

ОРИГИНАЛЬНАЯ СТАТЬЯ

УДК 519.862.6

<https://doi.org/10.26907/2541-7746.2025.4.627-640>**О разрешимости оптимизационной задачи построения
вполне интерпретируемых линейных регрессий****М.П. Базилевский***Иркутский государственный университет путей сообщения, г. Иркутск, Россия**mik2178@yandex.ru***Аннотация**

Статья посвящена совершенствованию технологии построения интерпретируемых регрессионных моделей, параметры которых оцениваются с помощью метода наименьших квадратов. Введено определение вполне интерпретируемой линейной регрессии. К ней предъявлены требования согласованности знаков оценок параметров содержательному смыслу переменных, значимости оценок, низкой степени мультиколлинеарности и высокого качества аппроксимации. Принадлежность модели к классу вполне интерпретируемых регрессий зависит от уровня значимости. В терминах аппарата частично-булевого линейного программирования, прогрессирующего за последние годы в вычислительном плане, сформулирована оптимизационная задача построения вполне интерпретируемых линейных регрессий с довольно большим количеством линейных ограничений. Доказана разрешимость этой задачи при определенных условиях. Предлагаемый математический аппарат может успешно применяться для обработки больших данных, поскольку число ограничений в сформулированной задаче, в отличие от существующих зарубежных аналогов, не зависит от объема выборки.

Ключевые слова: линейная регрессия, метод наименьших квадратов, интерпретируемость, мультиколлинеарность, значимость, задача частично-булевого линейного программирования, разрешимость

Для цитирования: Базилевский М.П. О разрешимости оптимизационной задачи построения вполне интерпретируемых линейных регрессий // Учен. зап. Казан. ун-та. Сер. Физ.-матем. науки. 2025. Т. 167, кн. 4. С. 627–640. <https://doi.org/10.26907/2541-7746.2025.4.627-640>.

ORIGINAL ARTICLE

<https://doi.org/10.26907/2541-7746.2025.4.627-640>

On the solvability of the optimization problem for constructing quite interpretable linear regressions

M.P. Bazilevskiy*Irkutsk State Transport University, Irkutsk, Russia**mik2178@yandex.ru***Abstract**

This article is devoted to improving the technique for constructing interpretable regression models in which parameters are estimated using the ordinary least squares method. A definition of quite interpretable linear regressions is provided. The main requirements for such regressions include the consistency between the signs of the parameter estimates and the substantive meanings of the variables, the significance of the estimates, the low degree of multicollinearity, and the high quality of approximation. Whether a model belongs to the class of quite interpretable regressions or not depends on its significance level. In terms of mixed 0-1 integer linear programming, which has made progress in recent years due to computational advances, an optimization problem is formulated for constructing quite interpretable linear regressions with a fairly large number of linear constraints. The problem is proved to be solvable under certain conditions. The proposed mathematical framework can be successfully applied to processing big data, as the number of constraints in the formulated problem does not depend on the sample size, unlike existing foreign analogues.

Keywords: linear regression, ordinary least squares method, interpretability, multicollinearity, significance, mixed 0-1 integer linear programming problem, solvability

For citation: Bazilevskiy M.P. On the solvability of the optimization problem for constructing quite interpretable linear regressions. *Uchenye Zapiski Kazanskogo Universiteta. Seriya Fiziko-Matematicheskie Nauki*, 2025, vol. 167, no. 4, pp. 627–640. <https://doi.org/10.26907/2541-7746.2025.4.627-640>. (In Russian)

Введение

В настоящее время в научных исследованиях (см., например, работы [1, 2]) выделилось новое направление, связанное с тем, что при построении моделей машинного обучения важно обеспечивать не только их высокое качество, но и интерпретируемость. Среди моделей машинного обучения наиболее просто интерпретируются модели линейной регрессии, чему посвящен отдельный раздел в монографии [1]. Однако даже при построении простейшей линейной регрессии с помощью метода наименьших квадратов (МНК) возникают две проблемы. Во-первых, еще до оценивания неизвестных параметров модели требуется определить состав входящих в нее объясняющих переменных. Во-вторых, нет гарантии,

что модель, построенную с помощью МНК, можно будет адекватно интерпретировать. Выделим монографию [3], в которой предложен алгоритм построения интерпретируемой регрессионной модели с использованием анализа последовательностей. Но в нем для выбора наиболее существенных регрессоров в линейной регрессии использован метод всех возможных регрессий [4], требующий больших объемов вычислений.

Вместо метода всех возможных регрессий для построения линейной регрессии с помощью МНК в зарубежных исследованиях применяют аппарат частично-булевого квадратичного программирования (ЧБКП). Считается, что начало этому положила статья [5] 2009 года авторов Х. Конно и Р. Ямамото. Благодаря этому вышли работы [6, 7], посвященные отбору в линейной регрессии оптимального числа регрессоров по критериям Акаике, Шварца и Мэллоуза, а также по скорректированному коэффициенту детерминации. В [8, 9] исследованы вопросы снижения при моделировании эффекта мультиколлинеарности. Разработке новых эффективных алгоритмов решения задач ЧБКП для выбора структуры линейной регрессии посвящены статьи [10, 11].

Однако в 2018 году в отечественной работе [12] появилась новая формализация задачи построения линейной регрессии, но уже в терминах аппарата частично-булевого линейного программирования (ЧБЛП). Своевременность такого подхода обусловлена тем, что, как отмечено в [13], в среднем с 2001 по 2020 год компьютерное оборудование для решения задач ЧБЛП стало примерно в двадцать раз быстрее, а алгоритмы – в пятьдесят. Задача, предложенная в [12], за последние годы была существенно расширена. Так, в ней появились ограничения на мультиколлинеарность [14], значимость оценок по t -критерию Стьюдента [15], значимость модели по F -критерию Фишера [16] и т. д. В зарубежных исследованиях параллельным ходом стало принято делать так, чтобы задача ЧБКП не просто осуществляла отбор переменных в линейной регрессии, но и по возможности гарантировала значимость оценок, выполнение условий теоремы Гаусса – Маркова и т. д. Например, в [17] для этого использовано понятие регрессионной диагностики, а в [18] – целостной регрессии.

Цель данной статьи состоит в слиянии рассмотренных в работах [12, 14–16] и в других манускриптах автора задач ЧБЛП в единую задачу построения интерпретируемой линейной регрессии, а также в исследовании предложенной задачи на разрешимость.

1. Вполне интерпретируемая линейная регрессия

Рассмотрим модель множественной линейной регрессии, включающую объясняемую переменную y и l объясняющих переменных $x_1, x_2, \dots, x_l$:

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

где $\alpha_0, \alpha_1, \dots, \alpha_l$ – неизвестные параметры; n – объем выборки; $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ – ошибки регрессии.

Приведем кратко некоторые предпосылки, позволившие сформулировать определение вполне интерпретируемой линейной регрессии. Во-первых, если в оцененной с помощью МНК модели (1) отсутствует мультиколлинеарность, то недопустимо, чтобы знаки оценок $\tilde{\alpha}_j, j = \overline{1, l}$, противоречили смыслу стоящих при них переменных. Во-вторых, степень

мультиколлинеарности, негативно влияющей на интерпретацию оценок регрессии, должна быть низкой, причем как пары объясняющих переменных не должны сильно коррелировать, так и линейные зависимости каждой переменной с остальными предикторами должны быть неадекватны. Для получения выводов о наличии или отсутствии значимых связей между переменными мы предлагаем использовать классический аппарат проверки статистических гипотез, предполагающий на начальном этапе выбор уровня значимости α . В-третьих, интерпретация модели (1) будет сомнительной, если среди ее оценок будут незначимые по t -критерию Стьюдента. К тому же недопустимо, чтобы относительные вклады переменных в общую детерминацию R^2 , которые находятся по формулам $C_{x_j}^{abs} = \frac{\sigma_{x_j}}{\sigma_y} r_{yx_j} \tilde{\alpha}_j$, $j = \overline{1, l}$, где $\sigma_y, \sigma_{x_1}, \dots, \sigma_{x_l}$ – стандартные отклонения переменных, r_{yx_j} , $j = \overline{1, l}$, – коэффициенты корреляции переменной y с предикторами, были не положительными. В-четвертых, интерпретация недопустима, если коэффициент детерминации R^2 слишком низок либо модель не значима в целом по F -критерию Фишера.

Определение 1. Линейная регрессия (1), оцененная с помощью МНК, называется вполне интерпретируемой (ВИЛинР) для выбранного уровня значимости α и нижней границы θ абсолютных вкладов $C_{x_j}^{abs}$ переменных в общую детерминацию R^2 , если она удовлетворяет следующим условиям:

- 1) знаки коэффициентов корреляции r_{yx_j} , $j = \overline{1, l}$, соответствуют содержательному смыслу решаемой задачи;
- 2) знаки всех оценок $\tilde{\alpha}_j$ согласуются со знаками соответствующих коэффициентов корреляции r_{yx_j} , т.е. $\tilde{\alpha}_j r_{yx_j} > 0$, $j = \overline{1, l}$;
- 3) $C_{x_j}^{abs} \geq \theta$, $j = \overline{1, l}$;
- 4) коэффициенты интеркорреляций $r_{x_i x_j}$, $i = \overline{1, l-1}$, $j = \overline{i+1, l}$, не значимы по t -критерию Стьюдента;
- 5) все вспомогательные зависимости каждой объясняющей переменной x_j на остальные предикторы не значимы по F -критерию Фишера;
- 6) оценки $\tilde{\alpha}_j$, $j = \overline{1, l}$, значимы по t -критерию Стьюдента;
- 7) модель в целом значима по F -критерию Фишера;
- 8) $R^2 \geq 0.8$.

Отметим, что для выполнения условия 1) ВИЛинР еще до построения модели необходимо привлекать эксперта из соответствующей предметной области, который способен выявить объясняющие переменные, по-настоящему влияющие по смыслу на отклик y , а также направления их влияния.

Как видно из определения, относится ли построенная модель (1) к ВИЛинР или не относится, зависит от параметров α и θ . Уровень значимости α традиционно можно принять равным 0.1, 0.05 или 0.01, а нижнюю границу вкладов θ выбрать из промежутка $[0, 1]$, причем чем ниже α и выше θ , тем жестче требования к ВИЛинР, а значит, ниже шанс, что такая модель для конкретной выборки вообще существует. Выбранное нами пороговое значение 0.8 для R^2 в условии 8) ВИЛинР также может быть скорректировано в соответствии с предпочтением исследователя.

2. Постановка оптимизационной задачи

Упростить расчеты определения оценок неизвестных параметров модели (1) позволяет стандартизация (нормирование) переменных [19]:

$$y_i^\bullet = \frac{y_i - \bar{y}}{\sigma_y}, \quad x_{i1}^\bullet = \frac{x_{i1} - \bar{x}_1}{\sigma_{x_1}}, \dots, x_{il}^\bullet = \frac{x_{il} - \bar{x}_l}{\sigma_{x_l}}, \quad i = \overline{1, n},$$

где $\bar{y}, \bar{x}_1, \dots, \bar{x}_l$ – средние значения переменных. Тогда уравнение линейной регрессии (1) в стандартизованном масштабе [19] приобретет вид

$$y_i^\bullet = \sum_{j=1}^l \beta_j x_{ij}^\bullet + \varepsilon_i^\bullet, \quad i = \overline{1, n}, \quad (2)$$

где $\beta_1, \dots, \beta_l$ – неизвестные стандартизованные коэффициенты; $\varepsilon_1^\bullet, \dots, \varepsilon_n^\bullet$ – ошибки стандартизованной регрессии.

МНК-оценки стандартизованных коэффициентов регрессии (2) находятся в результате решения системы линейных алгебраических уравнений вида

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

$$R_{xx} = \begin{pmatrix} 1 & r_{x_1 x_2} & s & r_{x_1 x_l} \\ r_{x_1 x_2} & 1 & s & r_{x_2 x_l} \\ \vdots & \vdots & \ddots & \vdots \\ r_{x_1 x_l} & r_{x_2 x_l} & s & 1 \end{pmatrix}, \quad R_{yx} = \begin{pmatrix} r_{yx_1} \\ r_{yx_2} \\ \vdots \\ r_{yx_l} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_l \end{pmatrix},$$

а коэффициент детерминации R^2 модели – по формуле

$$R^2 = R_{yx}^T \beta,$$

где T – операция транспонирования.

С использованием этих формул в [12] была сформулирована следующая задача ЧБЛП:

$$R^2 = \sum_{j=1}^l r_{yx_j} \beta_j \rightarrow \max; \quad (3)$$

$$-(1 - \delta_j) M \leq \sum_{k=1}^l r_{x_j x_k} \beta_k - r_{yx_j} \leq (1 - \delta_j) M, \quad j = \overline{1, l}; \quad (4)$$

$$-\delta_j M \leq \beta_j \leq \delta_j M, \quad j = \overline{1, l}; \quad (5)$$

$$\delta_j \in \{0, 1\}, \quad j = \overline{1, l}; \quad (6)$$

$$\sum_{j=1}^l \delta_j = m, \quad (7)$$

где m – число наиболее существенных регрессоров, причем $m \leq \min\{l, n - 1\}$; M – достаточно большое положительное число; $\delta_j, j = \overline{1, l}$, – булевы переменные, заданные по правилу

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

Важно, что в задаче (3)–(7) число M на всей области допустимых решений должно удовлетворять неравенству $M \geq \max_{j=1, \overline{l}} \{ |r_{x_j x_k} \beta_k - r_{yx_j}|, |\beta_j| \}$, иначе можно «упустить» оптимальное решение задачи. Будем считать, что число $M \rightarrow \infty$. Тогда решение задачи (3)–(7) приводит к выбору наилучшей по критерию R^2 линейной регрессии с фиксированным числом m регрессоров.

Используя результаты работ [12, 14–16], сформулируем единую задачу построения ВИЛинР. Для выполнения условия 2) ВИЛинР вместо (5) введем ограничения

$$0 \leq \beta_j \leq \delta_j M, \quad j \in J^+; \quad (8)$$

$$-\delta_j M \leq \beta_j \leq 0, \quad j \in J^-, \quad (9)$$

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

для выполнения условия 3) ВИЛинР используем ограничения

$$r_{yx_j} \beta_j \geq \theta \delta_j, \quad j = \overline{1, l}; \quad (10)$$

для выполнения условия 4) ВИЛинР – ограничения

$$|r_{x_i x_j}| (\delta_i + \delta_j - 1) \leq r, \quad i = \overline{1, l-1}, \quad j = \overline{i+1, l}, \quad (11)$$

в которых для обеспечения незначимости коэффициентов интеркорреляций по t -критерию Стьюдента для выбранного уровня значимости α число $r = \frac{t_{\text{крит}}(\alpha, n-2)}{\sqrt{n-2 + t_{\text{крит}}^2(\alpha, n-2)}}$;

для выполнения условия 5) ВИЛинР – ограничения

$$\sum_{j=1}^{l-1} r_{x_{q_{kj}} x_k} \beta_{kj} - (1 - \delta_k) M - (1 - \rho_i) M \leq r_i^*, \quad k = \overline{1, l}, \quad i = \overline{1, m^*}; \quad (12)$$

$$-(1 - \delta_{q_{ki}}) M \leq \sum_{j=1}^{l-1} r_{x_{q_{ki}} x_{q_{kj}}} \beta_{kj} - r_{x_{q_{ki}} x_k} \leq (1 - \delta_{q_{ki}}) M, \quad k = \overline{1, l}, \quad i = \overline{1, l-1}; \quad (13)$$

$$-\delta_{q_{ki}} M \leq \beta_{ki} \leq \delta_{q_{ki}} M, \quad k = \overline{1, l}, \quad i = \overline{1, l-1}; \quad (14)$$

$$-M(1 - \rho_j) \leq \sum_{k=1}^l \delta_k - j \leq M(1 - \rho_j), \quad j = \overline{1, m^*}; \quad (15)$$

$$\rho_j \in \{0, 1\}, \quad j = \overline{1, m^*}; \quad (16)$$

$$\sum_{j=1}^{m^*} \rho_j = 1, \quad (17)$$

где q_{ij} – элементы матрицы $Q_{l \times (l-1)}$, полученной путем вычеркивания главной диагонали

из квадратной матрицы $\begin{pmatrix} 1 & 2 & s & l \\ 1 & 2 & s & l \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 2 & s & l \end{pmatrix}$; β_{kj} , $k = \overline{1, l}$, $j = \overline{1, l-1}$, – стандартизованные коэффициенты вспомогательных регрессий для каждой переменной $x_j^\bullet$ на все остальные

ные предикторы; $m^* = \min\{l, n-1\}$; числа $r_1^* = 1$, $r_j^* = \frac{F_{\text{крит}}(\alpha, j-1, n-j-2)}{\frac{n-j-2}{j-1} + F_{\text{крит}}(\alpha, j-1, n-j-2)}$, $j = \overline{2, m^*}$; ρ_j , $j = \overline{1, m^*}$, – булевы переменные, заданные по правилу

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

для выполнения условия 6) ВИЛинР – ограничения

$$\sum_{j=1}^l r_{yx_j} \beta_j - \sum_{j=1}^{l-1} r_{yx_{q_{kj}}} \beta_{kj}^{\bullet} \geq (1 - \sum_{j=1}^l r_{yx_j} \beta_j) T_i - M(1 - \delta_k) - M(1 - \rho_i), \quad k = \overline{1, l}, \quad i = \overline{1, m^*}; \quad (18)$$

$$-(1 - \delta_{q_{ki}}) M \leq \sum_{j=1}^{l-1} r_{x_{q_{ki}} x_{q_{kj}}} \beta_{kj}^{\bullet} - r_{yx_{q_{ki}}} \leq (1 - \delta_{q_{ki}}) M, \quad k = \overline{1, l}, \quad i = \overline{1, l-1}; \quad (19)$$

$$-\delta_{q_{ki}} M \leq \beta_{ki}^{\bullet} \leq \delta_{q_{ki}} M, \quad k = \overline{1, l}, \quad i = \overline{1, l-1}, \quad (20)$$

где $\beta_{kj}^{\bullet}$, $k = \overline{1, l}$, $j = \overline{1, l-1}$, – стандартизованные коэффициенты вспомогательных регрессий переменной $y^{\bullet}$ на все остальные предикторы без $x_j^{\bullet}$; числа $T_j = \frac{t_{\text{крит}}^2(\alpha, n-j-1)}{n-j-1}$, $j = \overline{1, m^*}$;

для выполнения условия 7) ВИЛинР – ограничения

$$R^2 \geq G_j - M(1 - \rho_j), \quad j = \overline{1, m^*}, \quad (21)$$

где числа $G_j = \frac{F_{\text{крит}}(\alpha, j, n-j-1)}{\frac{n-j-1}{j} + F_{\text{крит}}(\alpha, j, n-j-1)}$, $j = \overline{1, m^*}$.

Как видно, сформулированная задача ЧБЛП (3), (4), (6), (8)–(21) содержит довольно большое количество линейных ограничений, поэтому возникает вопрос, будет ли эта задача разрешима, т. е. будет ли она иметь хотя бы одно решение в зависимости от назначенных параметров α и θ .

3. Исследование оптимизационной задачи на разрешимость

Теорема. Пусть из всех оцененных с помощью МНК парных линейных регрессий y от x_j , $j = \overline{1, l}$, $l \geq 2$, наибольшее значение $\theta^{\#}$, $\theta^{\#} < 1$, коэффициента детерминации R^2 имеет модель, значимая по F -критерию Фишера для заданного уровня значимости $\alpha^{\#}$. Тогда задача ЧБЛП (3), (4), (6), (8)–(21) при $M \rightarrow \infty$ разрешима для любого уровня значимости $\alpha \in [\alpha^{\#}, 1)$ и числа $\theta \in [0, \theta^{\#}]$.

Доказательство. Для удобства будем считать, что из всех объясняющих переменных наиболее тесно коррелирует с откликом y , например, переменная x_1 . Тогда по условию теоремы по F -критерию Фишера для уровня $\alpha^{\#}$ значима стандартизованная парная линейная регрессия

$$y^{\bullet} = \beta_1 x_1^{\bullet} + \varepsilon^{\bullet}. \quad (22)$$

Последнее равносильно значимости для уровня $\alpha^\#$ коэффициента β_1 регрессии (22), а также коэффициента корреляции r_{yx_1} по t -критерию Стьюдента.

МНК-оценка параметра β_1 модели (22) равна r_{yx_1} , а коэффициент детерминации $R^2 = r_{yx_1}^2$, откуда $\theta^\# = r_{yx_1}^2$.

Покажем, что при $M \rightarrow \infty$ всем ограничениям задачи для любого уровня значимости $\alpha \in [\alpha^\#, 1)$ и числа $\theta \in [0, \theta^\#]$ удовлетворяет следующее решение:

$$\begin{aligned}\beta_1 &= r_{yx_1}, \quad \beta_j = 0, \quad j = \overline{2, l}; \\ \delta_1 &= 1, \quad \delta_j = 0, \quad j = \overline{2, l}; \\ \beta_{k1} &= r_{x_1 x_k}, \quad k = \overline{2, l}, \quad \beta_{1i} = 0, \quad i = \overline{1, l-1}, \quad \beta_{ki} = 0, \quad k = \overline{2, l}, \quad i = \overline{2, l-1}; \\ \rho_1 &= 1, \quad \rho_j = 0, \quad j = \overline{2, m^*}; \\ \beta_{k1}^\bullet &= r_{yx_1}, \quad k = \overline{2, l}, \quad \beta_{1i}^\bullet = 0, \quad i = \overline{1, l-1}, \quad \beta_{ki}^\bullet = 0, \quad k = \overline{2, l}, \quad i = \overline{2, l-1}.\end{aligned}$$

Подставим значения всех этих переменных в ограничения задачи. Тогда ограничения (4) примут вид

$$\begin{aligned}0 &\leq 0 \leq 0 \quad \text{при } j = 1; \\ -M &\leq r_{x_1 x_j} r_{yx_1} - r_{yx_j} \leq M, \quad j = \overline{2, l};\end{aligned}$$

ограничения (8), (9) при $r_{yx_1} > 0$:

$$\begin{aligned}0 &\leq r_{yx_1} \leq M \quad \text{при } j = 1; \\ 0 &\leq 0 \leq 0, \quad j \in J^+ \setminus \{1\}; \\ 0 &\leq 0 \leq 0, \quad j \in J^-;\end{aligned}$$

ограничения (8), (9) при $r_{yx_1} < 0$:

$$\begin{aligned}0 &\leq 0 \leq 0, \quad j \in J^+; \\ -M &\leq r_{yx_1} \leq 0 \quad \text{при } j = 1; \\ 0 &\leq 0 \leq 0, \quad j \in J^- \setminus \{1\};\end{aligned}$$

ограничения (10):

$$\begin{aligned}r_{yx_1}^2 &\geq \theta \quad \text{при } j = 1; \\ 0 &\geq 0, \quad j = \overline{2, l};\end{aligned} \tag{23}$$

ограничения (11):

$$0 \leq \frac{t_{\text{крит}}(\alpha, n-2)}{\sqrt{n-2+t_{\text{крит}}^2(\alpha, n-2)}}, \quad i = 1, \quad j = \overline{2, l}; \tag{24}$$

$$-|r_{x_i x_j}| \leq \frac{t_{\text{крит}}(\alpha, n-2)}{\sqrt{n-2+t_{\text{крит}}^2(\alpha, n-2)}}, \quad i = \overline{2, l-1}, \quad j = \overline{i+1, l}; \tag{25}$$

ограничения (12):

$$\begin{aligned}0 &\leq 1, \quad k = 1, \quad i = 1; \\ r_{x_1 x_k}^2 - M &\leq 1, \quad k = \overline{2, l}, \quad i = 1; \\ -M &\leq r_i^*, \quad k = 1, \quad i = \overline{2, m^*}; \\ r_{x_1 x_k}^2 - 2M &\leq r_i^*, \quad k = \overline{2, l}, \quad i = \overline{2, m^*};\end{aligned}$$

ограничения (13):

$$\begin{aligned} -M &\leq -r_{x_{q_{ki}}x_k} \leq M, \quad k = 1, \quad i = \overline{1, l-1}; \\ -M &\leq -r_{x_{q_{ki}}x_k} \leq M, \quad k = \overline{2, l}, \quad i = \overline{2, l-1}; \\ 0 &\leq 0 \leq 0, \quad k = \overline{2, l}, \quad i = 1; \end{aligned}$$

ограничения (14):

$$\begin{aligned} 0 &\leq 0 \leq 0, \quad k = 1, \quad i = \overline{1, l-1}; \\ 0 &\leq 0 \leq 0, \quad k = \overline{2, l}, \quad i = \overline{2, l-1}; \\ -M &\leq r_{x_1x_k} \leq M, \quad k = \overline{2, l}, \quad i = 1; \end{aligned}$$

ограничения (15):

$$\begin{aligned} 0 &\leq 0 \leq 0 \quad \text{при } j = 1; \\ -M &\leq 1 - j \leq M, \quad j = \overline{2, m^*}; \end{aligned}$$

ограничение (17):

$$1 = 1;$$

ограничения (18):

$$\begin{aligned} r_{yx_1}^2 &\geq (1 - r_{yx_1}^2) \frac{t_{\text{крит}}^2(\alpha, n-2)}{n-2}, \quad k = 1, \quad i = 1; \\ r_{yx_1}^2 &\geq (1 - r_{yx_1}^2) T_1 - M, \quad k = \overline{2, l}, \quad i = 1; \\ r_{yx_1}^2 &\geq (1 - r_{yx_1}^2) T_i - M, \quad k = 1, \quad i = \overline{2, m^*}; \\ r_{yx_1}^2 &\geq (1 - r_{yx_1}^2) T_i - 2M, \quad k = \overline{2, l}, \quad i = \overline{2, m^*}; \end{aligned} \tag{26}$$

ограничения (19):

$$\begin{aligned} -M &\leq -r_{yx_{q_{ki}}} \leq M, \quad k = 1, \quad i = \overline{1, l-1}; \\ -M &\leq -r_{yx_{q_{ki}}} \leq M, \quad k = \overline{2, l}, \quad i = \overline{2, l-1}; \\ 0 &\leq 0 \leq 0, \quad k = \overline{2, l}, \quad i = 1; \end{aligned}$$

ограничения (20):

$$\begin{aligned} 0 &\leq 0 \leq 0, \quad k = 1, \quad i = \overline{1, l-1}; \\ 0 &\leq 0 \leq 0, \quad k = \overline{2, l}, \quad i = \overline{2, l-1}; \\ -M &\leq r_{yx_1} \leq M, \quad k = \overline{2, l}, \quad i = 1; \end{aligned}$$

ограничения (21):

$$\begin{aligned} r_{yx_1}^2 &\geq \frac{F_{\text{крит}}(\alpha, 1, n-2)}{n-2 + F_{\text{крит}}(\alpha, 1, n-2)}, \quad j = 1; \\ r_{yx_1}^2 &\geq G_j - M, \quad j = \overline{2, m^*}. \end{aligned} \tag{27}$$

Очевидно, что все вышеприведенные неравенства, кроме (23)–(27), справедливы при $M \rightarrow \infty$. Также выполнены условия (6) и (16) булевости переменных.

Рассмотрим неравенство (23). Поскольку $\theta^\# = r_{yx_1}^2$, то оно справедливо для любого $\theta \in [0, \theta^\#]$. Поскольку $t_{\text{крит}}(\alpha, n-2) > 0$ для любого уровня значимости α , то неравенства (24) и (25) справедливы для любого α , в частности, для $\alpha \in [\alpha^\#, 1)$.

Поскольку коэффициент β_1 парной регрессии (22) значим для уровня $\alpha^\#$ по t -критерию Стьюдента, то он значим и при $\alpha \in [\alpha^\#, 1)$. Это означает, что для любого $\alpha \in [\alpha^\#, 1)$ справедливо неравенство

$$|t_{\text{набл}}| \geq t_{\text{крит}}(\alpha, n-2),$$

где наблюдаемое значение t -критерия Стьюдента находится по формуле

$$t_{\text{набл}} = \sqrt{\frac{r_{yx_1}^2}{1 - r_{yx_1}^2}} (n-2).$$

Отсюда следует справедливость неравенства (26) для любого $\alpha \in [\alpha^\#, 1)$.

Для парной линейной регрессии (22) $F_{\text{крит}}(\alpha, 1, n-2) = t_{\text{крит}}^2(\alpha, n-2)$, поэтому неравенства (26) и (27) равносильны. Отсюда неравенство (27) справедливо для любого $\alpha \in [\alpha^\#, 1)$.

Таким образом, при $M \rightarrow \infty$ для любых $\theta \in [0, \theta^\#]$ и $\alpha \in [\alpha^\#, 1)$ множество допустимых решений задачи (3), (4), (6), (8)–(21) не пусто. А поскольку все переменные ограничены сверху и снизу, то рассматриваемая задача разрешима.

Теорема доказана. $\square$

Решение задачи (3), (4), (6), (8)–(21) приводит к построению линейной регрессии с оптимальным по критерию R^2 числом регрессоров, удовлетворяющей первым семи условиям ВИЛинР. Выполнение условия 8) ВИЛинР проверяется после решения задачи ЧБЛП.

Сделаем два важных замечания касательно задачи (3), (4), (6), (8)–(21).

Замечание 1. Значения коэффициентов корреляции при решении этой задачи неизбежно подвергаются округлению. Если матрица R_{xx} плохо обусловлена [20], то из-за округлений решение задачи может оказаться неоптимальным. Поэтому значения коэффициентов корреляции следует брать с большей точностью, снижая ошибки округления.

Замечание 2. При решении задачи неминуемо придется ограничиваться каким-то конкретным значением числа M из-за невозможности обеспечения условия $M \rightarrow \infty$. Следует помнить, что число M должно быть достаточно большим, иначе оптимальное решение может быть «упущено».

Заключение

Дано определение вполне интерпретируемой линейной регрессии. Для ее построения в терминах аппарата частично-булевого линейного программирования сформулирована оптимизационная задача. Доказана разрешимость задачи при определенных условиях. Значимость разработанного математического аппарата состоит в том, что многие этапы построения линейной регрессионной модели удалось формализовать в виде единой задачи, которая может быть решена с использованием современных эффективных решателей CPLEX или Gurobi.

Нерешенным до конца остается вопрос, связанный с выбором достаточно больших чисел M в некоторых линейных ограничениях задачи. Дальнейшие исследования автора будут посвящены исследованию этого вопроса.

Конфликт интересов. Автор заявляет об отсутствии конфликта интересов.

Conflicts of Interest. The author declares no conflicts of interest.

Литература

1. *Molnar C.* Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. Lulu.com, 2020. 320 p.
2. *Doshi-Velez F., Kim B.* Towards a rigorous science of interpretable machine learning. Ver. 2. arXiv Preprint 1702.08608. 2017. <https://doi.org/10.48550/arXiv.1702.08608>.
3. *Горбач А.Н., Цейтлин Н.А.* Покупательское поведение: анализ спонтанных последовательностей и регрессионных моделей в маркетинговых исследованиях. Киев: Освіта України, 2011. 220 с.
4. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: Юнити, 1998. 1005 с.
5. *Konno H., Yamamoto R.* Choosing the best set of variables in regression analysis using integer programming // J. Global Optim. 2009. V. 44, No 2. P. 273–282. <https://doi.org/10.1007/s10898-008-9323-9>.
6. *Miyashiro R., Takano Y.* Mixed integer second-order cone programming formulations for variable selection in linear regression // Eur. J. Oper. Res. 2015. V. 247, No 3. P. 721–731. <https://doi.org/10.1016/j.ejor.2015.06.081>.
7. *Miyashiro R., Takano Y.* Subset selection by Mallows' C_p : A mixed integer programming approach // Expert Syst. Appl. 2015. V. 42, No 1. P. 325–331. <https://doi.org/10.1016/j.eswa.2014.07.056>.
8. *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 // J. Global Optim. 2019. V. 73, No 2. P. 431–446. <https://doi.org/10.1007/s10898-018-0713-3>.
9. *Tamura R., Kobayashi K., Takano Y., Miyashiro R., Nakata K., Matsui T.* Best subset selection for eliminating multicollinearity // J. Oper. Res. Soc. Jpn. 2017. V. 60, No 3. P. 321–336. <https://doi.org/10.15807/jorsj.60.321>.
10. *Bertsimas D., King A., Mazumder R.* Best subset selection via a modern optimization lens // Ann. Stat. 2016. V. 44, No 2. P. 813–852. <https://doi.org/10.1214/15-AOS1388>.
11. *Bertsimas D., King A.* OR forum—an algorithmic approach to linear regression // Oper. Res. 2016. V. 64, No 1. P. 2–16. <https://doi.org/10.1287/opre.2015.1436>.
12. *Базилевский М.П.* Сведение задачи отбора информативных регрессоров при оценивании линейной регрессионной модели по методу наименьших квадратов к задаче частично-булевого линейного программирования // МОИТ. 2018. Т. 6, № 1 (20). С. 108–117.
13. *Koch T., Berthold T., Pedersen J., Vanaret C.* Progress in mathematical programming solvers from 2001 to 2020 // EURO J. Comput. Optim. 2022. V. 10. Art. 100031. <https://doi.org/10.1016/j.ejco.2022.100031>.
14. *Базилевский М.П.* Отбор информативных регрессоров с учетом мультиколлинеарности между ними в регрессионных моделях как задача частично-булевого линейного программирования // МОИТ. 2018. Т. 6, № 2 (21). С. 104–118.

15. *Базилевский М.П.* Отбор значимых по критерию Стьюдента информативных регрессоров в оцениваемых с помощью МНК регрессионных моделях как задача частично-булевого линейного программирования // Вестн. ВГУ. Сер.: Сист. анализ и информ. технол. 2021. № 3. С. 5–16. <https://doi.org/10.17308/sait.2021.3/3731>.
16. *Базилевский М.П.* Оптимизационные задачи отбора информативных регрессоров в линейной регрессии с контролем ее значимости по критерию Фишера // Изв. Самарск. науч. центра РАН. 2024. Т. 26, № 6. С. 200–207.
17. *Chung S., Park Y.W., Cheong T.* A mathematical programming approach for integrated multiple linear regression subset selection and validation // Pattern Recognit. 2020. V. 108. Art. 107565. <https://doi.org/10.1016/j.patcog.2020.107565>.
18. *Bertsimas D., Li M.L.* Scalable holistic linear regression // Oper. Res. Lett. 2020. V. 48, No 3. P. 203–208. <https://doi.org/10.1016/j.orl.2020.02.008>.
19. *Ферстер Э., Рениц Б.* Методы корреляционного и регрессионного анализа. М.: Финансы и статистика, 1983. 303 с.
20. *Лебедева А.В., Рябов В.М.* О численном решении систем линейных алгебраических уравнений с плохо обусловленными матрицами // Вестн. СПбГУ. Матем. Механ. Астрон. 2019. Т. 6, № 4. С. 619–626. <https://doi.org/10.21638/11701/spbu01.2019.407>.

References

1. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Lulu.com, 2020. 320 p.
2. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning. Ver. 2. arXiv Preprint 1702.08608. 2017. <https://doi.org/10.48550/arXiv.1702.08608>.
3. Gorbach A.N., Tseitlin N.A. *Pokupatel'skoe povedenie: analiz spontannykh posledovatel'nykh i regressionnykh modelei v marketingovykh issledovaniyakh* [Buying Behavior: Analysis of Spontaneous Sequences and Regression Models in Marketing Research]. Kyiv, Osvita Ukrainy, 2011, 220 p. (In Russian)
4. Aivazyan S.A., Mkhitaryan V.S. *Prikladnaya statistika i osnovy ekonometriki* [Applied Statistics and Basics of Econometrics]. Moscow, Yuniti, 1998, 1005 p. (In Russian)
5. Konno H., Yamamoto R. Choosing the best set of variables in regression analysis using integer programming. *J. Global Optim.*, 2009, vol. 44, no. 2, pp. 273–282. <https://doi.org/10.1007/s10898-008-9323-9>.
6. Miyashiro R., Takano Y. Mixed integer second-order cone programming formulations for variable selection in linear regression. *Eur. J. Oper. Res.*, 2015, vol. 247, no. 3, pp. 721–731. <https://doi.org/10.1016/j.ejor.2015.06.081>.
7. Miyashiro R., Takano Y. Subset selection by Mallows' C_p : A mixed integer programming approach. *Expert Syst. Appl.*, 2015, vol. 42, no. 1, pp. 325–331. <https://doi.org/10.1016/j.eswa.2014.07.056>.
8. 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. *J. Global Optim.*, 2019, vol. 73, no. 2, pp. 431–446. <https://doi.org/10.1007/s10898-018-0713-3>.

9. Tamura R., Kobayashi K., Takano Y., Miyashiro R., Nakata K., Matsui T. Best subset selection for eliminating multicollinearity. *J. Oper. Res. Soc. Jpn.*, 2017, vol. 60, no. 3, pp. 321–336. <https://doi.org/10.15807/jorsj.60.321>.
10. Bertsimas D., King A., Mazumder R. Best subset selection via a modern optimization lens. *Ann. Stat.*, 2016, vol. 44, no. 2, pp. 813–852. <https://doi.org/10.1214/15-AOS1388>.
11. Bertsimas D., King A. OR forum—an algorithmic approach to linear regression. *Oper. Res.*, 2016, vol. 64, no. 1, pp. 2–16. <https://doi.org/10.1287/opre.2015.1436>.
12. Bazilevskiy M.P. 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. *Model., Optim. Inf. Technol.*, 2018, vol. 6, no. 1 (20), pp. 108–117. (In Russian)
13. Koch T., Berthold T., Pedersen J., Vanaret C. Progress in mathematical programming solvers from 2001 to 2020. *EURO J. Comput. Optim.*, 2022, vol. 10, art. 100031. <https://doi.org/10.1016/j.ejco.2022.100031>.
14. Bazilevskiy M.P. Subset selection in regression models with considering multicollinearity as a task of mixed 0-1 integer linear programming. *Model., Optim. Inf. Technol.*, 2018, vol. 6, no. 2 (21), pp. 104–118. (In Russian)
15. Bazilevskiy M.P. Selection of informative regressors significant by Student’s t-test in regression models estimated using OLS as a partial Boolean linear programming problem. *Proc. Voronezh State Univ. Series: Syst. Anal. Inf. Technol.*, 2021, no. 3, pp. 5–16. <https://doi.org/10.17308/sait.2021.3/3731>. (In Russian)
16. Bazilevskiy M.P. Optimization problems of subset selection in linear regression with control of its significance using F-test. *Izv. Samara Sci. Cent. Russ. Acad. Sci.*, 2024, vol. 26, no. 6, pp. 200–207. (In Russian)
17. Chung S., Park Y.W., Cheong T. A mathematical programming approach for integrated multiple linear regression subset selection and validation. *Pattern Recognit.*, 2020, vol. 108, art. 107565. <https://doi.org/10.1016/j.patcog.2020.107565>.
18. Bertsimas D., Li M.L. Scalable holistic linear regression. *Oper. Res. Lett.*, 2020, vol. 48, no. 3, pp. 203–208. <https://doi.org/10.1016/j.orl.2020.02.008>.
19. Foerster E., Renz B. *Metody korrelyatsionnogo i regressionnogo analiza* [Methods of Correlation and Regression Analysis]. Moscow, Finans. Stat., 1983. 303 p. (In Russian)
20. Lebedeva A.V., Ryabov V.M. On the numerical solution of system of linear algebraic equations with ill-conditioned matrices. *Vestn. St. Petersburg Univ. Math. Mech. Astron.*, 2019, vol. 6, no. 4, pp. 619–626. <https://doi.org/10.21638/11701/spbu01.2019.407>. (In Russian)

Информация об авторах

Михаил Павлович Базилевский, кандидат технических наук, доцент, доцент кафедры «Математика», Иркутский государственный университет путей сообщения

E-mail: mik2178@yandex.ru

ORCID: <https://orcid.org/0000-0002-3253-5697>

Author Information

Mikhail P. Bazilevskiy, Cand. Sci. (Engineering), Associate Professor, Department of Mathematics, Irkutsk State Transport University

E-mail: *mik2178@yandex.ru*

ORCID: <https://orcid.org/0000-0002-3253-5697>

Поступила в редакцию 12.10.2025

Принята к публикации 8.11.2025

Received October 12, 2025

Accepted November 8, 2025