Базилевский, М. П. Технология построения вполне интерпретируемых квазилинейных регрессионных моделей / М. П. Базилевский // Прикладная математика и вопросы управления. – 2024. – № 1. – С. 123–138. – DOI 10.15593/2499-9873/2024.1.08

Библиографическое описание согласно ГОСТ Р 7.0.100–2018
Базилевский, М. П. Технология построения вполне интерпретируемых квазилинейных регрессионных моделей / М. П. Базилевский. – Текст : непосредственный // Прикладная математика и вопросы управления / Applied Mathematics and Control Sciences. – 2024. – № 1. – С. 123–138. – DOI 10.15593/2499-9873/2024.1.08

пермский политех

ПРИКЛАДНАЯ МАТЕМАТИКА И ВОПРОСЫ УПРАВЛЕНИЯ
№ 1, 2024
https://ered.pstu.ru/index.php/amcs

Научная статья
DOI: 10.15593/2499-9873/2024.1.08
УДК 519.862.6

Технология построения вполне интерпретируемых квазилинейных регрессионных моделей

М.П. Базилевский

Иркутский государственный университет путей сообщения, Иркутск, Российская Федерация

О СТАТЬЕ

Получена: 28 января 2024
Одобрена: 22 апреля 2024
Принята к публикации:
27 апреля 2024

Финансирование
Исследование не имело спонсорской поддержки.
Конфликт интересов
Автор заявляет об отсутствии конфликта интересов.
Вклад автора
100 %.

Ключевые слова:
машинное обучение, большие данные, квазилинейная регрессия, интерпретируемость, отбор информативных регрессоров, критерий нелинейности, мультиколлинеарность, математическое программирование.

АННОТАЦИЯ

Рассматривается актуальная проблема поиска закономерностей в больших объемах статистических данных. Инструментом анализа данных выступает регрессионный анализ. При построении регрессионных моделей исследователи зачастую стремятся только к их высокому качеству аппроксимации. Но, как отмечено в современных научных работах, одной такой метрики недостаточно. Поэтому сегодня активно развивается интерпретируемое машинное обучение. Ранее автором было предложено определение вполне интерпретируемой линейной регрессии, а задача ее построения была формализована в виде задачи частично-булевого линейного программирования. Исследования выявили высокую эффективность разработанного математического аппарата при решении задач обработки больших данных. Поэтому было принято решение расширить предложенную технологию для построения квазилинейных регрессий. В статье дано определение вполне интерпретируемой квазилинейной регрессии, включающее 6 условий. Разработан алгоритм интерпретации влияния в оцененной квазилинейной регрессии монотонно преобразованных объясняющих переменных на зависимую переменную. Задача построения вполне интерпретируемой квазилинейной регрессии формализована в виде задачи частично-булевого линейного программирования. Показано, как в этой задаче выбирать допустимые границы параметра *M*. Для демонстрации работоспособности предложенного математического аппарата решена задача моделирования прочности бетона на сжатие по данным, содержащим более 1000 наблюдений. Для этого использовалась программа «ВИнтер-2». В построенную модель вошли следующие преобразованные переменные: цементно-водное отношение, шлак доменной печи, пластификатор и возраст бетона. Построенная регрессия оказалась лучше по качеству аппроксимации и проще по структуре существующей модели. Дана интерпретация построенной квазилинейной регрессии. Влияние объясняющих переменных на прочность бетона в ней согласуется как с содержательным смыслом задачи, так и с другими существующими математическими моделями. Предложенная в статье технология построения вполне интерпретируемых квазилинейных регрессий обладает высоким потенциалом для решения задач обработки больших данных в различных предметных областях.

© **Базилевский Михаил Павлович** – кандидат технических наук, доцент, доцент кафедры «Математика», e-mail: mik2178@yandex.ru, ORCID: 0000-0002-3253-5697.

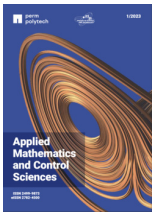
Perm Polytech Style: Bazilevskiy M.P. Technology for constructing quite interpretable quasilinear regression models. *Applied Mathematics and Control Sciences.* 2024, no. 1, pp. 123–138. DOI: 10.15593/2499-9873/2024.1.08

MDPI and ACS Style: Bazilevskiy, M.P. Technology for constructing quite interpretable quasilinear regression models. *Appl. Math. Control Sci.* **2024, 1**, 123–138. https://doi.org/10.15593/2499-9873/2024.1.08

Chicago/Turabian Style: Bazilevskiy, Mikhail P. 2024. "Technology for constructing quite interpretable quasilinear regression models". *Appl. Math. Control Sci.* no. 1: 123–138. https://doi.org/10.15593/2499-9873/2024.1.08

perm polytech

APPLIED MATHEMATICS AND CONTROL SCIENCES
№ 1, 2024
https://ered.pstu.ru/index.php/amcs



Article
DOI: 10.15593/2499-9873/2024.1.08
UDC 519.862.6

Technology for constructing quite interpretable quasilinear regression models

M.P. Bazilevskiy

Irkutsk State Transport University, Irkutsk, Russian Federation

ARTICLE INFO

Received: 28 January 2024
Approved: 22 April 2024
Accepted for publication:
27 April 2024

Funding
This research received no external funding.
Conflicts of Interest
The author declares no conflict of interest.
Author Contributions
100 %.

Keywords:
machine learning, big data, quasilinear regression, interpretability, subset selection, nonlinearity criterion, multicollinearity, mathematical programming.

ABSTRACT

This article is devoted to the current problem of searching for patterns in large volumes of statistical data. The tool for data analysis is regression analysis. When constructing regression models, researchers often strive only for their high quality of approximation. But, as noted in modern scientific works, such a metric alone is not enough. Therefore, interpretable machine learning is actively developing today. Previously, the author proposed a definition of a quite interpretable linear regression, and the problem of its construction was formalized as a mixed integer 0-1 linear programming problem. Research has revealed the high efficiency of the developed mathematical apparatus in solving problems of big data processing. Therefore, it was decided to expand the proposed technology for constructing quasilinear regressions. The article gives a definition of a quite interpretable quasilinear regression, which includes 6 conditions. An algorithm has been developed for interpreting the influence in the estimated quasilinear regression of monotonically transformed explanatory variables on the dependent variable. The problem of constructing a quite interpretable quasilinear regression is formalized as a mixed integer 0-1 linear programming problem. It is shown how to select the acceptable limits of the parameter M in this problem. To demonstrate the performance of the proposed mathematical apparatus, the problem of modeling the compressive strength of concrete using data containing more than 1000 observations was solved. For this purpose, the VInter-2 program was used. The constructed model included the following transformed variables: cement-water ratio, blast furnace slag, plasticizer and concrete age. The constructed regression turned out to be better in terms of the quality of approximation and simpler in the structure of the existing model. An interpretation of the constructed quasilinear regression is given. The influence of explanatory variables on the strength of concrete in it is consistent both with the substantive meaning of the problem and with other existing mathematical models. The technology proposed in the article for constructing quite interpretable quasilinear regressions has high potential for solving problems of big data processing in various subject areas.

© **Mikhail P. Bazilevskiy** – CSc in Technical Sciences, Associate Professor of Department of Mathematics, e-mail: mik2178@yandex.ru, ORCID: 0000-0002-3253-5697.

Введение

В настоящее время во всем мире чрезвычайно актуальны проблемы анализа и эффективной обработки больших объемов статистических данных (Big Data) [1; 2], накопленных в разных сферах человеческой деятельности. К одной из техник Big Data относится регрессионный анализ [3; 4]. С помощью регрессионных моделей успешно решается множество прикладных задач. Например, в [5] с помощью регрессионного анализа определены наиболее значимые факторы риска развития микрососудистых осложнений сахарного диабета второго типа, в [6] – получено значение предельно допустимого состава смесевого топлива для применения в тракторных дизелях. Традиционно много регрессионных моделей строится на основе экономической статистики (см, например, [7]). Такие модели называются эконометрическими.

При всем этом одной из главных проблем, связанных с построением регрессионной модели, считается выбор ее спецификации [8], т.е. общего вида модели, в том числе состава и формы входящих в нее связей. Выбор состава входящих в уравнение регрессии переменных формализуется в виде так называемой задачи отбора наиболее информативных регрессоров (ОИР) [9] на основе некоторого критерия качества. Для решения проблемы ОИР существует множество различных методов, описание большинства из которых можно найти в [10]. Единственным из них, который гарантирует оптимальное решение задачи, считается метод «всех регрессий», реализация которого предполагает перебор всех возможных комбинаций переменных в модели. Но этот метод и самый трудоемкий из всех существующих, поэтому неэффективен при решении задач Big Data при большом числе переменных.

Кроме того, построенную в результате реализации метода «всех регрессий» регрессионную модель бывает затруднительно или вовсе невозможно интерпретировать. Например, знаки ее оценок могут противоречить содержательному смыслу факторов, в ней может присутствовать мультиколлинеарность [11] и т.д. На сегодняшний день проблеме построения интерпретируемых моделей машинного обучения [12; 13] в научной литературе уделяется значительное внимание. Интерпретация построенной модели способствует выявлению и устранению ее «уязвимых» мест, что повышает доверие к модели у экспертов из данной предметной области. К сожалению, точного определения интерпретируемости модели не существует. Успешная попытка дать такое определение для регрессионных моделей была предпринята в работе [14], в которой рассмотрено необходимое и достаточное условие однозначной интерпретируемости регрессии. Однако для построения таких моделей в [14] предложен все тот же малоэффективный метод «всех регрессий».

За последние десятилетия была существенно развита технология решения задач частично-целочисленного программирования, поэтому появился новый более эффективный метод нахождения оптимального решения задачи ОИР. В зарубежной литературе задача ОИР при оценивании линейных регрессий с помощью метода наименьших квадратов (МНК) формализуется в виде задачи частично-булевого квадратичного программирования (ЧБКП) (см, например, [15–17]), число ограничений которой зависит от объема выборки. В работе [18] автору впервые удалось свести такую задачу к задаче частично-булевого линейного программирования (ЧБЛП), число ограничений которой не зависит от объема выборки. В дальнейшем эта задача эволюционировала, в ней появлялись новые ограничения. Например, в [19] была сформулирована задача ЧБЛП, решение которой приводит к построению линейной регрессии с оптимальным числом объясняющих переменных, абсолютные вклады которых в общую детерминацию не меньше, чем число θ. В той же работе [19] было предложено определение вполне интерпретируемой линейной регрессии (ВИЛинР). В исследовании

[20] определение было уточнено за счет ограничений на коэффициенты интеркорреляций. И в той же работе [20] проведены многочисленные вычислительные эксперименты, подтверждающие высокую эффективность предложенного метода для обработки Big Data. Например, удалось эффективно обработать выборку из 515345 наблюдений.

Целью данной работы является обобщение предложенного в работах [18–20] математического аппарата, эффективно справляющегося с обработкой Big Data, для построения вполне интерпретируемых квазилинейных регрессионных моделей.

1. Технология построения вполне интерпретируемых квазилинейных регрессий

Введем в рассмотрение элементарную квазилинейную регрессию (КЛинР) [9]:

$$y_i = \alpha_0 + \sum_{j=1}^{l} \sum_{k=1}^{\text{elem}} \alpha_{jk} \cdot f_k\left(x_{ij}\right) + \varepsilon_i, \quad i = \overline{1,n}, \tag{1}$$

где n – объем выборки; l – число объясняющих переменных; y_i – i-е значение объясняемой переменной; x_{ij} – i-е значение j-й объясняющей переменной; α_0, α_{jk}, $j = \overline{1,l}$, $k = \overline{1,\text{elem}}$ – неизвестные параметры; ε_i – i-я ошибка аппроксимации; $f_k\left(x_j\right)$ – k-е элементарное преобразование j-й переменной, выбранное из набора $\{f_1(x), f_2(x), ..., f_{\text{elem}}(x)\}$. Линейная регрессия является частным случаем КЛинР (1).

Достоинство КЛинР (1) в том, что они линейны по параметрам, поэтому их можно оценивать с помощью МНК. А недостаток заключается в проблематичности их интерпретации. Например, если оценена модель $\tilde{y} = 5 + 17x^2$, то затруднительно каким-либо образом объяснить оценку 17 при переменной x^2. В некоторых источниках, например, в [12], предлагается интерпретировать такую модель следующим образом: если квадрат переменной x увеличится на единицу, то значение переменной y увеличится на 17 единиц. На наш взгляд, такой подход к интерпретации не совсем корректен, поскольку не объясняется прямое влияние именно переменной x на y.

Введем определение вполне интерпретируемой квазилинейной регрессии (ВИКЛинР). Для удобства запишем КЛинР (1) в виде

$$y_i = \alpha_0 + \sum_{j=1}^{l} \sum_{k=1}^{\text{elem}} \alpha_{jk} \cdot w_{ijk} + \varepsilon_i, \quad i = \overline{1,n}, \tag{2}$$

где $w_{ijk} = f_k\left(x_{ij}\right)$, $i = \overline{1,n}$, $j = \overline{1,l}$, $k = \overline{1,\text{elem}}$.

Определение. КЛинР (2), оцененная с помощью МНК, называется вполне интерпретируемой, если:

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

2) знаки всех коэффициентов корреляции $r_{yw_{jk}}$, $j = \overline{1,l}$, $k = \overline{1,\text{elem}}$, удовлетворяют содержательному смыслу решаемой задачи;

3) знаки всех оценок $\tilde{\alpha}_{jk}$ согласуются со знаками соответствующих коэффициентов корреляции $r_{yw_{jk}}$, т.е. $\tilde{\alpha}_{jk} \cdot r_{yw_{jk}} > 0$, $j = \overline{1,l}$, $k = \overline{1,\text{elem}}$;

4) все абсолютные вклады переменных в общую детерминацию удовлетворяют неравенствам $C_{w_{jk}}^{\text{абс}} \geq 0,01$, $j=\overline{1,l}$, $k=\overline{1,\text{elem}}$;

5) все коэффициенты интеркорреляций $r_{w_{j_1k_1}w_{j_2k_2}} \leq 0,4$, $j_1=\overline{1,l-1}$, $k_1=\overline{1,\text{elem}}$, $j_2=\overline{j_1+1,l}$, $k_2=\overline{1,\text{elem}}$;

6) коэффициент детерминации модели $R^2 \geq 0,8$.

Кратко прокомментируем данное определение. В нем условие № 2 означает, что еще до оценивания КЛинР (2) требуется анализировать соответствие знаков коэффициентов корреляции преобразованных объясняющих переменных с *y* содержательному смыслу решаемой задачи. В случае выявления противоречий необходимо либо увеличить объем выборки, либо исключить соответствующую преобразованную переменную из рассмотрения. На этом этапе желательно привлекать экспертов в данной предметной области. Экспертам следует помнить, что элементарные преобразования переменных могу искажать направление влияния объясняющих переменных на *y*.

Стандартизованная регрессия для (2) записывается в виде

$$y_i^{\bullet} = \sum_{j=1}^{l}\sum_{k=1}^{\text{elem}} \beta_{jk}^{\#} \cdot w_{ijk}^{\bullet} + \xi_i, \qquad i=\overline{1,n},$$

где $y_i^{\bullet} = \dfrac{y_i - \overline{y}}{\sigma_y}$, $w_{ijk}^{\bullet} = \dfrac{w_{ijk} - \overline{w_{jk}}}{\sigma_{w_{jk}}}$, $i=\overline{1,n}$, $j=\overline{1,l}$, $k=\overline{1,\text{elem}}$; $\beta_{jk}^{\#}$, $j=\overline{1,l}$, $k=\overline{1,\text{elem}}$ – неизвестные стандартизованные коэффициенты; ξ_i, $i=\overline{1,n}$ – новые ошибки аппроксимации. Если выполняется условие № 3, то, как отмечено в [19], становятся справедливы формулы для абсолютных вкладов переменных в общую детерминацию R^2:

$$C_{w_{jk}}^{\text{абс}} = r_{yw_{jk}} \cdot \beta_{jk}^{\#}, \qquad j=\overline{1,l},\ k=\overline{1,\text{elem}}.$$

По этим критериям можно делать выводы о степени влияния преобразованных переменных на *y*. Условие № 4 означает, что каждая такая переменная должна вносить вклад в детерминацию не менее 0,01.

Условие № 5 означает ограничение на эффект мультиколлинеарности. Уровень 0,4, при котором модель считается приближенно интерпретируемой, взят из [14]. Также эту границу можно вычислить, например, по формуле $\dfrac{t_{\text{крит}}(\alpha, n-1)}{\sqrt{n-2+t_{\text{крит}}^2(\alpha, n-1)}}$, где α – заданный уровень значимости. В этом случае все коэффициенты интеркорреляций будут незначимы по *t*-критерию Стьюдента для уровня α.

Если коэффициент детерминации R^2 регрессии меньше 0,8, то вряд ли можно отнести такую модель ко вполне интерпретируемой, поэтому введено условие № 6.

Рассмотрим решение проблемы интерпретации квазилинейных регрессий. Предположим, что оцененная ВИКЛинР имеет вид:

$$\tilde{y} = \tilde{\alpha}_0 + \sum_{j=1}^{l} \tilde{\alpha}_j f_{\omega_j}(x_j), \tag{3}$$

где $1 \le \omega_j \le \text{elem}$ – элемент вектора $(\omega_1, \omega_2, ..., \omega_l)$, указывающий номер выбранного преобразования для *j*-й объясняющей переменной. Если все функции из набора $\{f_1(x), f_2(x), ..., f_{\text{elem}}(x)\}$ непрерывны и монотонны на отрезках $\left[x^j_{\min}, x^j_{\max}\right]$, где $x^j_{\min} = \min\{x_{1j}, x_{2j}, ..., x_{nj}\}$, $x^j_{\max} = \max\{x_{1j}, x_{2j}, ..., x_{nj}\}$, $j = \overline{1,l}$, то становятся справедливы следующие формулы критериев нелинейности [21] для каждой объясняющей переменной:

$$NC_j = \left| \frac{f_{\omega_j}(x^j_{\max}) + f_{\omega_j}(x^j_{\min})}{f_{\omega_j}(x^j_{\max}) - f_{\omega_j}(x^j_{\min})} - \frac{2\int\limits_{x^j_{\min}}^{x^j_{\max}} f_{\omega_j}(x_j)dx_j}{\left(x^j_{\max} - x^j_{\min}\right)\left(f_{\omega_j}(x^j_{\max}) - f_{\omega_j}(x^j_{\min})\right)} \right|, \; j = \overline{1,l}. \tag{4}$$

Каждый из критериев нелинейности (4) принимает значения от 0 до 1. Если $NC_j = 0$, то преобразование *j*-й объясняющей переменной является линейным. А если $NC_j \to 1$, то преобразование *j*-й объясняющей переменной в значительной степени нелинейно.

Сформулируем алгоритм интерпретации влияния *j*-й объясняющей переменной на *y* в квазилинейной регрессии (3) с монотонными функциями.

1. Вычислить критерий нелинейности NC_j для *j*-й переменной по формуле (4).

2. Если $NC_j \le 0,2$, то вместо $\tilde{\alpha}_j$ объясняется коэффициент $\tilde{\alpha}^{\bullet}_j = \tilde{\alpha}_j \dfrac{f_{\omega_j}(x^j_{\max}) - f_{\omega_j}(x^j_{\min})}{x^j_{\max} - x^j_{\min}}$.

3. Если $NC_j > 0,2$, то сначала по формуле Лагранжа $f_{\omega_j}(x^j_{\max}) - f_{\omega_j}(x^j_{\min}) = f_{\omega_j}{}'(c) \cdot (x^j_{\max} - x^j_{\min})$ находится точка c, а потом объясняются коэффициенты $\alpha^{\bullet\bullet}_j = \tilde{\alpha}_j \dfrac{f_{\omega_j}(c) - f_{\omega_j}(x^j_{\min})}{c - x^j_{\min}}$ и $\tilde{\alpha}^{\bullet\bullet\bullet}_j = \tilde{\alpha}_j \dfrac{f_{\omega_j}(x^j_{\max}) - f_{\omega_j}(c)}{x^j_{\max} - c}$ на отрезках $[x^j_{\min}, c]$ и $\left[c, x^j_{\max}\right]$ соответственно.

По аналогии с задачей построения вполне интерпретируемой линейной регрессии, рассмотренной, например, в [20], сформулируем следующую задачу ЧБЛП построения ВИКЛинР (2):

$$\sum_{j=1}^{l}\sum_{k=1}^{\text{elem}} r_{yw_{jk}} \cdot \beta^{\#}_{jk} \to \max, \tag{5}$$

$$-(1-\delta_{jk}) \cdot M \le \sum_{s_1=1}^{l}\sum_{s_2=1}^{\text{elem}} r_{w_{s_1 s_2} w_{jk}} \beta^{\#}_{jk} - r_{yw_{jk}} \le (1-\delta_{jk}) \cdot M, \; j = \overline{1,l}, \; k = \overline{1,\text{elem}}, \tag{6}$$

$$0 \le \beta^{\#}_{jk} \le \delta_{jk} \cdot M, \; (j,k) \in \left\{(s_1, s_2) \mid r_{yw_{s_1 s_2}} > 0\right\}, \tag{7}$$

$$-\delta_{jk} \cdot M \le \beta^{\#}_{jk} \le 0, \; (j,k) \in \left\{(s_1, s_2) \mid r_{yw_{s_1 s_2}} < 0\right\}, \tag{8}$$

$$\delta_{jk} \in \{0,1\}, \; j = \overline{1,l}, \; k = \overline{1,\text{elem}}, \tag{9}$$

$$r_{yw_{jk}} \cdot \beta^{\#}_{jk} \ge \theta \cdot \delta_{jk}, \; j = \overline{1,l}, \; k = \overline{1,\text{elem}}, \tag{10}$$

$$\left|r_{w_{j_1k_1}w_{j_2k_2}}\right|\cdot\left(\delta_{j_1k_1}+\delta_{j_2k_2}-1\right)\le r,\ j_1=\overline{1,l-1},\ k_1=\overline{1,\text{elem}},\ j_2=\overline{j_1+1,l},\ k_2=\overline{1,\text{elem}},\qquad(11)$$

$$\sum_{k=1}^{\text{elem}}\delta_{jk}\le 1,\ \ j=\overline{1,l},\qquad(12)$$

где символами r_{ab} обозначены коэффициенты корреляции между переменными a и b; M – большое положительное число; δ_{jk} – бинарная переменная, принимающая значение 1, если j-я объясняющая переменная с k-м преобразованием входит в регрессию, и 0, если не входит; $0<r\le 1$ – верхняя граница интеркорреляций; $0\le\theta<1$ – нижняя граница абсолютных вкладов переменных. Точность коэффициентов корреляций в этой задаче должна быть не менее 12 знаков.

В задаче (5)–(12) ограничения (12) обеспечивают выполнение условия № 1 в предложенном выше определении, ограничения (7), (8) – условия № 3, ограничения (10) при $\theta=0,01$ – условия № 4, ограничения (11) при $r=0,4$ – условия № 5. Условие № 2 проверяется до решения этой задачи, а условие № 6 – после решения. Ограничения (6) предназначены для включения/исключения уравнений в линейную систему, с помощью которой находятся МНК-оценки. Целевая функция (5) означает выбор регрессии с максимальным значением коэффициент детерминации R^2. Таким образом, решение задачи ЧБЛП (5)–(12) приводит к построению ВИКЛинР с оптимальным по критерию R^2 количеством регрессоров, в которой $\beta^{\#}_{jk}\cdot r_{yw_{jk}}>0$, вклады $C^{\text{абс}}_{w_{jk}}\ge\theta$, а интеркорреляции $\left|r_{w_{j_1k_1}w_{j_2k_2}}\right|\le r$.

Если исследователь желает контролировать в задаче (5)–(12) коэффициенты вздутия дисперсии VIF, то ее необходимо дополнить линейными ограничениями из работы [22], а если контролировать значимость оценок по t-критерию Стьюдента, то ограничениями из работы [23].

Выбирать допустимые границы параметра M в задаче (5)–(12) можно по аналогии с процедурой выбора таких границ в задаче построения вполне интерпретируемой линейной регрессии [19]. Для этого ограничения (6)–(8) следует заменить следующими выражениями:

$$(1-\delta_{jk})\cdot M^{-}_{u_{jk}}\le\sum_{s_1=1}^{l}\sum_{s_2=1}^{\text{elem}}r_{w_{s_1s_2}w_{jk}}\beta^{\#}_{jk}-r_{yw_{jk}}\le(1-\delta_{jk})\cdot M^{+}_{u_{jk}},\ j=\overline{1,l},\ k=\overline{1,\text{elem}},\qquad(13)$$

$$0\le\beta^{\#}_{jk}\le\delta_{jk}\cdot M_{\beta^{\#}_{jk}},\ (j,k)\in\left\{(s_1,s_2)\,|\,r_{yw_{s_1s_2}}>0\right\},\qquad(14)$$

$$\delta_{jk}\cdot M_{\beta^{\#}_{jk}}\le\beta^{\#}_{jk}\le 0,\ (j,k)\in\left\{(s_1,s_2)\,|\,r_{yw_{s_1s_2}}<0\right\},\qquad(15)$$

где $M_{\beta^{\#}_{jk}}=R^2_{\max}/r_{yw_{jk}}$, $j=\overline{1,l}$, $k=\overline{1,\text{elem}}$, $R^2_{\max}$ – значение коэффициента детерминации регрессии, построенной со всеми $l\cdot\text{elem}$ регрессорами. Для нахождения $M^{-}_{u_{jk}}$ нужно решить серию из $l\cdot\text{elem}$ задач линейного программирования с целевыми функциями $u_{jk}\to\min$ и линейными ограничениями

$$0\le\beta^{\#}_{jk}\le M_{\beta^{\#}_{jk}},\ (j,k)\in\left\{(s_1,s_2)\,|\,r_{yw_{s_1s_2}}>0\right\},$$
$$M_{\beta^{\#}_{jk}}\le\beta^{\#}_{jk}\le 0,\ (j,k)\in\left\{(s_1,s_2)\,|\,r_{yw_{s_1s_2}}<0\right\},$$

$$\sum_{s_1=1}^{l}\sum_{s_2=1}^{\text{elem}} r_{w_{s_1 s_2} w_{jk}} \beta_{jk}^{\#} - r_{y w_{jk}} = u_{jk},\ j=\overline{1,l},\ k=\overline{1,\text{elem}}.$$

Значения параметров $M_{u_{jk}}^{-}$ берутся как значения целевых функций этих задач в точках оптимума.

Для нахождения чисел $M_{u_{jk}}^{+}$ нужно решить серию из $l \cdot \text{elem}$ задач линейного программирования с теми же линейными ограничениями, но с целевыми функциями $u_{jk} \to \max$.

Оптимальные решения задач (5)–(12) и (5), (9)–(15) не отличаются.

2. Моделирование прочности бетона на сжатие

Для демонстрации работоспособности предложенного математического аппарата решалась задача обработки данных из строительной области. Как известно, одно из лидирующих мест среди строительных материалов во всем мире занимает бетон. Ключевой проблемой при его изготовлении считается подбор состава компонент смеси и их пропорций так, чтобы обеспечить максимальную его долговечность. Для решения этой проблемы, как и многих других, могут применяться методы математического моделирования.

Важный показатель, которым характеризуется бетон, – прочность на сжатие. Математическому моделированию прочности бетона посвящено множество научных работ. Из них хотелось бы выделить статью [24], в которой приведен весьма солидный обзор существующих математических моделей. Их анализ позволил сделать вывод, что во всех моделях фигурирует так называемое водоцементное отношение по объему, т.е. отношение массы воды к массе цемента. Так, например, более 100 лет назад Д. Абрамс сформулировал правило [25]:

$$y = \frac{K_1}{K_2^{\frac{\text{В}}{\text{Ц}}}},$$

где y – прочность бетона, K_1, K_2 – некоторые константы, $\frac{\text{В}}{\text{Ц}}$ – водоцементное отношение. Как видно, чем выше водоцементное отношение, тем ниже прочность на сжатие.

А, например, в работе М. Боломея [26] предложена следующая формула прочности бетона:

$$y = K_3\left(\frac{\text{Ц}}{\text{В}} - K_4\right),$$

где K_3, K_4 – некоторые константы, $\frac{\text{Ц}}{\text{В}}$ – цементно-водное отношение. Как видно, оно влияет на y в противоположном направлении.

Однако на прочность бетона может влиять не только водоцементное отношение, но и качество заполнителей, добавление пластификатора, метод замеса, температура воздуха и др. В связи с этим вызывает интерес статистическое моделирование прочности бетона на сжатие.

Построению статистических моделей для прогнозирования прочности бетона посвящена работа [24]. В ней использованы статистические данные из хранилища [27] по следующим переменным:

y – прочность бетона на сжатие (МПа);

x_1 – цемент (кг/м^3 смеси);

x_2 – шлак доменной печи (кг/м^3 смеси);

x_3 – зола (кг/м^3 смеси);

x_4 – вода (кг/м^3 смеси);

x_5 – пластификатор (кг/м^3 смеси);

x_6 – щебень (кг/м^3 смеси);

x_7 – песок (кг/м^3 смеси);

x_8 – возраст бетона (дни).

Как отмечено в [24], выборка из 920 наблюдений была разделена на обучающую, объемом 700, и контрольную, объемом 220. На сегодняшний день на сайте [27] объем этой выборки составляет уже 1030 наблюдений.

В [24] авторами предложены следующие спецификации нелинейных регрессионных моделей:

$$y = \alpha_0 + \alpha_1 \frac{x_1}{x_4} + \alpha_2 \frac{x_6}{x_1} + \alpha_3 \frac{x_7}{x_1} + \alpha_4 \frac{x_2}{x_1} + \alpha_5 \lg(x_8) + \varepsilon, \tag{16}$$

$$y = \alpha_0 + \alpha_1 \frac{x_1}{x_4} + \alpha_2 \frac{x_6}{x_1} + \alpha_3 \frac{x_7}{x_1} + \alpha_4 \frac{x_2}{x_1} + \alpha_5 \frac{x_3}{x_1} + \alpha_6 \lg(x_8) + \varepsilon, \tag{17}$$

$$y = \alpha_0 + \alpha_1 \frac{x_1}{x_4} + \alpha_2 \frac{x_6}{x_1} + \alpha_3 \frac{x_7}{x_1} + \alpha_4 \frac{x_2}{x_1} + \alpha_5 \frac{x_3}{x_1} + \alpha_6 \frac{x_5}{x_1} + \alpha_7 \lg(x_8) + \varepsilon. \tag{18}$$

Коэффициенты детерминации R^2, оцененные с помощью МНК регрессий (16)–(18), составили 0,721669, 0,772118 и 0, 778409 соответственно. С использованием эконометрического пакета Gretl модели (16)–(18), построенные по выборке объема 700, были переоценены по новой выборке объема 1030. Новые значения R^2 составили 0,713829, 0,754206 и 0,762094. Как видно, за счет новой информации качество аппроксимации моделей несколько ухудшилось. Лучшая из этих трех регрессий, для которой $R^2 = 0,762094$, может быть использована для прогнозирования. Но ее никак нельзя считать вполне интерпретируемой, поскольку она не удовлетворяет ни одному из шести условий, рассмотренных выше. Так, во-первых, некоторые преобразованные переменные входят в наилучшую модель более одного раза, например, x_1; во-вторых, не проведен анализ согласованности знаков преобразованных переменных, например, $\frac{x_5}{x_1}$, содержательному смыслу задачи; в-третьих, знаки некоторых оценок не согласуются со знаками соответствующих коэффициентов корреляции, например, оценка при переменной $\frac{x_3}{x_1}$ составляет 7,83, а коэффициент корреляции между $\frac{x_3}{x_1}$ и y равен -0,18. В-четвертых, не работают формулы для абсолютных вкладов переменных в общую детерминацию; в-пятых, интеркорреляции между некоторыми преобразованными переменными близки к единице по абсолютной величине, например, между $\frac{x_6}{x_1}$ и $\frac{x_7}{x_1}$ такой коэффициент равен 0,941, что свидетельствует о наличии мультиколлинеарности; в-шестых, коэффициент детерминации модели оказался ниже 0,8.

Была поставлена цель по данным, расположенным на сайте [27], построить ВИКЛинР прочности бетона на основе решения задачи ЧБЛП (5), (9)–(15). Для этого была использована программа «ВИнтер-2», подробное описание которой можно найти в [28]. «ВИнтер-2» позволяет в зависимости от выбранных пользователем начальных параметров автоматически формулировать для решателя LPSolve задачи ЧБЛП для построения, в частности, ВИКЛинР.

Поскольку цементно-водное отношение фигурирует в большинстве известных математических моделей, было принято решение использовать одну переменную $\frac{x_1}{x_4}$ вместо двух переменных x_1 и x_4.

Предварительно был проведен анализ соответствия знаков коэффициентов корреляции объясняющих переменных с y содержательному смыслу задачи. Оказалось, что $r_{y,\frac{x_1}{x_4}} = 0,559$, $r_{yx_2} = 0,135$, $r_{yx_3} = -0,106$, $r_{yx_5} = 0,366$, $r_{yx_6} = -0,165$, $r_{yx_7} = -0,167$, $r_{yx_8} = 0,329$. Экспертами было принято решение о согласованности знаков абсолютно всех этих коэффициентов смыслу задачу.

Поскольку переменные x_2, x_3, x_5, x_8 содержат нулевые значения, что не позволяет использовать такие элементарные преобразования, как $\ln x$, x^{-1} и др., было решено увеличить их на 1.

Для формирования задачи ЧБЛП (5), (9)–(15) для решателя LPSolve в «ВИнтер-2» были выбраны следующие начальные параметры:

1) объясняющие переменные – $\frac{x_1}{x_4}$, x_2+1, x_3+1, x_5+1, x_6, x_7, x_8+1;

2) элементарные преобразования – x^{-2}, $x^{-1,5}$, x^{-1}, $x^{-0,5}$, $x^{0,5}$, x, $x^{1,5}$, x^2, $\ln(x)$;

3) точность вещественных чисел – 12 знаков;

4) наименьший абсолютный вклад $\theta = 0,01$;

5) наибольшая интеркорреляция $r = 0,3$.

Запустив «ВИнтер-2», сначала автоматически были сформированы все возможные комбинации преобразованных переменных, общее число которых составило 63. Затем «ВИнтер-2» сама проверила согласованность знаков коэффициентов корреляции преобразованных переменных с y содержательному смыслу задачи. В итоге все 63 переменных проверку прошли. После чего была автоматически сформирована задача ЧБЛП (5), (9)–(15). Значения больших чисел M в ограничениях (13)–(15) программа также осуществила самостоятельно, для чего была решена необходимая серия задач линейного программирования.

Далее сформированная задача ЧБЛП вручную была перенесена в решатель LPSolve, в котором была построена следующая КЛинР:

$$\tilde{y} = -45,716 + \underset{(47,17)}{\overset{(0,366)}{43,613}}\sqrt{\frac{x_1}{x_4}} + \underset{(23,83)}{\overset{(0,044)}{0,064}}(x_2+1) - \underset{(-19,86)}{\overset{(0,077)}{9,319}}\frac{1}{(x_5+1)^2} + \underset{(44,85)}{\overset{(0,331)}{8,393}}\ln(x_8+1). \qquad (19)$$

В уравнении (19) в скобках под коэффициентами указаны наблюдаемые значения t-критерия Стьюдента, а над ними – абсолютные вклады переменных в общую детерминацию.

Для КЛинР (19) $R^2 = 0,818199$, т.е. по этому критерию качество ее аппроксимации выше, чем у оцененной модели (18) на 0,0561. При этом спецификация регрессии (19), у которой 5 неизвестных параметров, проще спецификации (18) с восьмью параметрами.

Модель (19) удовлетворяет всем шести вышеперечисленным требованиям, поэтому ее справедливо можно считать ВИКЛинР. Дополнительно к этому все ее оценки значимы по *t*-критерию Стьюдента для уровня значимости $\alpha = 0,01$, а коэффициенты вздутия дисперсии VIF, равные 1,094, 1,085, 1,007 и 1,004 соответственно, говорят об отсутствии мультиколлинеарности.

Для интерпретации ВИКЛинР (19) сначала были найдены значения критериев нелинейности преобразованных переменных по формулам (4):

$$NC_{\sqrt{\frac{x_1}{x_4}}} = 0,1509,\ NC_{x_2+1} = 0,\ NC_{\frac{1}{(x_5+1)^2}} = 0,9415,\ NC_{\ln(x_8+1)} = 0,6665.$$

Затем проводилась интерпретация влияния каждой объясняющей переменной на *y* на основе предложенного выше алгоритма. Для наглядности на рисунке сплошными линиями выделены входящие в уравнение (19) нелинейные функции, а пунктиром – заменяющие их при интерпретации прямые и ломаные.

1. Поскольку $NC_{\sqrt{\frac{x_1}{x_4}}} < 0,2$, то разница между кривой $f_1\left(\frac{x_1}{x_4}\right) = 46,613\sqrt{\frac{x_1}{x_4}}$ и прямой линией $g_2\left(\frac{x_1}{x_4}\right) = 23,093 + 16,368\frac{x_1}{x_4}$ несущественна (рисунок). Тогда справедлива следующая интерпретация: с увеличением цементно-водного отношения x_1 / x_4 на одну единицу (при неизменных значениях остальных переменных) прочность бетона на сжатие *y* увеличивается примерно на 16,368 МПа.

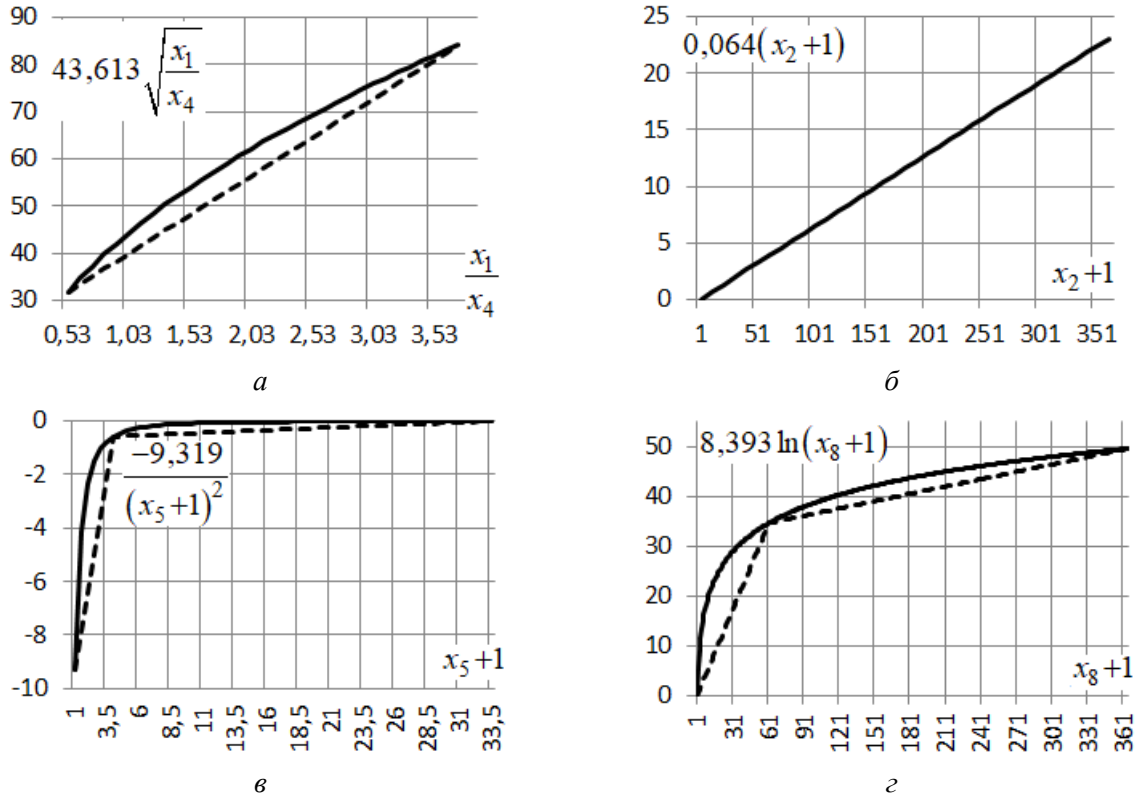


Рис. Прямые и ломаные, заменяющие нелинейные функции (авторские результаты)

2. Так как $NC_{x_2+1}=0$, т.е. графики кривой и заменяющей ее прямой совпадают (рисунок, *б*), то можно сделать следующий вывод: с увеличением шлака доменной печи x_2 на 1 кг/м^3 смеси (при неизменных значениях остальных переменных) прочность бетона на сжатие *y* увеличивается в среднем на 0,064 МПа. Полученный результат о влиянии шлака доменной печи на прочность бетона подтверждают, например, исследования [29; 30].

3. Ввиду того, что $NC_{\frac{1}{(x_5+1)^2}}>0,2$, кривая $f_3(x_5+1)=\dfrac{-9,319}{(x_5+1)^2}$ заменена ломаной, у которой уравнение первого звена $g_{3,1}(x_5+1)=-12,224+2,904(x_5+1)$ при $x_5+1\in[1,4]$, а уравнение второго звена $g_{3,2}(x_5+1)=-0,658+0,0196(x_5+1)$ при $x_5+1\in(4,33.2]$ (рисунок, *в*). Тогда можно дать следующую интерпретацию: если $x_5\le 3$, то с увеличением пластификатора x_5 на 1 кг/м^3 смеси (при неизменных значениях остальных переменных) прочность бетона на сжатие *y* увеличивается в среднем на 2,904 МПа, а если $x_5>3$, то на 0,0196 МПа.

4. На том основании, что $NC_{\ln(x_8+1)}>0,2,$ кривая $f_4(x_8+1)=8,393\ln(x_8+1)$ заменена ломаной, у которой уравнение первого звена $g_{4,1}(x_8+1)=-0,57+0,57(x_8+1)$ при $x_8+1\in[1,\ 61,696]$, а уравнение второго звена $g_{4,2}(x_8+1)=31,566+0,0492(x_8+1)$ при $x_8+1\in(61,696,\ 365]$ (рисунок, *г*). Тогда справедлива следующая интерпретация: если $x_8\le 60,696,$ то с увеличением возраста бетона x_8 на один день (при неизменных значениях остальных переменных) прочность бетона на сжатие *y* увеличивается в среднем на 0,57 МПа, а если $x_8>60.696$, то на 0,0492 МПа.

Заключение

1. Достоинство сформулированной задачи ЧБЛП (5)–(12) построения ВИКЛинР состоит в том, что число ограничений в ней не зависит от объема выборки. Так, если бы в данных, по которым моделировалась прочность бетона, было бы не 1030 наблюдений, а, например, более миллиона, то можно предположить, что время построения модели по-прежнему осталось бы в допустимых пределах. Тем самым предложенная в статье технология построения ВИКЛинР обладает потенциалом в области решения задач обработки Big Data.

2. Недостаток задачи ЧБЛП (5)–(12) состоит в том, что не до конца ясно, как выбирать большие числа *M* в ограничениях (6)–(8). Например, при моделировании прочности бетона на компьютере с процессором AMD Ryzen 3 4300U (2,70 GHz) и объемом оперативной памяти 16 ГБ на решение задачи при использовании ограничений (13)–(15) вместо (6)–(8) ушло 611 секунд. Было установлено, что если в этой задаче выбрать в ограничениях (6)–(8) число $M=500$, то время ее решения составит уже 302 секунды, т.е. примерно в 2 раза меньше. Это классическая проблема, решение которой ищут как отечественные, так и зарубежные ученые.

3. Несмотря на то, что структура построенной ВИКЛинР (19) была выбрана автоматически, влияние объясняющих переменных на *y* в ней согласуется как с содержательным смыслом задачи, так и с другими существующими математическими моделями. К сожалению, пока в программе «ВИнтер-2» реализовано только 9 элементарных функций, поэтому за счет использования более широкого набора преобразований качество аппроксимации модели (19) может быть серьезно улучшено. Тем не менее построенную регрессию (19) можно использовать для прогнозирования прочности бетона в зависимости от входящих в нее факторов.

4. С помощью предложенной технологии можно успешно решать задачи выявления скрытых нелинейных зависимостей в больших наборах данных абсолютно в любой предметной области.

Список литературы

1. Big Data technologies: A survey / A. Oussous, F.Z. Benjelloun, A.A. Lahcen, S. Belfkih / Journal of King Saud University - Computer and Information Sciences. – 2018. – Vol. 30, № 4. – P. 431–448. DOI: 10.1016/j.jksuci.2017.06.001

2. A survey on deep learning for big data / Q. Zhang, L.T. Yang, Z. Chen, P. Li / Information Fusion. – 2018. – Vol. 42. – P. 146–157. DOI: 10.1016/j.inffus.2017.10.006

3. Gunst, R.F. Regression analysis and its application: a data-oriented approach / R.F. Gunst, R.L. Mason. – CRC Press, 2018.

4. Montgomery, D.C. Introduction to linear regression analysis / D.C. Montgomery, E.A. Peck, G.G. Vining. – John Wiley & Sons, 2021.

5. Невзорова, Е.В. Многофакторный регрессионный анализ факторов риска развития микрососудистых осложнений сахарного диабета 2 типа / Е.В. Невзорова, А.К. Засядько, О.Н. Загуменнова // Медицина и физическая культура: наука и практика. – 2020. – Т. 2, № 2. – С. 58–67.

6. Бузиков, Ш.В. Оптимизация состава смесевого топлива для применения в тракторных дизелях / Ш.В. Бузиков, С.А. Плотников, И.С. Козлов // Труды НАМИ. – 2021. – № 1. – С. 16–24.

7. Chang, J.J. Temperature and GDP: A review of climate econometrics analysis / J.J. Chang, Z. Mi, Y.M. Wei // Structural Change and Economic Dynamics. – 2023. – Vol. 66. – P. 383–392. DOI: 10.1016/j.strueco.2023.05.009

8. Айвазян, С.А. Методы эконометрики / С.А. Айвазян. – М.: Магистр: ИНФРА-М, 2010. – 512 с.

9. Носков, С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных / С.И. Носков. – Иркутск: РИЦ ГП «Облинформпечать», 1996. – 320 с.

10. Стрижов, В.В. Методы выбора регрессионных моделей / В.В. Стрижов, Е.А. Крымова. – М.: ВЦ РАН, 2010. – 60 с.

11. Shrestha, N. Detecting multicollinearity in regression analysis / N. Shrestha // American Journal of Applied Mathematics and Statistics. – 2020. – Vol. 8, № 2. – P. 39–42. DOI: 10.12691/ajams-8-2-1

12. Molnar, C. Interpretable machine learning / C. Molnar. – Lulu. Com, 2020.

13. Doshi-Velez, F. Towards a rigorous science of interpretable machine learning / F. Doshi-Velez, B. Kim // arXiv preprint arXiv:1702.08608. – 2017.

14. Горбач, А.Н. Покупательское поведение: анализ спонтанных последовательностей и регрессионных моделей в маркетинговых исследованиях / А.Н. Горбач, Н.А. Цейтлин. – Київ: Освіта України, 2011. – 220 с.

15. Konno, H. Choosing the best set of variables in regression analysis using integer programming / H. Konno, R. Yamamoto // Journal of Global Optimization. – 2009. – Vol. 44. – P. 273–282. DOI: 10.1007/s10898-008-9323-9

16. 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. DOI: 10.1016/j.patcog.2020.107565

17. Bertsimas, D. Scalable holistic linear regression / D. Bertsimas, M.L. Li // Operations Research Letters. – 2020. – Vol. 48, № 3. – P. 203–208. DOI: 10.1016/j. orl.2020.02.008

18. Базилевский, М.П. Сведение задачи отбора информативных регрессоров при оценивании линейной регрессионной модели по методу наименьших квадратов к задаче частично-булевого линейного программирования / М.П. Базилевский // Моделирование, оптимизация и информационные технологии. – 2018. – Т. 6, № 1 (20). – С. 108–117.

19. Базилевский, М.П. Построение вполне интерпретируемых линейных регрессионных моделей с помощью метода последовательного повышения абсолютных вкладов переменных в общую детерминацию / М.П. Базилевский // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2022. – № 2. – С. 5–16.

20. Базилевский, М.П. Сравнительный анализ эффективности методов построения вполне интерпретируемых линейных регрессионных моделей / М.П. Базилевский // Моделирование и анализ данных. – 2023. – Т. 13, № 4. – С. 59–83.

21. Базилевский, М.П. Критерии нелинейности квазилинейных регрессионных моделей / М.П. Базилевский // Моделирование, оптимизация и информационные технологии. – 2018. – Т. 6, № 4 (23). – С. 185–195.

22. Базилевский, М.П. Отбор информативных регрессоров с учетом 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.

23. Базилевский, М.П. Отбор значимых по критерию Стьюдента информативных регрессоров в оцениваемых с помощью МНК регрессионных моделях как задача частично-булевого линейного программирования / М.П. Базилевский // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2021. – № 3. – С. 5–16.

24. Михайлова, Н.А. Множественные регрессионные модели прочности бетона на сжатие / Н.А. Михайлова, И.В. Стефаненко // Вестник Волгоградского государственного архитектурно-строительного университета. Серия: Строительство и архитектура. – 2017. – Т. 49, № 68. – С. 30–42.

25. Abrams, D.A. Design of concrete mixtures / D.A. Abrams // Structural Materials Research Laboratory, Lewis Institute. – 1918. – Vol. 1.

26. Bolomey, J. Deformation elastigues, plastigues et de retrait de guelgues betons / J. Bolomey // Bulleten technique de la Suisse Romande. – 1942. – Ann. 68. – № 15. – 80 p.

27. Concrete Compressive Strength: сайт [Электронный ресурс] / UC Irvine Machine Learning Repository. – URL: https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength (дата обращения: 25.01.2024).

28. Базилевский, М.П. Программа построения вполне интерпретируемых элементарных и неэлементарных квазилинейных регрессионных моделей / М.П. Базилевский // Труды ИСП РАН. – 2023. – Т. 35, вып. 4. – С. 129–144. DOI: /10.15514/ISPRAS–2023–35(4)–7

29. Голик, В.И. Влияние параметров подготовки заменителей цемента на прочность бетонных смесей / В.И. Голик, С.Г. Страданченко, С.А. Масленников // Технологии бетонов. – 2016. – № 9–10. – С. 21–25.

30. Кузнецов, Д.В. Влияние молотого доменного гранулированного шлака ПАО «Северсталь» на прочность бетона / Д.В. Кузнецов, Н.Н. Калиновская, К.С. Аль-Мусави // Технологии бетонов. – 2021. – № 2. – С. 33–36.

References

1. Oussous A., Benjelloun F.Z., Lahcen A.A., Belfkih S. Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, 2018, vol. 30, no. 4, pp. 431–448. DOI: https://doi.org/10.1016/j.jksuci.2017.06.001

2. Zhang Q., Yang L.T., Chen Z., Li P. A survey on deep learning for big data. *Information Fusion*, 2018, vol. 42, pp. 146–157. DOI: https://doi.org/10.1016/j.inffus.2017.10.006

3. Gunst R.F., Mason R.L. Regression analysis and its application: a data-oriented approach. CRC Press, 2018.

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

5. Nevzorova E.V., Zasiad'ko A.K., Zagumennova O.N. Mnogofaktornyi regressionnyi analiz faktorov riska razvitiia mikrososudistykh oslozhnenii sakharnogo diabeta 2 tipa [Multi-factor regression analysis of risk factors of developing microvascular complications of 2 type diabetes mellitus]. *Medicine and Physical Education: Science and Practice*, 2020, vol. 2, no. 2, pp. 58–67.

6. Buzikov Sh.V., Plotnikov S.A., Kozlov I.S. Optimizatsiia sostava smesevogo topliva dlia primeneniia v traktornykh dizeliakh [The blend fuel composition optimization to apply in tractor diesel engines]. *Trudy NAMI*, 2021, no. 1, pp. 16–24.

7. Chang J.J., Mi Z., Wei Y.M. Temperature and GDP: A review of climate econometrics analysis. *Structural Change and Economic Dynamics*, 2023, vol. 66, pp. 383–392. DOI: 10.1016/j.strueco.2023.05.009

8. Aivazian S.A. Metody ekonometriki [Methods of econometrics]. Moscow, Magistr: INFRA-M, 2010, 512 p.

9. Noskov S.I. Tekhnologiia modelirovaniia ob"ektov s nestabil'nym funktsionirovaniem i neopredelennost'iu v dannykh [Technology for modeling objects with unstable functioning and data uncertainty]. Irkutsk, RITs GP «Oblinformpechat'», 1996, 320 p.

10. Strizhov V.V., Krymova E.A. Metody vybora regressionnykh modelei [Methods for regression models selection]. Moscow, VTs RAN, 2010, 60 p.

11. Shrestha N. Detecting multicollinearity in regression analysis. *American Journal of Applied Mathematics and Statistics*, 2020, vol. 8, no. 2, pp. 39–42. DOI: 10.12691/ ajams-8-2-1

12. Molnar C. Interpretable machine learning. Lulu. Com, 2020.

13. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

14. Gorbach A.N., Tseitlin N.A. Pokupatel'skoe povedenie: analiz spontannykh posledovatel'nostei i regressionnykh modelei v marketingovykh issledovaniiakh [Buying behavior: analysis of spontaneous sequences and regression models in marketing research]. Kyiv, Osvita Ukraine, 2011, 220 p.

15. Konno H., Yamamoto R. Choosing the best set of variables in regression analysis using integer programming. *Journal of Global Optimization*, 2009, vol. 44, pp. 273– 282. DOI: https://doi.org/10.1007/s10898-008-9323-9

16. 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, p. 107565. DOI: https:// doi.org/10.1016/j.patcog.2020.107565

17. Bertsimas D., Li M.L Scalable holistic linear regression. *Operations Research Letters*, 2020, vol. 48, no. 3, pp. 203– 208. DOI: https://doi.org/10.1016/j. orl.2020.02.008

18. Bazilevskii M.P. Svedenie zadachi otbora informativnykh regressorov pri otsenivanii lineinoi regressionnoi modeli po metodu naimen'shikh kvadratov k zadache chastichno-bulevogo lineinogo programmirovaniia [Reduction the problem of selecting informative regressors when

estimating a linear regression model by the method of least squares to the problem of partial-Boolean linear programming]. *Modeling, Optimization and Information Technology*, 2018, vol. 6, no. 1 (20), pp. 108–117.

19. Bazilevskii M.P. Postroenie vpolne interpretiruemykh lineinykh regressionnykh modelei s pomoshch'iu metoda posledovatel'nogo povysheniia absoliutnykh vkladov peremennykh v obshchuiu determinatsiiu [Construction of quite interpretable linear regression models using the method of successive increase the absolute contributions of variables to the general determination]. *Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies*, 2022, no. 2, pp. 5–16.

20. Bazilevskii M.P. Sravnitel'nyi analiz effektivnosti metodov postroeniia vpolne interpretiruemykh lineinykh regressionnykh modelei [Comparative analysis of the effectiveness of methods for constructing quite interpretable linear regression models]. *Modeling and Data Analysis*, 2023, vol. 13, no. 4, pp. 59–83.

21. Bazilevskii M.P. Kriterii nelineinosti kvazilineinykh regressionnykh modelei [Nonlinear criteria of quasilinear regression models]. *Modeling, Optimization and Information Technology*, 2018, vol. 6, no. 4 (23), pp. 185–195.

22. Bazilevskii M.P. Otbor informativnykh regressorov s uchetom mul'tikollinearnosti mezhdu nimi v regressionnykh modeliakh kak zadacha chastichno-bulevogo lineinogo programmirovaniia [Subset selection in regression models with considering multicollinearity as a task of mixed 0-1 integer linear programming]. *Modeling, Optimization and Information Technology*, 2018, vol. 6, no. 2 (21), pp. 104–118.

23. Bazilevskii M.P. Otbor znachimykh po kriteriiu St'iudenta informativnykh regressorov v otsenivaemykh s pomoshch'iu MNK regressionnykh modeliakh kak zadacha chastichno-bulevogo lineinogo programmirovaniia [Selection of informative regressors sighnificant by Student's t-test in regression models estimated using OLS as a partial Boolean linear programming problem]. *Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies*, 2021, no. 3, pp. 5–16.

24. Mikhailova N.A., Stefanenko I.V. Mnozhestvennye regressionnye modeli prochnosti betona na szhatie [Multiple regression models of concrete compressive strength]. *Bulletin of Volgograd State University of Architecture and Civil Engineering. Series: Civil Engineering and Architecture*, 2017, vol. 49, no. 68, pp. 30–42.

25. Abrams D.A. Design of concrete mixtures. *Structural Materials Research Laboratory, Lewis Institute*, 1918, vol. 1.

26. Bolomey J. Deformation elastigues, plastigues et de retrait de guelgues betons. *Bulleten technique de la Suisse Romande*, 1942, ann. 68, no. 15, 80 p.

27. Concrete Compressive Strength [UC Irvine Machine Learning Repository], available at: https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength (Accessed 25 January 2024)

28. Bazilevskii M.P. Programma postroeniia vpolne interpretiruemykh elementarnykh i neelementarnykh kvazilineinykh regressionnykh modelei [Program for constructing quite interpretable elementary and non-elementary quasilinear regression models]. *Proceedings of the Institute for System Programming of the RAS*, 2023, vol. 35 (4), pp. 129–144. DOI: https://doi.org/10.15514/ISPRAS–2023–35(4)–7

29. Golik V.I., Stradanchenko S.G., Maslennikov S.A. Vliianie parametrov podgotovki zamenitelei tsementa na prochnost' betonnykh smesei [The influence of preparation parameters of cement substitutes on the strength of concrete mixtures]. *Concrete Technologies*, 2016, no. 9-10, pp. 21–25.

30. Kuznetsov D.V., Kalinovskaia N.N., Al'-Musavi K.S. Vliianie molotogo domennogo granulirovannogo shlaka PAO «Severstal'» na prochnost' betona [Influence of ground granulated blast furnace slag produced by the Severstal company of the strength of concrete]. *Concrete Technologies*, 2021, no. 2, pp. 33–36.