min\{x_2;0{,}000933x_{18}\}-\underset{(-3{,}361)}{\overset{(0{,}1129)}{0{,}0196}}\min\{x_5;0{,}006754x_{20}\}-$$
$$-\underset{(-2{,}182)}{\overset{(0{,}0843)}{0{,}0323}}\min\{x_8;0{,}11725x_{58}\}+\underset{(3{,}859)}{\overset{(0{,}1063)}{0{,}0254}}\max\{x_3;23{,}079x_{22}\}\,, \qquad (29)$$

для которой $R^2=0{,}946183$.

В уравнении (29) в скобках под коэффициентами указаны наблюдаемые *t*-статистики, а над коэффициентами – абсолютные вклады регрессоров в R^2.

Сначала было решено проверить, можно ли отнести модель (29) к ВИНЛинР, т.е. проверить выполнение всех девяти предложенных выше условий. Будем считать, что число $\theta=0{,}05$, а уровень значимости $\alpha=0{,}01$.

Очевидно, что для модели (29) выполняются условия 1, 4 и 9.

В работе [20] экспертами проведен подробный содержательный анализ знаков влияния объясняющих переменных на *y*, поэтому выполнены условие 2 и, как следствие, условие 3.

По критерию Стьюдента для уровня $\alpha = 0,01$ в модели (29) незначим только коэффициент при регрессоре $\min\{x_8; 0,11725x_{58}\}$, но уже для уровня $\alpha = 0,05$ он становится значимым. Поэтому условие 5 можно считать по большому счету выполненным.

Наблюдаемое значение критерия Фишера для модели (29) составляет 70,3619, что гораздо выше критического значения для уровня $\alpha = 0,01$, поэтому модель в целом значима, т.е. условие 8 выполнено.

Проверим наличие мультиколлинеарности в модели (29). Как отмечено выше, пороговое значение *r* для абсолютных величин интеркорреляций можно найти по формуле

$$r = \frac{t_{\text{крит}}\left(0,01;\ 21-2\right)}{\sqrt{21-2+t^2_{\text{крит}}\left(0,01;\ 21-2\right)}} = \frac{2,8609}{\sqrt{19+2,8609^2}} = 0,5487 .$$

Все абсолютные величины интеркорреляций между внешними регрессорами, кроме коэффициента корреляции между $\min\{x_5; 0,006754x_{20}\}$ и $\min\{x_8; 0,11725x_{58}\}$, равного 0,8576, не превышают значения 0,5487. Отсюда следует, что условие 6 нарушено. К тому же абсолютные величины многих интеркорреляций между внутренними переменными превышают значение 0,5487. Например, коэффициент корреляции между x_2 и x_3 равен 0,8199, между x_2 и x_5 равен –0,9017, между x_3 и x_5 равен –0,863 и др.

Две вспомогательные регрессии из четырех значимы по критерию Фишера для уровня $\alpha = 0,01$: зависимость $\min\{x_5; 0,006754x_{20}\}$ от $\min\{x_2; 0,000933x_{18}\}$, $\min\{x_8; 0,11725x_{58}\}$, $\max\{x_3; 23,079x_{22}\}$ и зависимость $\min\{x_8; 0,11725x_{58}\}$ от $\min\{x_2; 0,000933x_{18}\}$, $\min\{x_5; 0,006754x_{20}\}$, $\max\{x_3; 23,079x_{22}\}$. Тем самым условие 7 нарушено.

Таким образом, модель (29) нельзя отнести к ВИНЛинР из-за присутствия в ней мультиколлинеарности.

После этого с использованием тех же данных была предпринята попытка построить ВИНЛинР. Для автоматической формализации задачи ЧБЛП (5)–(28) по данным о грузовых ж/д перевозках была использована программа построения вполне интерпретируемых элементарных и неэлементарных квазилинейных регрессионных моделей (ВИнтер-2) [18]. Она позволяет автоматически формулировать нужную оптимизационную задачу для пакета LPSolve IDE. Начальные настройки в программе ВИнтер-2 остались те же, которые использовались в [20]. Параметры $\theta = 0,05$, $r = 0,5487$.

В результате с помощью пакета LPSolve была построена НЛинР вида

$$\tilde{y} = -77,893 + \underset{(9,889)}{\overset{(0,551)}{0,000888}} \min\{x_{18}; 1071,164x_{22}\} + \underset{(7,691)}{\overset{(0,372)}{1,379}} \max\{x_2; 0,047x_3\}, \qquad (30)$$

для которой $R^2 = 0,923181$.

Далее для модели (30) проверялись 9 условий ВИНЛинР. Нетрудно заметить, что условия 1, 2, 3, 4, 5 и 9 выполняются.

Величина интеркорреляции между внешними регрессорами $\min\{x_{18}; 1071,164x_{22}\}$ и $\max\{x_2; 0,047x_3\}$ составляет 0,283. А величины интеркорреляций между внутренними переменными x_{18} и x_2, x_{18} и x_3, x_{22} и x_2, x_{22} и x_3 составляют 0,0841, 0,0225, 0,3926 и –0,0013 соответственно. Все эти величины по модулю меньше, чем 0,5487, поэтому условие 6 выполнено.

Поскольку в модели (30) всего два внешних регрессора, коэффициент корреляции между которыми незначим, то из этого автоматически следует выполнение условия 7.

Наблюдаемое значение критерия Фишера для модели (29) составляет 107,933, что гораздо выше критического значения для уровня $\alpha = 0,01$, поэтому модель в целом значима и условие 8 выполнено.

Таким образом, все 9 условий выполняются, следовательно, модель (30) относится к ВИНЛинР. Это означает, что если представить ее в виде кусочно-заданной функции, то в ней, в отличие от кусочно-заданного представления регрессии (29), можно корректно интерпретировать отдельное влияние абсолютно любой объясняющей переменной на результативный признак *y*. В модели (29) этому препятствует мультиколлинеарность, из-за которой различать отдельное влияние объясняющих переменных на *y* затруднительно.

Заключение

Сформулировано строгое определение ВИНЛинР, содержащее девять условий. Среди них ограничения на количество вхождений объясняющих переменных в модель, на соответствие знаков оценок содержательному смыслу задачи, на значимость коэффициентов по *t*-критерию Стьюдента и величинам абсолютных вкладов переменных в общую детерминацию, на мультиколлинеарность, на адекватность регрессии по коэффициенту детерминации и критерию Фишера. Выполнение всех этих девяти условий означает, что в поставленной в соответствие регрессии кусочно-заданной функции можно корректно интерпретировать отдельное влияние любой объясняющей переменной на объясняемую переменную *y*. В ранее предложенную задачу ЧБЛП для выбора оптимальной структурной спецификации ВИНЛинР интегрированы линейные ограничения для контроля мультиколлинеарности как между внешними регрессорами, так и между внутренними переменными из разных регрессоров. Разработанный математический аппарат реализован в программе ВИнтер-2. С помощью программы перестроена полученная ранее модель железнодорожных перевозок в Иркутской области. При этом показано, что все условия ВИНЛинР для нее выполнены. Описанный в статье математический аппарат может успешно применяться в машинном обучении для поиска в статистических данных новых и хорошо интерпретируемых закономерностей.

Список источников

1. Joshi A., Raman B., Mohan C.K., Cenkeramaddi L.R. Application of a new machine learning model to improve earthquake ground motion predictions // Natural Hazards. 2024. V. 120 (1). P. 729–753. doi: 10.1007/s11069-023-06230-4
2. Zhou Y., Liu Y., Wang D., Liu X., Wang Y. A review on global solar radiation prediction with machine learning models in a comprehensive perspective // Energy Conversion and Management. 2021. V. 235. Art. 113960. doi: 10.1016/j.enconman.2021.113960
3. Sonkavde G., Dharrao D.S., Bongale A.M., Deokate S.T., Doreswamy D., Bhat S.K. Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications // International Journal of Financial Studies. 2023. V. 11 (3). Art. 94. doi: 10.3390/ijfs11030094
4. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning // arXiv preprint arXiv:1702.08608. 2017. doi: 10.48550/arXiv.1702.08608
5. Molnar C. Interpretable machine learning. URL: https://originalstatic.aminer.cn/misc/pdf/Molnar-interpretable-machine-learning_compressed.pdf (accessed: 10.06.2024).
6. Chen Z., Xiao F., Guo F., Yan J. Interpretable machine learning for building energy management: A state-of-the-art review // Advances in Applied Energy. 2023. V. 9. Art. 100123. doi: 10.1016/j.adapen.2023.100123
7. Esterhuizen J.A., Goldsmith B.R., Linic S. Interpretable machine learning for knowledge generation in heterogeneous catalysis // Nature catalysis. 2022. V. 5 (3). P. 175–184. doi: 10.1038/s41929-022-00744-z
8. Maulud D., Abdulazeez A.M. A review on linear regression comprehensive in machine learning // Journal of Applied Science and Technology Trends. 2020. V. 1 (2). P. 140–147. doi: 10.38094/jastt1457
9. Schober P., Vetter T.R. Logistic regression in medical research // Anesthesia & Analgesia. 2021. V. 132 (2). P. 365–366. doi: 10.1213/ANE.0000000000005247
10. Mejia A.F., Yue Y.R., Bolin D., Lindgren F., Lindquist M.A. A Bayesian general linear modeling approach to cortical surface fMRI data analysis // Journal of the American Statistical Association. 2020. V. 115. P. 501–520. doi: 10.1080/01621459.2019.1611582
11. Fasiolo M., Nedellec R., Goude Y., Wood S.N. Scalable visualization methods for modern generalized additive models // Journal of computational and Graphical Statistics. 2020. V. 29 (1). P. 78–86. doi: 10.1080/10618600.2019.1629942
12. Базилевский М.П. Построение вполне интерпретируемых линейных регрессионных моделей с помощью метода последовательного повышения абсолютных вкладов переменных в общую детерминацию // Вестник Воронежского государственного университета. Сер. Системный анализ и информационные технологии. 2022. № 2. С. 5–16. doi: 10.17308/sait/1995-5499/2022/2/5-16

13. Montgomery D.C., Peck E.A., Vining G.G. Introduction to linear regression analysis. John Wiley & Sons, 2021.
14. Miller A. Subset selection in regression. New York : CRC Press, 2002.
15. Kim S.H., Boukouvala F. Machine learning-based surrogate modeling for data-driven optimization: a comparison of subset selection for regression techniques // Optimization Letters. 2020. V. 14 (4). P. 989–1010. doi: 10.1007/s11590-019-01428-7
16. Chan J.Y.L., Leow S.M.H., Bea K.T., Cheng W.K., Phoong S.W., Hong Z.W., Chen Y.L. Mitigating the multicollinearity problem and its machine learning approach: a review // Mathematics. 2022. Vol. 10 (8). Art. 1283. doi: 10.3390/math10081283
17. Obite C.P., Olewuezi N.P., Ugwuanyim G.U., Bartholomew D.C. Multicollinearity effect in regression analysis: A feed forward artificial neural network approach // Asian J. Probab. Stat. 2020. V. 6 (1). P. 22–33. doi: 10.9734/AJPAS/2020/v6i130151
18. Базилевский М.П. Сравнительный анализ эффективности методов построения вполне интерпретируемых линейных регрессионных моделей // Моделирование и анализ данных. 2023. Т. 13, № 4. С. 59–83. doi: 10.17759/mda.2023130404
19. Базилевский М.П. Технология построения вполне интерпретируемых квазилинейных регрессионных моделей // Прикладная математика и вопросы управления. 2024. № 1. С. 123–138. doi: 10.15593/2499-9873/2024.1.08
20. Базилевский М.П. Метод построения неэлементарных линейных регрессий на основе аппарата математического программирования // Проблемы управления. 2022. № 4. С. 3–14. doi: 10.25728/pu.2022.4.1
21. Базилевский М.П. Построение вполне интерпретируемых неэлементарных линейных регрессионных моделей // Вестник Югорского государственного университета. 2022. № 4 (67). С. 105–114. doi: 10.18822/byusu202204105-114

References

1. Joshi, A., Raman, B., Mohan, C.K. & Cenkeramaddi, L.R. (2024) Application of a new machine learning model to improve earthquake ground motion predictions. *Natural Hazards*. 120(1). pp. 729–753. DOI: 10.1007/s11069-023-06230-4
2. Zhou, Y., Liu, Y., Wang, D., Liu, X. & Wang, Y. (2021) A review on global solar radiation prediction with machine learning models in a comprehensive perspective. *Energy Conversion and Management*. 235. p. 113960. DOI: 10.1016/j.enconman.2021.113960
3. Sonkavde, G., Dharrao, D.S., Bongale, A.M., Deokate, S.T., Doreswamy, D. & Bhat, S.K. (2023) Forecasting stock market prices using machine learning and deep learning models: A systematic review, performance analysis and discussion of implications. *International Journal of Financial Studies*. 11(3). p. 94. DOI: 10.3390/ijfs11030094
4. Doshi-Velez, F. & Kim, B. (2017) Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*. DOI: 10.48550/arXiv.1702.08608
5. Molnar, C. (2020) *Interpretable machine learning*. [Online] Available from: https://originalstatic.aminer.cn/misc/pdf/Molnar-interpretable-machine-learning_compressed.pdf (Accessed: 10th June 2024).
6. Chen, Z., Xiao, F., Guo, F. & Yan, J. (2023) Interpretable machine learning for building energy management: A state-of-the-art review. *Advances in Applied Energy*. 9. p. 100–123. DOI: 10.1016/j.adapen.2023.100123
7. Esterhuizen, J.A., Goldsmith, B.R. & Linic, S. (2022) Interpretable machine learning for knowledge generation in heterogeneous catalysis. *Nature catalysis*. 5(3). pp. 175–184. DOI: 10.1038/s41929-022-00744-z
8. Maulud, D. & Abdulazeez, A.M. (2020) A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*. 1(2). pp. 140–147. DOI: 10.38094/jastt1457
9. Schober, P. & Vetter, T.R. (2021) Logistic regression in medical research. *Anesthesia & Analgesia*. 132(2). pp. 365–366. DOI: 10.1213/ANE.0000000000005247
10. Mejia, A.F., Yue, Y.R., Bolin, D., Lindgren, F. & Lindquist, M.A. (2020) A Bayesian general linear modeling approach to cortical surface fMRI data analysis. *Journal of the American Statistical Association*. 115. pp. 501–520. DOI: 10.1080/01621459.2019.1611582
11. Fasiolo, M., Nedellec, R., Goude, Y. & Wood, S.N. (2020) Scalable visualization methods for modern generalized additive models. *Journal of Computational and Graphical Statistics*. 29(1). pp. 78–86. DOI: 10.1080/10618600.2019.1629942
12. Bazilevskiy, M.P. (2022) Postroenie vpolne interpretiruemykh lineynykh regressionnykh modeley s pomoshch'yu metoda posledovatel'nogo povysheniya absolyutnykh vkladov peremennykh v obshchuyu determinatsiyu [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. Ser. Sistemnyy analiz i informatsionnye tekhnologii – Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies*. 2. pp. 5–16. DOI: 10.17308/sait/1995-5499/2022/2/5-16
13. Montgomery, D.C., Peck, E.A. & Vining, G.G. (2021) *Introduction to Linear Regression Analysis*. John Wiley & Sons.
14. Miller, A. (2002) *Subset Selection in Regression*. Chapman and hall/CRC.
15. Kim, S.H. & Boukouvala, F. (2020) Machine learning-based surrogate modeling for data-driven optimization: a comparison of subset selection for regression techniques. *Optimization Letters*. 14(4). pp. 989–1010. DOI: 10.1007/s11590-019-01428-7
16. Chan, J.Y.L., Leow, S.M.H., Bea, K.T., Cheng, W.K., Phoong, S.W., Hong, Z.W. & Chen, Y.L. (2022) Mitigating the multicollinearity problem and its machine learning approach: a review. *Mathematics*. 10(8). p. 1283. DOI: 10.3390/math10081283
17. Obite, C.P., Olewuezi, N.P., Ugwuanyim, G.U. & Bartholomew, D.C. (2020) Multicollinearity effect in regression analysis: A feed forward artificial neural network approach. *Asian Journal of Probability and Statistics*. 6(1). pp. 22–33. DOI: 10.9734/AJPAS/2020/v6i130151
18. Bazilevskiy, M.P. (2023) Sravnitel'nyy analiz effektivnosti metodov postroeniya vpolne interpretiruemykh lineynykh regressionnykh modeley [Comparative analysis of the effectiveness of methods for constructing quite interpretable linear regression models]. *Modelirovanie i analiz dannykh – Modelling and Data Analysis*. 13(4). pp. 59–83. DOI: 10.17759/mda.2023130404

19. Bazilevskiy, M.P. (2024) Tekhnologiya postroeniya vpolne interpretiruemykh kvazilineynykh regressionnykh modeley [Technology for constructing quite interpretable quasilinear regression models]. *Prikladnaya matematika i voprosy upravleniya – Applied Mathematics and Control Sciences*. 1. pp. 123–138. DOI: 10.15593/2499-9873/2024.1.08
20. Bazilevskiy, M.P. (2022) Metod postroeniya neelementarnykh lineynykh regressiy na osnove apparata matematicheskogo programmirovaniya [A method for constructing nonelementary linear regressions based on mathematical programming]. *Problemy upravleniya*. 4. pp. 3–14. DOI: 10.25728/pu.2022.4.1
21. Bazilevskiy, M.P. (2022) Postroenie vpolne interpretiruemykh neelementarnykh lineynykh regressionnykh modeley [Construction of quite interpretable non-elementary linear regression models]. *Vestnik Yugorskogo gosudarstvennogo universiteta*. no. 4 (67). pp. 105–114. DOI: 10.18822/byusu202204105-114

***Информация об авторе*:**
Базилевский Михаил Павлович – доцент, кандидат технических наук, доцент кафедры «Математика» Иркутского государственного университета путей сообщения (Иркутск, Россия). E-mail: mik2178@yandex.ru

Автор заявляет об отсутствии конфликта интересов.

***Information about the author*:**
Bazilevskiy Mikhail P. (Candidate of Technical Sciences, Associate Professor of the Mathematics Department, Irkutsk State Transport University, Irkutsk, Russian Federation). E-mail: mik2178@yandex.ru

The author declares no conflicts of interests.

Поступила в редакцию 17.07.2024; принята к публикации 02.12.2024

Received 17.07.2024; accepted for publication 02.12.2024