2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

doi 10.26089/NumMet.v26r435 УДК 519.862.6

Метод построения вполне интерпретируемых линейных регрессий со статистически значимыми по критерию Стьюдента оценками и незначимыми коэффициентами интеркорреляции

М. П. Базилевский
Иркутский государственный университет путей сообщения,
Иркутск, Российская Федерация
ORCID: 0000-0002-3253-5697, e-mail: mik2178@yandex.ru

Аннотация: Статья посвящена актуальной проблеме построения интерпретируемых моделей машинного обучения, а именно моделей множественной линейной регрессии. Для оценки их неизвестных параметров использован метод наименьших квадратов. Сформулировано вероятностно-статистическое определение вполне интерпретируемой линейной регрессии. Ее построение связано с отбором оптимального по коэффициенту детерминации количества наиболее информативных регрессоров так, чтобы гарантировать согласованность знаков коэффициентов регрессии с содержательным смыслом переменных. Кроме того, обеспечивается значимость оценок и незначимость коэффициентов интеркорреляций по t-критерию Стьюдента. Для построения вполне интерпретируемых линейных регрессий предложен метод, основанный на применении аппарата частично-булевого линейного программирования. Рассмотрены его строгая и нестрогая разновидности. Проведены вычислительные эксперименты, показавшие в большинстве случаев эффективность предложенного метода по сравнению с методом всех возможных регрессий.

Ключевые слова: регрессионный анализ, вполне интерпретируемая линейная регрессия, метод наименьших квадратов, отбор информативных регрессоров, мультиколлинеарность, t-критерий Стьюдента, метод всех возможных регрессий, задача частично-булевого линейного программирования.

Для цитирования: Базилевский М.П. Метод построения вполне интерпретируемых линейных регрессий со статистически значимыми по критерию Стьюдента оценками и незначимыми коэффициентами интеркорреляции // Вычислительные методы и программирование. 2025. **26**, № 4. 534–547. doi 10.26089/NumMet.v26r435.

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

A method for constructing fully interpretable linear regressions with statistically significant estimates according to the Student's criterion and insignificant intercorrelation coefficients

Mikhail P. Bazilevskiy
Irkutsk State Transport University, Irkutsk, Russia
ORCID: 0000-0002-3253-5697, e-mail: mik2178@yandex.ru

Abstract: The article is devoted to the urgent problem of constructing interpretable machine learning models, namely, multiple linear regression models. To estimate their unknown parameters, ordinary least squares is used. A probabilistically-statistical definition of fully interpretable linear regression is formulated. Its construction involves selecting the optimal number of most informative regressors based on determination coefficients in order to ensure consistency between the signs of regression coefficients and the substantive meanings of variables. Additionally, estimates are ensured to be significant and intercorrelation coefficients insignificant by using Student's t-test. To construct fully interpretable regressions, we propose a method that uses mixed 0-1 integer linear programming. Both strict and non-strict versions of this method are considered. Computational experiments were carried out, which in most cases showed the effectiveness of the proposed method compared to the generating all subsets method.

Keywords: regression analysis, fully interpretable linear regression, ordinary least squares, subset selection, multicollinearity, Student's t-test, generating all subsets, mixed 0-1 integer linear programming.

For citation: M. P. Bazilevskiy, "A method for constructing fully interpretable linear regressions with statistically significant estimates according to the Student's criterion and insignificant intercorrelation coefficients," Numerical Methods and Programming. **26** (4), 534–547 (2025). doi 10.26089/NumMet.v26r435.

1. Введение. В настоящее время для обработки больших объемов данных с целью автоматизации процессов принятия решений в различных предметных областях активно применяются модели машинного обучения. Они позволяют выявлять сложные нелинейные зависимости в данных, что делает их незаменимым инструментом в задачах прогнозирования, классификации и кластеризации. В условиях широкого внедрения сложных моделей машинного обучения, таких как глубокие нейронные сети, возникает необходимость не только в их высокой предсказательной точности, но и в объяснимости механизмов их функционирования для человека [1]. Это особенно критично в таких областях человеческой деятельности, как медицина, финансы и юриспруденция, где ошибки при принятии решений могут иметь серьезные последствия. В этой связи на сегодняшний день актуальна проблема построения интерпретируемых моделей машинного обучения [2], вызывающих больше доверия у экспертов и повышающих прозрачность автоматизированных систем принятия решений.

В монографии [2] подчеркивается, что высокими интерпретационными качествами из-за простоты объяснения коэффициентов обладают модели множественной линейной регрессии, в которых объясняемая переменная связана с объясняющими переменными линейной функциональной зависимостью. Построение линейной регрессии связано с проблемой отбора для нее наиболее информативных регрессоров (ОИР) [3]. Для решения этой проблемы разработано довольно много методов, подробное описание которых можно найти в [4, 5]. Наиболее точным из них является метод всех возможных регрессий (МВВР) [3], поскольку он подразумевает проверку в модели абсолютно всех комбинаций предикторов. Вместе с тем МВВР требует больших объемов вычислений. Значительно быстрее, чем МВВР, работают методы включения, исключения и включения-исключения. Но они не гарантируют получение оптимального с точки зрения выбранного критерия качества состава переменных. Благодаря развитию за последние годы алгоритмического и программного обеспечения для решения задач целочисленного программирования [6] появился

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

еще один точный метод ОИР. При решении таких задач задействуется метод ветвей и границ, при котором отсекаются подмножества, заведомо не содержащие оптимального решения. Поэтому скорость решения формализованных с помощью аппарата целочисленного программирования задач ОИР оказывается существенно выше, чем при использовании МВВР.

Фундаментальной работой, на которую ссылаются многие зарубежные авторы, является статья [7], в которой задача ОИР в линейной регрессии, оцениваемой с помощью метода наименьших квадратов (МНК), сведена к задаче частично-булевого квадратичного программирования (ЧБКП). Впоследствии появились работа [8], в которой критериями качества в задаче ЧБКП для ОИР служат скорректированный коэффициент детерминации, критерий Акаике и критерий Шварца, и манускрипт [9], в котором используется критерий Мэллоуза. Абсолютно другой подход к формализации задачи ОИР в терминах целочисленного программирования предложен в отечественной работе [10]. В ней с помощью теории корреляций удалось свести задачу ОИР для МНК к задаче частично-булевого линейного программирования (ЧБЛП), в которой число ограничений не зависит от объема выборки.

К сожалению, представленные в работах [7–10] формализации, работающие на оптимизацию только одного критерия качества, не гарантируют интерпретируемость построенной линейной регрессии. Для этого нужно комплексно решить множество проблем, связанных с устранением явления мультиколлинеарности [11], исключением незначимых по критерию Стьюдента [12] оценок регрессии, согласованием знаков оценок с содержательным смыслом факторов и т.д. Безусловно, все эти проверки можно организовать при реализации МВВР. Например, так было сделано в монографии [13], в которой описана программа построения интерпретируемой регрессионной модели объекта маркетинговых исследований с использованием анализа последовательностей. Но такой подход представляется вычислительно затратным. Поэтому стало актуальным формализовывать задачи ОИР в виде задач целочисленного программирования, внедряя в них различные ограничения на мультиколлинеарность, гомоскедастичность, автокорреляцию и др. Так, предложенная в [7] задача ЧБКП получила развитие в [14, 15]. В [14] она дополнена отложенными ограничениями (lazy constraints) на значимость оценок по критерию Стьюдента, а в [15] — ограничениями на значимость оценок на основе теста асимптотической нормальности и на степень мультиколлинеарности. Однако в [14, 15] вопросы интерпретируемости полученной модели не исследуются.

Предложенная в [10] задача ЧБЛП также получила развитие в многочисленных научных работах. Причем ее эволюция привела к возникновению понятия “вполне интерпретируемая линейная регрессия” (ВИЛинР). Например, в [16] исследуется задача отбора оптимального числа переменных при построении ВИЛинР, оценки которой согласуются с содержательным смыслом стоящих при них факторов, коэффициенты интеркорреляции не превосходят установленного порогового значения, а абсолютные вклады регрессоров в детерминацию выше выбранного порога. Главный недостаток сформулированной в [16] задачи ЧБЛП в том, что в ней для оценки влияния предикторов на объясняемую переменную задействованы не имеющие статистического характера абсолютные вклады регрессоров. Цель данной работы состоит во внедрении в задачу ЧБЛП из [16] вместо ограничений на абсолютные вклады переменных ограничений на значимость оценок по критерию Стьюдента, а также в тестировании эффективности предложенного метода по сравнению с МВВР. Заметим, что в [17] рассмотрена задача ЧБЛП с контролем значимости оценок по критерию Стьюдента при отборе фиксированного числа регрессоров, тогда как в данной работе будет уделено внимание отбору их оптимального количества.

2. Описание метода. Пусть по имеющейся выборке статистических данных объема n с помощью МНК оцениваются неизвестные параметры $\alpha_0, \alpha_1, \ldots, \alpha_l$ модели множественной линейной регрессии вида

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

где $y_i,\ i = \overline{1,n}$ — известные значения зависимой (объясняемой) переменной y, $x_{ij},\ i = \overline{1,n},\ j = \overline{1,l}$ — известные значения независимых (объясняющих) переменных $x_1, x_2, \ldots, x_l$, $\varepsilon_i,\ i = \overline{1,n}$ — ошибки регрессии.

Упростить МНК-оценивание параметров линейной регрессии (1), как замечено в [18], позволяет стандартизация (нормирование) всех переменных.

При наличии линейной связи между всеми или некоторыми объясняющими переменными в линейной регрессии возникает явление мультиколлинеарности [11, 18]. Сильная корреляция между предикторами приводит к серьезному понижению точности оценки параметров регрессии, следствием чего является

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

ненадежность найденных коэффициентов и отчасти неприемлемость их использования для интерпретации. Известно, что для уменьшения мультиколлинеарности высоко коррелирующие объясняющие переменные можно устранить из регрессии. Часто рекомендуется исключать одну из двух переменных x_i или x_j, если абсолютное значение коэффициента корреляции $r_{x_ix_j}$ между ними превышает пороговую величину 0.8 [18]. Из известной схемы проверки значимости выборочного коэффициента корреляции по t-критерию Стьюдента следует, что все коэффициенты интеркорреляции между объясняющими переменными незначимы для уровня значимости α, если они удовлетворяют ограничениям

$$|r_{x_ix_j}| \leqslant \frac{t_{\text{крит}}(\alpha, n-2)}{\sqrt{n-2+t^2_{\text{крит}}(\alpha, n-2)}}, \qquad i = \overline{1, l-1}, \quad j = \overline{i+1, l}, \tag{2}$$

где $t_{\text{крит}}(\alpha, n-2)$ — критическое значение t-критерия Стьюдента.

В регрессионном анализе также принято проверять статистическую значимость всех полученных оценок линейной регрессии. Если имеются незначимые коэффициенты, то стоящие при них объясняющие переменные исключаются и модель перестраивается. В данном случае руководствуются принципом простоты — чем меньше будет переменных в модели, тем проще ее анализировать и интерпретировать. К тому же модель с незначимыми переменными может хуже обобщать новые данные. Из известной схемы [19] проверки значимости стандартизованного коэффициента регрессии по t-критерию Стьюдента следует, что все оцененные коэффициенты линейной регрессии значимы для уровня значимости α, если они удовлетворяют ограничениям

$$R^2 - R^2_{y|x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_l} \geqslant T \cdot (1 - R^2), \qquad j = \overline{1, l}, \tag{3}$$

где R^2 — коэффициент детерминации линейной регрессии со всеми предикторами, $R^2_{y|x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_l}$ — коэффициент детерминации линейной регрессии без включения в нее предиктора x_j, $T = \dfrac{t^2_{\text{крит}}(\alpha, n-l-1)}{n-l-1}$.

Предварительно введем следующее вероятностно-статистическое определение ВИЛинР.

Определение. Линейная регрессия (1), оцененная с помощью МНК, называется вполне интерпретируемой (ВИЛинР) для выбранного уровня значимости α, если она удовлетворяет следующим условиям:

1) знаки коэффициентов корреляции r_{yx_j}, $j = \overline{1, l}$, соответствуют содержательному смыслу факторов;
2) знаки всех оценок $\tilde{\alpha}_j$ согласуются со знаками соответствующих коэффициентов корреляции r_{yx_j}, т.е. $\tilde{\alpha}_j \cdot r_{yx_j} > 0$, $j = \overline{1, l}$;
3) коэффициенты интеркорреляции $r_{x_ix_j}$, $i = \overline{1, l-1}$, $j = \overline{i+1, l}$, незначимы по t-критерию Стьюдента;
4) оценки $\tilde{\alpha}_j$, $j = \overline{1, l}$, значимы по t-критерию Стьюдента.

Сформулируем следующую задачу ОИР: требуется из имеющихся l предикторов выбрать оптимальное число переменных так, чтобы коэффициент детерминации R^2 линейной регрессии (1) был максимальным и выполнялись все четыре условия ВИЛинР.

В [10] представлена следующая задача ЧБЛП для ОИР в линейной регрессии:

$$R^2 = \sum_{j=1}^{l} r_{yx_j} \cdot \beta_j \to \max, \tag{4}$$

$$-(1 - \delta_j) \cdot M \leqslant \sum_{k=1}^{l} r_{x_jx_k} \cdot \beta_k - r_{yx_j} \leqslant (1 - \delta_j) \cdot M, \qquad j = \overline{1, l}, \tag{5}$$

$$-\delta_j \cdot M \leqslant \beta_j \leqslant \delta_j \cdot M, \qquad j = \overline{1, l}, \tag{6}$$

$$\delta_j \in \{0, 1\}, \qquad j = \overline{1, l}, \tag{7}$$

$$\sum_{j=1}^{l} \delta_j = m, \tag{8}$$

где β_j, $j = \overline{1, l}$ — стандартизованные коэффициенты [18] регрессии, M — известное достаточно большое положительное число, m — известное число отбираемых предикторов; булевы переменные δ_j, $j = \overline{1, l}$,

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

удовлетворяют условию

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

Решение задачи (4)–(8) позволяет идентифицировать оптимальную по критерию R^2 линейную регрессию с фиксированным числом m регрессоров.

В [17] для контроля значимости оценок линейной регрессии задача (4)–(8) была расширена следующими ограничениями:

$$-(1-\delta_{q_{ki}})\cdot M \leqslant \sum_{j=1}^{l-1} r_{x_{q_{ki}}x_{q_{kj}}}\cdot\beta^{\bullet}_{kj} - r_{yx_{q_{ki}}} \leqslant (1-\delta_{q_{ki}})\cdot M, \qquad k=\overline{1,l},\ i=\overline{1,l-1}, \tag{9}$$

$$-\delta_{q_{ki}}\cdot M \leqslant \beta^{\bullet}_{ki} \leqslant \delta_{q_{ki}}\cdot M, \qquad k=\overline{1,l},\ i=\overline{1,l-1}, \tag{10}$$

$$\sum_{j=1}^{l} r_{yx_j}\cdot\beta_j - \sum_{j=1}^{l-1} r_{yx_{q_{kj}}}\cdot\beta^{\bullet}_{kj} \geqslant (1-\sum_{j=1}^{l} r_{yx_j}\cdot\beta_j)\cdot T^* - (1-\delta_k)\cdot M, \qquad k=\overline{1,l}, \tag{11}$$

где q_{ij} — элементы матрицы $Q_{l\times(l-1)}$, полученной путем вычеркивания главной диагонали из квадратной матрицы $\begin{pmatrix} 1 & 2 & \dots & l \\ 1 & 2 & \dots & l \\ \dots & \dots & \dots & \dots \\ 1 & 2 & \dots & l \end{pmatrix}$, $\beta^{\bullet}_{kj}$ — неизвестный j-й коэффициент в k-й стандартизованной линейной регрессии переменной $y^{\bullet}$ от всех предикторов, кроме $x^{\bullet}_k$, $T^* = \dfrac{t^2_{\text{крит}}(\alpha, n-m-1)}{n-m-1}$.

Решение задачи (4)–(11) позволяет идентифицировать оптимальную по критерию R^2 линейную регрессию с фиксированным числом m значимых для уровня значимости α регрессоров. Линейные ограничения (11) сформированы на основе (3).

Перейдем теперь к формализации задачи построения линейной регрессии не с фиксированным, а с оптимальным числом регрессоров. Сложность здесь состоит в том, что в ограничениях (11) число T^* должно меняться в зависимости от количества переменных в модели. Для решения этой проблемы воспользуемся приемом, предложенным в [20] для контроля значимости регрессии в целом по критерию Фишера.

Введем дополнительные булевы переменные $\rho_j,\ j=\overline{1,m^*}$, по правилу:

$$\rho_j = \begin{cases} 1, & \text{если } \sum_{k=1}^{l}\delta_k = j, \\ 0, & \text{в противном случае,} \end{cases}$$

где $m^* = \min\{l, n-1\}$. Тем самым только одна из этих булевых переменных равна “1” — та, у которой индекс совпадает с числом отобранных предикторов. В этой связи будут справедливы следующие линейные ограничения:

$$-(1-\rho_j)\cdot M \leqslant \sum_{k=1}^{l}\delta_k - j \leqslant (1-\rho_j)\cdot M, \qquad j=\overline{1,m^*}, \tag{12}$$

$$\rho_j \in \{0,1\}, \qquad j=\overline{1,m^*}, \tag{13}$$

$$\sum_{j=1}^{m^*}\rho_j = 1. \tag{14}$$

Пусть $T_i = \dfrac{t^2_{\text{крит}}(\alpha, n-i-1)}{n-i-1}$, $i=\overline{1,m^*}$ — известные числа, найденные с использованием таблицы критических значений t-критерия Стьюдента. Тогда скорректируем ограничения (11), заменив в них фиксированное число T^* числами T_i и добавив компонент $(1-\rho_i)\cdot M$:

$$\sum_{j=1}^{l} r_{yx_j}\cdot\beta_j - \sum_{j=1}^{l-1} r_{yx_{q_{kj}}}\cdot\beta^{\bullet}_{kj} \geqslant (1-\sum_{j=1}^{l} r_{yx_j}\cdot\beta_j)\cdot T_i - (1-\delta_k)\cdot M - (1-\rho_i)\cdot M, \quad k=\overline{1,l},\ i=\overline{1,m^*}. \tag{15}$$

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

Как можно видеть, проверка значимости МНК-оценки в линейной регрессии включается только в одном случае — если в (15) $\rho_i = 1$ и $\delta_k = 1$, т.е. если число отобранных предикторов равно i и среди них имеется переменная под номером k. В противном случае неравенства (15) выполняются для любых значений коэффициентов стандартизованных регрессий.

Решение задачи (4)–(7), (9), (10), (12)–(15) позволяет идентифицировать оптимальную по критерию R^2 структуру линейной регрессии со всеми значимыми для уровня значимости α регрессорами.

Для согласованности знаков оценок линейной регрессии со знаками соответствующих коэффициентов корреляции r_{yx_j} заменим в задаче ЧБЛП, как это сделано в [16], условия (6) следующими линейными ограничениями:

$$0 \leqslant \beta_j \leqslant \delta_j \cdot M, \qquad j \in J^+, \tag{16}$$

$$-\delta_j \cdot M \leqslant \beta_j \leqslant 0, \qquad j \in J^-, \tag{17}$$

где J^+, J^- — индексные множества, элементы которых удовлетворяют условиям $r_{yx_j} > 0$ и $r_{yx_j} < 0$ соответственно.

Для контроля мультиколлинеарности введем следующие ограничения [16]:

$$|r_{x_i x_j}| \cdot (\delta_i + \delta_j - 1) \leqslant r, \qquad i = \overline{1, l-1},\ j = \overline{i+1, l}, \tag{18}$$

где число r для незначимости всех коэффициентов интеркорреляции, как следует из (2), нужно принять равным $\dfrac{t_{\text{крит}}(\alpha, n-2)}{\sqrt{n-2+t^2_{\text{крит}}(\alpha, n-2)}}$.

Пусть экспертами из конкретной предметной области предварительно были согласованы по смыслу знаки всех коэффициентов корреляции r_{yx_j}, $j = \overline{1, l}$, с направлением влияния предикторов на объясняемую переменную y. Тогда решение задачи ЧБЛП (4), (5), (7), (9), (10), (12)–(18) позволяет идентифицировать оптимальную по критерию R^2 структуру линейной регрессии с согласованными по направлению влияния на y оценками со всеми значимыми для уровня значимости α регрессорами и незначимыми коэффициентами интеркорреляции. Тем самым реализуется метод построения линейной регрессии, удовлетворяющей сформулированному выше вероятностно-статистическому определению ВИЛинР.

Задача (4), (5), (7), (9), (10), (12)–(18) содержит $(l + m^*)$ булевых переменных, параметры r и T_i, $i = \overline{1, m^*}$, зависящие от уровня значимости α, а также параметр M. Ее решение гарантирует, что все оценки линейной регрессии будут строго значимы для уровня α. Будем говорить далее, что эта задача реализует строгий метод построения ВИЛинР. Тогда введем задачу (4), (5), (7), (9)–(11), (16)–(18), реализующую нестрогий метод построения ВИЛинР. Она включает всего l булевых переменных и параметры r, M и $T^* = \dfrac{t^2_{\text{крит}}(\alpha, n-2)}{n-2}$. Решение этой задачи с ослабленными ограничениями на значимость оценок гарантирует лишь то, что для заданного уровня значимости α все наблюдаемые значения t-критерия по модулю будут не меньше, чем его критическое значение для модели с одним регрессором (при $m = 1$). При большом числе степеней свободы $(n - m - 1)$ критические значения t-критерия мало различаются для конкретного уровня α, поэтому строгий и нестрогий методы должны часто давать одинаковый результат.

Задачи ЧБЛП, реализующие строгий и нестрогий метод построения ВИЛинР, будут разрешимы, если выполняется условие $\alpha \geqslant \alpha^{\#}$, где $\alpha^{\#}$ — выбранный уровень, подтверждающий значимость наибольшего по модулю коэффициента корреляции из набора r_{yx_j}, $j = \overline{1, l}$.

3. Вычислительные эксперименты. Целью вычислительных экспериментов было тестирование эффективности предложенного метода (строгого и нестрогого) по сравнению с МВВР. Для проведения экспериментов использовались встроенные в пакет Gretl статистические данные *data7-11* о продаже 59-ти домов в Ла-Хойе (северо-западный район калифорнийского города Сан-Диего) и Юниверсити Сити. Эти данные включают следующие переменные:

price — цена дома в тысячах долларов;
age — срок эксплуатации дома в годах;
aircon — наличие кондиционера в доме (“1” — да, “0” — нет);
baths — количество ванных комнат;
bedrms — количество спален;
cond — оценка состояния дома от плохого (1) до отличного (6);
corner — расположение дома на угловом участке (“1” — да, “0” — нет);

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

culd — расположение дома в тупике (“1” — да, “0” — нет);
dish — наличие в доме посудомоечной машины (“1” — да, “0” — нет);
fence — наличие у дома забора (“1” — да, “0” — нет);
firepl — количество каминов;
floors — количество этажей;
garage — количество мест для машин в гараже;
irreg — наличие у земельного участка нестандартной геометрической формы (“1” — да, “0” — нет);
lajolla — расположение дома в Ла-Хойе (“1” — да, “0” — нет);
lndry — наличие отдельного пространства для стирки и сушки белья (“1” — да, “0” — нет);
patio — количество отдельных патио (внутренних дворов);
pool — наличие в доме бассейна (“1” — да, “0” — нет);
rooms — количество комнат, не считая спален и ванных;
sprink — наличие в доме системы полива (“1” — да, “0” — нет);
sqft — жилая площадь в кв. футах;
view — наличие с территории дома живописного обзора (“1” — да, “0” — нет);
yard — площадь двора в кв. футах.

Для удобства обозначим переменную *price* как y, *age* — x_1, *aircon* — x_2, *baths* — x_3, *bedrms* — x_4, *cond* — x_5, *corner* — x_6, *culd* — x_7, *dish* — x_8, *fence* — x_9, *firepl* — x_{10}, *floors* — x_{11}, *garage* — x_{12}, *irreg* — x_{13}, *lajolla* — x_{14}, *lndry* — x_{15}, *patio* — x_{16}, *pool* — x_{17}, *rooms* — x_{18}, *sprink* — x_{19}, *sqft* — x_{20}, *view* — x_{21}, *yard* — x_{22}.

Предварительно был проведен анализ значений коэффициентов корреляции объясняемой переменной y с предикторами. Была выявлена некоторая противоречивость в связях цены y с x_1 и x_9, поскольку $r_{yx_1} = 0.0293$, а $r_{yx_9} = -0.2382$. Ошибочно утверждать, что чем выше срок эксплуатации дома x_1, тем выше его цена. Новые дома обычно продаются дорого, но исторический особняк в хорошем районе может обойтись дороже. Также наличие забора x_9 не однозначно приводит к снижению цены дома. Несмотря на это, было принято решение ни одну переменную не исключать, чтобы не снизить вычислительную сложность задач ОИР.

Для проведения экспериментов использовался компьютер с процессором AMD Ryzen 3 4300U with Radeon Graphics (2.70 ГГц) и 16 Гб оперативной памяти.

Для решения задач ОИР с помощью МВВР на языке программирования hansl эконометрического пакета Gretl была разработана специальная программа. Она функционирует по алгоритму 1.

Алгоритм 1. Алгоритм решения задачи ОИР с помощью МВВР
Algorithm 1. Algorithm for solving the subset selection problem using generating all subsets method

for $m = 1 \dots 22$ **do**
  формирование матрицы *Glavn* сочетаний переменных
  включение счетчика времени *time*
  for $j = 1 \dots C_{22}^m$ **do**
    формирование матрицы R_{xx} и вектора R_{yx}
    if $|r_{x_{j_1}x_{j_2}}| > r$ **then** переход к следующему j
    вычисление оценок $\beta = R_{xx}^{-1} \cdot R_{yx}$
    if $r_{yx_k} \cdot \beta_k < 0$ **then** переход к следующему j
    вычисление коэффициента детерминации $R^2 = R_{yx}^T \cdot \beta$
    if $|t_{\beta_k}| < T_m$ **then** переход к следующему j
  end for
  выбор лучшей модели с m регрессорами по R^2
  выключение счетчика по времени *time*
end for

В алгоритме 1 наблюдаемое значение t-критерия k-го коэффициента регрессии находится по формуле $t_{\beta_k} = \beta_k \Big/ \sqrt{\dfrac{1-R^2}{59-(m+1)} \cdot (R_{xx}^{-1})_{kk}}$. Для работы с программой предварительно нужно установить уровень значимости α_T для проверки значимости оценок регрессии и уровень α_r для проверки значимости коэффициентов интеркорреляции. Зависящие от них значения параметров r и T_m, $m = \overline{1,22}$, определяются в программе автоматически с помощью встроенной функции *critical*. В результате работы программы

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

находится оптимальная структура линейной регрессии с заданными свойствами и время, затраченное на ее поиск. Заметим, что в алгоритме 1 не фиксируется время, затраченное на формирование матриц сочетаний переменных. Главная особенность алгоритма 1 в том, что при проверке требований к текущей регрессии в случае возникновения хотя бы одного нарушения сразу осуществляется прерывание цикла и начинается проверка следующей модели. Это значительно сокращает общее время решения задачи.

Для автоматического формирования по выборке данных предложенных задач ЧБЛП для строгого и нестрогого метода была усовершенствована и оснащена новыми возможностями программа ВИнтер-2 [16]. В ней в обоих случаях предварительно нужно задать достаточно большое число M, параметр r и выбрать один из пяти уровней значимости α_T для контроля значимости оценок — 0.2, 0.15, 0.1, 0.05 или 0.01. Критические значения t-критерия при числе степеней свободы менее 500 уже встроены в программу. Во всех остальных случаях критическое значение берется как при числе степеней свободы 500. Далее при выборе в ВИнтер-2 метода построения ВИЛинР формируется задача ЧБЛП для решателя COPT [21], управление которым осуществляется средствами языка программирования Python.

Стоит отметить, что в ВИнтер-2 большое число M устанавливается пользователем только для ограничений (9), (10)–(12) и (15). Достаточные значения этих чисел для ограничений (5), (16) и (17) находятся в программе автоматически по алгоритму, описанному в [16].

Вычислительные эксперименты проводились для пяти категорий уровня значимости α_T (0.2, 0.15, 0.1, 0.05 и 0.01) и для тех же категорий уровня значимости α_r. Каждая из 25 задач ОИР решалась тремя методами — МВВР, предложенными строгим и нестрогим методами. При настройке ВИнтер-2 параметр M был взят равным 100. Результаты проведенных экспериментов представлены в табл. 1. В ней каждому решению ставится в соответствие: состав отобранных переменных, значение R^2, минимальное по модулю наблюдаемое значение $|t_\beta|_{\min}$ t-критерия отобранных коэффициентов регрессии и его критическое значение $t_{\text{крит}}$, максимальное по модулю значение $|r_{xx}|_{\max}$ коэффициентов интеркорреляции отобранных предикторов и верхняя граница r его значимости, время t решения задачи в секундах.

Таблица 1. Результаты вычислительных экспериментов по выборке объема $n = 59$

Table 1. Results of computational experiments on a sample of size $n = 59$

№	α_T	α_r	Метод Method	Состав Components	R^2	$\lvert t_\beta\rvert_{\min}$	$t_{\text{крит}}$	$\lvert r_{xx}\rvert_{\max}$	r	t, с
1	0.2	0.2	МВВР	$x_3, x_{13}, x_{14}, x_{17}$	0.65423	1.886	1.2974	0.1038	0.1692	100.7
			Строгий	$x_3, x_{13}, x_{14}, x_{17}$	0.65423	1.886	1.2974	0.1038	0.1692	34.62
			Нестрогий	$x_3, x_{13}, x_{14}, x_{17}$	0.65423	1.886	1.2974	0.1038	0.1692	23.21
2	0.2	0.15	МВВР	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	1.2974	0.1697	0.1897	106.33
			Строгий	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	1.2974	0.1697	0.1897	17.36
			Нестрогий	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	1.2974	0.1697	0.1897	11.00
3	0.2	0.1	МВВР	$x_3, x_{13}, x_{14}, x_{17}, x_{19}$	0.67297	1.472	1.2977	0.1982	0.2162	117.36
			Строгий	$x_3, x_{13}, x_{14}, x_{17}, x_{19}$	0.67297	1.472	1.2977	0.1982	0.2162	4.41
			Нестрогий	$x_3, x_{13}, x_{14}, x_{17}, x_{19}$	0.67297	1.472	1.2977	0.1982	0.2162	5.50
4	0.2	0.05	МВВР	$x_4, x_{10}, x_{13}, x_{14}, x_{16}, x_{17}, x_{19}$	0.69619	1.591	1.2983	0.2221	0.2563	128.44
			Строгий	$x_4, x_{10}, x_{13}, x_{14}, x_{16}, x_{17}, x_{19}$	0.69619	1.591	1.2983	0.2221	0.2563	21.51
			Нестрогий	$x_4, x_{10}, x_{13}, x_{14}, x_{16}, x_{17}, x_{19}$	0.69619	1.591	1.2983	0.2221	0.2563	9.40
5	0.2	0.01	МВВР	$x_3, x_{10}, x_{13}, x_{14}, x_{17}, x_{19}, x_{22}$	0.74581	1.702	1.2983	0.3158	0.3328	151.75
			Строгий	$x_3, x_{10}, x_{13}, x_{14}, x_{17}, x_{19}, x_{22}$	0.74581	1.702	1.2983	0.3158	0.3328	43.62
			Нестрогий	$x_3, x_{10}, x_{13}, x_{14}, x_{15}, x_{17}, x_{21}, x_{22}$	0.74870	1.250	1.2987	0.3158	0.3328	26.49
6	0.15	0.2	МВВР	$x_3, x_{13}, x_{14}, x_{17}$	0.65423	1.886	1.4603	0.1038	0.1692	100.91
			Строгий	$x_3, x_{13}, x_{14}, x_{17}$	0.65423	1.886	1.4603	0.1038	0.1692	34.82
			Нестрогий	$x_3, x_{13}, x_{14}, x_{17}$	0.65423	1.886	1.4603	0.1038	0.1692	23.45

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

Таблица 1. Результаты вычислительных экспериментов по выборке объема $n = 59$

Table 1. Results of computational experiments on a sample of size $n = 59$

№	α_T	α_r	Метод Method	Состав Components	R^2	$\lvert t_\beta \rvert_{\min}$	$t_{\text{крит}}$	$\lvert r_{xx} \rvert_{\max}$	r	t, с
7	0.15	0.15	МВВР	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	1.4603	0.1697	0.1897	105.79
			Строгий	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	1.4603	0.1697	0.1897	17.36
			Нестрогий	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	1.4603	0.1697	0.1897	3.84
8	0.15	0.1	МВВР	$x_3, x_{13}, x_{14}, x_{17}, x_{19}$	0.67297	1.472	1.4607	0.1982	0.2162	117.55
			Строгий	$x_3, x_{13}, x_{14}, x_{17}, x_{19}$	0.67297	1.472	1.4607	0.1982	0.2162	4.19
			Нестрогий	$x_3, x_{13}, x_{14}, x_{17}, x_{19}$	0.67297	1.472	1.4607	0.1982	0.2162	5.25
9	0.15	0.05	МВВР	$x_4, x_{10}, x_{13}, x_{14}, x_{16}, x_{17}, x_{19}$	0.69619	1.591	1.4615	0.2221	0.2563	128.73
			Строгий	$x_4, x_{10}, x_{13}, x_{14}, x_{16}, x_{17}, x_{19}$	0.69619	1.591	1.4615	0.2221	0.2563	19.08
			Нестрогий	$x_4, x_{10}, x_{13}, x_{14}, x_{16}, x_{17}, x_{19}$	0.69619	1.591	1.4615	0.2221	0.2563	10.20
10	0.15	0.01	МВВР	$x_3, x_{10}, x_{13}, x_{14}, x_{17}, x_{19}, x_{22}$	0.74581	1.702	1.4615	0.3158	0.3328	151.99
			Строгий	$x_3, x_{10}, x_{13}, x_{14}, x_{17}, x_{19}, x_{22}$	0.74581	1.702	1.4615	0.3158	0.3328	44.91
			Нестрогий	$x_3, x_{10}, x_{13}, x_{14}, x_{17}, x_{19}, x_{22}$	0.74581	1.702	1.4615	0.3158	0.3328	25.30
11	0.1	0.2	МВВР	$x_3, x_{13}, x_{14}, x_{17}$	0.65423	1.886	1.6735	0.1038	0.1692	102.02
			Строгий	$x_3, x_{13}, x_{14}, x_{17}$	0.65423	1.886	1.6735	0.1038	0.1692	34.55
			Нестрогий	$x_3, x_{13}, x_{14}, x_{17}$	0.65423	1.886	1.6735	0.1038	0.1692	23.32
12	0.1	0.15	МВВР	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	1.6735	0.1697	0.1897	106.23
			Строгий	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	1.6735	0.1697	0.1897	18.29
			Нестрогий	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	1.6735	0.1697	0.1897	4.65
13	0.1	0.1	МВВР	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	1.6735	0.1697	0.2162	117.73
			Строгий	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	1.6735	0.1697	0.2162	4.53
			Нестрогий	$x_4, x_{10}, x_{13}, x_{14}, x_{17}, x_{19}$	0.66832	1.666	1.6746	0.1982	0.2162	5.58
14	0.1	0.05	МВВР	$x_4, x_{10}, x_{14}, x_{16}, x_{17}, x_{19}$	0.68108	2.145	1.6746	0.2221	0.2563	129.16
			Строгий	$x_4, x_{10}, x_{14}, x_{16}, x_{17}, x_{19}$	0.68108	2.145	1.6746	0.2221	0.2563	15.91
			Нестрогий	$x_4, x_{10}, x_{13}, x_{14}, x_{16}, x_{17}, x_{19}$	0.69619	1.592	1.6775	0.2221	0.2563	12.37
15	0.1	0.01	МВВР	$x_3, x_{10}, x_{13}, x_{14}, x_{17}, x_{19}, x_{22}$	0.74581	1.702	1.6752	0.3158	0.3328	152.08
			Строгий	$x_3, x_{10}, x_{13}, x_{14}, x_{17}, x_{19}, x_{22}$	0.74581	1.702	1.6752	0.3158	0.3328	39.90
			Нестрогий	$x_3, x_{10}, x_{13}, x_{14}, x_{17}, x_{19}, x_{22}$	0.74581	1.702	1.6752	0.3158	0.3328	25.48
16	0.05	0.2	МВВР	$x_8, x_{10}, x_{14}, x_{19}$	0.63760	2.033	2.0048	0.1574	0.1692	100.63
			Строгий	$x_8, x_{10}, x_{14}, x_{19}$	0.63760	2.033	2.0048	0.1574	0.1692	34.28
			Нестрогий	$x_8, x_{10}, x_{14}, x_{19}$	0.63760	2.033	2.0048	0.1574	0.1692	23.18
17	0.05	0.15	МВВР	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	2.0048	0.1697	0.1897	107.21
			Строгий	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	2.0048	0.1697	0.1897	17.65
			Нестрогий	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	2.0048	0.1697	0.1897	5.37
18	0.05	0.1	МВВР	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	2.0048	0.1697	0.2162	118.43
			Строгий	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	2.0048	0.1697	0.2162	4.02
			Нестрогий	$x_3, x_{14}, x_{17}, x_{19}$	0.65960	2.113	2.0048	0.1697	0.2162	6.01
19	0.05	0.05	МВВР	$x_4, x_{10}, x_{14}, x_{16}, x_{17}, x_{19}$	0.68108	2.145	2.0066	0.2221	0.2563	129.1
			Строгий	$x_4, x_{10}, x_{14}, x_{16}, x_{17}, x_{19}$	0.68108	2.145	2.0066	0.2221	0.2563	17.72
			Нестрогий	$x_4, x_{10}, x_{14}, x_{16}, x_{17}, x_{19}$	0.68108	2.145	2.0066	0.2221	0.2563	13.83

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

Таблица 1. Результаты вычислительных экспериментов по выборке объема $n = 59$

Table 1. Results of computational experiments on a sample of size $n = 59$

№	α_T	α_r	Метод Method	Состав Components	R^2	$\lvert t_\beta\rvert_{\min}$	$t_{\text{крит}}$	$\lvert r_{xx}\rvert_{\max}$	r	t, с
20	0.05	0.01	МВВР	$x_3, x_{10}, x_{13}, x_{14}, x_{17}, x_{22}$	0.73137	2.014	2.0066	0.3158	0.3328	153.26
			Строгий	$x_3, x_{10}, x_{13}, x_{14}, x_{17}, x_{22}$	0.73137	2.014	2.0066	0.3158	0.3328	55.67
			Нестрогий	$x_3, x_{10}, x_{13}, x_{14}, x_{17}, x_{22}$	0.73137	2.014	2.0066	0.3158	0.3328	51.95
21	0.01	0.2	МВВР	x_3, x_{14}	0.59646	3.944	2.6665	0.0656	0.1692	100.31
			Строгий	x_3, x_{14}	0.59646	3.944	2.6665	0.0656	0.1692	34.42
			Нестрогий	x_3, x_{14}	0.59646	3.944	2.6665	0.0656	0.1692	21.59
22	0.01	0.15	МВВР	x_3, x_{14}	0.59646	3.944	2.6665	0.0656	0.1897	105.74
			Строгий	x_3, x_{14}	0.59646	3.944	2.6665	0.0656	0.1897	7.20
			Нестрогий	x_3, x_{14}	0.59646	3.944	2.6665	0.0656	0.1897	5.64
23	0.01	0.1	МВВР	x_3, x_{14}	0.59646	3.944	2.6665	0.0656	0.2162	117.38
			Строгий	x_3, x_{14}	0.59646	3.944	2.6665	0.0656	0.2162	6.42
			Нестрогий	x_3, x_{14}	0.59646	3.944	2.6665	0.0656	0.2162	5.98
24	0.01	0.05	МВВР	x_{10}, x_{14}, x_{15}	0.62011	2.680	2.6682	0.2354	0.2563	128.14
			Строгий	x_{10}, x_{14}, x_{15}	0.62011	2.680	2.6682	0.2354	0.2563	39.31
			Нестрогий	x_{10}, x_{14}, x_{15}	0.62011	2.680	2.6682	0.2354	0.2563	32.66
25	0.01	0.01	МВВР	x_{14}, x_{20}	0.66830	5.573	2.6665	0.2635	0.3328	151.76
			Строгий	x_{14}, x_{20}	0.66830	5.573	2.6665	0.2635	0.3328	220.24
			Нестрогий	x_{14}, x_{20}	0.66830	5.573	2.6665	0.2635	0.3328	168.08

По табл. 1 можно сделать следующие выводы.

1. На решение всех 25 задач с помощью МВВР ушло в общем 3039.86 с, а с помощью строгого метода — 778.11 с. При этом результаты решения задач обоими методами полностью совпадают, для всех построенных регрессий справедливы неравенства $|t_\beta|_{\min} > t_{\text{крит}}$ и $|r_{xx}|_{\max} < r$. Только в одном случае из 25 (задача № 25 при $\alpha_T = 0.01$ и $\alpha_r = 0.01$) строгий метод оказался медленнее, чем МВВР. В остальных 24 задачах строгий метод оказался быстрее на 63.7%–96.6%.
2. На решение всех 25 задач с помощью нестрогого метода ушло еще меньше времени — 550.18 с. При этом в трех случаях из 25 результаты решения расходятся с МВВР и строгим методом — задачи № 5, № 13 и № 14. Так, в задаче № 5 (при $\alpha_T = 0.2$ и $\alpha_r = 0.01$) минимальное по модулю наблюдаемое значение $|t_\beta|_{\min}$ t-критерия меньше критического ($1.250 < 1.2987$), в задаче № 13 — $1.666 < 1.6746$, а в задаче № 14 — $1.592 < 1.6775$. Тем самым с формальной стороны в этих трех линейных регрессиях присутствуют незначимые коэффициенты, но наблюдаемые значения их t-критерия настолько близки к области отклонения нулевой гипотезы, что эти коэффициенты можно считать практически значимыми. К тому же нестрогий метод в 21-м случае из 25 оказался эффективнее строго метода и в 24 случаях, кроме той же задачи № 25, быстрее МВВР.

Стоит подчеркнуть, что время решения предложенных задач ЧБЛП зависит от выбранного значения параметра M. Например, при изменении его значения 100 на 120 время построения ВИЛинР нестрогим методом при $\alpha_T = 0.01$ и $\alpha_r = 0.01$ уменьшится со 168.08 с до 141.08 с, т.е. станет меньше 151.76 с — времени на реализацию МВВР. Однако каких-либо строгих правил выбора параметра M для снижения времени решения задач ЧБЛП в науке пока не существует.

Далее было решено протестировать эффективность предложенного в данной статье метода на примере обработки массива данных существенно большего объема. Для этого была использована хранящаяся в открытом доступе выборка [22] о годах выпуска песен, содержащая $n = 515345$ наблюдений и $l = 90$ объясняющих переменных. К сожалению, построенная по этим данным линейная регрессия имеет низкое значение критерия R^2 — всего 0.237. Поэтому было решено использовать в качестве зависимой переменной одну из 90 объясняющих переменных. Выбор четвертой объясняющей переменной обеспечил значение критерия R^2 линейной регрессии, построенной на оставшихся 89 переменных, равное 0.86579. Заметим, что не ставилась цель получить интерпретируемые модели, поэтому анализ корректности связей между переменными не проводился.

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

Особенность используемой выборки в том, что число степеней свободы $(n - m - 1)$ чрезвычайно велико для любого m. Поэтому величина коэффициента T^* в ограничениях (11) будет близка к нулю для любой из пяти категорий уровня значимости α_T (0.2, 0.15, 0.1, 0.05 и 0.01). В этой связи ограничения (11) практически всегда будут выполняться, следовательно, эффективнее будет ими, а также ограничениями (9) и (10), пренебречь, решая задачу ЧБЛП (4), (5), (7), (16)–(18).

Предварительно 89 объясняющих переменных были упорядочены по убыванию модулей их коэффициентов корреляции с зависимой переменной. Вычислительные эксперименты проводились для наборов из l объясняющих переменных (30, 40, 50, 60, 70, 80, 89) и четырех значений параметра r (0.1, 0.15, 0.2, 0.25). В каждом случае решалась задача ЧБЛП (4), (5), (7), (16)–(18). Ограничение на время решения задач в COPT было установлено 1800 с. Результаты проведенных экспериментов представлены в табл. 2. В ней каждому решению ставится в соответствие: число отобранных переменных, значение R^2, время t решения задачи в секундах. Оценки всех представленных в табл. 2 моделей оказались значимыми по t-критерию Стьюдента для уровня $\alpha = 0.01$.

Таблица 2. Результаты вычислительных экспериментов по выборке объема $n = 515345$
Table 2. Results of computational experiments on a sample of size $n = 515345$

l	r			
	0.1	0.15	0.2	0.25
30	4 0.37498 0.29	7 0.42180 1.03	8 0.56136 1.86	10 0.56467 4.30
40	5 0.38543 0.67	8 0.42785 4.44	10 0.56281 8.99	15 0.60896 31.80
50	6 0.39259 0.80	12 0.45097 12.78	12 0.56497 50.93	19 0.62982 271.16
60	7 0.39766 2.27	13 0.45251 23.11	14 0.56829 313.44	24 0.63964 1800
70	7 0.39766 6.59	13 0.45251 74.89	16 0.56869 1758.32	28 0.63377 1800
80	7 0.39766 31.06	14 0.45889 486.55	17 0.57120 1800	30 0.64102 1800
89	7 0.39766 69.14	15 0.46159 1253.02	25 0.54459 1800	32 0.60748 1800

По табл. 2 можно сделать следующие выводы.

1. В 6-ти случаях из 28-ми не удалось получить оптимальное решение задачи ЧБЛП (4), (5), (7), (16)–(18) в установленный лимит времени. Но при этом были найдены субоптимальные решения с довольно высоким для выборки объема $n = 515345$ значением критерия R^2.
2. Для каждого l зафиксирован рост времени решения задач ЧБЛП при увеличении верхней границы r модулей коэффициентов интеркорреляций. Также для каждого r зафиксирован рост времени решения задач ЧБЛП при увеличении числа l исходных объясняющих переменных.

4. Заключение. Таким образом, в данной работе предложены две вариации нового метода построения ВИЛинР — строгая и нестрогая. Для реализации метода была усовершенствована и оснащена новыми возможностями программа ВИнтер-2. К тому же разработана программа, реализующая построение ВИЛинР с помощью МВВР. В результате проведенных вычислительных экспериментов новый метод в большинстве случаев оказался эффективнее, чем МВВР. При этом линейные регрессии, полученные нестрогим методом, в основном практически не отличаются по структуре от моделей, построенных, главным образом за большее время, по строгому методу.

Подчеркнем следующие важные аспекты.

1. Разработанное программное обеспечение можно использовать для построения ВИЛинР при обработке данных из самых разных предметных областей, контролируя структуру модели, согласованность ее коэффициентов со смыслом факторов, значимость оценок и мультиколлинеарность.
2. Требование незначимости всех коэффициентов интеркорреляции для входящих в линейную регрессию предикторов по t-критерию Стьюдента логично, но не гарантирует обязательную ликвидацию мультиколлинеарности. Поэтому число r в ограничениях (18) можно выбирать из диапазона [0,1] на свое усмотрение. На разрешимости задач ЧБЛП это никак не сказывается.
3. От выбора большого числа M зависит скорость решения предложенных задач ЧБЛП. И хотя строгих правил его выбора нет, есть некоторые специальные приемы, например техника дополнения задач отложенными ограничениями, которая может позитивно влиять на эффективность решения. Этот вопрос требует дополнительных исследований.

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

Список литературы

1. *Doshi-Velez F., Kim B.* Towards a rigorous science of interpretable machine learning // arXiv preprint. 2017. doi 10.48550/arXiv.1702.08608.
2. *Molnar C.* Interpretable machine learning. A Guide for Making Black Box Models Explainable. 2020. https://christophm.github.io/interpretable-ml-book/. (Дата обращения: 21 ноября 2025).
3. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998.
4. *Miller A.* Subset selection in regression. New York: Chapman and Hall/CRC, 2002. doi 10.1201/9781420035933.
5. *Стрижов В.В., Крымова Е.А.* Методы выбора регрессионных моделей. М.: Вычислительный центр РАН, 2010.
6. *Koch T., Berthold T., Pedersen J., Vanaret C.* Progress in mathematical programming solvers from 2001 to 2020 // EURO Journal on Computational Optimization. 2022. **10**. Article Number 100031. doi 10.1016/j.ejco.2022.100031.
7. *Konno H., Yamamoto R.* Choosing the best set of variables in regression analysis using integer programming // Journal of Global Optimization. 2009. **44**, N 2. 273–282. doi 10.1007/s10898-008-9323-9.
8. *Miyashiro R., Takano Y.* Mixed integer second-order cone programming formulations for variable selection in linear regression // European Journal of Operational Research. 2015. **247**, N 3. 721–731. doi 10.1016/j.ejor.2015.06.081.
9. *Miyashiro R., Takano Y.* Subset selection by Mallows' C_p: a mixed integer programming approach // Expert Systems with Applications. 2015. **42**, N 1. 325–331. doi 10.1016/j.eswa.2014.07.056.
10. *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. **6**, № 1. 118–127. https://moitvivt.ru/ru/journal/pdf?id=434. (Дата обращения: 21 ноября 2025).
11. *Shrestha N.* Detecting multicollinearity in regression analysis // American Journal of Applied Mathematics and Statistics. 2020. **8**, N 2. 39–42. doi 10.12691/ajams-8-2-1.
12. *Aslam M.* The t–test of a regression coefficient for imprecise data // Hacettepe Journal of Mathematics and Statistics. 2024. **53**, N 4. 1130–1140. doi 10.15672/hujms.1342344.
13. *Горбач А.Н., Цейтлин Н.А.* Покупательское поведение: анализ спонтанных последовательностей и регрессионных моделей в маркетинговых исследованиях. Киев: Освита Украины, 2011.
14. *Chung S., Park Y.W., Cheong T.* A mathematical programming approach for integrated multiple linear regression subset selection and validation // Pattern Recognition. 2020. **108**. Article Number 107565. doi 10.1016/j.patcog.2020.107565.
15. *Bertsimas D., Li M.L.* Scalable holistic linear regression // Operations Research Letters. 2020. **48**, N 3. 203–208. doi 10.1016/j.orl.2020.02.008.
16. *Базилевский М.П.* Сравнительный анализ эффективности методов построения вполне интерпретируемых линейных регрессионных моделей // Моделирование и анализ данных. 2023. **13**, № 4. 59–83. https://psyjournals.ru/journals/mda/archive/2023_n4/mda_2023_n4_Bazilevskiy.pdf. (Дата обращения: 21 ноября 2025).
17. *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. 5–16. https://journals.vsu.ru/sait/article/view/3731/3801. (Дата обращения: 21 ноября 2025).
18. *Ферстер Э., Ренц Б.* Методы корреляционного и регрессионного анализа. М.: Финансы и статистика, 1983.
19. *Елисеева И.И., Курышева С.В., Костеева Т.В. и др.* Эконометрика. М.: Финансы и статистика, 2007.
20. *Базилевский М.П.* Оптимизационные задачи отбора информативных регрессоров в линейной регрессии с контролем ее значимости по критерию Фишера // Известия Самарского научного центра Российской академии наук. 2024. **26**, № 6. 200–207. https://ssc.smr.ru/media/journals/izvestia/2024/2024_6_200_207.pdf. (Дата обращения: 21 ноября 2025).

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

21. *Ge D., Huangfu Q., Wang Z., Wu J., Ye Y.* Cardinal Optimizer (COPT) user guide. https://guide.coap.online/copt/en-doc. (Дата обращения: 21 ноября 2025).

22. UCI Machine Learning Repository. doi 10.24432/C50K61.

Получена *13 сентября 2025 г.*

Принята *18 ноября 2025 г.*

Опубликована *1 декабря 2025 г.*

Информация об авторе

Михаил Павлович Базилевский — к.т.н., доцент; Иркутский государственный университет путей сообщения, ул. Чернышевского, 15, 664074, Иркутск, Российская Федерация.

References

1. F. Doshi-Velez and B. Kim, "Towards a Rigorous Science of Interpretable Machine Learning," arXiv preprint. (2017). doi 10.48550/arXiv.1702.08608.
2. C. Molnar, *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. (2020). https://christophm.github.io/interpretable-ml-book/. Cited November 21, 2025.
3. S. A. Aivazjan and V. S. Mhitarjan, *Applied Statistics and Basics of Econometrics* (YUNITI, Moscow, 1998) [in Russian].
4. A. Miller, *Subset Selection in Regression* (Chapman and Hall/CRC, New York, 2002). doi 10.1201/9781420035933.
5. V. V. Strizhov and E. A. Krymova, *Methods Selection of Regression Models* (Comp. Cent. of RAS, Moscow, 2010) [in Russian].
6. T. Koch, T. Berthold, J. Pedersen, and C. Vanaret, "Progress in Mathematical Programming Solvers from 2001 to 2020," EURO J. Comp. Opt. **10**, Article Number 100031 (2022). doi 10.1016/j.ejco.2022.100031.
7. H. Konno and R. Yamamoto, "Choosing the Best Set of Variables in Regression Analysis Using Integer Programming," J. Glob. Opt. **44** (2), 273–282 (2009). doi 10.1007/s10898-008-9323-9.
8. R. Miyashiro and Y. Takano, "Mixed Integer Second-Order Cone Programming Formulations for Variable Selection in Linear Regression," Europ. J. Oper. Res. **247** (3), 721–731 (2015). doi 10.1016/j.ejor.2015.06.081.
9. R. Miyashiro and Y. Takano, "Subset Selection by Mallows' C_p: A Mixed Integer Programming Approach," Exp. Syst. Appl. **42** (1), 325–331 (2015). doi 10.1016/j.eswa.2014.07.056.
10. M. P. Bazilevskiy, "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," Mod. Opt. Inf. Tech. **6** (1), 118–127 (2018). https://moitvivt.ru/ru/journal/pdf?id=434. Cited November 21, 2025.
11. N. Shrestha, "Detecting Multicollinearity in Regression Analysis," Amer. J. Appl. Math. Stat. **8** (2), 39–42 (2020). doi 10.12691/ajams-8-2-1.
12. M. Aslam, "The T–Test of a Regression Coefficient for Imprecise Data," Hac. J. Math. Stat. **53** (4), 1130–1140 (2024). doi 10.15672/hujms.1342344.
13. A. N. Gorbach and N. A. Tseytlin, *Buying Behavior: Analysis of Spontaneous Sequences and Regression Models in Marketing Research* (Education of Ukraine, Kyiv, 2011) [in Russian].
14. S. Chung, Y. W. Park, and T. Cheong, "A Mathematical Programming Approach for Integrated Multiple Linear Regression Subset Selection and Validation," Pat. Recogn. **108**, Article Number 107565 (2020). doi 10.1016/j.patcog.2020.107565.
15. D. Bertsimas and M. L. Li, "Scalable Holistic Linear Regression," Oper. Res. Let. **48** (3), 203–208 (2020). doi 10.1016/j.orl.2020.02.008.
16. M. P. Bazilevskiy, "Comparative Analysis of the Effectiveness of Methods for Constructing Quite Interpretable Linear Regression Models," Mod. D. Anal. **13** (4), 59–83 (2023). https://psyjournals.ru/journals/mda/archive/2023_n4/mda_2023_n4_Bazilevskiy.pdf. Cited November 21, 2025.
17. M. P. Bazilevskiy, "Selection of Informative Regressors Significant by Student's T–Test in Regression Models Estimated Using OLS as a Partial Boolean Linear Programming Problem," Proc. VSU. Ser.: Syst. Anal. Inform. Tech. N 3, 5–16 (2021). https://journals.vsu.ru/sait/article/view/3731/3801. Cited November 21, 2025.

2025, **26** (4), 534–547. doi 10.26089/NumMet.v26r435

18. E. Ferster and B. Rentz, *Methods of Correlation and Regression Analysis* (Finance and Statistics, Moscow, 1983) [in Russian].
19. I. I. Eliseeva, S. V. Kurysheva, T. V. Kosteeva, et al., *Econometrics* (Finance and Statistics, Moscow, 2007) [in Russian].
20. M. P. Bazilevskiy, "Optimization Problems of Subsets Selection in Linear Regression with Control of Its Significance Using F–Test," Izv. RAS SamSC. **26** (6), 200–207 (2024). https://ssc.smr.ru/media/journals/izvestia/2024/2024_6_200_207.pdf. Cited November 21, 2025.
21. D. Ge, Q. Huangfu, Z. Wang, et al., Cardinal Optimizer (COPT) User Guide. https://guide.coap.online/copt/en-doc. Cited November 21, 2025.
22. UCI Machine Learning Repository. doi 10.24432/C50K61.

Received *September 13, 2025*
Accepted *November 18, 2025*
Published *December 1, 2025*

Information about the author

Mikhail P. Bazilevskiy — Ph.D., Associate Professor; Irkutsk State Transport University, ulitsa Chernyshevskogo, 15, 664074, Irkutsk, Russia.