Моделирование и анализ данных
2023. Том 13. № 4. С. 59–83
DOI: https://doi.org/10.17759/mda.2023130404
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2023. Vol. 13, no. 4, pp. 59–83
DOI: https://doi.org/10.17759/mda.2023130404
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

◇◇◇◇◇◇◇ МЕТОДЫ ОПТИМИЗАЦИИ ◇◇◇◇◇◇◇

УДК 519.862.6

Сравнительный анализ эффективности методов построения вполне интерпретируемых линейных регрессионных моделей

Базилевский М.П.*
Иркутский государственный университет путей сообщения
(ФГБОУ ВО ИрГУПС), г. Иркутск, Российская Федерация
ORCID: https://orcid.org/0000-0002-3253-5697
e-mail: mik2178@yandex.ru

Ранее автору удалось свести задачу построения вполне интерпретируемой линейной регрессии, оцениваемой с помощью метода наименьших квадратов, к задаче частично-булевого линейного программирования. В таких моделях знаки оценок соответствуют содержательному смыслу факторов, абсолютные вклады переменных в общую детерминацию существенны, а степень мультиколлинеарности мала. Оптимальное решение сформулированной задачи также может быть найдено методом полного перебора регрессий. Цель статьи заключается в проведении сравнительного анализа эффективности этих двух подходов. Для проведения вычислительных экспериментов использовано 5 наборов реальных статистических данных различных объемов. В результате с помощью пакета LPSolve в разных условиях было решено более 550 различных частично-булевых задач. Параллельно оценена эффективность решения подобных им задач методом полного перебора в пакете Gretl. Во всех экспериментах предложенный нами метод оказался многократно эффективнее метода полного перебора. Самая высокая эффективность была достигнута при решении задач выбора оптимального числа регрессоров из 103 переменных, для решения каждой из которых методом перебора потребовалось бы оценить примерно 2103 (10,1 нониллиона) моделей, с чем обычный компьютер не справился бы и за 1000 лет. В LPSolve каждая из этих задач была решена за 32–191 секунду. Предложенным методом за приемлемое время удалось обработать выборку данных большого объема, содержащую 40 объясняющих переменных и 515345 наблюдений, что подтверждает независимость его эффективности от объема выборки. Выявлено, что ужесточение в линейных ограничениях задачи требований на мультиколлинеарность и абсолютные вклады переменных практически всегда снижает скорость её решения.

Ключевые слова: линейная регрессия, метод наименьших квадратов, интерпретируемость, задача частично-булевого линейного программирования, метод полного перебора, вклады переменных в детерминацию, мультиколлинеарность, эффективность.

Для цитаты:
Базилевский М.П. Сравнительный анализ эффективности методов построения вполне интерпретируемых линейных регрессионных моделей // Моделирование и анализ данных. 2023. Том 13. № 4. С. 59–83. DOI: https://doi.org/10.17759/mda.2023130404

Базилевский Михаил Павлович, кандидат технических наук, доцент кафедры математики, Иркутский государственный университет путей сообщения (ФГБОУ ВО ИрГУПС), г. Иркутск, Российская Федерация, ORCID: https://orcid.org/0000-0002-3253-5697, e-mail: mik2178@yandex.ru

1. ВВЕДЕНИЕ

Процесс построения линейной регрессионной модели условно можно разделить на два этапа: 1) выбор структурной спецификации; 2) оценивание неизвестных параметров регрессии. Выбор структурной спецификации, т.е. математической формы связи между переменными, в линейной модели означает решение задачи отбора m наиболее информативных регрессоров (ОИР) [1,2] из общего их числа l на основе некоторого оптимизационного критерия. Точным методом, гарантирующим оптимальное решение задачи ОИР, считается метод полного перебора [1], алгоритм которого предполагает оценивание C_l^m моделей-претендентов. Например, если $l = 50$, $m = 10$, то потребуется оценить 10272278170 моделей. Таким образом, метод полного перебора и самый трудоёмкий из всех методов решения задач ОИР. Если же число отбираемых регрессоров неизвестно, то трудоёмкость метода становится ещё больше, поскольку своей оценки при достаточном объеме выборки n требуют уже $2^l - 1$ моделей. Самым простым методом оценивания линейных регрессий является метод наименьших квадратов (МНК), в рамках которого разработано множество различных статистических тестов.

Для решения задач ОИР в линейных регрессиях (см., например, [3–11]) в настоящее время успешно применяется аппарат математического программирования, что гораздо эффективнее, чем использование переборных процедур. При этом также гарантируется оптимальность построенной модели. В зарубежной литературе задачи ОИР при использовании МНК принято в основном формулировать в виде задач частично-булевого квадратичного программирования (ЧБКП) [3], скорость решения которых зависит от объема выборки n. Результаты анализа научных статей по данной тематике представлены в табл. 1, во втором столбце которой приводятся фамилии учёных и год исследования; в третьем – краткая характеристика решаемой задачи ОИР; в четвёртом – максимальный объем выборки n_{max} и максимальное число объясняющих переменных l_{max}, обработанных в результате исследования; в пятом – название решателя задачи математического программирования; в шестом – информация о системе, в которой проводились вычислительные эксперименты.

Таблица 1

Результаты анализа научных статей

№	Авторы (год)	Особенность задачи	(n_{max}, l_{max})	Решатель	Оборудование
1	Конно, Ямамото (2009) [3]	ЧБКП, отбор m регрессоров, эвристический алгоритм	(1000, 70)	CPLEX 10.1	Xeon, 3.73 GHz
2	Мияширо, Такано (2015) [4]	ЧБКП, отбор регрессоров по критериям AIC, BIC и скорректированному коэффициенту детерминации	(1933, 100)	CPLEX	Xeon
3	Мияширо, Такано (2015) [5]	ЧБКП, отбор регрессоров по критерию Меллоуза	(1993, 100)	CPLEX 12.5	Intel Xeon W5590, 3.33 GHz×2, 24 GB RAM
4	Парк, Клабжан (2020) [6]	ЧБКП, отбор регрессоров по разным критериям MAE, MSE и mRMR (минимальной избыточности и максимальной релевантности), итерационный эвристический алгоритм	(506, 103)	CPLEX	Xeon, 2.8 GHz, 15 GB RAM
5	Тамура и др. (2019) [7]	ЧБКП, отбор регрессоров с контролем мультиколлинеарности по критерию VIF	(1993, 100)	CPLEX	Intel Core i7-4770, 3.40 GHz, 8 GB RAM
6	Тамура и др. (2017) [8]	Полуопределенная ЧБКП, отбор регрессоров с контролем мультиколлинеарности по числу обусловленности корреляционной матрицы	(1066, 65)	Gurobi 5.6, SCIP-SDP-2.0.0	Intel Core2 Quad, 2.66 GHz, 4 GB RAM
7	Бертсимас, Ли (2020) [9]	ЧБКП, отбор значимых регрессоров с контролем мультиколлинеарности, holistic regression	(9358, 125)	Gurobi 8.0	i7-5820k 6-core CPU, 16 GB DRAM
8	Чанг, Парк, Чонг (2020) [10]	ЧБКП, regression diagnostics	(1599, 124)	Gurobi 9.0	Intel Core i7-8700 CPU, 3.40 GHz (8 CPUs), 32 GB RAM
9	Такано, Мияширо (2020) [11]	ЧБКП, отбор регрессоров в ридж-регрессии с помощью критерия кросс-валидации	(100, 25)	CPLEX 12.8	Intel Core i7-4790 MCU, 3.60 GHz, 16 GB

По табл. 1 видно, что на сегодняшний день существует множество формулировок задач ОИР в терминах ЧБКП, позволяющих контролировать в процессе решения самые разные характеристики линейных регрессий – качество аппроксимации, мультиколлинеарность, значимость коэффициентов и пр. Также видно, что исследо-

ватели избегают решения задач большой размерности, когда объемы выборок более 10000 наблюдений, а число объясняющих переменных более 150 штук. При этом вычислительные эксперименты практически всегда проводятся с использованием дорогостоящих пакетов CPLEX и Gurobi на довольно мощных компьютерах.

В [12] автору удалось свести задачу ОИР в линейной регрессии, оцениваемой с помощью МНК, к задаче частично-булевого линейного программирования (ЧБЛП). Целевой функцией в ней выступает коэффициент детерминации, а количество линейных ограничений, в отличие от формулировок [3–11], не зависит об объема выборки n. В дальнейшем формализованная задача дополнилась ограничениями на коэффициенты вздутия дисперсии VIF [13], на t-критерии Стьюдента [14] и пр. На данный момент в [15] приведена самая последняя формулировка задачи ЧБЛП, решение которой приводит к построению вполне интерпретируемой линейной регрессии с оптимальным по коэффициенту детерминации количеством регрессоров, в которой знаки МНК-оценок согласованы со знаками соответствующих коэффициентов корреляции с y, абсолютные вклады переменных в общую детерминации не меньше заданного числа θ, а величины интеркорреляций по модулю не больше заданного числа r. Тестирование сформулированной в [15] задачи на реальных выборках большого объема никогда ещё не проводилось.

Цель работы заключается в проведении на основе реальных данных различных объемов сравнительного анализа эффективности решения задачи построения вполне интерпретируемых линейных регрессий методом полного перебора и методом решения специальным образом сформулированной задачи ЧБЛП.

2. ЗАДАЧА ЧБЛП И ПРОГРАММА ДЛЯ ЕЁ АВТОМАТИЧЕСКОГО ФОРМИРОВАНИЯ

Пусть в распоряжении исследователя имеется выборка данных объема n для зависимой (объясняемой) переменной y и l независимых (объясняющих) переменных x_1, x_2, ..., x_l. Составим матрицу коэффициентов интеркорреляций

$$R_{xx} = \begin{pmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_l} \\ r_{x_1x_2} & 1 & \dots & r_{x_2x_l} \\ \dots & \dots & \dots & \dots \\ r_{x_1x_l} & r_{x_2x_l} & \dots & 1 \end{pmatrix},$$

и вектор $R_{yx} = \begin{pmatrix} r_{yx_1} & r_{yx_2} & \dots & r_{yx_l} \end{pmatrix}^T$ корреляций объясняющих переменных с y.

Сформулируем задачу ЧБЛП для ОИР в линейной регрессии так, как это сделано в [15]:

$$R^2 = \sum_{j=1}^{l} r_{yx_j} \cdot \beta_j \to \max, \tag{1}$$

$$-\left(1-\delta_j\right)\cdot M \le \sum_{k=1}^{l} r_{x_j x_k}\cdot\beta_k - r_{yx_j} \le \left(1-\delta_j\right)\cdot M\ ,\quad j=\overline{1,l}\ , \tag{2}$$

$$0 \le \beta_j \le \delta_j \cdot M\ ,\quad j\in J^+\ , \tag{3}$$

$$-\delta_j \cdot M \le \beta_j \le 0\ ,\quad j\in J^-\ , \tag{4}$$

$$\delta_j \in \{0,1\}\ ,\quad j=\overline{1,l}\ , \tag{5}$$

$$C_{x_j}^{\text{абс}} = r_{yx_j}\cdot\beta_j \ge \theta\cdot\delta_j\ ,\quad j=\overline{1,l}\ , \tag{6}$$

$$\left|r_{x_i x_j}\right|\left(\delta_i+\delta_j-1\right)\le r\ ,\quad (i,j)\in\left\{(s_1,s_2):\left|r_{x_{s1}x_{s2}}\right|\ge r\right\}, \tag{7}$$

где β_1, β_2, ..., β_l – неизвестные параметры линейной регрессии в стандартизованном виде; R^2 – коэффициент детерминации; δ_1, δ_2, ..., δ_l – бинарные переменные, которые определяются по правилу

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

M – большое положительное число; J^+, J^- – индексные подмножества, элементы которых удовлетворяют условиям $r_{yx_j}>0$ и $r_{yx_j}<0$; $C_{x_j}^{\text{àáñ}} = r_{yx_j}^{\text{абс}}\cdot\beta_j$ – абсолютный вклад j-й переменной в общую детерминацию R^2; параметр $\theta\ge 0$ – наименьшая величина абсолютных вкладов входящих в модель переменных; параметр $0\le r\le 1$ – наибольшая величина коэффициентов интеркорреляций входящих в модель переменных.

Решение задачи ЧБЛП (1) – (7) приводит к построению линейной регрессии с оптимальным по критерию R^2 количеством объясняющих переменных, в которой $r_{yx_j}\cdot\beta_j>0$, $j\in\Phi$, вклады $C_{x_j}^{\text{абс}}\ge\theta$, $j\in\Phi$, а интеркорреляции $\left|r_{x_i x_j}\right|\le r$, $i,j\in\Phi$, $i<j$, где Φ – множество номеров отобранных объясняющих переменных. Заметим, что в построенной регрессии некоторые коэффициенты могут оказаться незначимыми по t-критерию Стьюдента, либо может возникнуть мультиколлинеарность сразу между несколькими переменными. Избежать этого можно, дополнив задачу (1) – (7) линейными ограничениями из работ [13,14].

Как известно, эффективность решения задачи ЧБЛП (1) – (7) зависит от выбора большого положительного числа M: если оно слишком велико, то процесс решения может замедлиться, а если мало, то найденное решение может оказаться неоптимальным. Выбор границ параметра M для задачи (1) – (7) обсуждается в [16]. Сначала определяются параметры M_{β_j}, $j=\overline{1,l}$ для линейных ограничений (3) и (4) по формулам:

$$M_{\beta_j} = \frac{R_{\max}}{r_{yx_j}}\ ,\quad j=\overline{1,l}\ , \tag{8}$$

где $R_{\max}$ – коэффициент детерминации регрессии, построенной со всеми l объясняющими переменными. Для удобства можно всегда брать $R_{\max} = 1$.

Затем находятся параметры $M^-_{u_j}$, $j = \overline{1,l}$ для линейных ограничений (2). Для этого нужно решить серию из l задач линейного программирования при $p = 1, 2, ..., l$:

$$M^-_{u_p} = \sum_{k=1}^{l} r_{x_p x_k} \cdot \beta_k - r_{yx_p} \to \min, \tag{9}$$

$$0 \le \beta_j \le M_{\beta_j}, \quad j \in J^+, \tag{10}$$

$$M_{\beta_j} \le \beta_j \le 0, \quad j \in J^-, \tag{11}$$

$$\sum_{j=1}^{l} r_{yx_j} \cdot \beta_j \le R_{\max}. \tag{12}$$

После чего аналогично для линейных ограничений (2) находятся параметры $M^+_{u_j}$, $j = \overline{1,l}$. Для этого решается та же серия задач (9) – (12), но с целевыми функциями на максимум.

И, наконец, в задаче ЧБЛП (1) – (7) ограничения (2) – (4) нужно заменить на следующие:

$$\left(1-\delta_j\right) \cdot M^-_{u_j} \le \sum_{k=1}^{l} r_{x_j x_k} \cdot \beta_k - r_{yx_j} \le \left(1-\delta_j\right) \cdot M^+_{u_j}, \quad j = \overline{1,l}, \tag{13}$$

$$0 \le \beta_j \le \delta_j \cdot M_{\beta_j}, \quad j \in J^+, \tag{14}$$

$$\delta_j \cdot M_{\beta_j} \le \beta_j \le 0, \quad j \in J^-. \tag{15}$$

Формировать задачу ЧБЛП (1), (5) – (7), (13) – (15) по реальной выборке данных для программы-решателя вручную не представляется возможным, особенно, если эта выборка большого объема. Для этого была разработана программа построения вполне интерпретируемых элементарных и неэлементарных квазилинейных регрессионных моделей (ВИнтер-2). Она позволяет в зависимости от выбранных пользователем начальных параметров автоматически формировать для решателя LPSolve IDE задачи ЧБЛП для построения различных, в частности, линейных, регрессионных моделей. Для формирования задачи (1), (5) – (7), (13) – (15) нужно выбрать: 1) число отбираемых регрессоров (при необходимости); 2) число знаков после запятой в действительных числах; 3) параметр θ; 4) параметр r. Выбор больших чисел M осуществляется автоматически. Сформированная задача представляет собой следующую последовательность блоков: 1) целевая функция (1); 2) левые части двойных неравенств (13); 3) правые части двойных неравенств (13); 4) левые и правые части двойных неравенств (14) и (15), содержащие бинарные переменные; 5) левые и правые части двойных неравенств (14) и (15), не содержащие бинарных переменных; 6) ограничения (6); 7) ограничения (7); 8) ограничения типа b12>=-Inf, указывающие на то, что переменные β_1, β_2, ..., β_l могут быть не только неотрицательными, но и отрицательными; 9) ограничения типа d10<=1, указывающие верхние границы

целочисленных переменных; 10) ограничение типа int d1,d2,d3, указывающее на бинарность целочисленных переменных.

Сформированную задачу нужно вручную открыть в решателе LPSolve, после запуска которого наблюдать за процессом её решения.

3. ОПИСАНИЕ ДАННЫХ И ИХ ПРЕДВАРИТЕЛЬНАЯ ОБРАБОТКА

Для проведения вычислительных экспериментов были использованы статистические данные, представленные в табл. 2. В третьем столбце указан объем выборки n и количество объясняющих переменных l. Как видно по табл. 2, в каждом наборе данных его объем n превосходит число объясняющих переменных l. Поэтому предварительно, чтобы убедиться в корректности данных, в пакете Gretl по каждому набору с помощью МНК оценивались модели множественной линейной регрессии. С наборами данных Data1, Data3 и Data5 проблем не возникло, были получены модели без совершенной коллинеарности с коэффициентами детерминации 0.620157, 0.737267 и 0.237001 соответственно.

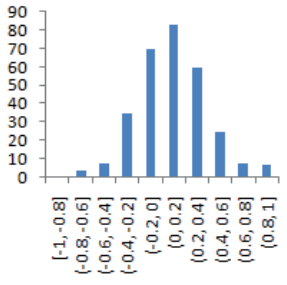
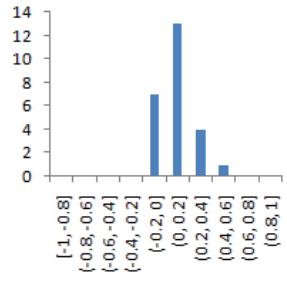
При построении модели по набору Data2 была выявлена функциональная зависимость:

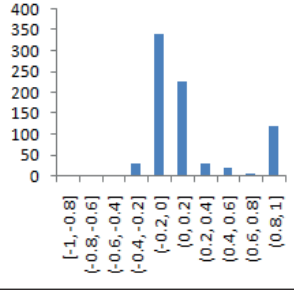
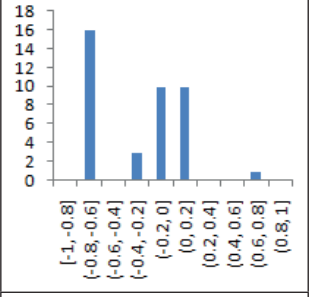
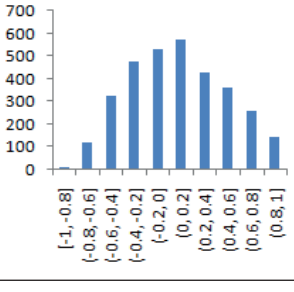
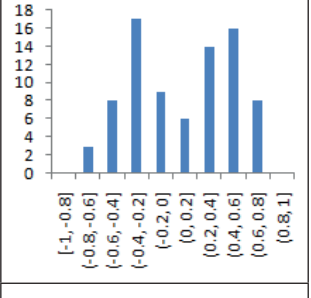
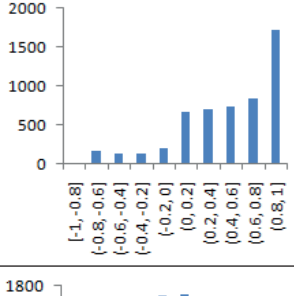
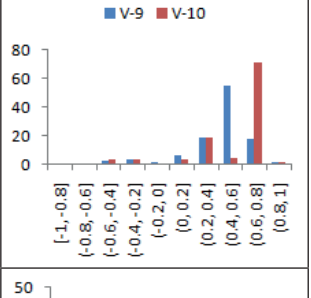
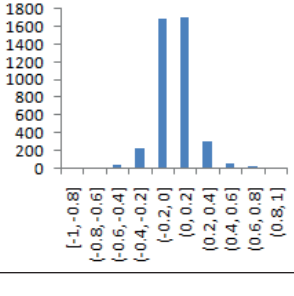
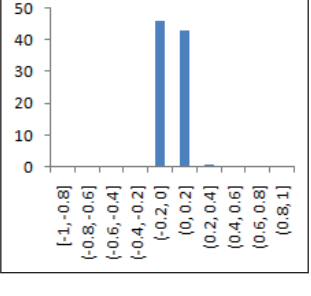
$$diffSeTime8 = SeTime6 - SeTime7 + 0.5SeTime8 - 0.5SeTime9 - 0.5diffSeTime6,$$

приводящая к совершенной коллинеарности. При этом установлено, что переменные *diffSeTime2* и *diffSeTime8* содержат только 2 отличных от нуля значения (-0,001 и 0,001). Поэтому было принято решение исключить эти факторы. В результате осталось 38 объясняющих переменных. В оцененной по новому набору Data2* линейной регрессии нет совершенной коллинеарности, а её коэффициент детерминации составил 0,819211.

Таблица 2

Описание данных и их характеристики

№	Название (источник); описание; зависимые переменные	(n, l)	Распределение коэффициентов интеркорреляций	Распределение коэффициентов корреляции с y
1	**Data1** (пакет Gretl, встроенный файл data7–20.gdt); данные о зарплатах игроков НБА; SALARY	(56, 25)		

№	Название (источник); описание; зависимые переменные	(n, l)	Распределение коэффициентов интеркорреляций	Распределение коэффициентов корреляции с y
2	**Data2** (сайт [17]); данные об элеронах самолёта F16; Goal	(13750, 40)		
3	**Data3** (сайт [18]); данные о критической температуре сверх-проводников; critical_temp	(21263, 81)		
4	**Data4** (сайт [19]); данные о стоимостях строительства и ценах продаж квартир в Иране; V-9, V-10	(372, 103)		
5	**Data5** (сайт [20]); данные о годах выпуска песен; year	(515345, 90)		

Набор данных Data4 содержит две зависимых переменных – V-9 и V-10. Поэтому выборку с зависимой переменной V-9 будем называть Data4a, а с V-10 – Data4b. В модели, оцененной по набору Data4a, выявлена совершенная коллинеарность, поэто-

му Gretl автоматически исключил 29 переменных. Коэффициент детерминации итоговой модели составил 0,987584. Аналогично, по набору Data4b было исключено 29 факторов, а для итоговой регрессии $R^2 = 0,982121$. Для проведения вычислительных экспериментов было принято решение в наборах Data4a и Data4b оставить список из 103 переменных в полном составе.

4. ОЦЕНКА СКОРОСТИ РЕШЕНИЯ ЗАДАЧ ОИР МЕТОДОМ ПОЛНОГО ПЕРЕБОРА

Получить решение задачи ЧБЛП (1), (5) – (7), (13) – (15) можно также тривиальным способом, организовав полный перебор всех возможных вариантов регрессионных моделей. Для того чтобы была возможность сравнивать эффективность двух этих подходов, требовалось найти статистическую зависимость скорости v решения задач ОИР полным перебором от заданного числа регрессоров m. Для этого был разработан специальный скрипт для пакета Gretl, реализующий процедуру отбора ровно m регрессоров в линейной регрессии по следующему алгоритму:

1. формируется матрица всех возможных комбинаций регрессоров, содержащая C_l^m строк и m столбцов;
2. по выборке находятся корреляционные матрицы R_{xx} и R_{yx};
3. с помощью матрицы комбинаций в цикле с помощью матриц R_{xx} и R_{yx} находятся стандартизованные оценки линейной регрессии и коэффициенты детерминации R^2;
4. выбирается лучшая модель с наибольшей величиной R^2.

С помощью этого скрипта на персональном компьютере с процессором Intel Core i5-4670 CPU (3.40 GHz) и объемом оперативной памяти 8 GB RAM было проведено два эксперимента – по набору данных Data3 и Data5. Число m задавалось равным 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 20, 25, 30, 35, 36, 37, 40, 45, 50, 55, 58, 64, 70 (при $m = 1$ задача решается практически мгновенно). Для каждого m фиксировалось общее количество обработанных моделей и время решения задачи ОИР (без учёта времени формирования матрицы комбинаций) в секундах. В результате были найдены скорости v (моделей/сек), как отношения общего числа моделей ко времени. Сразу было установлено, что при переходе с $m = 36$ на $m = 37$ в обоих экспериментах произошел солидный скачок скорости v в сторону уменьшения. Поэтому было принято решение строить статистические зависимости отдельно при $m \in [2, 36]$ и при $m \geq 37$.

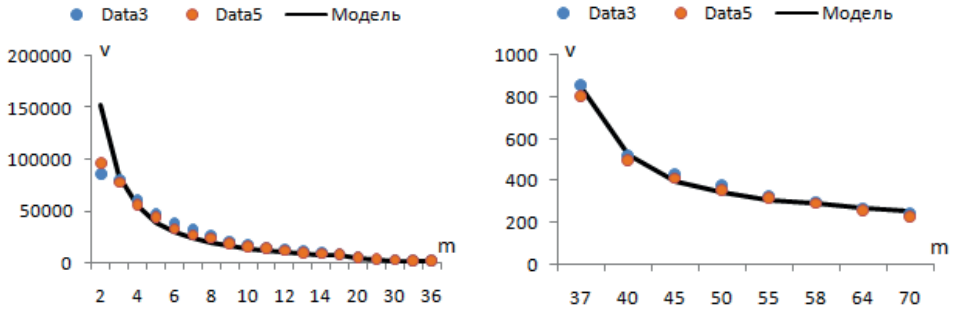
На рис. 1 синими точками указаны скорости, полученные при обработке выборки Data3, а оранжевыми – Data5. Как видно, эти скорости при равных m практически не отличаются. Кроме того, зависимость v от m носит нелинейных характер.

Для построения зависимостей в качестве объясняемой переменной была использована средняя скорость $\overline{v}$. С помощью МНК оценивалась степенная функция $\overline{v} = \alpha_0 m^{\alpha_1}$ при $m \in [2, 36]$ и $\overline{v} = \alpha_0 (m - 36)^{\alpha_1}$ при $m \geq 37$. Коэффициенты детерминации оцененных моделей в прологарифмированном виде составили 0,9799 и 0,9856, что подтверждает их высокое качество. Итоговая зависимость средней скорости

$\overline{v}$ решения задач ОИР полным перебором (на конкретном персональном компьютере) имеет вид:

$$\overline{v} = \begin{cases} e^{12,9507} m^{-1,45763}, \text{ при } 2 \le m \le 36, \\ e^{6,74197} (m-36)^{-0,341061}, \text{ при } m \ge 37. \end{cases} \quad (16)$$

На рис. 1 черным цветом указаны расчетные по формуле (16) скорости, которые практически совпадают с фактическими.



(а) $m \in [2,36]$ (б) $m \ge 37$

Рис. 1. Зависимости скоростей от числа m

5. ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ

Вычислительные эксперименты проводились на персональном компьютере с процессором Intel Core i5-4670 CPU (3.40 GHz) и объемом оперативной памяти 8 GB RAM. Решались задачи ЧБЛП (1), (5) – (7), (13) – (15) по наборам данных Data1, Data2*, Data3, Data4a, Data4b и Data5 в зависимости от заданных параметров r и θ. Параметр r для всех выборок задавался равным от 0,1 до 1 с шагом 0,1. Наименьшее значение параметра θ равно 0, а наибольшее $\theta_{\max}$ выбиралось так, чтобы при $\theta = \theta_{\max}$ и $r = 0,1$ происходил отбор одного или двух регрессоров. Для формирования задач ЧБЛП была использована программа ВИнтер-2, точность действительных чисел – 12 знаков после запятой. Большие числа M в ограничениях (13) – (15) ВИнтер-2 определяет автоматически по формулам (8) при $R_{\max} = 1$ и как результат решения серий задач линейного программирования типа (9) – (12). Решателем задач ЧБЛП выступает пакет LPSolve IDE с настройками по умолчанию. Лимит времени на решение задачи составляет 1800 секунд (полчаса). Все эксперименты проводились в предположении, что знаки коэффициентов корреляции объясняющих переменных с y согласуются с содержательным смыслом факторов.

Результаты вычислительных экспериментов по данным Data1 представлены в табл. 3.

Таблица 3

Результаты вычислительных экспериментов по набору данных Data1

θ \ r	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**1**
0	4 0,2653 0,476	5 0,2655 1,241	7 0,3106 2,791	8 0,3236 4,650	10 0,3899 5,601	10 0,3899 12,041	11 0,4124 10,716	11 0,4335 28,438	11 0,4335 35,705	14 0,4446 53,507
0,001	4 0,2653 0,505	4 0,2653 1,168	6 0,3097 2,467	7 0,3231 4,080	8 0,3869 5,119	8 0,3869 9,790	9 0,4118 10,741	9 0,4335 24,513	9 0,4335 33,431	11 0,4439 30,999
0,005	4 0,2653 0,466	4 0,2653 0,900	6 0,3097 1,632	6 0,3225 2,120	6 0,3819 2,604	6 0,3819 4,492	8 0,4116 4,580	8 0,4314 9,258	8 0,4314 12,051	8 0,4314 17,781
0,01	2 0,2495 0,400	2 0,2495 0,737	3 0,2925 1,220	4 0,3062 1,588	5 0,3807 1,919	5 0,3807 3,010	6 0,4044 3,189	7 0,4256 5,875	7 0,4256 7,675	7 0,4256 10,781
0,05	1 0,2416 0,239	1 0,2416 0,310	1 0,2416 0,607	1 0,2416 0,794	2 0,2674 0,910	2 0,2674 1,390	2 0,2674 1,511	2 0,2674 2,588	2 0,2674 3,284	2 0,2674 4,400

В табл.3 для каждой пары значений параметров r и θ указано количество отобранных переменных m, коэффициент детерминации R^2 найденной модели, время t решения задачи в секундах. По этой таблице видно, что, во-первых, во всех 50 случаях получено оптимальное решение в установленный получасовой лимит. Во-вторых, больше всего времени (53,507 с) ушло на решение задачи при $\theta = 0$ и $r = 1$ (при полном отсутствии требований на вклады переменных и мультиколлинеарность), а меньше всего (0,239 с) – при $\theta = 0,05$ и $r = 0,1$ (при самых жестких требованиях на вклады переменных и мультиколлинеарность). В целом по табл. 3 можно наблюдать, что увеличение θ (ужесточение требования на вклады) и уменьшение r (ужесточение требования на мультиколлинеарность) практически всегда снижает время решения задачи.

Решить любую задачу из табл. 3 можно методом полного перебора. Причём, при $\theta = 0$, 0.001, 0.005, 0.01 для этого пришлось бы оценить $2^{25} - 1 = 33554431$ моделей. С помощью формулы (16) установлено, что на это потребовалось бы (без учёта времени на формирование матриц комбинаций и проверки условий) примерно 3206,46 с. Отсюда следует, что эффективность решения задач ОИР представленным способом в LPSolve оказалась в 59,9–8016,1 раз выше, чем методом полного перебора в Gretl. Поскольку значение R^2 для модели со всеми 25 переменными составляет 0,620157, то при $\theta = 0,05$ максимальное число регрессоров $m = \left[\frac{0,620157}{0,05}\right] = 12$, поэтому пришлось бы оценить только $\sum_{i=1}^{12} C_{25}^{i} = 16777215$ моделей, на что потребуется примерно 1232,81 с. В этом случае эффективность решения задач ОИР нашим методом оказалась в 280,2–5158,2 раз выше, чем методом полного перебора.

Результаты вычислительных экспериментов по данным Data2* представлены в табл. 4.

Таблица 4

Результаты вычислительных экспериментов по набору данных Data2*

θ \ r	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0	5 0,5893 12,863	9 0,6996 33,198	13 0,7159 72,437	11 0,8129 103,799	11 0,8129 149,199	14 0,8129 150,915	15 0,8147 204,339	15 0,8147 248,128	15 0,8147 247,745	18 0,8148 365,958
0,0005	3 0,5887 8,019	7 0,6995 17,779	10 0,7154 29,021	6 0,8128 40,145	6 0,8128 53,399	7 0,8128 55,803	7 0,8143 73,818	7 0,8143 80,246	7 0,8143 79,919	11 0,8145 88,751
0,001	3 0,5887 6,959	5 0,6970 15,336	8 0,7127 22,617	5 0,8122 31,016	5 0,8122 41,017	6 0,8123 43,394	7 0,8139 56,285	7 0,8139 60,167	7 0,8139 59,634	8 0,8142 57,594
0,01	2 0,5857 5,513	4 0,6956 8,093	4 0,6956 9,199	4 0,8104 10,514	4 0,8104 15,558	4 0,8104 16,499	4 0,8104 18,942	4 0,8104 19,330	4 0,8104 19,161	6 0,8107 21,441
0,05	2 0,5857 1,982	3 0,6435 2,840	3 0,6435 2,758	3 0,7514 3,836	3 0,7514 5,242	3 0,7514 5,455	3 0,7514 5,508	3 0,7514 5,509	3 0,7514 5,418	5 0,7518 5,257

Как видно, снова во всех 50 случаях получено оптимальное решение в установленный получасовой лимит. И вновь в большинстве случаев чем выше θ и ниже r, тем меньше время решения задачи. Больше всего времени (365,958 с) ушло на решение задачи при $\theta = 0$ и $r = 1$, а меньше всего (1,982 с) – при $\theta = 0,05$ и $r = 0,1$. Замечено, что, например, при $\theta = 0$ с ростом r число отобранных регрессоров может убывать.

При $\theta = 0$, 0.0005, 0.001, 0.01 методом перебора пришлось бы оценить $2^{38} - 1 = 274877906943$ (275 миллиардов) моделей, на что потребовалось бы примерно 48138036,2 с или 557 суток непрерывной работы компьютера. Таким образом, эффективность решения задач ОИР представленным способом в LPSolve оказалась в 131540–8731732 раз выше, чем методом полного перебора в Gretl. При $\theta = 0,05$ максимальное число регрессоров $m = \left[\frac{0,819211}{0,05}\right] = 16$, поэтому пришлось бы оценить $\sum_{i=1}^{16} C_{38}^{i} = 57407177581$ моделей, на что потребуется примерно 6903730,63 с. В этом случае эффективность решения задач ОИР нашим методом в 1253173–3483214 раз выше, чем методом полного перебора.

Вычислительные эксперименты по набору данных Data3 проводились в четырех разных условиях: 1) прямой порядок объясняющих переменных от 1 до 81; 2) обратный порядок объясняющих переменных от 81 до 1; 3) обратный порядок объясняющих переменных и задача ЧБЛП (1) – (7) с параметрами M=50; 4) обратный

порядок объясняющих переменных, задача ЧБЛП (1) – (7) с параметрами M=50 и дополнительными ограничениями на R^2:

$$\sum_{j=1}^{l} r_{yx_j} \cdot \beta_j \le 1, \ \sum_{j=1}^{l} r_{yx_j} \cdot \beta_j \ge r^*, \tag{17}$$

где r^* – нижняя граница коэффициента детерминации. Это значение всегда можно брать равным максимальному из коэффициентов детерминации однофакторных линейных регрессий. В данном случае в условиях № 4 таблица результатов формировалась построчно слева направо, начиная с нижнего левого угла ($\theta = 0,1$, $r = 0,1$). При этом значение r^* для активных $\theta_{\text{акт}}$ и $r_{\text{акт}}$ выбиралось равным максимальному из коэффициентов детерминации моделей, полученных на предыдущих шагах при $\theta > \theta_{\text{акт}}$ и $r < r_{\text{акт}}$. Точность величины r^* составляла 6 знаков после запятой.

Результаты вычислительных экспериментов по данным Data3 представлены в табл. 5.

Таблица 5

Результаты вычислительных экспериментов по набору данных Data3

θ \ r	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Условия № 1. Прямой порядок переменных										
0	2 0,5397 197,767	4 0,5666 1621,51	8 0,5994 1800	11 0,6165 1800	13 0,6554 1800	14 0,6453 1800	17 0,6446 1800	19 0,6437 1800	22 0,6185 1800	– – 1800
0,01	2 0,5397 201,172	5 0,5665 1273,26	6 0,5894 1800	9 0,6022 1800	9 0,6022 1800	11 0,5787 1800	13 0,6184 1800	15 0,6121 1800	15 0,6121 1800	16 0,6314 1800
0,05	1 0,5202 133,713	1 0,5202 502,107	5 0,5817 1268,24	4 0,5881 1800	7 0,5376 1800	– – 1800	– – 1800	– – 1800	– – 1800	– – 1800
0,07	1 0,5202 110,147	1 0,5202 332,869	2 0,5582 631,762	5 0,5749 1103,64	4 0,6113 1467,54	4 0,6119 1704,19	4 0,6119 1800	4 0,6119 1800	4 0,6119 1800	4 0,6119 1800
0,10	1 0,5202 80,43	1 0,5202 175,849	1 0,5202 304,19	1 0,5202 454,936	3 0,5835 574,821	4 0,5983 617,37	4 0,5983 629,754	4 0,5983 618,929	4 0,5983 728,914	4 0,5983 655,219
Условия № 2. Обратный порядок переменных										
0	2 0,5397 188,246	4 0,5666 1679,87	8 0,5994 1800	12 0,6176 1800	13 0,6552 1800	14 0,6392 1800	18 0,6367 1800	– – 1800	– – 1800	– – 1800
0,01	2 0,5397 167,225	5 0,5665 1209,17	5 0,5948 1800	9 0,6297 1800	11 0,6180 1800	12 0,6341 1800	14 0,6420 1800	14 0,6420 1800	– – 1800	– – 1800
0,05	1 0,5202 93,31	1 0,5202 385,996	5 0,5817 666,43	4 0,5881 878,796	5 0,6160 826,987	5 0,6273 715,434	7 0,6393 677,096	7 0,6393 695,793	7 0,6393 878,9	7 0,6393 903,362

θ \ r	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,07	1 0,5202 81,226	1 0,5202 241,379	2 0,5582 431,757	5 0,5749 457,4	4 0,6113 407,74	4 0,6119 393,263	4 0,6119 384,443	4 0,6119 396,190	4 0,6119 434,534	4 0,6119 443,542
0,10	1 0,5202 59,355	1 0,5202 174,296	1 0,5202 269,955	1 0,5202 267,461	3 0,5835 223,619	4 0,5983 203,14	4 0,5983 192,821	4 0,5983 198,769	4 0,5983 216,876	4 0,5983 211,203
Условия № 3. Обратный порядок переменных, M=50										
0	2 0,5397 122,945	4 0,5666 1350,71	8 0,5994 1800	12 0,6176 1800	12 0,6199 1800	14 0,6392 1800	18 0,6430 1800	– – 1800	– – 1800	16 0,5957 1800
0,01	2 0,5397 97,617	5 0,5665 884,315	5 0,5948 1800	9 0,6297 1800	11 0,6180 1800	12 0,6341 1800	14 0,6420 1800	14 0,6420 1800	– – 1800	– – 1800
0,05	1 0,5202 58,057	1 0,5202 262,76	5 0,5817 401,332	4 0,5881 575,673	5 0,6160 586,709	5 0,6273 471,719	7 0,6393 430,101	7 0,6393 432,044	7 0,6393 534,787	7 0,6393 522,983
0,07	1 0,5202 51,048	1 0,5202 159,867	2 0,5582 287,328	5 0,5749 354,384	4 0,6113 372,046	4 0,6119 355,241	4 0,6119 324,892	4 0,6119 318,005	4 0,6119 332,508	4 0,6119 346,67
0,10	1 0,5202 44,343	1 0,5202 133,396	1 0,5202 224,789	1 0,5202 283,247	3 0,5835 294,821	4 0,5983 260,734	4 0,5983 249,054	4 0,5983 247,104	4 0,5983 276,976	4 0,5983 271,1
Условия № 4. Обратный порядок переменных, M=50, доп. ограничения на R^2										
0	2 0,539710 108,714	4 0,566696 772,040	8 0,599489 1800	– – 1800	13 0,655216 1800	– – 1800	18 0,662074 1800	18 0,662192 1800	16 0,666679 1800	– – 1800
0,01	2 0,539710 99,664	5 0,566547 1085,624	6 0,589429 1800	9 0,629783 1800	10 0,650646 1800	11 0,654328 1800	12 0,655761 1800	12 0,655761 1800	11 0,659170 1800	– – 1800
0,05	1 0,520231 54,256	1 0,520231 249,392	5 0,581710 335,111	4 0,588120 462,939	5 0,616071 358,06	5 0,627385 318,455	7 0,639385 288,019	7 0,639385 1800	7 0,639385 1800	7 0,639385 1800
0,07	1 0,520231 40,110	1 0,520231 112,673	2 0,558250 118,839	5 0,574943 185,496	4 0,611381 123,995	4 0,611953 107,23	4 0,611953 132,089	4 0,611953 125,089	4 0,611953 150,584	4 0,611953 144,08
0,10	1 0,520231 26,751	1 0,520231 48,796	1 0,520231 60,380	1 0,520231 71,449	3 0,583520 35,280	4 0,598350 35,910	4 0,598350 32,866	4 0,598350 30,728	4 0,598350 31,091	4 0,598350 30,568

В табл. 5 для наглядности серым цветом выделены ячейки, для которых за полчаса либо не доказана оптимальность решения задачи, либо вообще не получено решение. Для сравнения результатов вычислительных экспериментов, проведенных в четырех условиях, была составлена табл. 6. Во втором её столбце для каждого условия указан процент оптимальных решений 50 задач, в третьем – процент задач без решения в установленный лимит времени, в четвертом – общее время решения 50-ти задач,

в пятом – среднее время поиска оптимального решения, в шестом – среднее значение R^2 для моделей с доказанной и не доказанной оптимальностью.

Таблица 6

Показатели эффективности решенных по данным Data3 задач в зависимости от условий

Условие	Оптимальные решения, %	Решения без результата за 30 мин, %	Общее время решения 50-ти задач, с	Среднее время на поиск оптимального решения, с	Среднее качество моделей по R^2
1	46	12	63988,33	669,95	0,584779
2	68	10	44455,58	460,45	0,591822
3	68	8	40719,31	350,56	0,591276
4	62	8	39976,28	186,33	0,597076

По табл. 6 можно сделать следующие выводы.

1. Порядок следования объясняющих переменных в исходной выборке может существенно влиять на скорость решения задачи ЧБЛП. Оказалось, что изменение порядка следования переменных с прямого (от 1 до 81) на обратный (от 81 до 1) увеличило процент оптимальных решений на 22 %, снизило процент решений без результата на 2 %, уменьшило общее время решения 50-ти задач на 19533 с (примерно на 5,5 часов) и среднее время на поиск оптимального решения на 209,5 с, а также увеличило среднее качество моделей по R^2 на 0,007043.
2. На скорость решения задачи ЧБЛП могут существенно влиять выбранные значения больших чисел M. Получилось, что по выборке с обратным порядком следования переменных изменение в задаче (1), (5) – (7), (13) – (15) всех больших чисел на $M = 50$ уменьшило процент решений без результата на 2 %, общее время решения 50-ти задач на 3736,27 с, среднее время на поиск оптимального решения на 109,89. При этом процент оптимальных решений не изменился, а среднее качество моделей по R^2 снизилось на 0,000546.
3. Внедрение ограничений (17) может существенно влиять на скорость решения задачи ЧБЛП. Последовательное проведение экспериментов в условиях № 4 оказалось самым эффективным. По сравнению с экспериментами в условиях № 3 уменьшилось общее время решения 50-ти задач на 743,03 с, среднее время на поиск оптимального решения на 164,23 с, увеличилось среднее качество моделей по R^2 на 0,0058. Процент решений без результата не изменился, а процент оптимальных решений снизился на 6 % из-за задач при $\theta = 0,05$, $r = 0.8, 0.9, 1$.

Для сравнения с методом перебора была взяты результаты вычислительных экспериментов (только оптимальные решения), полученные в условиях № 3. Тут снова прослеживается снижение времени решения задач при больших θ и малых r.

При $\theta = 0$ методом перебора пришлось бы оценить $2^{81} - 1 = 2417851639229258349412351$ (2,4 септиллиона) моделей примерно за $4,27 \cdot 10^{21}$ с; при $\theta = 0,01$ – примерно то же самое; при $\theta = 0,05$ –

$\sum_{i=1}^{14} C_{81}^i = 2283695211364530$ (2,3 квадриллиона) моделей примерно за $2,47 \cdot 10^{11}$ с (7853 года); при $\theta = 0,07$ – $\sum_{i=1}^{10} C_{81}^i = 2175273626013$ (2,1 триллионов) моделей примерно за 144911622 с (4,6 лет); при $\theta = 0,1$ – $\sum_{i=1}^{7} C_{81}^i = 3829130793$ (3,8 миллиарда) моделей примерно за 151981 с (42,2 часов).

Таким образом, эффективность решения задач ОИР нашим методом в LPSolve при $\theta = 0$ в $3,16 \cdot 10^{18}$ – $3,47 \cdot 10^{19}$ раз, при $\theta = 0,01$ в $4,82 \cdot 10^{18}$ – $4,37 \cdot 10^{19}$ раз, при $\theta = 0,05$ в 420992349–4254439602 раз, при $\theta = 0,07$ в 389499,2–2838732,6 раз, при $\theta = 0,1$ в 515,5–3427,4 раз выше, чем методом полного перебора в Gretl. Подтверждают эффективность и дополнительные эксперименты при $\theta = 0,01$ и $r = 0.3, 0.4, 0.5, 0.6$. Эти задачи были решены за 3278,48 с, 7383,34 с, 11866,6 с и 12386,7 с соответственно. Количества отобранных переменных и коэффициенты детерминации построенных моделей оказались следующие: (5, 0.5948), (9, 0.6297), (10, 0.6520) и (11, 0.6543).

Вычислительные эксперименты по набору данных Data4a проводились в двух разных условиях: 1) без ограничений на R^2; 2) с ограничениями (17) на R^2, $r^* = 0,95$. Результаты представлены в табл. 7.

Таблица 7

Результаты вычислительных экспериментов по набору данных Data4a

θ \ r	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Условия № 1а. Без ограничений на R^2										
0	3 0,9586 254,802	6 0,9590 1044,10	7 0,9613 284,86	9 0,9621 363,74	10 0,9636 534,326	10 0,96412 450,241	10 0,96412 753,755	11 0,96416 727,38	11 0,96416 903,586	11 0,96416 771,539
0,005	2 0,9586 204,809	2 0,9586 1574,24	4 0,9609 545,858	6 0,9615 960,526	9 0,9634 1477,09	10 0,5124 1800	13 0,5194 1800	14 0,5188 1800	14 0,5188 1800	14 0,5188 1800
0,007	2 0,9586 190,126	2 0,9586 1189,97	3 0,9605 810,467	3 0,9605 1375,36	9 0,6942 1800	10 0,6982 1800	12 0,7092 1800	11 0,7128 1800	14 0,7133 1800	15 0,7134 1800
0,01	2 0,9586 167,112	2 0,9586 868,475	2 0,9586 968,963	2 0,9586 1557,69	10 0,6981 1800	9 0,6965 1800	10 0,7090 1800	11 0,6836 1800	12 0,7131 1800	13 0,7132 1800
0,015	1 0,9534 174,815	1 0,9534 554,243	1 0,9534 788,331	1 0,9534 1169,25	4 0,9564 1423,2	8 0,6822 1800	8 0,7031 1800	12 0,6817 1800	10 0,6743 1800	8 0,5689 1800
Условия № 2а. С ограничениями на R^2										
0	3 0,9586 31,949	6 0,9590 47,057	7 0,9613 73,006	9 0,9621 101,18	10 0,9636 160,98	10 0,96412 160,675	10 0,96412 191,21	11 0,96416 173,251	11 0,96416 165,566	11 0,96416 117,857
0,005	2 0,9586 58,102	2 0,9586 83,709	4 0,9609 106,131	6 0,9615 115,674	9 0,9634 108,645	8 0,9638 108,432	8 0,9638 115,711	9 0,9638 103,331	9 0,9638 94,869	9 0,9638 64,817

θ \ r	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,007	2 0,9586 55,209	2 0,9586 76,711	3 0,9605 89,198	3 0,9605 101,247	8 0,9632 90,868	7 0,9635 90,318	8 0,9636 91,7	8 0,9636 79,838	8 0,9636 71,328	8 0,9636 48,881
0,01	2 0,9586 46,585	2 0,9586 66,94	2 0,9586 75,757	2 0,9586 83,938	6 0,9620 75,291	6 0,9632 76,675	6 0,9633 77,701	6 0,9633 67,239	6 0,9633 59,494	6 0,9633 40,978
0,015	1 0,9534 38,231	1 0,9534 57,707	1 0,9534 70,649	1 0,9534 88,668	4 0,9564 70,492	5 0,9584 68,855	5 0,9584 69,151	5 0,9584 59,99	5 0,9584 52,417	5 0,9584 35,984

По табл. 7 видно, что в обоих условиях процент решений без результата составляет 0 %. Оказалось, что введение ограничений на R^2 увеличило процент оптимальных решений с 56 % до 100 %, снизило общее время решения 50-ти задач с 61688,85 с до 4260,19 с (примерно в 14,5 раз) и среднее время на поиск оптимального решения с 788,88 с до 85,2 с (примерно в 9,2 раза), увеличило среднее качество моделей по R^2 с 0,824478 до 0,960828.

Для сравнения с методом перебора была взяты результаты вычислительных экспериментов, полученные в условиях № 2a. Снова прослеживается снижение времени решения задач при больших θ и малых r. Но в этот раз оно менее выраженное.

При $\theta = 0, 0.005, 0.007$ методом перебора пришлось бы оценить $2^{103} - 1 = 10141204801825835211973625643007$ (10,1 нониллиона) моделей примерно за $3 \cdot 10^{28}$ с (без учёта времени на формирование матриц комбинаций и проверки условий); при $\theta = 0.01, 0.015$ – примерно то же самое.

Таким образом, эффективность решения задач ОИР нашим методом в LPSolve в $1,569 \cdot 10^{26} - 9,39 \cdot 10^{26}$ раз выше, чем методом полного перебора в Gretl.

Вычислительные эксперименты по набору данных Data4b проводились в двух разных условиях: 1) без ограничений на R^2; 2) с ограничениями (17) на R^2, параметр $M = 60$, значение r^* с шестью знаками после запятой выбиралось так же, как при экспериментировании по выборке Data3 при условиях № 4. Результаты представлены в табл. 8.

Таблица 8

Результаты вычислительных экспериментов по набору данных Data4b

θ \ r	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
Условия № 1b. Без ограничений на R^2										
0	1 0,9279 476,549	4 0,9579 1456,01	6 0,9589 1800	9 0,9606 1800	10 0,9615 1800	8 0,9630 1800	11 0,9660 1800	13 0,9665 1800	14 0,9665 1800	19 0,9666 1800
0,01	1 0,9279 407,725	2 0,9576 1223,05	4 0,8503 1800	9 0,8445 1800	10 0,8418 1800	9 0,8834 1800	13 0,8891 1800	10 0,8883 1800	12 0,8908 1800	15 0,8932 1800

θ \ r	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1
0,02	1 0,9279 262,429	2 0,9576 925,852	2 0,9576 1732,88	10 0,8666 1800	10 0,8636 1800	10 0,9060 1800	11 0,9072 1800	10 0,8507 1800	10 0,8519 1800	11 0,8524 1800
0,03	1 0,9279 244,271	2 0,9576 761,42	2 0,9576 1260,54	2 0,9576 1800	9 0,8738 1800	9 0,9089 1800	8 0,8958 1800	8 0,8900 1800	7 0,7390 1800	8 0,7099 1800
0,04	1 0,9279 204,977	2 0,9576 496,809	2 0,9576 762,148	2 0,9576 1033,84	2 0,9576 1446,00	2 0,9576 1384,48	9 0,9152 1800	9 0,8881 1800	9 0,8881 1800	8 0,8878 1800
Условия № 2b. M=60, с ограничениями на R^2										
0	1 0,927989 46,401	4 0,957929 1800	6 0,958979 1800	9 0,960627 1800	10 0,961561 1800	– – 1800	11 0,967157 1800	13 0,967616 1800	– – 1800	19 0,967702 1800
0,01	1 0,927989 52,337	2 0,957639 98,446	2 0,957639 234,152	3 0,958815 262,537	7 0,960867 265,578	8 0,963793 281,362	8 0,966249 526,623	10 0,967416 947,019	10 0,967416 1033,387	12 0,967487 759,165
0,02	1 0,927989 36,52	2 0,957639 69,071	2 0,957639 100,474	2 0,957639 123,127	2 0,957639 122,061	4 0,960368 116,364	6 0,964694 183,657	7 0,966500 164,036	7 0,966500 145,346	8 0,966540 110,548
0,03	1 0,927989 31,049	2 0,957639 54,54	2 0,957639 71,537	2 0,957639 84,363	2 0,957639 81,779	4 0,959505 83,223	4 0,961506 128,814	4 0,961806 123,908	4 0,961806 105,228	4 0,961806 76,214
0,04	1 0,927989 25,434	2 0,957639 46,274	2 0,957639 52,513	2 0,957639 62,761	2 0,957639 66,221	2 0,957639 72,432	3 0,961018 97,052	3 0,961018 85,535	3 0,961018 77,604	3 0,961018 55,490

По табл. 8 было установлено, что в условиях № 2b процент решений без результата увеличился с 0 % до 4 %, процент оптимальных решений увеличился с 32 % до 82 %, общее время решения 50-ти задач уменьшилось с 75279 с до 23360,2 с, среднее время на поиск оптимального решения уменьшилось с 879,9 с до 174,6 с, среднее качество моделей по R^2 увеличилось с 0,91034 до 0,957776.

Для сравнения с методом перебора была взяты результаты вычислительных экспериментов, полученные в условиях № 2b. Судить о влиянии θ и r на время решения задач не правомерно, поскольку все они были решены при разных r^*.

При $\theta = 0$ методом перебора пришлось бы оценить $2^{103} - 1$ моделей примерно за $3 \cdot 10^{28}$ с; при $\theta = 0.01, 0.02$ – примерно то же самое; при $\theta = 0,03$ – $\sum_{i=1}^{32} C_{103}^{i} = 7,76988 \cdot 10^{26}$ (0,777 октиллиона) моделей примерно за $2,7909 \cdot 10^{23}$ с; при $\theta = 0,04$ – $\sum_{i=1}^{24} C_{103}^{i} = 2,52555 \cdot 10^{23}$ (0,252 септиллиона) моделей примерно за $6,0142 \cdot 10^{19}$ с.

Таким образом, эффективность решения задач ОИР нашим методом в LPSolve при $\theta = 0, 0.01, 0.02$ в $2,903 \cdot 10^{25} - 8,215 \cdot 10^{26}$ раз, при $\theta = 0,03$ в $2,166 \cdot 10^{21} - 8,989 \cdot 10^{21}$

раз, при $\theta = 0,04$ в $6,197 \cdot 10^{17} - 2,364 \cdot 10^{18}$ раз выше, чем методом полного перебора в Gretl.

Вычислительные эксперименты по самому крупному набору данных Data5 предварительно проводились по ненулевым значениям θ и низким значениям r. В установленные получасовые лимиты были получены следующие результаты: 1) при $\theta = 0,0075$ и $r = 0,1$ отобрано 3 регрессора, $R^2 = 0,06$, оптимальность не доказана; 2) при $\theta = 0,0075$ и $r = 0,2$ нет результата; 3) при $\theta = 0,01$ и $r = 0,1$ отобрано 2 регрессора, $R^2 = 0,0525$, оптимальность не доказана; 4) при $\theta = 0,01$ и $r = 0,2$ нет результата.

Таким образом, установленного получасового лимита оказалось недостаточно для решения задач по выборке Data5. Тем не менее, в некоторых случаях довольно быстро получаются близкие к оптимальным, а, возможно, и оптимальные, решения.

Далее было принято решение упорядочить объясняющие переменные по убыванию модулей их коэффициентов корреляции с y и из первых 40 факторов сформировать новую выборку Data5*. Для линейной регрессии со всеми 40 переменными $R^2 = 0,179946$. Результаты вычислительных экспериментов по набору данных Data5* представлены в табл. 9.

Таблица 9

Результаты вычислительных экспериментов по набору данных Data5

θ \ r	**0,1**	**0,2**	**0,3**	**0,4**	**0,5**	**0,6**	**0,7**	**0,8**	**0,9**	**1**
0	8 0,0729 52,382	17 0,1001 1335,40	19 0,1359 1800	19 0,1482 1800	20 0,1690 1800	20 0,1691 1800	20 0,1691 1800	20 0,1691 1800	21 0,1694 1800	21 0,1694 1800
0,0025	6 0,0708 38,411	12 0,0990 564,321	13 0,1328 1493,81	14 0,1465 1761,19	12 0,1663 1417,78	12 0,1663 1635,9	12 0,1663 1655,63	12 0,1663 1663,39	12 0,1663 1681,36	12 0,1663 1661,53
0,005	4 0,0654 26,289	9 0,0918 214,816	7 0,1233 432,443	8 0,1363 595,44	7 0,1566 609,035	7 0,1566 667,344	7 0,1566 672,059	7 0,1566 695,564	7 0,1566 671,565	7 0,1566 670,639
0,0075	3 0,0600 22,405	4 0,0717 150,41	4 0,1073 286,088	5 0,1166 440,173	5 0,1493 440,781	5 0,1493 484,955	5 0,1493 487,463	5 0,1493 504,33	5 0,1493 483,014	5 0,1493 503,118
0,01	2 0,0525 21,002	3 0,0665 128,272	3 0,1032 239,305	4 0,1103 350,11	4 0,1394 358,752	4 0,1394 397,330	4 0,1394 396,758	4 0,1394 404,456	4 0,1394 397,752	4 0,1394 397,8

По табл. 9 можно сделать вывод, что оптимальных решений – 84 %, решений без результата – 0 %, общее время решения 50-ти задач – 41510 с, среднее время на поиск оптимального решения – 645,5 с, среднее качество моделей по R^2 – 0,1339. Значения в табл. 9 подтверждают, что с ростом θ и уменьшением r скорость решения задач практически всегда возрастает.

При $\theta = 0$ методом перебора пришлось бы оценить $2^{40} - 1$ моделей примерно за 207409219 с (6,57 лет); при $\theta = 0.0025, 0.005$ – примерно то же самое; при $\theta = 0,0075$ – $\sum_{i=1}^{23} C_{40}^{i} = 9,52 \cdot 10^{11}$ моделей примерно за 168933865 с (5,35 лет); при $\theta = 0,01$ – $\sum_{i=1}^{17} C_{40}^{i} = 2,36 \cdot 10^{11}$ моделей примерно за 31076852 с (1 год).

Таким образом, эффективность решения задач ОИР нашим методом в LPSolve при $\theta = 0, 0.0025, 0.005$ в 117766–7889582 раз, при $\theta = 0,0075$ в 334967–7540007 раз, при $\theta = 0,01$ в 76836–1479709 раз выше, чем методом полного перебора в Gretl.

6. ЗАКЛЮЧЕНИЕ

Подчеркнем основные результаты, полученные в данной работе.

Экспериментально доказано, что построение вполне интерпретируемой линейной регрессии предложенным методом, состоящим в решении задачи ЧБЛП (1), (5) – (7), (13) – (15), многократно эффективнее метода полного перебора. Для набора данных Data1 наш подход оказался эффективнее в 59,9–8016,1 раз; для Data2* – в 131540–8731732 раз; для Data3 – в 515,5 – $4,37 \cdot 10^{19}$ раз; для Data4a – в $1,569 \cdot 10^{26} - 9,39 \cdot 10^{26}$ раз; для Data4b – в $6,19 \cdot 10^{17} - 8,21 \cdot 10^{26}$ раз; для Data5* – в 76836–7889582 раз.

Среди полученных за получасовой промежуток времени 550-ти решений, представленных в табл. 3, 4, 5, 7, 8, 9, оптимальных оказалось 399 (72,6 %), обычных – 130 (23,6 %), отсутствующих – 21 (3,8 %). Таким образом, за приемлемое время в большинстве случаев были найдены оптимальные или близкие к ним решения.

Эффективность предложенного метода не зависит от объема выборки n, поэтому удалось обработать набор данных Data5*, содержащий 40 объясняющих переменных и 515345 наблюдений. Так, например, в работах [3–11] зарубежных авторов максимальный объем выборки для решения подобных задач составил всего 9358 наблюдений, т.е. примерно в 55 раз меньше.

Замечено, что выбор параметров θ и r может существенно влиять на скорость решения задачи ЧБЛП. Причём, с увеличением θ (ужесточение требований на вклады переменных) и уменьшением r (ужесточение требований на мультиколлинеарность) время решения задачи в большинстве случаев снижается, т.е. вполне интерпретируемая линейная регрессия строится быстрее, чем реализуется обычный ОИР без ограничений на θ и r.

Установлено, что на эффективность решения задачи ЧБЛП может существенно влиять порядок следования объясняющих переменных в выборке, параметр M, а также дополнительные ограничения (17) на коэффициент детерминации. Механизм влияния этих и других параметров на скорость решения задачи ЧБЛП требует дальнейших исследований.

Литература

1. *Стрижов В.В., Крымова Е.А.* Методы выбора регрессионных моделей. М.: Вычислительный центр им. А.А. Дородницына РАН, 2010. 60 с.

2. *Miller A.* Subset selection in regression. CRC Press, 2002.
3. *Konno H., Yamamoto R.* Choosing the best set of variables in regression analysis using integer programming // Journal of Global Optimization. 2009. Vol. 44. P. 273–282. DOI:10.1007/s10898-008-9323-9.
4. *Miyashiro R., Takano Y.* Mixed integer second-order cone programming formulations for variable selection in linear regression // European Journal of Operational Research. 2015. Vol. 247. P. 721–731. DOI:10.1016/j.ejor.2015.06.081.
5. *Miyashiro R., Takano Y.* Subset selection by Mallows' Cp: A mixed integer programming approach // Expert Systems with Applications. 2015. Vol. 42. P. 325–331. DOI:10.1016/j.eswa.2014.07.056.
6. *Park Y.W., Klabjan D.* Subset selection for multiple linear regression via optimization // Journal of Global Optimization. 2020. Vol. 77. P. 543–574. DOI:10.1007/s10898-020-00876-1.
7. *Tamura R., Kobayashi K., Takano Y., Miyashiro R., Nakata K., Matsui T.* Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor // Journal of Global Optimization. 2019. Vol. 73. P. 431–446. DOI:10.1007/s10898-018-0713-3.
8. *Tamura R., Kobayashi K., Takano Y., Miyashiro R., Nakata K., Matsui T.* Best subset selection for eliminating multicollinearity // Journal of the Operations Research Society of Japan. 2017. Vol. 60(3). P. 321–336. DOI:10.15807/jorsj.60.321.
9. *Bertsimas D., Li M.L.* Scalable holistic linear regression // Operations Research Letters. 2020. Vol. 48, Is. 3. P. 203–208. DOI:10.1016/j.orl.2020.02.008.
10. *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. DOI:10.1016/j.patcog.2020.107565.
11. *Takano Y., Miyashiro R.* Best subset selection via cross-validation criterion. Top. 2020. Vol. 28, Is. 2. P. 475–488. DOI: 10.1007/s11750-020-00538-1.
12. *Базилевский М.П.* Сведение задачи отбора информативных регрессоров при оценивании линейной регрессионной модели по методу наименьших квадратов к задаче частично-булевого линейного программирования // Моделирование, оптимизация и информационные технологии. 2018. Т. 6. № 1 (20). С. 108–117.
13. *Базилевский М.П.* Отбор информативных регрессоров с учётом мультиколлинеарности между ними в регрессионных моделях как задача частично-булевого линейного программирования // Моделирование, оптимизация и информационные технологии. 2018. Т. 6. № 2 (21). С. 104–118.
14. *Базилевский М.П.* Отбор значимых по критерию Стьюдента информативных регрессоров в оцениваемых с помощью МНК регрессионных моделях как задача частично-булевого линейного программирования // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2021. № 3. С. 5–16.
15. *Базилевский М.П.* Формализация процесса отбора информативных регрессоров в линейной регрессии в виде задачи частично-булевого линейного программирования с ограничениями на коэффициенты интеркорреляций // Современные наукоёмкие технологии. 2023. № 8. С. 10–14.
16. *Базилевский М.П.* Способ определения параметра M в задаче частично-булевого линейного программирования для отбора регрессоров в линейной регрессии // Вестник Технологического университета. 2022. Т. 25. № 2. С. 62–66.
17. Knowledge Extraction based on Evolutionary Learning [Электронный ресурс]. – Режим доступа: https://sci2s.ugr.es/keel/dataset.php?cod=93 (дата обращения 04.10.2023).
18. UCI Machine Learning Repository [Электронный ресурс]. – Режим доступа: https://archive.ics.uci.edu/dataset/464/superconductivty+data (дата обращения 04.10.2023).

19. UCI Machine Learning Repository [Электронный ресурс]. – Режим доступа: https://archive.ics.uci.edu/dataset/437/residential+building+data+set (дата обращения 04.10.2023).
20. UCI Machine Learning Repository [Электронный ресурс]. – Режим доступа: https://archive.ics.uci.edu/dataset/203/yearpredictionmsd (дата обращения 04.10.2023).

Comparative Analysis of the Effectiveness of Methods for Constructing Quite Interpretable Linear Regression Models

Mikhail P. Bazilevskiy*
Irkutsk State Transport University (ISTU), Irkutsk, Russia,
ORCID: https://orcid.org/0000-0002-3253-5697
e-mail: mik2178@yandex.ru

Previously, the author managed to reduce the problem of constructing a quite interpretable linear regression, estimated using ordinary least squares method, to a mixed-integer 0–1 linear programming problem. In such models, the signs of the estimates correspond to the substantive meaning of the factors, the absolute contributions of the variables to the overall determination are significant, and the degree of multicollinearity is small. The optimal solution to the formulated problem can also be found by generating all subsets method. The purpose of this article is to conduct a comparative analysis of the effectiveness of these two approaches. To conduct computational experiments, 5 sets of real statistical data of various volumes were used. As a result, more than 550 different mixed-integer 0–1 problems were solved using the LPSolve package under different conditions. At the same time, the efficiency of solving similar problems using the generating all subsets method in the Gretl package was assessed. In all experiments, our proposed method turned out to be many times more effective than the generating all subsets method. The highest efficiency was achieved in solving the subset selection problem from 103 variables, solving each of which by generating all subsets would require estimating approximately 2103 (10.1 nonillion) models, which a conventional computer would not have been able to cope with in 1000 years. In LPSolve, each of these problems was solved in 32–191 seconds. The proposed method was able to process a large data sample containing 40 explanatory variables and 515,345 observations in an acceptable time, which confirms the independence of its effectiveness from the sample size. It has been revealed that tightening the requirements for multicollinearity and absolute contributions of variables in the linear constraints of the problem almost always reduces the speed of its solution.

Keywords: linear regression, ordinary least squares method, interpretability, mixed-integer 0–1 linear programming problem, generating all subsets method, contributions of variables to determination, multicollinearity, efficiency.

For citation:
Bazilevskiy M.P. Comparative Analysis of the Effectiveness of Methods for Constructing Quite Interpretable Linear Regression Models. *Modelirovanie i analiz dannykh = Modelling and Data Analysis,* 2023. Vol. 13, no. 4, pp. 59–83. DOI: https://doi.org/10.17759/mda.2023130404 (In Russ.,abstr. in Engl.).

Mikhail P. Bazilevskiy, PhD (Engineering), Associate Professor, Department of Mathematics, Irkutsk State Transport University (ISTU), Irkutsk, Russia, ORCID: https://orcid. org/0000-0002-3253-5697, e-mail: mik2178@yandex.ru

References

1. Strizhov V.V., Krymova E.A. *Metody vybora regressionnykh modeley* [Methods for selecting regression models]. Moscow, Vychislitel'nyy tsentr im. A.A. Dorodnitsyna RAN, 2010. 60 p.
2. Miller A. *Subset selection in regression.* CRC Press, 2002.
3. 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:10.1007/s10898-008-9323-9.
4. Miyashiro R., Takano Y. Mixed integer second-order cone programming formulations for variable selection in linear regression, *European Journal of Operational Research*, 2015, vol. 247, pp. 721–731. DOI:10.1016/j.ejor.2015.06.081.
5. Miyashiro R., Takano Y. Subset selection by Mallows' Cp: A mixed integer programming approach, *Expert Systems with Applications*, 2015, vol. 42, pp. 325–331. DOI:10.1016/j.eswa.2014.07.056.
6. Park Y.W., Klabjan D. Subset selection for multiple linear regression via optimization, *Journal of Global Optimization*, 2020, vol. 77, pp. 543–574. DOI:10.1007/s10898-020-00876-1.
7. Tamura R., Kobayashi K., Takano Y., Miyashiro R., Nakata K., Matsui T. Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor, *Journal of Global Optimization*, 2019, vol. 73, pp. 431–446. DOI:10.1007/s10898-018-0713-3.
8. Tamura R., Kobayashi K., Takano Y., Miyashiro R., Nakata K., Matsui T. Best subset selection for eliminating multicollinearity, *Journal of the Operations Research Society of Japan*, 2017, vol. 60(3), pp. 321–336. DOI:10.15807/jorsj.60.321.
9. Bertsimas D., Li M.L. Scalable holistic linear regression, *Operations Research Letters*, 2020, vol. 48, is. 3, pp. 203–208. DOI:10.1016/j.orl.2020.02.008.
10. 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. DOI:10.1016/j.patcog.2020.107565.
11. Takano Y., Miyashiro R. Best subset selection via cross-validation criterion, *Top*, 2020, vol. 28, is. 2, pp. 475–488. DOI: 10.1007/s11750-020-00538-1.
12. Bazilevskiy M.P. Svedenie zadachi otbora informativnykh regressorov pri otsenivanii lineynoy regressionnoy modeli po metodu naimen'shikh kvadratov k zadache chastichno-bulevogo lineynogo programmirovaniya [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.
13. Bazilevskiy M.P. Otbor informativnykh regressorov s uchetom mul'tikollinearnosti mezhdu nimi v regressionnykh modelyakh kak zadacha chastichno-bulevogo lineynogo programmirovaniya [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.
14. Bazilevskiy M.P. Otbor znachimykh po kriteriyu St'yudenta informativnykh regressorov v otsenivaemykh s pomoshch'yu MNK regressionnykh modelyakh kak zadacha chastichno-bulevogo lineynogo programmirovaniya [Selection of informative regressors significant 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.
15. Bazilevskiy M.P. Formalizatsiya protsessa otbora informativnykh regressorov v lineynoy regressii v vide zadachi chastichno-bulevogo lineynogo programmirovaniya s ogranicheniyami na koeffitsienty interkorrelyatsiy [Formalization the subset selection process in linear regression as

a mixed integer 0–1 linear programming problem with constraints on intercorrelation coefficients], *Modern High Technologies*, 2023, no. 8, pp. 10–14.
16. Bazilevskiy M.P. Sposob opredeleniya parametra M v zadache chastichno-bulevogo lineynogo programmirovaniya dlya otbora regressorov v lineynoy regressii [Method for the M parameter determination in 0–1 mixed-integer linear programming problem for subset selection in linear regression], *Bulletin of the Technological University*, 2022, vol. 25, no. 2, pp. 62–66.
17. Knowledge Extraction based on Evolutionary Learning [Elektronnyy resurs]. URL https://sci2s.ugr.es/keel/dataset.php?cod=93 (Accessed 04.10.2023).
18. UCI Machine Learning Repository [Elektronnyy resurs]. URL https://archive.ics.uci.edu/dataset/464/superconductivty+data (Accessed 04.10.2023).
19. UCI Machine Learning Repository [Elektronnyy resurs]. URL https://archive.ics.uci.edu/dataset/437/residential+building+data+set (Accessed 04.10.2023).
20. UCI Machine Learning Repository [Elektronnyy resurs]. URL https://archive.ics.uci.edu/dataset/203/yearpredictionmsd (Accessed 04.10.2023).

Получена 30.10.2023 Received 30.10.2023
Принята в печать 15.11.2023 Accepted 15.11.2023