ВЕСТНИК ЮГОРСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА
2022 г. Выпуск 4 (67). С. 105–114

УДК 519.862.6
DOI: 10.18822/byusu202204105-114

ПОСТРОЕНИЕ ВПОЛНЕ ИНТЕРПРЕТИРУЕМЫХ НЕЭЛЕМЕНТАРНЫХ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Базилевский Михаил Павлович
кандидат технических наук,
доцент кафедры «Математика»,
ФГБОУ ВО «Иркутский государственный университет путей сообщения»
Иркутск, Россия
E-mail: mik2178@yandex.ru

Предмет исследования: задача частично-булевого линейного программирования, предназначенная для выбора оптимальных структур неэлементарных линейных регрессионных моделей.

Цель исследования: интегрировать в задачу частично-булевого линейного программирования дополнительные ограничения, которые будут гарантировать построение вполне интерпретируемых неэлементарных линейных регрессий.

Методы исследования: регрессионный анализ, математическое программирование, метод последовательного повышения абсолютных вкладов переменных в общую детерминацию.

Объект исследования: неэлементарные линейные регрессионные модели.

Основные результаты исследования: в задачу частично-булевого линейного программирования, предназначенную для построения неэлементарных регрессий, интегрированы дополнительные линейные ограничения на абсолютные вклады переменных в общую детерминацию, позволяющие контролировать в модели как сами вклады, так и мультиколлинеарность. Показано, каким образом необходимо регулировать эти ограничения, чтобы полученная в результате решения задачи неэлементарная линейная регрессия была вполне интерпретируемой. Предложенный математический аппарат был использован для моделирования железнодорожных грузоперевозок Тюменской области. Дана интерпретация полученной высокоточной и вполне интерпретируемой неэлементарной линейной регрессии.

Ключевые слова: неэлементарная линейная регрессия, вполне интерпретируемая регрессия, задача частично-булевого линейного программирования, метод наименьших квадратов, мультиколлинеарность, абсолютные вклады переменных в общую детерминацию, железнодорожные грузоперевозки Тюменской области.

CONSTRUCTION OF QUITE INTERPRETABLE NON-ELEMENTARY LINEAR REGRESSION MODELS

Mikhail P. Bazilevskiy
Candidate of Technical Sciences,
Associate Professor of the Department of Mathematics,
Irkutsk State Transport University
Irkutsk, Russia
E-mail: mik2178@yandex.ru

Subject of research: mixed-integer 0-1 linear programming problem for choosing optimal structures for non-elementary linear regression models.

Purpose of research: integrate into the mixed-integer 0-1 linear programming problem additional constraints that will guarantee the construction of quite interpretable non-elementary linear regressions.

Methods of research: regression analysis, mathematical programming, method of successive increase the absolute contributions of variables to the general determination.

Object of research: non-elementary linear regression models.

Main results of research: in the mixed-integer 0-1 linear programming problem, designed to construct non-elementary regressions, additional linear constraints on the absolute contributions of variables to the general determination are integrated, allowing you to control both the contributions themselves and multicollinearity in the model. It is shown how it is necessary to regulate these constraints so that the non-elementary linear regression obtained as a result of solving the problem is quite interpretable. The proposed mathematical apparatus was used to model railroad freight transportation in the Tyumen region. An interpretation of the obtained high-precision and quite interpretable non-elementary linear regression is given.

Keywords: non-elementary linear regression, quite interpretable regression, mixed-integer 0-1 linear programming problem, ordinary least squares method, multicollinearity, absolute contributions of variables to the general determination, railway freight transportation in the Tyumen region.

Введение

Актуальным в настоящее время направлением в науке считается интерпретируемое машинное обучение [1–4]. Как отмечено в [2], со ссылкой на [3], интерпретируемость дает возможность моделям машинного обучения представлять свое поведение в понятных людям терминах. С точки зрения конечных пользователей, интерпретируемость повышает доверие к модели машинного обучения, поскольку им становится ясно и понятно, как именно она работает. С точки зрения разработчиков, интерпретируемость помогает лучше понять проблему, как устроены данные и причины неточной работы модели, что в конечном итоге приводит к повышению её точности.

Среди моделей машинного обучения высокой степенью интерпретируемости обладают регрессионные модели [5, 6]. Среди них самыми простыми закономерно следует считать линейные регрессии, в которых каждый коэффициент трактуется как величина изменения зависимой переменной при изменении соответствующей ему объясняющей переменной на одну условную единицу. Однако даже при построении линейной регрессии может оказаться так, что у неё будут искажены знаки коэффициентов при объясняющих переменных. Причина такого искажения – мультиколлинеарность, означающая наличие сильной корреляционной связи между объясняющими переменными. Таким образом, мультиколлинеарность негативно сказывается на интерпретируемости регрессионных моделей.

На сегодняшний день ведется активная работа по созданию новых эффективных форм связи между переменными в регрессионных моделях. Так, например, в работах [7–10] исследуются так называемые кусочно-линейные регрессии, для оценки неизвестных параметров которых авторы используют метод наименьших модулей. В то же время в работах [11–14] предложены неэлементарные регрессионные модели, являющиеся обобщением линейных регрессий и содержащие в своем составе помимо объясняющих переменных все возможные комбинации их пар, преобразованных с помощью бинарных операций min и max. Для оценки таких моделей используется метод наименьших квадратов (МНК). В [14] предложен метод построения неэлементарных линейных регрессий (НЛР) на основе аппарата математического программирования и продемонстрированы их высокие интерпретационные свойства. Однако вопрос о том, каким образом следует контролировать мультиколлинеарность в НЛР, а значит, и их интерпретируемость, решен не был.

Работа [15] посвящена построению вполне интерпретируемых линейных регрессионных моделей. Вполне интерпретируемая регрессия удовлетворяет трем условиям:

1) её спецификация изначально выбрана так, что после оценивания можно объяснить любой коэффициент модели или некоторый его аналог, за исключением, быть может, свободного члена;
2) все знаки коэффициентов модели соответствуют содержательному смыслу решаемой задачи;
3) эффект мультиколлинеарности незначителен.

Для построения вполне интерпретируемых линейных регрессий в работе [15] был предложен метод последовательного повышения абсолютных вкладов переменных в общую детерминацию. Целью данной работы является интеграция в задачу построения НЛР дополнительных ограничений, позволяющих контролировать абсолютные вклады и мультиколлинеарность, что позволит сформулировать алгоритм, гарантирующий получение вполне интерпретируемых НЛР.

Результаты и обсуждение

НЛР [14] имеет вид

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \sum_{j=1}^{C_l^2} \alpha_{j+l} \min\{x_{i,\mu_{j1}}, \lambda_{j1} x_{i,\mu_{j2}}\} + \\ + \sum_{j=1}^{C_l^2} \alpha_{j+l+C_l^2} \max\{x_{i,\mu_{j1}}, \lambda_{j2} x_{i,\mu_{j2}}\} + \varepsilon_i, \quad i = \overline{1,n}, \tag{1}$$

где n – объем выборки; l – количество объясняющих переменных; y_i – i-е значение объясняемой переменной y; $x_{ij} > 0$ – i-е значение j-й объясняющей переменной; ε_i – i-я ошибка аппроксимации; α_0, α_1, ..., $\alpha_{l+2C_l^2}$, λ_{11}, λ_{21}, ..., $\lambda_{C_l^2,1}$, λ_{12}, λ_{22}, ..., $\lambda_{C_l^2,2}$ – неизвестные параметры; μ_{j1}, μ_{j2} – элементы j-й строки индексной матрицы M размера $C_l^2 \times 2$, содержащей в строках всевозможные комбинации пар индексов переменных.

Придавая каждому из параметров λ_{11}, λ_{21}, ..., $\lambda_{C_l^2,1}$, λ_{12}, λ_{22}, ..., $\lambda_{C_l^2,2}$ p значений так, как это сделано в [14], можно перейти к регрессии

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \sum_{j=1}^{C_l^2} \sum_{k=1}^{p} \alpha_{jk}^{-} z_{ijk}^{-} + \sum_{j=1}^{C_l^2} \sum_{k=1}^{p} \alpha_{jk}^{+} z_{ijk}^{+} + \varepsilon_i, \quad i = \overline{1,n}, \tag{2}$$

где α_{jk}^{-}, α_{jk}^{+}, $j = \overline{1, C_l^2}$, $k = \overline{1,p}$ – параметры для регрессоров с бинарной операцией min и max соответственно; $z_{ijk}^{-} = \min\left\{x_{i,\mu_{j1}}, \lambda_{jk}^{*} x_{i,\mu_{j2}}\right\}$, $z_{ijk}^{+} = \max\left\{x_{i,\mu_{j1}}, \lambda_{jk}^{*} x_{i,\mu_{j2}}\right\}$, $i = \overline{1,n}$, $j = \overline{1, C_l^2}$, $k = \overline{1,p}$.

Проведя нормирование всех переменных в (2) по известному правилу [14], вместо (2) можно получить стандартизованную регрессию с неизвестными коэффициентами β_j, $j = \overline{1,l}$ и β_{jk}^{-}, β_{jk}^{+}, $j = \overline{1, C_l^2}$, $k = \overline{1,p}$. А с помощью неё можно сформулировать следую-

щую задачу частично-булевого линейного программирования (ЧБЛП) построения модели НЛР:

$$R^2 = \sum_{j=1}^{l} r_{yx_j} \cdot \beta_j + \sum_{j=1}^{C_l^2} \sum_{k=1}^{p} r_{yz_{jk}^-} \cdot \beta_{jk}^- + \sum_{j=1}^{C_l^2} \sum_{k=1}^{p} r_{yz_{jk}^+} \cdot \beta_{jk}^+ \to \max, \tag{3}$$

$$-\left(1-\delta_j\right)M \le \sum_{k=1}^{l} r_{x_j x_k} \cdot \beta_k + \sum_{s=1}^{C_l^2} \sum_{k=1}^{p} r_{x_j z_{sk}^-} \cdot \beta_{sk}^- +$$

$$+\sum_{s=1}^{C_l^2} \sum_{k=1}^{p} r_{x_j z_{sk}^+} \cdot \beta_{sk}^+ - r_{yx_j} \le (1-\delta_j)M,\ j=\overline{1,l}, \tag{4}$$

$$-\left(1-\delta_{jk}^-\right)M \le \sum_{s=1}^{l} r_{x_s z_{jk}^-} \cdot \beta_s + \sum_{s_1=1}^{C_l^2} \sum_{s_2=1}^{p} r_{z_{s_1 s_2}^- z_{jk}^-} \cdot \beta_{s_1 s_2}^- +$$

$$+\sum_{s_1=1}^{C_l^2} \sum_{s_2=1}^{p} r_{z_{s_1 s_2}^+ z_{jk}^-} \cdot \beta_{s_1 s_2}^+ - r_{yz_{jk}^-} \le (1-\delta_{jk}^-)M,\ j=\overline{1,C_l^2},\ k=\overline{1,p}, \tag{5}$$

$$-\left(1-\delta_{jk}^+\right)M \le \sum_{s=1}^{l} r_{x_s z_{jk}^+} \cdot \beta_s + \sum_{s_1=1}^{C_l^2} \sum_{s_2=1}^{p} r_{z_{s_1 s_2}^- z_{jk}^+} \cdot \beta_{s_1 s_2}^- +$$

$$+\sum_{s_1=1}^{C_l^2} \sum_{s_2=1}^{p} r_{z_{s_1 s_2}^+ z_{jk}^+} \cdot \beta_{s_1 s_2}^+ - r_{yz_{jk}^+} \le (1-\delta_{jk}^+)M,\ j=\overline{1,C_l^2},\ k=\overline{1,p}, \tag{6}$$

$$0 \le \beta_j \le \left(r_{yx_j}\right)^{-1} \delta_j,\ j \in \left\{s \mid r_{yx_s} > 0\right\}, \tag{7}$$

$$\left(r_{yx_j}\right)^{-1} \delta_j \le \beta_j \le 0,\ j \in \left\{s \mid r_{yx_s} < 0\right\}, \tag{8}$$

$$0 \le \beta_{jk}^- \le \left(r_{yz_{jk}^-}\right)^{-1} \delta_{jk}^-,\ (j,k) \in \left\{(s_1,s_2) \mid r_{yz_{s_1 s_2}^-} > 0\right\}, \tag{9}$$

$$\left(r_{yz_{jk}^-}\right)^{-1} \delta_{jk}^- \le \beta_{jk}^- \le 0,\ (j,k) \in \left\{(s_1,s_2) \mid r_{yz_{s_1 s_2}^-} < 0\right\}, \tag{10}$$

$$0 \le \beta_{jk}^+ \le \left(r_{yz_{jk}^+}\right)^{-1} \delta_{jk}^+,\ (j,k) \in \left\{(s_1,s_2) \mid r_{yz_{s_1 s_2}^+} > 0\right\}, \tag{11}$$

$$\left(r_{yz_{jk}^+}\right)^{-1} \delta_{jk}^+ \le \beta_{jk}^+ \le 0,\ (j,k) \in \left\{(s_1,s_2) \mid r_{yz_{s_1 s_2}^+} < 0\right\}, \tag{12}$$

$$\delta_j \in \{0,1\},\ j=\overline{1,l};\ \delta_{jk}^- \in \{0,1\},\ \delta_{jk}^+ \in \{0,1\},\ j=\overline{1,C_l^2},\ k=\overline{1,p}, \tag{13}$$

где R^2 – коэффициент детерминации модели; символом r_{XY} обозначены коэффициенты парной корреляции между переменными X и Y; δ_j, $j=\overline{1,l}$ – булевы переменные, заданные по правилу

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

δ_{jk}^{-}, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$ – булевы переменные, заданные по правилу

$$\delta_{jk}^{-}=\begin{cases}1, \text{ если j-я бинарная операция минимум с k-м преобразованием}\\ \text{входит в регрессию,}\\ 0, \text{ в противном случае;}\end{cases}$$

δ_{jk}^{+}, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$ – булевы переменные, заданные по правилу

$$\delta_{jk}^{+}=\begin{cases}1, \text{ если j-я бинарная операция максимум с k-м преобразованием}\\ \text{входит в регрессию,}\\ 0, \text{ в противном случае;}\end{cases}$$

M – большое положительное число, возможный способ выбора которого подробно описан в [14].

Решение задачи ЧБЛП с целевой функцией (3) и с линейными ограничениями (4)–(13) приводит к выбору оптимальной структуры модели (2), в которой знаки оценок параметров будут согласованы со знаками соответствующих коэффициентов корреляции. В этой связи в полученной регрессии о значимости регрессоров можно судить по величинам абсолютных вкладов переменных в общую детерминацию R^2:

$$C_{x_j}^{\text{абс}}=r_{yx_j}\cdot\beta_j,\ j=\overline{1,l};\ C_{z_{jk}^{-}}^{\text{абс}}=r_{yz_{jk}^{-}}\cdot\beta_{jk}^{-},\ C_{z_{jk}^{+}}^{\text{абс}}=r_{yz_{jk}^{+}}\cdot\beta_{jk}^{+},\ j=\overline{1,C_l^2},\ k=\overline{1,p}.$$

Для того чтобы каждая объясняющая переменная входила в модель не более одного раза, необходимо ввести в задачу (3)–(13) следующие линейные ограничения:

$$\sum_{j=1}^{l}v_{ij}\delta_j+\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}v_{i,l+k+p(j-1)}\delta_{jk}^{-}+\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}v_{i,l+pC_l^2+k+p(j-1)}\delta_{jk}^{+}\le 1,\ i=\overline{1,l},\qquad(14)$$

где v_{ij} – элементы бинарной матрицы V размера $\left(l+2p\cdot C_l^2\right)\times l$, заданные по правилу

$$v_{ij}=\begin{cases}1, \text{ если j-я переменная входит в i-й регрессор модели (2),}\\ 0, \text{ в противном случае.}\end{cases}$$

В полученной в результате решения задачи ЧБЛП (3) – (14) НЛР может присутствовать мультиколлинеарность, а коэффициенты модели могут оказаться незначимыми. Для решения этой проблемы введем в задачу ограничения на абсолютные вклады переменных $C_{x_j}^{\text{абс}}$, $C_{z_{jk}^{-}}^{\text{абс}}$, $C_{z_{jk}^{+}}^{\text{абс}}$ в общую детерминацию R^2:

$$r_{yx_j}\cdot\beta_j\ge\theta\cdot\delta_j,\ j=\overline{1,l},\qquad(15)$$

$$r_{yz_{jk}^{-}}\cdot\beta_{jk}^{-}\ge\theta\cdot\delta_{jk}^{-},\ j=\overline{1,C_l^2},\ k=\overline{1,p},\qquad(16)$$

$$r_{yz_{jk}^{+}}\cdot\beta_{jk}^{+}\ge\theta\cdot\delta_{jk}^{+},\ j=\overline{1,C_l^2},\ k=\overline{1,p},\qquad(17)$$

где $\theta\ge 0$ – заданное минимальное значение вклада каждого регрессора в общую детерминацию. Очевидно, что с ростом числа θ будут увеличиваться вклады регрессоров в общую детерминацию и параллельно будет уменьшаться их количество, что приведет к снижению мультиколлинеарности.

Построение НЛР рекомендуется проводить методом последовательного повышения вкладов (МППВ) регрессоров по следующему алгоритму. Решить задачу (3) – (14). Для полученной НЛР вычислить абсолютные вклады регрессоров и оценить степень мультиколли-

неарности любым известным методом, например с помощью коэффициентов вздутия дисперсии. Если вклады достаточно высоки, а мультиколлинеарность слабая, то НЛР получена. В противном случае назначается величина θ, чуть большая, чем минимальный из текущих абсолютных вкладов, и решается задача (3) – (17). И так до тех пор, пока не будет получена НЛР со слабой мультиколлинеарностью и необходимыми абсолютными вкладами регрессоров в детерминацию. МППВ гарантирует построение вполне интерпретируемой НЛР.

Для решения сформулированных весьма трудоемких задач была разработана специальная программа «ВИнтер-2». Решателем задач ЧБЛП в этой программе выступает пакет LPSolve. Для того чтобы полученная структура НЛР гарантированно была интерпретируемой, «ВИнтер-2» по умолчанию исключает все регрессоры, не удовлетворяющие содержательному смыслу решаемой задачи (подробно эти условия описаны в [14]). Также в программе предусмотрена возможность контролировать число преобразованных переменных, исключая те из них, у которых коэффициент корреляции с y по абсолютной величине не превосходит некоторого числа $r \in [0,1)$, т. е. для которых не выполняются условия:

$$\left| r_{yz_{jk}^{-}} \right| \geq r\,,\ \left| r_{yz_{jk}^{+}} \right| \geq r\,,\ j = \overline{1, C_l^2}\,, k = \overline{1, p}\,. \tag{18}$$

Тюменская область, имеющая в своем составе Ханты-Мансийский автономный округ – Югру и Ямало-Ненецкий автономный округ, занимает третье место по площади среди субъектов Российской Федерации. Она представляет собой крупнейший нефтегазовый регион России, поэтому актуальной задачей является создание в Тюменской области современной и эффективной транспортной инфраструктуры. При этом с научной точки зрения актуальна задача моделирования железнодорожных грузовых перевозок в Тюменской области. Одно из решений этой задачи можно найти в работах [16, 17], в которых построена линейная регрессия

$$y = 31,6346 + 0,0007z_1 - 0,0001z_2 - 0,0003z_3\,,$$

где y – грузовые перевозки железнодорожного транспорта юга Тюменской области (млн тонн); z_1 – среднемесячная заработная плата работников магистрального ж/д транспорта (руб.); z_2 – экспорт в страны дальнего зарубежья минеральных продуктов (млрд долл. США); z_3 – среднесписочная численность работников (человек). Коэффициент детерминации R^2 этой линейной регрессии составил всего 0,67, поэтому вряд ли можно считать её адекватной.

Для построения вполне интерпретируемой НЛР железнодорожных грузоперевозок Тюменской области были использованы годовые статистические данные за период с 2000 по 2020 г. по следующим переменным:

y – отправление грузов железнодорожным транспортом общего пользования (миллионов тонн);

x_1 – численность рабочей силы (тысяч человек);

x_2 – число предприятий и организаций;

x_3 – производство электроэнергии (миллиард киловатт-часов);

x_4 – удельный вес автомобильных дорог с твердым покрытием в общей протяженности автомобильных дорог общего пользования (%);

x_5 – удельный вес автомобильных дорог с усовершенствованным покрытием в протяженности автомобильных дорог с твердым покрытием общего пользования (%);

x_6 – продукция сельского хозяйства (миллионов рублей);

x_7 – объем работ, выполненных по виду экономической деятельности «Строительство» (миллионов рублей);

x_8 – валовой региональный продукт (ВРП) (миллионов рублей).

Коэффициенты корреляции объясняющих переменных с *y* составляют:

$$r_{yx_1} = 0,840\,,\quad r_{yx_2} = 0,339\,,\quad r_{yx_3} = 0,896\,,\quad r_{yx_4} = -0,549\,,$$

$r_{yx_5} = -0,725$, $r_{yx_6} = 0,946$, $r_{yx_7} = 0,818$, $r_{yx_8} = 0,926$.

Как видно, знаки всех коэффициентов корреляции удовлетворяют содержательному смыслу задачи. Так, рост численности рабочей силы x_1, числа предприятий и организаций x_2, производства электроэнергии x_3, продукции сельского хозяйства x_6, объемов работ по виду «Строительство» x_7 и ВРП x_8 приводит к увеличению грузовых ж/д перевозок Тюменской области. А увеличение удельного веса автодорог с твердым покрытием x_4 и с усовершенствованным покрытием x_5 приводит, по логике, к повышению спроса на перевозки грузов автотранспортом, а следовательно, к снижению спроса на ж/д перевозки. Самая слабая корреляция наблюдается между y и x_2 (0,339), однако было принято решение не исключать переменную x_2 из рассмотрения, поскольку степень её влияния на y в НЛР может вырасти.

Сначала в программе «ВИнтер-2» были заданы следующие параметры построения НЛР: $p = 4$, $r = 0,2$. В результате было сформировано $p \cdot C_l^2 = 112$ пар переменных, преобразованных с помощью бинарной операции min, и столько же пар переменных, преобразованных с помощью max. Итого 224 преобразования. Затем из них были исключены те, для которых не выполнились условия (18). В итоге сформировался набор, содержащий 8 объясняющих переменных и 128 преобразований. По этому набору программа автоматически сформировала задачу ЧБЛП (3) – (14) на языке пакета LPSolve. Решение этой задачи было найдено за 153 секунды. В результате решения была выбрана следующая оптимальная структура НЛР:

$$\tilde{y} = 13,8072 + \underset{(1,146)}{0,000167}\min\{x_2, 0.9545x_7\} + \underset{(0,128)}{0,017}\min\{x_3, 0.0000709x_8\} + \underset{(4,229)}{0,0105}\max\{x_1, 0.04633x_6\} - \underset{(-1,706)}{0,403}\max\{x_4, 0.9475x_5\}. \quad (19)$$

Как видно, все восемь объясняющих переменных вошли в состав полученного уравнения. При этом для пар x_2 и x_7, x_3 и x_8 была идентифицирована бинарная операция min, а для пар x_1 и x_6, x_4 и x_5 – max.

Коэффициент детерминации R^2 НЛР (19) составил 0,951423, что подтверждает адекватность модели.

В скобках под коэффициентами регрессии (19) указаны значения t-критерия Стьюдента, по которым можно сделать вывод, что значимым для уровня значимости $\alpha = 0,1$ оказался только один регрессор – $\max\{x_1, 0.04633x_6\}$. В скобках над коэффициентами модели (19) указаны абсолютные вклады переменных в общую детерминацию, показывающие, что регрессор $\min\{x_3, 0.0000709x_8\}$ вносит слишком низкий вклад (0,0293), не превышающий 0,05.

Коэффициенты вздутия дисперсии регрессоров НЛР (19) составили 8,912, 23,039, 12,596 и 1,435 соответственно. Двое из этих коэффициентов превысили пороговое значение 10, из чего можно сделать вывод о присутствии в полученной модели мультиколлинеарности.

Перечисленные обстоятельства не позволяют отнести НЛР (19) к вполне интерпретируемым. Поэтому было принято решение перестроить модель, дополнив задачу ЧБЛП (3) – (14) ограничениями (15) – (17) на абсолютные вклады переменных. Поскольку минимальный из абсолютных вкладов регрессии (19) равен 0,0293, то величина параметра θ была выбрана равной 0,03. Начальные параметры p и r не менялись. Решение задачи (3) – (17) в LPSolve было найдено за 136 секунд. В результате автоматически определилась следующая структура НЛР:

$$\tilde{y} = 12,7415 + \underset{(3,827)}{0,0001866}\min\{x_2, 0.09879x_8\} + \underset{(13,14)}{0,01083}\max\{x_1, 0.04633x_6\} - \underset{(-1,751)}{0,4005}\max\{x_4, 0.9475x_5\}. \quad (20)$$

Как видно, переменные x_3 и x_7 не вошли в состав регрессии (20). При этом регрессоры $\max\{x_1, 0.04633x_6\}$ и $\max\{x_4, 0.9475x_5\}$ сохранились в уравнении, а регрессоры $\min\{x_2, 0.9545x_7\}$ и $\min\{x_3, 0.0000709x_8\}$ из (19) перегруппировались в модели (20) в регрессор $\min\{x_2, 0.09879x_8\}$.

Коэффициент детерминации R^2 НЛР (20) составил 0,951374, что меньше, чем у модели (19), всего лишь на 0,00005. Иными словами, исключение переменных x_3 и x_7 практически не изменило высокого качества аппроксимации НЛР.

Все коэффициенты регрессии (20) значимы по t-критерию Стьюдента для уровня значимости $\alpha = 0{,}1$, а минимальный абсолютный вклад в R^2 составляет 0,0661 для регрессора $\max\{x_4, 0.9475x_5\}$,что довольно существенно.

Коэффициенты вздутия дисперсии регрессоров НЛР (20) составили 1,035, 1,466 и 1,425 соответственно, откуда следует, что в модели мультиколлинеарности нет.

Таким образом, выполняются все необходимые условия, чтобы считать НЛР (20) вполне интерпретируемой регрессионной моделью.

Представим модель (20) в кусочно-заданной форме:

$$\tilde{y} = \begin{cases} 12{,}7415 + 0{,}00001843x_8 + 0{,}01083x_1 - 0{,}4005x_4, \text{ при } \frac{x_2}{x_8} \ge 0{,}09879,\ \frac{x_1}{x_6} \ge 0{,}04633,\ \frac{x_4}{x_5} \ge 0{,}9475; \\ 12{,}7415 + 0{,}00001843x_8 + 0{,}01083x_1 - 0{,}3795x_5, \text{ при } \frac{x_2}{x_8} \ge 0{,}09879,\ \frac{x_1}{x_6} \ge 0{,}04633,\ \frac{x_4}{x_5} < 0{,}9475; \\ 12{,}7415 + 0{,}00001843x_8 + 0{,}000502x_6 - 0{,}4005x_4, \text{ при } \frac{x_2}{x_8} \ge 0{,}09879,\ \frac{x_1}{x_6} < 0{,}04633,\ \frac{x_4}{x_5} \ge 0{,}9475; \\ 12{,}7415 + 0{,}00001843x_8 + 0{,}000502x_6 - 0{,}3795x_5, \text{ при } \frac{x_2}{x_8} \ge 0{,}09879,\ \frac{x_1}{x_6} < 0{,}04633,\ \frac{x_4}{x_5} < 0{,}9475; \\ 12{,}7415 + 0{,}0001866x_2 + 0{,}01083x_1 - 0{,}4005x_4, \text{ при } \frac{x_2}{x_8} < 0{,}09879,\ \frac{x_1}{x_6} \ge 0{,}04633,\ \frac{x_4}{x_5} \ge 0{,}9475; \\ 12{,}7415 + 0{,}0001866x_2 + 0{,}01083x_1 - 0{,}3795x_5, \text{ при } \frac{x_2}{x_8} < 0{,}09879,\ \frac{x_1}{x_6} \ge 0{,}04633,\ \frac{x_4}{x_5} < 0{,}9475; \\ 12{,}7415 + 0{,}0001866x_2 + 0{,}000502x_6 - 0{,}4005x_4, \text{ при } \frac{x_2}{x_8} < 0{,}09879,\ \frac{x_1}{x_6} < 0{,}04633,\ \frac{x_4}{x_5} \ge 0{,}9475; \\ 12{,}7415 + 0{,}0001866x_2 + 0{,}000502x_6 - 0{,}3795x_5, \text{ при } \frac{x_2}{x_8} < 0{,}09879,\ \frac{x_1}{x_6} < 0{,}04633,\ \frac{x_4}{x_5} < 0{,}9475. \end{cases}$$

Тогда НЛР (20) можно интерпретировать следующим образом.

1. Если показатель x_2/x_8 не меньше, чем 0,09879, то на отправление грузов ж/д транспортом y влияет ВРП x_8, а число предприятий и организаций x_2 не влияет. При этом с увеличением ВРП x_8 на 1 млн руб. (при неизменных значениях остальных переменных) y возрастает в среднем на 18,43 тонны. А если показатель x_2/x_8 меньше, чем 0,09879, то на y влияет число предприятий и организаций x_2, а ВРП x_8 не влияет. При этом с увеличением числа предприятий и организаций x_2 на 1 единицу (при неизменных значениях остальных переменных) y возрастает в среднем на 186,6 тонны.

2. Если показатель x_1/x_6 не меньше, чем 0,04633, то на отправление грузов ж/д транспортом y влияет численность рабочей силы x_1, а продукция сельского хозяйства x_6 не влияет. При этом с увеличением численности рабочей силы x_1 на 1 тыс. человек (при неизменных значениях остальных переменных) y возрастает в среднем на 10830 тонн. А если показатель x_1/x_6 меньше, чем 0,04633, то на y влияет продукция сельского хозяйства x_6, а численность рабочей

силы x_1 не влияет. При этом с увеличением продукции сельского хозяйства x_6 на 1 млн руб. (при неизменных значениях остальных переменных) y возрастает в среднем на 502 тонны.

Если показатель x_4/x_5 не меньше, чем 0,9475, то на отправление грузов ж/д транспортом y влияет удельный вес автодорог с твердым покрытием x_4, а удельный вес автодорог с усовершенствованным покрытием x_5 не влияет. При этом с увеличением удельного веса автодорог с твердым покрытием x_4 на 1 % (при неизменных значениях остальных переменных) y убывает в среднем на 0,4005 млн тонн. А если показатель x_4/x_5 меньше, чем 0,9475, то на y влияет удельный вес автодорог с усовершенствованным покрытием x_5, а удельный вес автодорог с твердым покрытием x_4 не влияет. При этом с увеличением удельного веса автодорог с усовершенствованным покрытием x_5 на 1 % (при неизменных значениях остальных переменных) y убывает в среднем на 0,3795 млн тонн.

Заключение и выводы

Таким образом, для построения НЛР в данной работе сформулирована задача ЧБЛП, позволяющая регулировать абсолютные вклады регрессоров в общую детерминацию и эффект мультиколлинеарности. Показано, что реализация метода последовательного повышения вкладов регрессоров гарантирует построение вполне интерпретируемых НЛР. С помощью предложенного математического аппарата решена задача моделирования ж/д грузовых перевозок Тюменской области. Автоматически полученная НЛР оптимальной структуры является адекватной по всем основным показателям, а также вполне интерпретируемой. Интерпретация модели позволила выявить новые закономерности функционирования ж/д транспорта Тюменской области, недоступные при использовании классических линейных регрессий. Построенная модель также может применяться для прогнозирования.

Литература

1. Molnar, C. Interpretable machine learning / C. Molnar. – Lulu. com, 2020.
2. Du, M. Techniques for interpretable machine learning / M. Du, N. Liu, X. Hu // Communications of the ACM. – 2019. – Vol. 63, №. 1. – P. 68–77.
3. Doshi-Velez, F. Towards a rigorous science of interpretable machine learning / F. Doshi-Velez, B. Kim // arXiv preprint arXiv:1702.08608. – 2017.
4. Definitions, methods, and applications in interpretable machine learning / W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, B. Yu // Proceedings of the National Academy of Sciences. – 2019. – Vol. 116, №. 44. – P. 22071–22080.
5. Montgomery, D. C. Introduction to linear regression analysis / D. C. Montgomery, E. A. Peck, G. G. Vining. – John Wiley & Sons, 2021.
6. Fox, J. Applied regression analysis and generalized linear models / J. Fox. – Sage Publications, 2015.
7. Носков, С. И. Программный комплекс построения некоторых типов кусочно-линейных регрессий / С. И. Носков, А. А. Хоняков. – Текст : непосредственный // Информационные технологии и математическое моделирование в управлении сложными системами. – 2019. – № 3 (4). – С. 47–55.
8. Носков, С. И. Кусочно-линейная производственная функция погрузки на железнодорожном транспорте / С. И. Носков. – Текст : непосредственный // Научные труды КубГТУ. – 2022. – № 4. – С. 72–79.
9. Носков, С. И. Построение кусочно-линейной регрессии с интервальной неопределенностью в данных для зависимой переменной / С. И. Носков. – Текст : непосредственный // Вестник кибернетики. – 2022. – № 2 (46). – С. 61–65.
10. Носков, С. И. Построение кусочно-линейной авторегрессионной модели произвольного порядка / С. И. Носков. – Текст : непосредственный // Вестник Югорского государственного университета. – 2022. – № 2 (65). – С. 89–94.

11. Базилевский, М. П. МНК-оценивание параметров специфицированных на основе функций Леонтьева двухфакторных моделей регрессии / М. П. Базилевский. – Текст : непосредственный // Южно-Сибирский научный вестник. – 2019. – № 2 (26). – С. 66–70.

12. Базилевский, М. П. Оценивание линейно-неэлементарных регрессионных моделей с помощью метода наименьших квадратов / М. П. Базилевский. – Текст : непосредственный // Моделирование, оптимизация и информационные технологии. – 2020. – Т. 8, № 4 (31). – С. 26–27.

13. Базилевский, М. П. Отбор информативных операций при построении линейно-неэлементарных регрессионных моделей / М. П. Базилевский. – Текст : непосредственный // International Journal of Open Information Technologies. – 2021. – Т. 9, № 5. – С. 30–35.

14. Базилевский, М. П. Метод построения неэлементарных линейных регрессий на основе аппарата математического программирования / М. П. Базилевский. – Текст : непосредственный // Проблемы управления. – 2022. – № 4. – С. 3–14.

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

16. Филимонова, Л. А. Особенности применения стохастических моделей в оценке и прогнозах уровня конкурентоспособности транспортно-логистической системы обслуживания грузопотоков на примере юга Тюменской области / Л. А. Филимонова, Д. А. Буткова, А В. Носырева. – Текст : непосредственный // Московский экономический журнал. – 2019. – № 1. – С. 354–364.

17. Мильчакова, Н. Н. Формирование системы оценки и прогноза эффективности транспортно-логистической системы обслуживания грузопотоков Тюменской области / Н. Н. Мильчакова, А. В. Носырева. – Текст : непосредственный // Вестник Сургутского государственного университета. – 2018. – № 2 (20). – С. 71–77.