УДК 519.862.6
DOI: 10.26102/2310-6018/2026.54.3.018

Алгоритм построения вполне интерпретируемых сегментированных линейных регрессий

М.П. Базилевский✉

Иркутский государственный университет путей сообщения, Иркутск, Российская Федерация

Резюме. Работа посвящена актуальному научному направлению – интерпретируемому машинному обучению. Ранее автором было введено понятие «вполне интерпретируемая линейная регрессия», построение которой осуществляется с помощью метода наименьших квадратов по всей выборке статистических данных. В статье это понятие обобщается на сегментированную линейную регрессию, при идентификации которой данные сначала разбиваются на сегменты, а затем на каждом из них строится своя линейная регрессия. Разработан алгоритм построения вполне интерпретируемых сегментированных линейных регрессий. Его особенность в том, что, во-первых, разбиение пространства предикторов на сегменты осуществляется с помощью логических функций активации аргументов бинарных операций min. Во-вторых, в каждом сегменте строится парная регрессия, что полностью решает проблему мультиколлинеарности. С помощью разработанного алгоритма по выборке объема 1030 наблюдений была построена сегментированная линейная регрессия прочности бетона на сжатие. Во всех ее восьми сегментах значения коэффициентов детерминации линейных регрессий не превосходят величины 0,8, что указывает на наличие неучтенных факторов, поэтому построенную модель нельзя отнести строго ко вполне интерпретируемым. Тем не менее, все остальные условия интерпретируемости выполнены. К тому же сегментированная модель по качеству аппроксимации в целом оказалась гораздо лучше простой линейной регрессии.

Ключевые слова: регрессионный анализ, интерпретируемость, сегментированная линейная регрессия, метод наименьших квадратов, мультиколлинеарность, значимость оценок.

Для цитирования: Базилевский М.П. Алгоритм построения вполне интерпретируемых сегментированных линейных регрессий. *Моделирование, оптимизация и информационные технологии.* 2026;14(3). URL: https://moitvivt.ru/ru/journal/article?id=2212 DOI: 10.26102/2310-6018/2026.54.3.018

Algorithm for constructing fully interpretable segmented linear regressions

M.P. Bazilevskiy✉

Irkutsk State Transport University, Irkutsk, the Russian Federation

Abstract. This article is devoted to the relevant scientific field – interpretable machine learning. Previously, the author introduced the concept of «fully interpretable linear regression», which is constructed using ordinary least squares for the entire set of statistical data. In this article, this concept is generalized to segmented linear regression, in which data is first divided into segments, and then its own linear regression is constructed on each of them. An algorithm for constructing fully interpretable segmented linear regressions has been developed. Its peculiarity is that, firstly, the division of the predictor space into segments is carried out using logical activation functions for the arguments of binary operations min. Secondly, paired regression is construct in each segment, which completely solves the problem of multicollinearity. Using the developed algorithm, a segmented linear regression of concrete compressive strength was constructed based on a sample of 1030 observations. In all its eight segments, the values of the linear regression determination coefficients do not exceed 0.8, which indicates the presence of unaccounted-for factors, so the constructed model cannot be strictly attributed to fully interpretable ones. However, all other interpretability conditions are met. In addition, the segmented model turned out to be much better in terms of approximation quality than simple linear regression.

Keywords: regression analysis, interpretability, segmented linear regression, ordinary least squares, multicollinearity, significance of estimates.

For citation: Bazilevskiy M.P. Algorithm for constructing fully interpretable segmented linear regressions. *Modeling, Optimization and Information Technology*. 2026;13(4). (In Russ.). URL: https://moitvivt.ru/ru/journal/article?id=2212 DOI: 10.26102/2310-6018/2026.54.3.018

Введение

В настоящее время активно развивается такое актуальное научное направление, как интерпретируемое машинное обучение [1, 2]. Как известно, наиболее просты для интерпретации модели линейной регрессии. Однако даже они по некоторым причинам могут обладать разным уровнем интерпретируемости. К таким причинам можно отнести, например, мультиколлинеарность [3, 4], из-за которой коэффициенты регрессии становятся неустойчивыми, что затрудняет их отдельную интерпретацию. К тому же из-за мультиколлинеарности входящие в модель переменные могут оказаться статистически незначимыми, что может привести в результате их исключения к потере важных признаков. Для четкого разделения хорошо и плохо интерпретируемых моделей в [5] введено вероятностно-статистическое определение вполне интерпретируемой линейной регрессии (ВИЛинР), оценки параметров которой находятся с помощью метода наименьших квадратов (МНК) [6]. К ВИЛинР предъявляются довольно жесткие требования, оформленные в виде 8-ми условий. Среди этих условий низкая мультиколлинеарность, согласованность знаков коэффициентов смыслу задачи, значимость оценок по t-критерию Стьюдента, довольно высокое качество аппроксимации и др. В [7] экспериментально доказана высокая эффективность метода построения ВИЛинР, основанного на решении задачи математического программирования.

На сегодняшний день определение ВИЛинР рассматривалось только для линейной регрессии, оцениваемой с помощью МНК по всей выборке данных. Однако мощным инструментом регрессионного моделирования уже давно стала сегментированная линейная регрессия, при которой данные разбиваются на сегменты, и на каждом из которых строится своя линейная регрессия. Очень качественный обзор методов сегментации данных сделан в [8]. Сегментированная линейная регрессия находит широкое применение при решении реальных прикладных задач. Так, например, в [9] построена сегментированная регрессия инвестиционного акселератора в сельском хозяйстве Орловской области, в [10, 11] с помощью программы Joinpoint Regression Program проведен сегментированный регрессионный анализ по медицинским данным, в [12] сегментированная регрессия применена для моделирования количества публикаций на основе традиционных и новых библиографических баз данных. Разработке нового метода сегментации данных в сегментированной регрессии посвящена статья [13].

Цель данной работы – ввести понятие вполне интерпретируемой сегментированной линейной регрессии, разработать основанный на переборных процедурах алгоритм ее построения и использовать его для оценки целесообразности применения сегментированных регрессий в практике интерпретируемого машинного обучения.

Материалы и методы

Пусть выборка данных объема n содержит значения y_i, $i=\overline{1,n}$, зависимой (объясняемой) переменной y и значения x_{ij}, $i=\overline{1,n}$, $j=\overline{1,l}$, независимых (объясняющих) переменных $x_1, x_2, \dots, x_l$. Тогда сегментированная линейная регрессия с s сегментами имеет вид:

$$y_i = \alpha_{k0} + \sum_{j=1}^{l} \alpha_{kj} x_{ij} + \varepsilon_i,\ i \in I_k,\ k = \overline{1,s}, \tag{1}$$

где $I_1, I_2, \dots, I_s$ – непересекающиеся индексные множества, для которых $I_1 \cap I_2 \cap \dots \cap I_s = \emptyset$ и $I_1 \cup I_2 \cup \dots \cup I_s = \{1,2,3,\dots,n\}$; α_{kj}, $k = \overline{1,s}$, $j = \overline{1,l}$ – неизвестный параметр линейной регрессии при j-й переменной в k-м сегменте; ε_i, $i = \overline{1,n}$ – ошибки регрессий.

Преимущество сегментированной линейной регрессии (1) перед классической линейной регрессией, обучаемой по всей выборке, заключается в способности первой моделировать нелинейные зависимости и структурные сдвиги в данных, оставаясь при этом простой и интерпретируемой. Сегментированная линейная регрессия (1) позволяет аппроксимировать сложную нелинейную зависимость набором линейных функций, что часто дает гораздо более точное предсказание, чем одна линейная функция.

Однако даже при построении простой линейной регрессии исследователи часто сталкиваются с проблемой ее интерпретации. Поэтому в ранее опубликованных автором статьях было введено понятие «вполне интерпретируемая линейная регрессия» (ВИЛинР), для построения которой применяется аппарат математического программирования. Это направление активно развивается в настоящее время. Например, в [5] предложено самое крупное на сегодняшний день определение ВИЛинР, состоящее из 8 условий. В ранних работах автора также рассмотрены методы построения вполне интерпретируемых квазилинейных регрессий и вполне интерпретируемых неэлементарных линейных регрессий. Разработано специализированное программное обеспечение для их построения – программа ВИнтер-2. С помощью программы ВИнтер-2 решено множество прикладных задач анализа данных. В текущей статье впервые формулируется идея распространения разработанного автором подхода для построения сегментированных линейных регрессий (1).

Предположим, что каждое уравнение сегментированной линейной регрессии (1) содержит только одну объясняющую переменную. Тогда в результате ее идентификации в каждом сегменте не будет мультиколлинеарности.

Определение 1. Сегментированная линейная регрессия (1), уравнения которой содержат одну объясняющую переменную и идентифицированы с помощью МНК, называется вполне интерпретируемой для выбранного уровня значимости α, если в каждом ее сегменте уравнение удовлетворяет следующим условиям: 1) знаки коэффициентов корреляции r_{yx_j}, $j = \overline{1,l}$ соответствуют содержательному смыслу решаемой задачи; 2) знаки всех оценок $\tilde{\alpha}_{kj}$ согласуются со знаками соответствующих коэффициентов корреляции r_{yx_j}, т. е. $\tilde{\alpha}_{kj} \cdot r_{yx_j} > 0$, $k = \overline{1,s}$, $j = \overline{1,l}$; 3) оценки $\tilde{\alpha}_{kj}$, $k = \overline{1,s}$, $j = \overline{1,l}$ значимы по t-критерию Стьюдента; 4) $R^2 \geq 0{,}8$.

Заметим, что в парных регрессиях значимость оценок $\tilde{\alpha}_{kj}$ по t-критерию Стьюдента равносильна значимости регрессии в целом по F-критерию Фишера.

Перейдем к формированию сегментов в модели (1). Введем еще одну переменную x_{l+1}, состоящую из единиц. Сформируем из индексов переменных $x_1, x_2, \dots, x_{l+1}$ матрицу M, содержащую по строкам в лексикографическом порядке все сочетания из $(l+1)$ элементов по 2. Пусть q – число строк в матрице M. Понятно, что $q = C_{l+1}^2$.

Пусть каждый сегмент модели (1) формируется из наблюдений, удовлетворяющих p линейным неравенствам. Сформируем матрицу Λ размера $C_q^p \times p$, содержащую по строкам в лексикографическом порядке все сочетания из q номеров строк матрицы M по p. И сформируем матрицу H размера $2^p \times p$, содержащую по строкам в лексикографическом порядке все размещения с повторениями из двух элементов (0 и 1, соответствующих знакам $\le$ и $>$ неравенств) по p. С помощью матриц M, Λ и H сформируем следующую совокупность из $s = 2^p$ сегментов:

$$I_1 = \left\{ i \middle| \left(act\left(min\left\{ x_{i,\mu_{\lambda_{j1},1}}, k_{\lambda_{j1}} \cdot x_{i,\mu_{\lambda_{j1},2}} \right\} \right) = h_{11} \right) \& \dots \& \left(act\left(min\left\{ x_{i,\mu_{\lambda_{jp},1}}, k_{\lambda_{jp}} \cdot x_{i,\mu_{\lambda_{jp},2}} \right\} \right) = h_{1p} \right) \right\},$$

$$I_2 = \left\{ i \middle| \left(act\left(min\left\{ x_{i,\mu_{\lambda_{j1},1}}, k_{\lambda_{j1}} \cdot x_{i,\mu_{\lambda_{j1},2}} \right\} \right) = h_{21} \right) \& \dots \& \left(act\left(min\left\{ x_{i,\mu_{\lambda_{jp},1}}, k_{\lambda_{jp}} \cdot x_{i,\mu_{\lambda_{jp},2}} \right\} \right) = h_{2p} \right) \right\},$$

$$\dots$$

$$I_s = \left\{ i \middle| \left(act\left(min\left\{ x_{i,\mu_{\lambda_{j1},1}}, k_{\lambda_{j1}} \cdot x_{i,\mu_{\lambda_{j1},2}} \right\} \right) = h_{s1} \right) \& \dots \& \left(act\left(min\left\{ x_{i,\mu_{\lambda_{jp},1}}, k_{\lambda_{jp}} \cdot x_{i,\mu_{\lambda_{jp},2}} \right\} \right) = h_{sp} \right) \right\},$$

$$j = \overline{1, C_q^p},$$

где min – бинарная операция, возвращающая минимальное значение двух аргументов; act – логическая функция активации [14] аргументов бинарной операций min, принимающая значение 0, если срабатывает первый аргумент, и 1, если второй; k_j, $j = \overline{1,q}$ – известные точки разделения (breakpoints).

Заметим, что

$$act\left(min\left\{ x_{i,\mu_{\lambda_{ju},1}}, k_{\lambda_{ju}} \cdot x_{i,\mu_{\lambda_{ju},2}} \right\} \right) = 0 \Leftrightarrow x_{i,\mu_{\lambda_{ju},1}} \le k_{\lambda_{ju}} \cdot x_{i,\mu_{\lambda_{ju},2}},$$

$$act\left(min\left\{ x_{i,\mu_{\lambda_{ju},1}}, k_{\lambda_{ju}} \cdot x_{i,\mu_{\lambda_{ju},2}} \right\} \right) = 1 \Leftrightarrow x_{i,\mu_{\lambda_{ju},1}} > k_{\lambda_{ju}} \cdot x_{i,\mu_{\lambda_{ju},2}}, \ u = \overline{1,p}.$$

При известных точках разделения k_j, $j = \overline{1,q}$, для идентификации параметров сегментированной модели (1) простым перебором необходимо оценить $N_1 = C_q^p \cdot 2^p$ линейных регрессий с назначенным списком объясняющих переменных. Затем выбрать такое разделение выборки, которое обеспечивает минимальное суммарное значение сумм квадратов остатков в сегментах.

Если точки разделения неизвестны, то идентификация параметров сегментированной модели (1) многократно усложняется в вычислительном плане. Нетрудно доказать [14], что неизвестные точки разделения для формируемых сегментов нужно искать внутри промежутков:

$$\left(k_j^{\text{нижн}}, k_j^{\text{верхн}} \right), \ j = \overline{1,q}, \tag{2}$$

где $k_j^{\text{нижн}} = min\left\{ \frac{x_{1,\mu_{j1}}}{x_{1,\mu_{j2}}}, \frac{x_{2,\mu_{j1}}}{x_{2,\mu_{j2}}}, \dots, \frac{x_{n,\mu_{j1}}}{x_{n,\mu_{j2}}} \right\}$, $k_j^{\text{верхн}} = max\left\{ \frac{x_{1,\mu_{j1}}}{x_{1,\mu_{j2}}}, \frac{x_{2,\mu_{j1}}}{x_{2,\mu_{j2}}}, \dots, \frac{x_{n,\mu_{j1}}}{x_{n,\mu_{j2}}} \right\}$.

Разобьем промежутки (2) равномерным образом r точками. Сформируем из этих точек матрицу K с элементами κ_{uj}, $u = \overline{1,r}$, $j = \overline{1,q}$ – u-я точка разделения j-го промежутка (2). Далее сформируем матрицу Z размера $r^p \times p$, содержащую по строкам в лексикографическом порядке все размещения с повторениями из r элементов (соответствующих номерам точек разделения) по p. Тогда с помощью матриц M, Λ, H, K и Z сформируем следующую совокупность из $s = 2^p$ сегментов:

$$I_1 = \left\{ i \middle| \left(act\left(min\left\{ x_{i,\mu_{\lambda_{j1},1}}, \kappa_{z_{u1},\lambda_{j1}} \cdot x_{i,\mu_{\lambda_{j1},2}} \right\} \right) = h_{11} \right) \& \dots \& \left(act\left(min\left\{ x_{i,\mu_{\lambda_{jp},1}}, \kappa_{z_{up},\lambda_{jp}} \cdot x_{i,\mu_{\lambda_{jp},2}} \right\} \right) = h_{1p} \right) \right\},$$

$$I_2 = \left\{ i \middle| \left(act\left(min\left\{ x_{i,\mu_{\lambda_{j1},1}}, \kappa_{z_{u1},\lambda_{j1}} \cdot x_{i,\mu_{\lambda_{j1},2}} \right\} \right) = h_{21} \right) \& \dots \& \left(act\left(min\left\{ x_{i,\mu_{\lambda_{jp},1}}, \kappa_{z_{up},\lambda_{jp}} \cdot x_{i,\mu_{\lambda_{jp},2}} \right\} \right) = h_{2p} \right) \right\},$$

$$\dots$$

$$I_s = \left\{ i \middle| \left(act\left(min\left\{ x_{i,\mu_{\lambda_{j1},1}}, \kappa_{z_{u1},\lambda_{j1}} \cdot x_{i,\mu_{\lambda_{j1},2}} \right\} \right) = h_{s1} \right) \& \dots \& \left(act\left(min\left\{ x_{i,\mu_{\lambda_{jp},1}}, \kappa_{z_{up},\lambda_{jp}} \cdot x_{i,\mu_{\lambda_{jp},2}} \right\} \right) = h_{sp} \right) \right\},$$

$$j = \overline{1, C_q^p}, \ u = \overline{1, r^p}.$$

При таком формировании сегментов для идентификации параметров сегментированной модели (1) простым перебором необходимо оценить $N_2 = C_q^p \cdot 2^p \cdot r^p$ линейных регрессий. Для снижения вычислительной сложности такой задачи предлагается формировать сегменты так, чтобы в образующие их p линейные неравенства каждая из l объясняющих переменных входила не более одного раза. В таком случае для идентификации сегментированной модели (1) количество линейных регрессий для оценки составит

$$N_3 = \left(\sum_{k=0}^{p} C_l^k \cdot \prod_{j=1}^{p-k} C_{l-k-2(j-1)}^2 \cdot \left((p-k)!\right)^{-1}\right) \cdot 2^p \cdot r^p.$$

Для обеспечения высокого уровня интерпретируемости предлагается в каждом сегменте по заданному списку переменных оценивать с помощью МНК не одну множественную регрессию, а совокупность парных линейных регрессий. Если такой список состоит из m переменных, то для идентификации модели (1) количество парных регрессий, а фактически, коэффициентов корреляции между объясняющими переменными и y, для оценки составит

$$N_4 = \left(\sum_{k=0}^{p} C_l^k \cdot \prod_{j=1}^{p-k} C_{l-k-2(j-1)}^2 \cdot \left((p-k)!\right)^{-1}\right) \cdot 2^p \cdot r^p \cdot m. \quad (3)$$

Алгоритм построения вполне интерпретируемой сегментированной линейной регрессии, удовлетворяющей условиям определения 1, представлен на Рисунке 1.

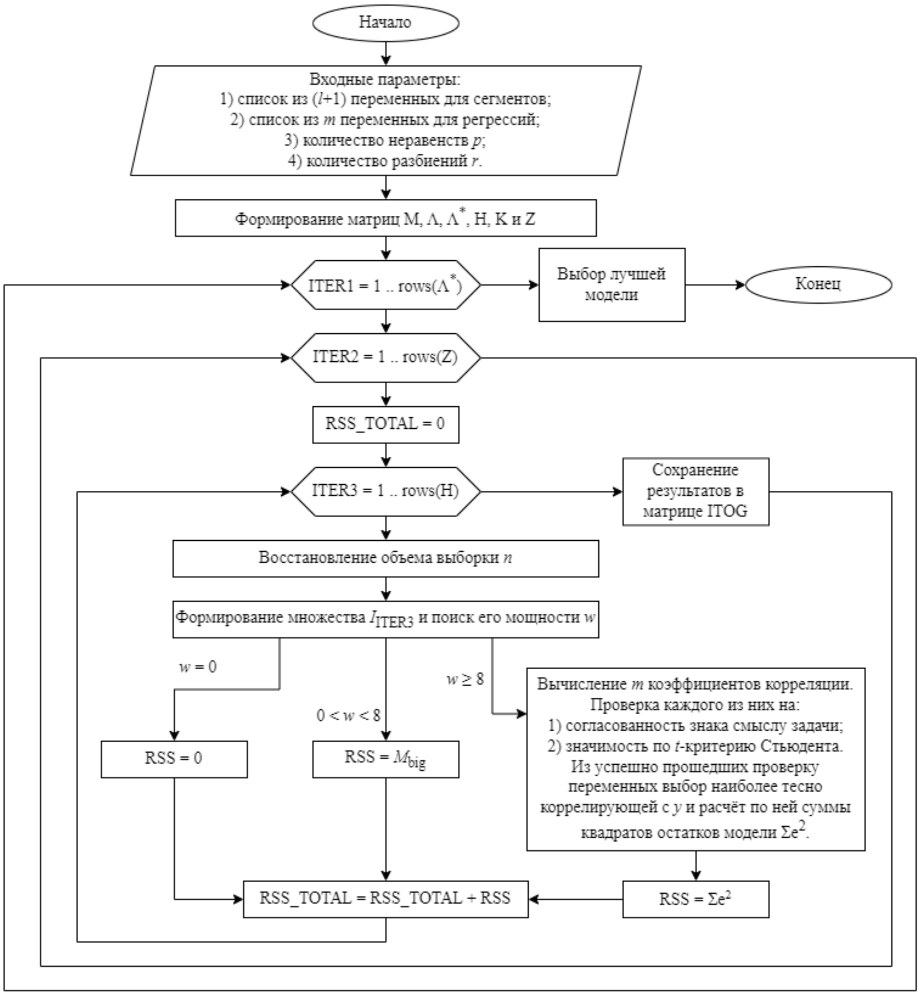

Рисунок 1 – Алгоритм построения сегментированной линейной регрессии
Figure 1 – Algorithm for constructing segmented linear regression

На Рисунке 1 матрица Λ^* формируется только из тех строк матрицы Λ, в которых номера объясняющих переменных встречаются не более одного раза. В зависимости от мощности w множества наблюдений некоторого сегмента меняется вычислительный сценарий. Если $0 < w < 8$, то такая сегментированная регрессия из-за малого объема выборки в сегменте исключается из рассмотрения за счет присвоения критерию качества довольно большого числа M_{big}. Если $w \geq 8$, то осуществляется вычисление и проверка значений коэффициентов корреляции. Если в регрессии есть пустые сегменты ($w = 0$), то сегментированная регрессия не исключается.

Результаты

Для моделирования были использованы статистические данные из хранилища[1] по следующим переменным: y – прочность бетона на сжатие (МПа); x_1 – цемент (кг/м^3 смеси); x_2 – шлак доменной печи (кг/м^3); x_3 – зола (кг/м^3); x_4 – вода (кг/м^3); x_5 – пластификатор (кг/м^3); x_6 – щебень (кг/м^3); x_7 – песок (кг/м^3); x_8 – срок твердения бетона (дни). Объем выборки $n = 1030$ наблюдений.

Представленный на Рисунке 1 алгоритм был реализован в виде скрипта для эконометрического пакета Gretl. Моделирование проводилось на персональном компьютере с процессором AMD Ryzen 3 4300U (2.7 ГГц) и 16 Гб оперативной памяти. Для построения сегментированной модели в программе были выбраны следующие параметры поиска: список переменных для сегментов – x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9 (x_9 – единичная переменная); список переменных для регрессий – $z = x_1/x_4$, x_2, x_3, x_5, x_6, x_7, x_8; количество неравенств $p = 3$; количество разбиений $r = 5$.

Большое число $M_{\text{big}} = 10^{10}$. Поскольку переменные x_2, x_3, x_5 и x_8 могут принимать нулевые значения, что приводит к делению на 0 при вычислении промежутков (2), то было решено использовать вместо них показатели x_2+1, x_3+1, x_5+1 и x_8+1. После построения модели осуществлялся возврат к исходным переменным. Выбор переменной $z = x_1/x_4$ (цементно-водного соотношения) продиктован описанными в [15] формулами Д. Абрамса и М. Боломея.

Сначала сегментированная линейная регрессия строилась без учета сформулированных в определении 1 условий. Для этого, как следует из (3), требовалось оценить 12152000 парных регрессий. На решение этой задачи ушло примерно 88 минут. Результаты решения представлены в Таблице 1. В ней в первом столбце указан номер сегмента, во втором – количество наблюдений в сегменте, в третьем – уравнение в сегменте, в четвертом – правило формирования сегмента, в пятом и шестом – значение коэффициента детерминации R^2 и суммы квадратов остатков Σe^2.

Таблица 1 – Характеристики построенной сегментированной линейной регрессии
Table 1 – Characteristics of the constructed segmented linear regression

№	***w***	**Уравнение**	**Система неравенств**	***R*²**	**Σ*e*²**
1	122	$\tilde{y} = 2{,}561 + 9{,}492z$	$(z \leq 1{,}9683)\&(x_5 \leq 3{,}0666)\&(x_8 \leq 17{,}333)$	0,337	3188,246
2	197	$\tilde{y} = 81{,}754 - 0{,}06178x_7$	$(z \leq 1{,}9683)\&(x_5 \leq 3{,}0666)\&(x_8 > 17{,}333)$	0,324	12856,52
3	108	$\tilde{y} = 15{,}411 + 1{,}0728x_8$	$(z \leq 1{,}9683)\&(x_5 > 3{,}0666)\&(x_8 \leq 17{,}333)$	0,390	5648,069
4	359	$\tilde{y} = 8{,}679 + 23{,}33z$	$(z \leq 1{,}9683)\&(x_5 > 3{,}0666)\&(x_8 > 17{,}333)$	0,414	36870,68
5	30	$\tilde{y} = -27{,}41 + 23{,}661z$	$(z > 1{,}9683)\&(x_5 \leq 3{,}0666)\&(x_8 \leq 17{,}333)$	0,610	1867,064
6	45	$\tilde{y} = -16{,}293 + 26{,}563z$	$(z > 1{,}9683)\&(x_5 \leq 3{,}0666)\&(x_8 > 17{,}333)$	0,782	1873,828
7	64	$\tilde{y} = 25{,}495 + 3{,}228x_8$	$(z > 1{,}9683)\&(x_5 > 3{,}0666)\&(x_8 \leq 17{,}333)$	0,437	3431,983
8	105	$\tilde{y} = 55{,}688 + 0{,}0779x_2$	$(z > 1{,}9683)\&(x_5 > 3{,}0666)\&(x_8 > 17{,}333)$	0,315	7426,632

[1] Yeh I.-Ch. *Concrete Compressive Strength.* UC Irvine Machine Learning Repository. URL: https://archive.ics.uci.edu/dataset/165/concrete+compressive+strength (дата обращения: 14.01.2026).

Наблюдаемые значения t-критерия Стьюдента для коэффициентов, стоящих при переменных в уравнениях в Таблице 1, равны 7,815, −9,670, 8,246, 15,89, 6,621, 12,43, 6,944 и 6,884 соответственно. Как видно, эти коэффициенты во всех сегментах значимы по t-критерию Стьюдента для уровня значимости α = 0,01. Знаки этих коэффициентов согласуются с направлениями влияния переменных на прочность бетона. Тем самым, выполняются все условия определения 1, кроме последнего. К сожалению, коэффициенты детерминации уравнений в каждом сегменте меньше 0,8, что не позволяет считать построенную модель вполне интерпретируемой.

Затем при тех же параметрах поиска сегментированная линейная регрессия строилась с учетом сформулированных в определении 1 условий. Однако ни одна спецификация не прошла проверку всех условий.

Для сравнения с помощью МНК по всей выборке была оценена простая линейная регрессия переменной y от z, x_2, x_3, x_5, x_6, x_7, x_8, для которой $R^2 = 0{,}6062$, $\Sigma e^2 = 113087{,}9$. Проверка показала, что эта модель не относится к ВИЛинР из-за низкого качества и наличия эффекта мультиколлинеарности. Для приведенной в Таблице 1 сегментированной линейной регрессии общая сумма квадратов остатков $\Sigma e^2 = 73163{,}02$, что на 35,3 % ниже, чем у простой линейной регрессии.

Дополнительно проводилось моделирование по встроенным в пакет Gretl данным «data7-12» о ценах на автомобили и их характеристиках. Были взяты первые 40 наблюдений по следующим переменным: y^* – цена автомобиля; x_1^* – колесная база, x_2^* – длина, x_3^* – ширина, x_4^* – высота, x_5^* – вес, x_6^* – количество цилиндров в двигателе, x_7^* – объем двигателя, x_8^* – расход топлива. Результаты моделирования представлены в Таблице 2.

Таблица 2 – Характеристики построенной сегментированной линейной регрессии
Table 2 – Characteristics of the constructed segmented linear regression

№	*w*	Уравнение	Система неравенств	R^2	Σe^2
1	18	$\widetilde{y^*} = 2{,}793 + 5{,}766x_7^*$	$(x_1^* \le 103{,}9)\&(x_3^* \le 70{,}75)$	0,803	112,3
2	6	$\widetilde{y^*} = 414{,}573 - 7{,}576x_4^*$	$(x_1^* \le 103{,}9)\&(x_3^* \ge 70{,}75)$	0,882	264,675
3	7	$\widetilde{y^*} = -603{,}82 + 5{,}927x_1^*$	$(x_1^* \ge 103{,}9)\&(x_3^* \le 70{,}75)$	0,917	13,995
4	9	$\widetilde{y^*} = -35{,}965 + 9{,}515x_6^*$	$(x_1^* \ge 103{,}9)\&(x_3^* \ge 70{,}75)$	0,878	100,218

Наблюдаемые значения t-критерия Стьюдента для коэффициентов, стоящих при переменных в уравнениях в Таблице 2, равны 8,073, −5,460, 7,445 и 7,113. Все коэффициенты в сегментах значимы по t-критерию Стьюдента для уровня значимости α = 0,01. Знаки этих коэффициентов согласуются с направлениями влияния переменных на цену автомобиля. Тем самым, выполняются все условия определения 1, поэтому построенную модель можно считать вполне интерпретируемой.

Для сравнения с помощью МНК по всей выборке была оценена простая линейная регрессия переменной y^* от x_1^*, x_2^*, x_3^*, x_4^*, x_5^*, x_6^*, x_7^*, x_8^*, для которой $R^2 = 0{,}6968$, $\Sigma e^2 = 1475{,}75$. Проверка показала, что эта модель не относится к ВИЛинР из-за низкого качества и наличия эффекта мультиколлинеарности. Для приведенной в Таблице 2 сегментированной линейной регрессии общая сумма квадратов остатков $\Sigma e^2 = 491{,}188$, что на 66,716 % ниже, чем у простой линейной регрессии.

Заключение

В статье впервые введено понятие «вполне интерпретируемая сегментированная линейная регрессия» и предложен основанный на переборных процедурах алгоритм ее построения. Построенную с помощью этого алгоритма модель прочности бетона нельзя считать строго вполне интерпретируемой, поскольку во всех ее сегментах значения R^2 линейных регрессий не превосходят 0,8. Тем не менее, по величине суммы квадратов остатков она оказалась на 35,3 % лучше, чем простая линейная регрессия, поэтому ее коэффициенты в сегментах, значимые по t-критерию Стьюдента, все же можно интерпретировать, помня о наличии неучтенных факторов. В целом, применение сегментированной линейной регрессии в практике интерпретируемого машинного обучения оправдывает себя, но выделенное в данной статье направление требует дополнительных исследований.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Molnar Ch. *Interpretable Machine Learning*. Lulu.com; 2020. 320 p.
2. Carter A., Imtiaz S., Naterer G.F. Review of Interpretable Machine Learning for Process Industries. *Process Safety and Environmental Protection*. 2023;170:647–659. https://doi.org/10.1016/j.psep.2022.12.018
3. Shrestha N. Detecting Multicollinearity in Regression Analysis. *American Journal of Applied Mathematics and Statistics*. 2020;8(2):39–42. https://doi.org/10.12691/ajams-8-2-1
4. Aslam M., Ahmad Sh. The Modified Liu-ridge-type Estimator: A New Class of Biased Estimators to Address Multicollinearity. *Communications in Statistics – Simulation and Computation*. 2022;51(11):6591–6609. https://doi.org/10.1080/03610918.2020.1806324
5. Базилевский М.П. О разрешимости оптимизационной задачи построения вполне интерпретируемых линейных регрессий. *Учёные записки Казанского университета. Серия Физико-математические науки*. 2025;167(4):627–640. https://doi.org/10.26907/2541-7746.2025.4.627-640
   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;167(4):627–640. (In Russ.). https://doi.org/10.26907/2541-7746.2025.4.627-640
6. Голиков Р.Ю. Кусочно-линейная аппроксимация формы сильно зашумленного сигнала методом наименьших квадратов. *Прикладная информатика*. 2022;17(5):116–124. https://doi.org/10.37791/2687-0649-2022-17-5-116-124
   Golikov R.Y. Piecewise linear approximation of a highly noisy signal waveform using least squares method. *Journal of Applied Informatics*. 2022;17(5):116–124. (In Russ.). https://doi.org/10.37791/2687-0649-2022-17-5-116-124
7. 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. https://doi.org/10.26089/NumMet.v26r435
   Bazilevskiy M.P. 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*. 2025;26(4):534–547. (In Russ.). https://doi.org/10.26089/NumMet.v26r435
8. Лебедев И.С. Адаптивное построение регрессионных моделей на основе анализа функционала качества обработки сегментов последовательности. *Информатика и автоматизация*. 2025;24(2):363–394. https://doi.org/10.15622/ia.24.2.1
   Lebedev I. Adaptive Regression Model Construction based on the Functional Quality Analysis of the Sequence Segment Processing. *Informatics and Automation*. 2025;24(2):363–394. (In Russ.). https://doi.org/10.15622/ia.24.2.1

9. Шестаков Р.Б., Ловчикова Е.И. Инвестиционный акселератор сельскохозяйственного производства. *Экономика региона*. 2019;15(3):908–923. https://doi.org/10.17059/2019-3-21
Shestakov R.B., Lovchikova E.I. Investment Accelerator of Agricultural Production. *Economy of Region*. 2019;15(3):908–923. (In Russ.). https://doi.org/10.17059/2019-3-21
10. Дубовиченко Д.М., Вальков М.Ю., Карпунов А.А., Панкратьева А.Ю. Популяционная оценка динамики заболеваемости и стадийной структуры рака прямой кишки в условиях реализации мероприятий Национального проекта «Здоровье» и диспансеризации определенных групп взрослого населения в Архангельской области (итоги предварительного исследования). *Research'n Practical Medicine Journal*. 2017;4(3):23–32. https://doi.org/10.17709/2409-2231-2017-4-3-3
Dubovichenko D.M., Valkov M.Y., Karpunov A.A., Pankrat'eva A.Yu. Population-based Assessment of the Rectal Cancer Stage Structure and Incidence after Implementation of the National Project "Health" and All-national Dispensarization in the Arkhangelsk region, Russia (The results of the preliminary study). *Research'n Practical Medicine Journal*. 2017;4(3):23–32. (In Russ.). https://doi.org/10.17709/2409-2231-2017-4-3-3
11. Макарова М.В., Вальков М.Ю. Сегментированный анализ динамических рядов официальных статистических показателей остеоартрита в 1994–2018 гг. в России, Северо-Западном федеральном округе и Архангельской области. *Научно-практическая ревматология*. 2021;59(5):584–591. https://doi.org/10.47360/1995-4484-2021-584-591
Makarova M.V., Valkov M.Yu. Segmented Analysis of Official Statistical Indicators Dynamic Series for Osteoarthritis in 1994–2018 in Russia, the North-Western Federal district and the Arkhangelsk region. *Rheumatology Science and Practice*. 2021;59(5):584–591. (In Russ.). https://doi.org/10.47360/1995-4484-2021-584-591
12. Bornmann L., Haunschild R., Mutz R. Growth Rates of Modern Science: A Latent Piecewise Growth Curve Approach to Model Publication Numbers from Established and New Literature Databases. *Humanities and Social Sciences Communications*. 2021;8. https://doi.org/10.1057/s41599-021-00903-w
13. Al-Azzeh J., Mesleh A., Zaliskyi M., Odarchenko R., Kuzmin V. A Method of Accuracy Increment Using Segmented Regression. *Algorithms*. 2022;15(10). https://doi.org/10.3390/a15100378
14. Базилевский М.П. Построение двухфакторных неэлементарных линейных регрессий с логическими функциями. *International Journal of Open Information Technologies*. 2023;11(6):113–117.
Bazilevskiy M.P. Construction of Two-factor Non-elementary Linear Regressions with Boolean Functions. *International Journal of Open Information Technologies*. 2023;11(6):113–117. (In Russ.).
15. Михайлова Н.А., Стефаненко И.В. Множественные регрессионные модели прочности бетона на сжатие. *Вестник Волгоградского государственного архитектурно-строительного университета. Серия: Строительство и архитектура*. 2017;(49):30–42.
Mikhailova N.A., Stefanenko I.V. Multiple Regression Models of Concrete Compressive Strength. *Bulletin of the Volgograd State University of Architecture and Civil Engineering. The Construction and Architecture series.* 2017;(49):30–42. (In Russ.).

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Базилевский Михаил Павлович, кандидат технических наук, доцент, доцент кафедры «Математика», Иркутский государственный университет путей сообщения, Иркутск, Российская Федерация.
e-mail: mik2178@yandex.ru
ORCID: 0000-0002-3253-5697

Mikhail P. Bazilevskiy, Candidate of Engineering Sciences, Docent, Associate Professor at the Department of Mathematics, Irkutsk State Transport University, Irkutsk, the Russian Federation.

Статья поступила в редакцию 02.02.2026; одобрена после рецензирования 19.03.2026; принята к публикации 26.03.2026.

The article was submitted 02.02.2026; approved after reviewing 19.03.2026; accepted for publication 26.03.2026.