УДК 519.862.6 DOI: http://doi.org/10.25728/pu.2022.4.1

МЕТОД ПОСТРОЕНИЯ НЕЭЛЕМЕНТАРНЫХ ЛИНЕЙНЫХ РЕГРЕССИЙ НА ОСНОВЕ АППАРАТА МАТЕМАТИЧЕСКОГО ПРОГРАММИРОВАНИЯ

М.П. Базилевский

Аннотация. Рассматривается проблема построения неэлементарных линейных регрессий, состоящих из объясняющих переменных и всевозможных комбинаций их пар, преобразованных с помощью бинарных операций минимум и максимум. Задача построения таких моделей формализована в виде задачи частично-булевого линейного программирования. Регулируя в ней ограничения на бинарные переменные, можно контролировать структурную спецификацию неэлементарной линейной регрессии, а именно количество входящих в нее регрессоров, их типы и состав объясняющих переменных. При этом оценки параметров модели находятся приближенно с помощью метода наименьших квадратов. К достоинствам сформулированной задачи относится то, что число ее ограничений не зависит от объема выборки, а знаки оценок при объясняющих переменных согласуются со знаками коэффициентов их корреляции с зависимой переменной. Показано, как на начальном этапе отсекать регрессоры, чтобы сократить время решения задачи и сделать модель вполне интерпретируемой. Построена неэлементарная линейная регрессия для моделирования железнодорожных грузоперевозок в Иркутской области и дана ее интерпретация.

Ключевые слова: неэлементарная линейная регрессия, метод наименьших квадратов, задача частично-булевого линейного программирования, отбор информативных регрессоров, коэффициент детерминации, интерпретация, железнодорожные грузоперевозки.

ВВЕДЕНИЕ

При проведении регрессионного анализа [1, 2] на основе экономических данных особое внимание уделяется построению производственных функций (ПФ), представляющих собой математические зависимости между объемами выпуска продукции и факторами производства. Теории, методам и применению ПФ целиком посвящена монография [3], выпущенная еще в 1986 г. В ней рассмотрены следующие ПФ: линейная, многорежимная, Кобба – Дугласа, Леонтьева, Аллена, CES (*Constant Elasticity of Substitution* – с постоянной эластичностью замены факторов), LES (*Linear Elasticity of Substitution* – с линейной эластичностью замены факторов), Солоу. В настоящее время появляются и исследуются новые модификации ПФ, которые активно применяются в эконометрических исследованиях [4–6]. В данной статье речь пойдет о построении неэлементарных регрессионных моделей, специфицированных на основе известной ПФ Леонтьева:

$$y_i = \min\{\alpha_1 x_{i1}, \alpha_2 x_{i2}, ..., \alpha_l x_{il}\} + \varepsilon_i, \quad i = \overline{1, n}, \quad (1)$$

где n – объем выборки; l – количество объясняющих переменных; y_i, $i = \overline{1, n}$, – значения объясняемой переменной y; x_{ij}, $i = \overline{1, n}$, $j = \overline{1, l}$, – значения объясняющих переменных x_1, x_2, ..., x_l; α_j, $j = \overline{1, l}$, – неизвестные параметры; ε_i, $i = \overline{1, n}$, – ошибки аппроксимации. С позиции теории ПФ переменная y в уравнении (1) трактуется как объем выпуска продукции, а x_1, ..., x_n – как показатели факторов производства.

Отметим, что в монографии [3] выделена еще и «параллельная» функция Леонтьева

$$y_i = \min\{\alpha_{11} x_{i1}, \alpha_{12} x_{i2}, ..., \alpha_{1l} x_{il}\} + \ldots$$

$$+ \min\{\alpha_{k1} x_{i1}, \alpha_{k2} x_{i2}, ..., \alpha_{kl} x_{il}\} + \varepsilon_i, \quad i = \overline{1, n},$$

отражающая процесс, в котором объем выпуска складывается из выпусков k параллельных производственных процессов с фиксированными про-

порциями факторов, использующих общие ресурсы. Для двух факторов производства x_1 и x_2 «параллельная» функция Леонтьева называется функцией линейного программирования.

В монографии [7] отмечается, что для нахождения оценок параметров ПФ Леонтьева (1) можно применять методы негладкой оптимизации [8–10], которые, как правило, являются труднореализуемыми. Поэтому в работе [7] задача точного оценивания ПФ (1) с помощью метода наименьших модулей (МНМ) сведена к задаче частично-булевого линейного программирования (ЧБЛП). Вместе с тем в монографии [7] предложен способ приближенного оценивания ПФ Леонтьева, основанный на переборе оценок из предварительно сформированной области определения.

В статье [11] предложена функция, противоположная по смыслу ПФ (1):

$$y_i = \max\{\alpha_1 x_{i1}, \alpha_2 x_{i2},..., \alpha_l x_{il}\} + \varepsilon_i, \quad i = \overline{1, n}, \quad (2)$$

а в статье [12] рассмотрен симбиоз функций (1) и (2):

$$\begin{aligned} y_i = &\min\{\alpha_1 x_{i1}, \alpha_2 x_{i2},..., \alpha_l x_{il}\} + \\ &+\max\{\beta_1 x_{i1}, \beta_2 x_{i2},..., \beta_l x_{il}\} + \varepsilon_i, \ i = \overline{1, n}. \end{aligned} \quad (3)$$

Задачи точного оценивания параметров регрессий (2) и (3) с помощью МНМ сведены в статьях [11] и [12] к соответствующим задачам ЧБЛП. Повышенное внимание к построению регрессионных моделей с использованием аппарата математического программирования в современной научной литературе (см., например, работы [13–15]) объясняется тем, что за последние годы была существенно развита технология решения задач ЧБЛП.

Данная статья посвящена оцениванию специфицированных на основе ПФ Леонтьева регрессионных моделей с помощью метода наименьших квадратов (МНК) [1, 2]. Впервые такая задача была сформулирована автором в работе [16] для регрессии (1) с двумя объясняющими переменными. А в статье [17] была предложена неэлементарная линейная регрессия (НЛР) вида

$$\begin{aligned} y_i = &\alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \\ &+\sum_{j=1}^{C_l^2} \alpha_{j+l} \min\{x_{i,\mu_{j1}}, \lambda_j x_{i,\mu_{j2}}\} + \varepsilon_i, \ i = \overline{1, n}, \end{aligned} \quad (4)$$

где μ_{j1} и μ_{j2}, $j = \overline{1, C_l^2}$ – элементы первого и второго столбца индексной матрицы $\mathbf{M}_{C_l^2 \times 2}$, содержащей по строкам всевозможные комбинации пар индексов переменных; α_j, $j = \overline{0, l + C_l^2}$, λ_j, $j = \overline{1, C_l^2}$, – неизвестные параметры. Считается, что значения всех переменных в уравнении (4) строго положительны.

Как видно, НЛР относится к классу нелинейных по параметрам моделей. Но если придать всем параметрам λ_j, $j = \overline{1, C_l^2}$, определенные значения, то регрессия становится линейной и нахождение оценок ее параметров α_j, $j = \overline{0, l + C_l^2}$, с помощью МНК не вызывает трудностей. В статье [17] установлено, что оптимальные с точки зрения МНК оценки параметров λ_j, $j = \overline{1, C_l^2}$, для НЛР принадлежат промежуткам

$$\lambda_j \in \left(\lambda_{\min}^{(j)}, \lambda_{\max}^{(j)}\right), \quad j = \overline{1, l}, \quad (5)$$

где $\lambda_{\min}^{(j)} = \min\left\{\frac{x_{1,\mu_{j1}}}{x_{1,\mu_{j2}}}, \frac{x_{2,\mu_{j1}}}{x_{2,\mu_{j2}}},..., \frac{x_{n,\mu_{j1}}}{x_{n,\mu_{j2}}}\right\}$, $\lambda_{\max}^{(j)} = \max\left\{\frac{x_{1,\mu_{j1}}}{x_{1,\mu_{j2}}}, \frac{x_{2,\mu_{j1}}}{x_{2,\mu_{j2}}},..., \frac{x_{n,\mu_{j1}}}{x_{n,\mu_{j2}}}\right\}$. Точки $\lambda_j = \lambda_{\min}^{(j)}$ и $\lambda_j = \lambda_{\max}^{(j)}$ нельзя использовать из-за возникновения совершенной мультиколлинеарности переменных.

Благодаря этим свойствам, в работе [17] был предложен способ приближенного МНК-оценивания НЛР (4), основанный на переборе значений параметров λ_j, $j = \overline{1, C_l^2}$, из промежутков (5).

К сожалению, при построении НЛР (4) с ростом числа объясняющих переменных l существенно возрастает общее количество ее регрессоров. Поэтому появляется необходимость в решении задачи отбора некоторого числа наиболее «информативных» регрессоров (ОИР) [7]. Специально для этого в работе [18] были разработаны две стратегии. Каждая из них предполагает формирование по указанному алгоритму множества альтернативных вариантов регрессий, для каждой из которых реализуется описанный в статье [17] способ приближенного МНК-оценивания, а затем выбирается модель с наименьшей величиной суммы квадратов остатков. Главным недостатком предложенного в работе [18] подхода к построению НЛР является то, что он основан на методе полного перебора всех возможных альтернатив, поэтому на решение задачи ОИР может уходить слишком много времени. Более перспективным выглядит следующий подход с использованием аппарата ЧБЛП.

В статье [19] задача ОИР при оценивании линейной регрессии с помощью МНК была сведена к задаче ЧБЛП. При этом нерешенной оставалась проблема выбора большого положительного числа M, влияющего как на скорость, так и на само решение задачи, пока не появилась работа [20]. В ней сформулирована задача ЧБЛП, которая приводит к построению линейной регрессии с заданным числом объясняющих переменных, в которой знаки МНК-оценок согласованы со знаками коэффициентов корреляции между переменными y и x_j, $j=\overline{1,l}$. В ходе вычислительных экспериментов был подтвержден сделанный в статье [21] вывод, что такая задача с ограничениями на знаки коэффициентов решается на порядок быстрее, чем без них. Главной целью данной работы является сведение задачи построения НЛР к эффективно решаемой задаче ЧБЛП, рассмотренной в статье [20].

1. МЕТОД ПОСТРОЕНИЯ НЕЭЛЕМЕНТАРНЫХ ЛИНЕЙНЫХ РЕГРЕССИЙ

Уравнение НЛР (4) содержит только одну бинарную операцию – минимум. Здесь и далее под бинарной операцией минимум (максимум) понимается математическая операция, принимающая два аргумента и возвращающая их минимум (максимум). Дополним эту регрессионную модель регрессорами с бинарной операцией максимум:

$$y_i=\alpha_0+\sum_{j=1}^{l}\alpha_j x_{ij}+\sum_{j=1}^{C_l^2}\alpha_{j+l}\min\{x_{i,\mu_{j1}},\lambda_j x_{i,\mu_{j2}}\}+ \\ +\sum_{j=1}^{C_l^2}\alpha_{j+l+C_l^2}\max\{x_{i,\mu_{j1}},\lambda_j x_{i,\mu_{j2}}\}+\varepsilon_i,\ i=\overline{1,n}, \tag{6}$$

Общее число регрессоров в уравнении (6) стало гораздо больше, чем в уравнении (4), и составляет $l+2C_l^2$.

Уравнение НЛР вида (6) вводится впервые, поэтому ставится задача формализовать процесс построения этой модели в виде задачи ЧБЛП. Это можно сделать следующим образом.

Для каждого параметра λ_j, $j=\overline{1,C_l^2}$, из уравнения (6) определим промежутки значений по формулам (5). Затем равномерно разобьем каждый из этих промежутков p точками и сформируем матрицу $\Lambda=\left(\lambda_{jk}^*\right)$, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$, элемент λ_{jk}^* которой показывает k-е значение параметра λ_j для j-й пары переменных. Заменив в уравнении (6) неизвестные параметры λ_j на известные элементы матрицы Λ, получим:

$$y_i=\alpha_0+\sum_{j=1}^{l}\alpha_j x_{ij}+\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\alpha_{jk}^{-}\min\left\{x_{i,\mu_{j1}},\lambda_{jk}^* x_{i,\mu_{j2}}\right\}+ \\ +\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\alpha_{jk}^{+}\max\left\{x_{i,\mu_{j1}},\lambda_{jk}^* x_{i,\mu_{j2}}\right\}+\varepsilon_i,\ i=\overline{1,n}, \tag{7}$$

где α_{jk}^{-}, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$, – неизвестные параметры для регрессоров с бинарной операцией минимум, а α_{jk}^{+}, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$, – неизвестные параметры для регрессоров с бинарной операцией максимум. В модели (7) общее число регрессоров еще больше, чем в модели (6), и равно $l+2pC_l^2$. Например, если общее число переменных $l=100$, а число разбиений $p=10$, то регрессия (7) будет иметь 99 100 регрессоров.

Сделаем в уравнении (7) замену $z_{ijk}^{-}=\min\left\{x_{i,\mu_{j1}},\lambda_{jk}^* x_{i,\mu_{j2}}\right\}$, $z_{ijk}^{+}=\max\left\{x_{i,\mu_{j1}},\lambda_{jk}^* x_{i,\mu_{j2}}\right\}$, $i=\overline{1,n}$, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$. Получим модель множественной линейной регрессии:

$$y_i=\alpha_0+\sum_{j=1}^{l}\alpha_j x_{ij}+\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\alpha_{jk}^{-}z_{ijk}^{-}+ \\ \sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\alpha_{jk}^{+}z_{ijk}^{+}+\varepsilon_i,\ i=\overline{1,n}, \tag{8}$$

Далее, как это сделано в статье [19], сведем задачу ОИР для линейной регрессии (8), оцениваемой с помощью МНК, к задаче ЧБЛП. Для этого предварительно проведем нормирование всех переменных из уравнения (8) по известному правилу, вычитая из каждого значения переменной ее среднее арифметическое и деля результат на стандартное отклонение.

Составим для модели (8) уравнение стандартизованной регрессии

$$w_i=\sum_{j=1}^{l}\beta_j q_{ij}+\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\beta_{jk}^{-}h_{ijk}^{-}+ \\ +\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\beta_{jk}^{+}h_{ijk}^{+}+\xi_i,\ i=\overline{1,n}, \tag{9}$$

где w – нормированная переменная y; q_j, $j=\overline{1,l}$, – нормированные переменные x_j, $j=\overline{1,l}$; h_{jk}^{-}, h_{jk}^{+}, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$, – нормированные переменные z_{jk}^{-}, z_{jk}^{+}, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$; β_j, $j=\overline{1,l}$, и β_{jk}^{-}, β_{jk}^{+}, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$, – неизвестные стандартизованные коэффициенты; ξ_i, $i=\overline{1,n}$, – новые ошибки аппроксимации.

Для модели (9) МНК-оценки находятся по формуле

$$\tilde{\beta} = R_{XX}^{-1} \cdot R_{YX}, \qquad (10)$$

где $R_{XX} = \begin{pmatrix} R_{xx} & R_{xz^-} & R_{xz^+} \\ R_{z^-x} & R_{z^-z^-} & R_{z^-z^+} \\ R_{z^+x} & R_{z^+z^-} & R_{z^+z^+} \end{pmatrix}$ – корреляционная блочная матрица размера $(l+2pC_l^2)\times(l+2pC_l^2)$, составленная из блоков

$$R_{xx} = \left(r_{x_j x_k}\right),\ j=\overline{1,l},\ k=\overline{1,l};$$

$$R_{xz^-} = \left(r_{x_s z^-_{jk}}\right),\ s=\overline{1,l},\ j=\overline{1,C_l^2},\ k=\overline{1,p};$$

$$R_{xz^+} = \left(r_{x_s z^+_{jk}}\right),\ s=\overline{1,l},\ j=\overline{1,C_l^2},\ k=\overline{1,p};$$

$$R_{z^-x} = \left(r_{z^-_{jk} x_s}\right),\ j=\overline{1,C_l^2},\ k=\overline{1,p},\ s=\overline{1,l};$$

$$R_{z^-z^-} = \left(r_{z^-_{s_1s_2} z^-_{kj}}\right),\ s_1=C_l^2,\ s_2=\overline{1,p},\ j=\overline{1,C_l^2},\ k=\overline{1,p};$$

$$R_{z^-z^+} = \left(r_{z^-_{s_1s_2} z^+_{kj}}\right),\ s_1=C_l^2,\ s_2=\overline{1,p},\ j=\overline{1,C_l^2},\ k=\overline{1,p};$$

$$R_{z^+x} = \left(r_{z^+_{jk} x_s}\right),\ j=\overline{1,C_l^2},\ k=\overline{1,p},\ s=\overline{1,l};$$

$$R_{z^+z^-} = \left(r_{z^+_{s_1s_2} z^-_{kj}}\right),\ s_1=C_l^2,\ s_2=\overline{1,p},\ j=\overline{1,C_l^2},\ k=\overline{1,p};$$

$$R_{z^+z^+} = \left(r_{z^+_{s_1s_2} z^+_{kj}}\right),\ s_1=C_l^2,\ s_2=\overline{1,p},\ j=\overline{1,C_l^2},\ k=\overline{1,p};$$

$R_{YX} = \left(R_{yx} \quad R_{yz^-} \quad R_{yz^+}\right)^{\mathrm{T}}$ – корреляционный блочный вектор размера $\left(l+2pC_l^2\right)\times 1$, составленный из блоков $R_{yx} = \left(r_{yx_j}\right)$, $j=\overline{1,l}$; $R_{yz^-} = \left(r_{yz^-_{jk}}\right)$, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$; $R_{yz^+} = \left(r_{yz^+_{jk}}\right)$, $j=\overline{1,C_l^2}$, $k=\overline{1,p}$.

Коэффициент детерминации модели (9) находится по формуле

$$R^2 = \sum_{j=1}^{l} r_{yx_j}\beta_j + \sum_{j=1}^{C_l^2}\sum_{k=1}^{p} r_{yz^-_{jk}}\beta^-_{jk} + \sum_{j=1}^{C_l^2}\sum_{k=1}^{p} r_{yz^+_{jk}}\beta^+_{jk}. \qquad (11)$$

Тогда с использованием формул (10) и (11) сформулируем задачу ОИР для линейной регрессии (8):

$$R^2 \to \max, \qquad (12)$$

$$-\left(1-\delta_j\right)M \le \sum_{k=1}^{l} r_{x_j x_k}\beta_k + \sum_{s=1}^{C_l^2}\sum_{k=1}^{p} r_{x_j z^-_{sk}}\beta^-_{sk} + \sum_{s=1}^{C_l^2}\sum_{k=1}^{p} r_{x_j z^+_{sk}}\beta^+_{sk} - r_{yx_j} \le (1-\delta_j)M,\ j=\overline{1,l}, \qquad (13)$$

$$-\left(1-\delta^-_{jk}\right)M \le \sum_{s=1}^{l} r_{x_s z^-_{jk}}\beta_s + \sum_{s_1=1}^{C_l^2}\sum_{s_2=1}^{p} r_{z^-_{s_1s_2} z^-_{jk}}\beta^-_{s_1s_2} + \sum_{s_1=1}^{C_l^2}\sum_{s_2=1}^{p} r_{z^+_{s_1s_2} z^-_{jk}}\beta^+_{s_1s_2} - r_{yz^-_{jk}} \le (1-\delta^-_{jk})M, \qquad (14)$$

$$j=\overline{1,C_l^2},\ k=\overline{1,p},$$

$$-\left(1-\delta^+_{jk}\right)M \le \sum_{s=1}^{l} r_{x_s z^+_{jk}}\beta_s + \sum_{s_1=1}^{C_l^2}\sum_{s_2=1}^{p} r_{z^-_{s_1s_2} z^+_{jk}}\beta^-_{s_1s_2} + \sum_{s_1=1}^{C_l^2}\sum_{s_2=1}^{p} r_{z^+_{s_1s_2} z^+_{jk}}\beta^+_{s_1s_2} - r_{yz^+_{jk}} \le (1-\delta^+_{jk})M, \qquad (15)$$

$$j=\overline{1,C_l^2},\ k=\overline{1,p},$$

$$-\delta_j M \le \beta_j \le \delta_j M,\ j=\overline{1,l}, \qquad (16)$$

$$-\delta^-_{jk} M \le \beta^-_{jk} \le \delta^-_{jk} M,\ j=\overline{1,C_l^2},\ k=\overline{1,p}, \qquad (17)$$

$$-\delta^+_{jk} M \le \beta^+_{jk} \le \delta^+_{jk} M,\ j=\overline{1,C_l^2},\ k=\overline{1,p}, \qquad (18)$$

$$\delta_j \in \{0,1\},\ j=\overline{1,l}, \qquad (19)$$

$$\delta^-_{jk} \in \{0,1\},\ j=\overline{1,C_l^2},\ k=\overline{1,p}, \qquad (20)$$

$$\delta^+_{jk} \in \{0,1\},\ j=\overline{1,C_l^2},\ k=\overline{1,p}, \qquad (21)$$

$$\sum_{j=1}^{l}\delta_j + \sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\delta^-_{jk} + \sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\delta^+_{jk} = m, \qquad (22)$$

где m – заданное число регрессоров; δ_j, $j=\overline{1,l}$, – булевы переменные, заданные по правилу

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

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

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

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

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

M – большое положительное число.

Достоинством задачи ЧБЛП (12)–(22) является то, что число ее ограничений не зависит от объема выборки n.

В задаче ЧБЛП (12)–(22) стратегия построения НЛР регулируется ограничениями на бинарные переменные. Рассмотрим следующие стратегии.

Стратегия 1. Отбор m регрессоров в линейной регрессии (7).

Для этого просто нужно решить задачу (12)–(22). В этом случае итоговая модель может содержать несколько регрессоров с одинаковой бинарной операцией и с одинаковой парой переменных, но с разными значениями параметра λ_j.

Стратегия 2. Приближенное оценивание НЛР (6) с помощью МНК (без отбора регрессоров).

Для этого нужно решить задачу с целевой функцией (12), ограничениями (13)–(21) и

$$\sum_{k=1}^{p}\delta_{jk}^{-} = 1,\ \sum_{k=1}^{p}\delta_{jk}^{+} = 1,\ j = \overline{1, C_l^2},$$

которые отвечают за вхождение в модель каждой бинарной операции только с одним значением параметра λ_j для каждой пары переменных.

Стратегия 3. Отбор m регрессоров в НЛР (6).

Для этого нужно решить задачу с целевой функцией (12), ограничениями (13)–(22) и

$$\sum_{k=1}^{p}\delta_{jk}^{-} \le 1,\ \sum_{k=1}^{p}\delta_{jk}^{+} \le 1,\ j = \overline{1, C_l^2}. \qquad (23)$$

Заметим, что регулируя ограничения на бинарные переменные, можно контролировать тип входящих в НЛР (6) регрессоров. Так, например, если добавить в задачу (12)–(22) ограничения $\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\delta_{jk}^{-} = 0$, $\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\delta_{jk}^{+} = 0$, то получим задачу ОИР для линейной регрессии. Если добавить ограничения $\sum_{j=1}^{l}\delta_j = 0$, $\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\delta_{jk}^{+} = 0$, то получим задачу ОИР для регрессии только с бинарными операциями минимум, а если ограничения $\sum_{j=1}^{l}\delta_j = 0$, $\sum_{j=1}^{C_l^2}\sum_{k=1}^{p}\delta_{jk}^{-} = 0$ – то задачу ОИР для регрессии только с бинарными операциями максимум.

Помимо этого можно контролировать состав входящих в модель переменных. Для этого введем бинарную матрицу $V = \{v_{ij}\}$, $i = \overline{1, l + 2pC_l^2}$, $j = \overline{1, l}$, в которой

$$v_{ij} = \begin{cases} 1, \text{ если } j\text{-я переменная входит} \\ \text{в } i\text{-й регрессор модели (7),} \\ 0, \text{ в противном случае.} \end{cases}$$

Тогда интеграция в задачу (12)–(22) линейных ограничений

$$\sum_{j=1}^{l} v_{ij}\delta_j + \sum_{j=1}^{C_l^2}\sum_{k=1}^{p} v_{i,l+k+p(j-1)}\delta_{jk}^{-} + \sum_{j=1}^{C_l^2}\sum_{k=1}^{p} v_{i,l+pC_l^2+k+p(j-1)}\delta_{jk}^{+} \le 1,\ i = \overline{1, l}, \qquad (24)$$

позволяет строить НЛР с m регрессорами, в которую каждая объясняющая переменная входит не более одного раза. В этом случае автоматически выполняются условия (23).

К сожалению, для задачи (12)–(22) не до конца ясно, как задавать большие числа M. Для решения этой проблемы поступим так, как предложено в работе [20]. Заменим ограничения (13)–(18) на следующие:

$$-\left(1-\delta_j\right)M_{u_j}^{-} \le \sum_{k=1}^{l} r_{x_j x_k}\,\beta_k + \sum_{s=1}^{C_l^2}\sum_{k=1}^{p} r_{x_j z_{sk}^{-}}\,\beta_{sk}^{-} + \sum_{s=1}^{C_l^2}\sum_{k=1}^{p} r_{x_j z_{sk}^{+}}\,\beta_{sk}^{+} - r_{yx_j} \le (1-\delta_j)M_{u_j}^{+},\ j = \overline{1, l}, \qquad (25)$$

$$-\left(1-\delta_{jk}^{-}\right)M_{u_{jk}^{-}}^{-} \le \sum_{s=1}^{l} r_{x_s z_{jk}^{-}}\,\beta_s + \sum_{s_1=1}^{C_l^2}\sum_{s_2=1}^{p} r_{z_{s_1s_2}^{-} z_{jk}^{-}}\,\beta_{s_1s_2}^{-} + \sum_{s_1=1}^{C_l^2}\sum_{s_2=1}^{p} r_{z_{s_1s_2}^{+} z_{jk}^{-}}\,\beta_{s_1s_2}^{+} - r_{yz_{jk}^{-}} \le (1-\delta_{jk}^{-})M_{u_{jk}^{-}}^{+}, \qquad (26)$$

$$j = \overline{1, C_l^2},\ k = \overline{1, p},$$

$$-\left(1-\delta_{jk}^{+}\right)M_{u_{jk}^{+}}^{-} \le \sum_{s=1}^{l} r_{x_s z_{jk}^{+}}\,\beta_s + \sum_{s_1=1}^{C_l^2}\sum_{s_2=1}^{p} r_{z_{s_1s_2}^{-} z_{jk}^{+}}\,\beta_{s_1s_2}^{-} + \sum_{s_1=1}^{C_l^2}\sum_{s_2=1}^{p} r_{z_{s_1s_2}^{+} z_{jk}^{+}}\,\beta_{s_1s_2}^{+} - r_{yz_{jk}^{+}} \le (1-\delta_{jk}^{+})M_{u_{jk}^{+}}^{+}, \qquad (27)$$

$$j = \overline{1, C_l^2},\ k = \overline{1, p},$$

$$0 \le \beta_j \le \delta_j M_{\beta_j},\ j \in J_{\beta}^{+}, \qquad (28)$$

$$\delta_j M_{\beta_j} \le \beta_j \le 0,\ j \in J_{\beta}^{-}, \qquad 29)$$

$$0 \le \beta_{jk}^{-} \le \delta_{jk}^{-} M_{\beta_{jk}^{-}},\ j, k \in J_{\beta^{-}}^{+}, \qquad (30)$$

$$\delta_{jk}^{-} M_{\beta_{jk}^{-}} \le \beta_{jk}^{-} \le 0,\ j, k \in J_{\beta^{-}}^{-}, \tag{31}$$

$$0 \le \beta_{jk}^{+} \le \delta_{jk}^{+} M_{\beta_{jk}^{+}},\ j, k \in J_{\beta^{+}}^{+}, \tag{32}$$

$$\delta_{jk}^{+} M_{\beta_{jk}^{+}} \le \beta_{jk}^{+} \le 0,\ j, k \in J_{\beta^{+}}^{-}, \tag{33}$$

где J_{β}^{+} и J_{β}^{-} – индексные множества, построенные из множества $\{1, 2,..., l\}$, элементы которых удовлетворяют условиям $r_{yx_j} > 0$ и $r_{yx_j} < 0$ соответственно; $J_{\beta^{-}}^{+}$ и $J_{\beta^{-}}^{-}$ – индексные множества, построенные из множества $\{\{1, 2\},..., \{1, p\}, \{2, 1\},..., \{2, p\},..., \{C_l^2, 1\},..., \{C_l^2, p\}\}$, элементы которых удовлетворяют условиям $r_{yz_{jk}^{-}} > 0$ и $r_{yz_{jk}^{-}} < 0$; $J_{\beta^{+}}^{+}$ и $J_{\beta^{+}}^{-}$ – индексные множества, построенные из множества $\{\{1, 2\},..., \{1, p\}, \{2, 1\},..., \{2, p\},..., \{C_l^2, 1\},..., \{C_l^2, p\}\}$, элементы которых удовлетворяют условиям $r_{yz_{jk}^{+}} > 0$ и $r_{yz_{jk}^{+}} < 0$; $M_{\beta_j} = 1 / r_{yx_j}$, $j = \overline{1, l}$; $M_{\beta_{jk}^{-}} = 1 / r_{yz_{jk}^{-}}$, $M_{\beta_{jk}^{+}} = 1 / r_{yz_{jk}^{+}}$, $j = \overline{1, C_l^2}$, $k = \overline{1, p}$.

Для нахождения чисел $M_{u_j}^{-}$ в ограничениях (25) нужно решить серию из l задач линейного программирования с целевыми функциями $u_j \to \min$ при ограничениях

$$0 \le \beta_j \le M_{\beta_j},\ j \in J_{\beta}^{+}, \tag{34}$$

$$M_{\beta_j} \le \beta_j \le 0,\ j \in J_{\beta}^{-}, \tag{35}$$

$$0 \le \beta_{jk}^{-} \le M_{\beta_{jk}^{-}},\ j, k \in J_{\beta^{-}}^{+}, \tag{36}$$

$$M_{\beta_{jk}^{-}} \le \beta_{jk}^{-} \le 0,\ j, k \in J_{\beta^{-}}^{-}, \tag{37}$$

$$0 \le \beta_{jk}^{+} \le M_{\beta_{jk}^{+}},\ j, k \in J_{\beta^{+}}^{+}, \tag{38}$$

$$M_{\beta_{jk}^{+}} \le \beta_{jk}^{+} \le 0,\ j, k \in J_{\beta^{+}}^{-}, \tag{39}$$

$$\sum_{k=1}^{l} r_{x_j x_k} \beta_k + \sum_{s=1}^{C_l^2} \sum_{k=1}^{p} r_{x_j z_{sk}^{-}} \beta_{sk}^{-} + \sum_{s=1}^{C_l^2} \sum_{k=1}^{p} r_{x_j z_{sk}^{+}} \beta_{sk}^{+} - r_{yx_j} = u_j,\ j = \overline{1, l}, \tag{40}$$

$$\sum_{s=1}^{l} r_{x_s z_{jk}^{-}} \beta_s + \sum_{s_1=1}^{C_l^2} \sum_{s_2=1}^{p} r_{z_{s_1 s_2}^{-} z_{jk}^{-}} \beta_{s_1 s_2}^{-} + \sum_{s_1=1}^{C_l^2} \sum_{s_2=1}^{p} r_{z_{s_1 s_2}^{+} z_{jk}^{-}} \beta_{s_1 s_2}^{+} - r_{yz_{jk}^{-}} = u_{jk}^{-},\ j = \overline{1, C_l^2},\ k = \overline{1, p}, \tag{41}$$

$$\sum_{s=1}^{l} r_{x_s z_{jk}^{+}} \beta_s + \sum_{s_1=1}^{C_l^2} \sum_{s_2=1}^{p} r_{z_{s_1 s_2}^{-} z_{jk}^{+}} \beta_{s_1 s_2}^{-} + \sum_{s_1=1}^{C_l^2} \sum_{s_2=1}^{p} r_{z_{s_1 s_2}^{+} z_{jk}^{+}} \beta_{s_1 s_2}^{+} - r_{yz_{jk}^{+}} = u_{jk}^{+},\ j = \overline{1, C_l^2},\ k = \overline{1, p}, \tag{42}$$

$$\sum_{j=1}^{l} r_{yx_j} \beta_j + \sum_{j=1}^{C_l^2} \sum_{k=1}^{p} r_{yz_{jk}^{-}} \beta_{jk}^{-} + \sum_{j=1}^{C_l^2} \sum_{k=1}^{p} r_{yz_{jk}^{+}} \beta_{jk}^{+} \le 1. \tag{43}$$

Для нахождения чисел $M_{u_j}^{+}$ нужно решить серию из l задач линейного программирования с целевыми функциями $u_j \to \max$ при ограничениях (34)–(43). Аналогично находятся числа $M_{u_{jk}^{-}}^{-}$, $M_{u_{jk}^{-}}^{+}$, $M_{u_{jk}^{+}}^{-}$, $M_{u_{jk}^{+}}^{+}$ решением серии из pC_l^2 задач линейного программирования с целевыми функциями $u_{jk}^{-} \to \min$, $u_{jk}^{-} \to \max$, $u_{jk}^{+} \to \min$, $u_{jk}^{+} \to \max$ соответственно при ограничениях (34)–(43).

Таким образом, решение задачи ЧБЛП с целевой функцией (12) и ограничениями (19)–(22), (25)–(33) приводит к построению линейной регрессии (7) с m регрессорами, в которой знаки оценок β-параметров согласованы со знаками соответствующих коэффициентов корреляции регрессоров с переменной y, т. е. справедливы неравенства $\beta_j r_{yx_j} > 0$, $j = \overline{1, l}$; $\beta_{jk}^{-} r_{yz_{jk}^{-}} > 0$, $\beta_{jk}^{+} r_{yz_{jk}^{+}} > 0$, $j = \overline{1, C_l^2}$, $k = \overline{1, p}$. Стратегия построения НЛР в этой задаче по-прежнему регулируется, например, ограничениями (23) и (24) на бинарные переменные.

В работах [20, 21] экспериментально установлено, что задача ЧБЛП (12), (19)–(22), (25)–(33) решается на порядок быстрее, чем задача (12)–(22). Кроме того, из-за согласованности знаков оценок β-параметров и знаков соответствующих коэффициентов корреляции для полученной регрессии становятся справедливыми формулы для абсолютных вкладов переменных в общую детерминацию R^2:

$$C_{x_j}^{\text{абс}} = r_{yx_j} \beta_j,\ j = \overline{1, l},\ C_{z_{jk}^{-}}^{\text{абс}} = r_{yz_{jk}^{-}} \beta_{jk}^{-},\ C_{z_{jk}^{+}}^{\text{абс}} = r_{yz_{jk}^{+}} \beta_{jk}^{+},\ j = \overline{1, C_l^2},\ k = \overline{1, p}, \tag{44}$$

по которым можно судить о степени влияния каждого регрессора на переменную y.

Сделаем два важных замечания относительно решения задачи (12), (19)–(22), (25)–(33).

Замечание 1. Поскольку в результате решения задачи знаки оценок β-параметров согласуются со знаками соответствующих коэффициентов корре-

ляции, то предварительно необходимо позаботиться о том, чтобы все знаки коэффициентов корреляции r_{yx_j} были согласованы с физическим смыслом переменных. Для этого можно привлекать экспертов из соответствующей предметной области. Несогласованные переменные следует исключать из рассмотрения. В противном случае полученную регрессию будет проблематично интерпретировать.

Замечание 2. Пусть, например, модель (8) имеет при параметре α_{11}^- регрессор $z_{11}^- = \min\{x_1, 8x_2\}$. Тогда при переходе к кусочно-заданному представлению при параметре α_{11}^- будет либо переменная x_1, либо $8x_2$. Если окажется, что $r_{yz_{11}^-} > 0$, то оценка параметра α_{11}^- гарантированно будет положительной и переменные x_1 и $8x_2$ будут влиять на y со знаком «плюс». В этом случае оба коэффициента корреляции r_{yx_1} и r_{yx_2} должны быть положительны, иначе возникает проблема с интерпретацией модели. А если окажется, что $r_{yz_{11}^-} < 0$, то оценка параметра α_{11}^- гарантированно будет отрицательной и переменные x_1 и $8x_2$ будут влиять на y со знаком «минус». В таком случае оба коэффициента корреляции r_{yx_1} и r_{yx_2} должны быть отрицательны. Из всего этого следует, что после согласования с экспертами знаков коэффициентов корреляции r_{yx_j}, $j = \overline{1, l}$, необходимо сформировать переменные z_{jk}^-, z_{jk}^+, $j = \overline{1, C_l^2}$, $k = \overline{1, p}$, найти их коэффициенты корреляции с переменной y и исключить те из них, для которых не выполняются условия

$$(r_{yz_{jk}^-} > 0 \text{ и } r_{yx_{\mu j1}} > 0 \text{ и } r_{yx_{\mu j2}} > 0)$$
$$\text{или } (r_{yz_{jk}^-} < 0 \text{ и } r_{yx_{\mu j1}} < 0 \text{ и } r_{yx_{\mu j2}} < 0), \qquad (45)$$
$$j = \overline{1, C_l^2},\ k = \overline{1, p},$$

$$(r_{yz_{jk}^+} > 0 \text{ и } r_{yx_{\mu j1}} > 0 \text{ и } r_{yx_{\mu j2}} > 0)$$
$$\text{или } (r_{yz_{jk}^+} < 0 \text{ и } r_{yx_{\mu j1}} < 0 \text{ и } r_{yx_{\mu j2}} < 0), \qquad (46)$$
$$j = \overline{1, C_l^2},\ k = \overline{1, p}.$$

Исключение противоречивых переменных естественным образом уменьшит время решения задачи построения НЛР. Это время можно еще значительно уменьшить, если дополнить выражения (45) и (46) условиями

$$\left|r_{yz_{jk}^-}\right| \geq r,\ \left|r_{yz_{jk}^+}\right| \geq r,\ j = \overline{1, C_l^2},\ k = \overline{1, p}, \qquad (47)$$

где r – выбранное из промежутка [0, 1) число. Чем больше число r, тем меньше становится количество переменных и время решения задачи.

2. МОДЕЛИРОВАНИЕ

Для построения НЛР были собраны ежегодные статистические данные за период с 2000 по 2020 г. для зависимой переменной y – отправление грузов железнодорожным транспортом общего пользования в Иркутской области (млн руб.), и шестидесяти двух переменных x_1, x_2, ..., x_{62}, предположительно влияющих на y. Сначала из этого списка было исключено шесть переменных, у которых значение коэффициента корреляции с y по абсолютной величине не превышало 0,2. Затем значения коэффициентов корреляции для оставшихся пятидесяти шести переменных были переданы двум экспертам, представляющим Управление Восточно-Сибирской железной дороги. Их задачей было исключить те переменные, для которых знаки коэффициентов корреляции с y не соответствуют экономическому смыслу решаемой задачи. В результате работы экспертов осталось восемь факторов:

x_2 – процент трудоспособного населения от общей численности;

x_3 – численность рабочей силы (тыс. чел.);

x_5 – численность пенсионеров (тыс. чел.);

x_8 – число собственных легковых автомобилей на 1000 человек населения (шт.);

x_{18} – число предприятий и организаций;

x_{20} – кредиторская задолженность организаций (млн руб.);

x_{22} – производство электроэнергии (млрд кВт·ч);

x_{58} – тарифы на грузовые перевозки (железнодорожный транспорт), усл.ед.

Значение переменной x_{58} за 2001 г. было назначено равным 1000 усл. ед. С его помощью по известным индексам тарифов были найдены оставшиеся значения переменной x_{58}.

Значения коэффициентов корреляции с переменной y для отобранных переменных составляют соответственно $r_{yx_2} = 0{,}785$, $r_{yx_3} = 0{,}543$, $r_{yx_5} = -0{,}483$, $r_{yx_8} = -0{,}446$, $r_{yx_{18}} = 0{,}538$, $r_{yx_{20}} = -0{,}204$, $r_{yx_{22}} = 0{,}476$, $r_{yx_{58}} = -0{,}465$.

Влияние выбранных переменных на переменную y можно обосновать следующим образом:

• рост численнсти рабочей силы x_2 и x_3, а также числа предприятий x_{18} и количества производимой электроэнергии x_{22}, приводит к увеличению объемов производимой регионом продукции, что влечет за собой повышение спроса на грузовые перевозки ж/д транспортом, в то время как рост значения переменной x_5 тормозит развитие экономики, снижая спрос на перевозки;

• избыток собственных автомобилей x_8 у населения снижает спрос как на пассажирские, так и на грузовые перевозки ж/д транспортом;

• рост суммарного объема кредиторской задолженности организаций x_{20} негативно сказывается на экономике региона, поскольку, например, может повлечь за собой наложение различных штрафных санкций;

• увеличение тарифов на грузовые перевозки x_{58} естественно снижает спрос на перевозки грузов по железной дороге.

Затем для каждой пары отобранных переменных по формулам (5) были определены промежутки для значений параметров λ_j. После этого для формирования матрицы Λ каждый промежуток был равномерно разбит четырьмя точками. В результате удалось сформировать $4C_8^2 = 112$ переменных z_{jk}^-, $j = \overline{1, 28}$, $k = \overline{1, 4}$, преобразованных с помощью бинарной операции минимум, и столько же переменных z_{jk}^-, $j = \overline{1, 28}$, $k = \overline{1, 4}$, преобразованных с помощью операции максимум. Далее из этих 224 переменных были исключены те, для которых не выполняются условия (45)–(47) при $r = 0{,}2$. Таких переменных оказалось 140. В итоге к построению НЛР мы подошли, имея в наличии 92 переменных, из которых 8 объясняющих и 84 преобразованных с помощью функций min и max.

Построение НЛР осуществлялось на основе решения задачи ЧБЛП с целевой функцией (12) и ограничениями (19)–(21), (25)–(33). Подчеркнем, что ограничение (22) на число входящих в модель регрессоров не ставилось. Для того чтобы в итоговую модель каждая объясняющая переменная входила не более одного раза, были учтены ограничения (24). Для решения задач ЧБЛП использовался решатель LPSolve IDE, а для формирования математических моделей задач для этого решателя была разработана специальная программа в среде программирования Delphi. Сначала с помощью этой программы были найдены неизвестные числа в ограничениях (25)–(27). Для этого было решено 184 задачи линейного программирования с соответствующими целевыми функциями и линейными ограничениями (34)–(43). Затем с использованием найденных чисел и разработанной программы для решателя LPSolve была сформулирована задача ЧБЛП (12), (19)–(21), (24)–(33), содержащая 284 ограничения, 92 вещественных и 92 бинарных переменных. Решение осуществлялось на персональном компьютере с процессором Intel Core i5 (3.40 ГГц, 4 ядра) и объемом оперативной памяти 8 ГБ. В результате примерно за 30 секунд была построена следующая НЛР:

$$\begin{aligned}\tilde{y} = -24{,}5274 &+ \underset{(13{,}98)}{\overset{(0{,}6427)}{1{,}1895}}\min\{x_2,\ 0{,}000933x_{18}\} - \\ &- \underset{(-3{,}361)}{\overset{(0{,}1129)}{0{,}0196}}\min\{x_5,\ 0{,}006754x_{20}\} - \\ &- \underset{(-2{,}182)}{\overset{(0{,}0843)}{0{,}0323}}\min\{x_8,\ 0{,}11725x_{58}\} + \\ &+ \underset{(3{,}859)}{\overset{(0{,}1063)}{0{,}0254}}\max\{x_3,\ 23{,}079x_{22}\},\end{aligned} \tag{48}$$

где в скобках под коэффициентами указаны значения *t*-критерия Стьюдента, а над коэффициентами – абсолютные вклады регрессоров в общую детерминацию, найденные по формулам (44). Оказалось, что все регрессоры значимы по критерию Стьюдента для уровня значимости $\alpha = 0{,}05$.

Предложенный в данной статье математический аппарат пока не позволяет контролировать в процессе построения НЛР значимость ее коэффициентов ни по критерию Стьюдента, ни по абсолютным вкладам переменных. Для этого в дальнейшем планируется интеграция в сформулированную задачу ЧБЛП специальных линейных ограничений.

Коэффициент детерминации НЛР (48) $R^2 = 0{,}946183$, что говорит о высоком качестве построенной модели.

Значения коэффициентов вздутия дисперсии для регрессоров модели (48) не превышают 10, что говорит об отсутствии в ней мультиколлинеарности. Стоит отметить, что контролировать мультиколлинеарность в сформулированной задаче ЧБЛП пока тоже нельзя.

Таким образом, выполняются все условия, чтобы отнести построенную модель (48) к вполне интерпретируемым.

Модель (48) в кусочно-заданном виде представлена в таблице.

Уравнения модели (48) для различных диапазонов значений переменных

Уравнение НЛР	Диапазоны значений переменных
$\tilde{y} = -24,527 + 0,0011x_{18} - 0,00013x_{20} - 0,0038x_{58} + 0,0254x_3$	$\frac{x_2}{x_{18}} \geq 0,000933, \frac{x_5}{x_{20}} \geq 0,00675, \frac{x_8}{x_{58}} \geq 0,117, \frac{x_3}{x_{22}} \geq 23,08$
$\tilde{y} = -24,527 + 0,0011x_{18} - 0,00013x_{20} - 0,0038x_{58} + 0,5857x_{22}$	$\frac{x_2}{x_{18}} \geq 0,000933, \frac{x_5}{x_{20}} \geq 0,00675, \frac{x_8}{x_{58}} \geq 0,117, \frac{x_3}{x_{22}} < 23,08$
$\tilde{y} = -24,527 + 0,0011x_{18} - 0,00013x_{20} - 0,0323x_8 + 0,0254x_3$	$\frac{x_2}{x_{18}} \geq 0,000933, \frac{x_5}{x_{20}} \geq 0,00675, \frac{x_8}{x_{58}} < 0,117, \frac{x_3}{x_{22}} \geq 23,08$
$\tilde{y} = -24,527 + 0,0011x_{18} - 0,00013x_{20} - 0,0323x_8 + 0,5857x_{22}$	$\frac{x_2}{x_{18}} \geq 0,000933, \frac{x_5}{x_{20}} \geq 0,00675, \frac{x_8}{x_{58}} < 0,117, \frac{x_3}{x_{22}} < 23,08$
$\tilde{y} = -24,527 + 0,0011x_{18} - 0,0196x_5 - 0,0038x_{58} + 0,0254x_3$	$\frac{x_2}{x_{18}} \geq 0,000933, \frac{x_5}{x_{20}} < 0,00675, \frac{x_8}{x_{58}} \geq 0,117, \frac{x_3}{x_{22}} \geq 23,08$
$\tilde{y} = -24,527 + 0,0011x_{18} - 0,0196x_5 - 0,0038x_{58} + 0,5857x_{22}$	$\frac{x_2}{x_{18}} \geq 0,000933, \frac{x_5}{x_{20}} < 0,00675, \frac{x_8}{x_{58}} \geq 0,117, \frac{x_3}{x_{22}} < 23,08$
$\tilde{y} = -24,527 + 0,0011x_{18} - 0,0196x_5 - 0,0323x_8 + 0,0254x_3$	$\frac{x_2}{x_{18}} \geq 0,000933, \frac{x_5}{x_{20}} < 0,00675, \frac{x_8}{x_{58}} < 0,117, \frac{x_3}{x_{22}} \geq 23,08$
$\tilde{y} = -24,527 + 0,0011x_{18} - 0,0196x_5 - 0,0323x_8 + 0,5857x_{22}$	$\frac{x_2}{x_{18}} \geq 0,000933, \frac{x_5}{x_{20}} < 0,00675, \frac{x_8}{x_{58}} < 0,117, \frac{x_3}{x_{22}} < 23,08$
$\tilde{y} = -24,527 + 1,1895x_2 - 0,00013x_{20} - 0,0038x_{58} + 0,0254x_3$	$\frac{x_2}{x_{18}} < 0,000933, \frac{x_5}{x_{20}} \geq 0,00675, \frac{x_8}{x_{58}} \geq 0,117, \frac{x_3}{x_{22}} \geq 23,08$
$\tilde{y} = -24,527 + 1,1895x_2 - 0,00013x_{20} - 0,0038x_{58} + 0,5857x_{22}$	$\frac{x_2}{x_{18}} < 0,000933, \frac{x_5}{x_{20}} \geq 0,00675, \frac{x_8}{x_{58}} \geq 0,117, \frac{x_3}{x_{22}} < 23,08$
$\tilde{y} = -24,527 + 1,1895x_2 - 0,00013x_{20} - 0,0323x_8 + 0,0254x_3$	$\frac{x_2}{x_{18}} < 0,000933, \frac{x_5}{x_{20}} \geq 0,00675, \frac{x_8}{x_{58}} < 0,117, \frac{x_3}{x_{22}} \geq 23,08$
$\tilde{y} = -24,527 + 1,1895x_2 - 0,00013x_{20} - 0,0323x_8 + 0,5857x_{22}$	$\frac{x_2}{x_{18}} < 0,000933, \frac{x_5}{x_{20}} \geq 0,00675, \frac{x_8}{x_{58}} < 0,117, \frac{x_3}{x_{22}} < 23,08$
$\tilde{y} = -24,527 + 1,1895x_2 - 0,0196x_5 - 0,0038x_{58} + 0,0254x_3$	$\frac{x_2}{x_{18}} < 0,000933, \frac{x_5}{x_{20}} < 0,00675, \frac{x_8}{x_{58}} \geq 0,117, \frac{x_3}{x_{22}} \geq 23,08$
$\tilde{y} = -24,527 + 1,1895x_2 - 0,0196x_5 - 0,0038x_{58} + 0,5857x_{22}$	$\frac{x_2}{x_{18}} < 0,000933, \frac{x_5}{x_{20}} < 0,00675, \frac{x_8}{x_{58}} \geq 0,117, \frac{x_3}{x_{22}} < 23,08$
$\tilde{y} = -24,527 + 1,1895x_2 - 0,0196x_5 - 0,0323x_8 + 0,0254x_3$	$\frac{x_2}{x_{18}} < 0,000933, \frac{x_5}{x_{20}} < 0,00675, \frac{x_8}{x_{58}} < 0,117, \frac{x_3}{x_{22}} \geq 23,08$
$\tilde{y} = -24,527 + 1,1895x_2 - 0,0196x_5 - 0,0323x_8 + 0,5857x_{22}$	$\frac{x_2}{x_{18}} < 0,000933, \frac{x_5}{x_{20}} < 0,00675, \frac{x_8}{x_{58}} < 0,117, \frac{x_3}{x_{22}} < 23,08$

Из таблицы видно, что в зависимости от выполненных условий меняется состав влияющих на y переменных, а оценки параметров $\lambda^{-}_{4,1} = 0,000933$, $\lambda^{-}_{16,2} = 0,00675$, $\lambda^{-}_{22,2} = 0,117$, $\lambda^{+}_{12,3} = 23,08$ играют роль точек переключения для следующих четырех автоматически сформированных показателей:

– отношение процента трудоспособного населения (x_2) к числу предприятий и организаций (x_{18});

– отношение численности пенсионеров (x_5) к кредиторской задолженности (x_{20});

– отношение числа собственных легковых автомобилей (x_8) к тарифам на грузовые ж/д перевозки (x_{58});

– отношение численности рабочей силы (x_3) к объему производства электроэнергии (x_{22}).

Тогда справедлива следующая интерпретация.

• Если значение показателя x_2/x_{18} не меньше, чем 0,000933, то на отправление грузов ж/д транспортом влияет число предприятий и организаций x_{18}, а процент трудоспособного населения x_2 не влияет. При этом с увеличением числа предприятий и организаций x_{18} на одну единицу (при неиз-

менных значениях остальных переменных) отправление грузов y возрастает в среднем на 0,0011 млн руб. А если значение показателя x_2/x_{18} меньше, чем 0,000933, то на отправление грузов влияет процент трудоспособного населения x_2, а число предприятий и организаций x_{18} не влияет. При этом с увеличением процента трудоспособного населения x_2 на одну единицу (при неизменных значениях остальных переменных) отправление грузов y возрастает в среднем на 1,1895 млн руб.

• Если значение показателя x_5/x_{20} не меньше, чем 0,00675, то на отправление грузов ж/д транспортом влияет кредиторская задолженность x_{20}, а численность пенсионеров x_5 не влияет. При этом с увеличением кредиторской задолженности x_{20} на 1 млн руб. (при неизменных значениях остальных переменных) отправление грузов y убывает в среднем на 0,00013 млн руб. А если значение показателя x_5/x_{20} меньше, чем 0,00675, то на отправление грузов влияет численность пенсионеров x_5, а кредиторская задолженность x_{20} не влияет. При этом с увеличением численности пенсионеров x_5 на 1 тыс. человек (при неизменных значениях остальных переменных) отправление грузов y убывает в среднем на 0,0196 млн руб.

• Если значение показателя x_8/x_{58} не меньше, чем 0,117, то на отправление грузов ж/д транспортом влияют тарифы на грузовые ж/д перевозки x_{58}, а число собственных легковых автомобилей x_8 не влияет. При этом с увеличением тарифов на грузовые ж/д перевозки x_{58} на одну условную единицу (при неизменных значениях остальных переменных) отправление грузов y убывает в среднем на 0,0038 млн руб. А если значение показателя x_8/x_{58} меньше, чем 0,117, то на отправление грузов влияет число собственных легковых автомобилей x_8, а тарифы на грузовые ж/д перевозки x_{58} не влияют. При этом с увеличением числа собственных легковых автомобилей на 1000 человек населения x_8 на 1 шт. (при неизменных значениях остальных переменных) отправление грузов y убывает в среднем на 0,0323 млн руб.

• Если значение показателя x_3/x_{22} не меньше, чем 23,08, то на отправление грузов ж/д транспортом влияет численность рабочей силы x_3, а объемы производства электроэнергии x_{22} не влияют. При этом с увеличением численности рабочей силы x_3 на 1 тыс. человек (при неизменных значениях остальных переменных) отправление грузов y возрастает в среднем на 0,0254 млн руб. А если значение показателя x_3/x_{22} меньше, чем 23,08, то на отправление грузов влияют объемы производства электроэнергии x_{22}, а численность рабочей силы x_3 не влияет. При этом с увеличением объемов производства электроэнергии x_{22} на 1 млрд кВт·ч (при неизменных значениях остальных переменных) отправление грузов y возрастает в среднем на 0,5857 млн руб.

Таким образом, интерпретационные характеристики НЛР представляются богаче и разнообразнее тех же характеристик традиционной линейной регрессионной модели. При этом аппроксимационные характеристики НЛР в зависимости от выбранной стратегии построения должны в большинстве случаев превосходить те же характеристики линейных регрессий, являющихся лишь частным случаем НЛР. Ценность предложенных НЛР заключается в том, что их можно использовать не только для прогнозирования, но и для извлечения новых интерпретируемых математических закономерностей, призванных повысить эффективность управленческих решений в различных отраслях экономики.

Также отметим, что сама по себе НЛР лучше подходит для моделирования в условиях мультиколлинеарности, чем традиционная линейная регрессия. Дело в том, что чем больше в НЛР бинарных операций, тем выше число ее степеней свободы по сравнению с линейной регрессией. Это означает, что НЛР позволяет «вместить» в себя больше переменных при меньшем числе регрессоров, чем линейная регрессия. Например, НЛР (48) содержит только четыре регрессора, но при этом восемь переменных, поэтому шанс на то, что в ней будет присутствовать мультиколлинеарность, априори ниже, чем шанс столкнуться с мультиколлинеарностью в линейной регрессии со всеми восемью переменными.

ЗАКЛЮЧЕНИЕ

В работе рассмотрена НЛР, в которой помимо бинарной операции минимум используется еще и бинарная операция максимум. Предложен метод построения НЛР, в основе которого лежит решение задачи ЧБЛП. В результате решения этой задачи определяется структурная спецификация НЛР и ее приближенные МНК-оценки. Показано, как с помощью регулирования ограничений на бинарные переменные можно контролировать структурную спецификацию НЛР. Продемонстрировано, каким образом нужно исключать на начальном этапе противоречивые переменные, чтобы уменьшить время решения задачи и гарантировать интерпретируемость НЛР. С помощью предложенного метода построена модель, позволившая выявить новые закономерности функционирования железнодорожного транспорта в Иркутской области,

недоступные при использовании классической линейной регрессии.

Предложенный в работе метод является универсальным и может применяться для построения НЛР в любой предметной области при наличии статистических данных, состоящих только из положительных переменных. Проводимое в его рамках разбиение параметров приводит к формированию задачи ЧБЛП, оптимальное решение которой при достаточно большом числе разбиений дает оценки, практически не отличающиеся от оптимальных МНК-оценок рассматриваемой НЛР. Естественно, что с ростом числа разбиений будет расти и время решения задачи. Тем не менее, как продемонстрировано в работах [20, 21] на примере линейной регрессии, такая задача ЧБЛП решается на порядок быстрее, чем при использовании стандартных переборных процедур. Тестирование скорости построения НЛР по выборкам разных объемов с помощью предложенного метода будет проведено в последующих работах автора.

ЛИТЕРАТУРА

1. *Arkes, J.* Regression Analysis: A Practical Introduction. – Routledge, 2019. – 362 p.
2. *Westfall, P.H., Arias, A.L.* Understanding Regression Analysis: A Conditional Distribution Approach. – Chapman and Hall/CRC, 2020. – 514 p.
3. *Клейнер Г.Б.* Производственные функции: Теория, методы, применение. – М.: Финансы и статистика, 1986. – 239 с. [*Kleiner, G.B.* Proizvodstvennye funktsii: Teoriya, metody, primenenie. – Moscow: Finansy i statistika, 1986. – 239 s. (In Russian)]
4. *Onalan, O., Basegmez, H.* Estimation of Economic Growth Using Grey Cobb-Douglas Production Function: An Application for US Economy // Journal of Business Economics and Finance. – 2018. – Vol. 7, no. 2. – P. 178–190.
5. *Yankovyi, O., Koval, V., Lazorenko, L.,* et al. Modeling Sustainable Economic Development Using Production Functions // Studies of Applied Economics. – 2021. – Vol. 39, no. 5.
6. *Ishikawa, A.* Why Does Production Function Take the Cobb–Douglas Form? // Statistical Properties in Firms' Large-scale Data. – Springer, Singapore, 2021. – P. 113–135.
7. *Носков С.И.* Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. – Иркутск: РИЦ ГП «Облинформпечать», 1996. – 320 с. [*Noskov, S.I.* Tekhnologiya modelirovaniya ob"ektov s nestabil'nym funktsionirovaniem i neopredelennost'yu v dannykh. – Irkutsk: RITS GP «Oblinformpechat'», 1996. – 320 s. (In Russian)]
8. *Шор Н.З.* Методы минимизации недифференцируемых функций и их приложения. – Киев: Наук. думка, 1979. – 200 с. [*Shor, N.Z.* Metody minimizatsii nedifferentsiruemykh funktsii i ikh prilozheniya. – Kiev: Nauk. dumka, 1979. – 200 s. (In Russian)]
9. *Scaman, K., Bach, F., Bubeck, S.,* et al. Optimal Algorithms for Non-smooth Distributed Optimization in Networks // Advances in Neural Information Processing Systems. – 2018. – Vol. 31.
10. *Khamaru, K., Wainwright, M. J.* Convergence Guarantees for a Class of Non-convex and Non-smooth Optimization Problems // Journal of Machine Learning Research. – 2019. – Vol. 20, no. 154. – P. 1–52.
11. *Иванова Н.К., Лебедева С.А., Носков С.И.* Идентификация параметров некоторых негладких регрессий // Информационные технологии и проблемы математического моделирования сложных систем. – 2016. – № 17. – С. 107–110. [*Ivanova, N.K., Lebedeva, S.A., Noskov, S.I.* Identifikatsiya parametrov nekotorykh negladkikh regressii // Informatsionnye tekhnologii i problemy matematicheskogo modelirovaniya slozhnykh sistem. – 2016. – No. 17. – P. 107–110. (In Russian)]
12. *Носков С.И., Хоняков А.А.* Программный комплекс построения некоторых типов кусочно-линейных регрессий // Информационные технологии и математическое моделирование в управлении сложными системами. – 2019. – № 3 (4). – С. 47–55. [*Noskov, S.I., Khonyakov, A.A.* Programmnyi kompleks postroeniya nekotorykh tipov kusochno-lineinykh regressii // Informa-tsionnye tekhnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami. – 2019. – No. 3 (4). – P. 47–55. (In Russian)]
13. *Park, Y.W., Klabjan, D.* Subset Selection for Multiple Linear Regression via Optimization // Journal of Global Optimization. – 2020. – Vol. 77. – P. 543–574.
14. *Chung, S., Park, Y.W., Cheong, T.* A Mathematical Programming Approach for Integrated Multiple Linear Regression Subset Selection and Validation // Pattern Recognition. – 2020. – Vol. 108. – P. 107565.
15. *Bertsimas, D., Li, M.L.* Scalable Holistic Linear Regression // Operations Research Letters. – 2020. – Vol. 48, no. 3. – P. 203–208.
16. *Базилевский М.П.* МНК-оценивание параметров специфицированных на основе функций Леонтьева двухфакторных моделей регрессии // Южно-Сибирский научный вестник. – 2019. – № 2 (26). – С. 66–70. [*Bazilevskii, M.P.* MNK-otsenivanie parametrov spetsifitsirovannykh na osnove funktsii Leont'eva dvukhfaktornykh modelei regressii // Yuzhno-Sibirskii nauchnyi vestnik. – 2019. – No. 2 (26). – P. 66–70. (In Russian)]
17. *Базилевский М.П.* Оценивание линейно-неэлементарных регрессионных моделей с помощью метода наименьших квадратов // Моделирование, оптимизация и информационные технологии. – 2020. – Т. 8. – № 4 (31). [*Bazilevskii, M.P.* Otsenivanie lineino-neehlementarnykh regressionnykh modelei s pomoshch'yu metoda naimen'shikh kvadratov // Modelirovanie, optimizatsiya i informatsionnye tekhnologii. – 2020. – Vol. 8, – no. 4 (31). (In Russian)]
18. *Базилевский М.П.* Отбор информативных операций при построении линейно-неэлементарных регрессионных моделей // International Journal of Open Information Technologies. – 2021. – Т. 9. – № 5. – С. 30–35. [*Bazilevskii, M.P.* Otbor informativnykh operatsii pri postroenii lineino-neehlementarnykh regressionnykh modelei // International Journal of Open Information Technologies. – 2021. – Vol. 9, no. 5. – P. 30–35. (In Russian)]
19. *Базилевский М.П.* Сведение задачи отбора информативных регрессоров при оценивании линейной регрессионной модели по методу наименьших квадратов к задаче частично-булевого линейного программирования // Моделирование, оптимизация и информационные технологии. – 2018. – Т. 6. – № 1 (20). – С. 108–117. [*Bazilevskii, M.P.* Svedenie zadachi

otbora informativnykh regressorov pri otsenivanii lineinoi regressionnoi modeli po metodu naimen'shikh kvadratov k zadache chastichno-bulevogo lineinogo programmirovaniya // Modelirovanie, optimizatsiya i informatsionnye tekhnologii. – 2018. – Vol. 6, no. 1 (20). – P. 108–117. (In Russian)]

20. *Базилевский М.П.* Способ определения параметра M в задаче частично-булевого линейного программирования для отбора регрессоров в линейной регрессии // Вестник Технологического университета. – 2022. – Т. 25. – № 2. – С. 62–66. [*Bazilevskii, M.P.* Sposob opredeleniya parametra M v zadache chastichno-bulevogo lineinogo programmirovaniya dlya otbora regressorov v lineinoi regressii // Vestnik Tekhnologicheskogo universiteta. – 2022. – Vol. 25, no. 2. – P. 62–66. (In Russian)]

21. *Konno, H., Yamamoto, R.* Choosing the Best Set of Variables in Regression Analysis Using Integer Programming // Journal of Global Optimization. – 2009. – Vol. 44. – P. 273– 282.

Статья представлена к публикации руководителем регионального редсовета М.И. Гераськиным.

Поступила в редакцию 23.04.2022,
после доработки 3.08.2022.
Принята к публикации 31.08.2022.

Базилевский Михаил Павлович – канд. техн. наук., Иркутский государственный университет путей сообщения, г. Иркутск, ✉ mik2178@yandex.ru.

A METHOD FOR CONSTRUCTING NONELEMENTARY LINEAR REGRESSIONS BASED ON MATHEMATICAL PROGRAMMING

M.P. Bazilevskiy

Irkutsk State Transport University, Irkutsk, Russia

✉ mik2178@yandex.ru

Abstract. This paper is devoted to constructing nonelementary linear regressions consisting of explanatory variables and all possible combinations of their pairs transformed using binary minimum and maximum operations. Such models are formalized through a 0-1 mixed integer linear programming problem. By adjusting the constraints on binary variables, we control the structural specification of a nonelementary linear regression, namely, the number of regressors, their types, and the composition of explanatory variables. In this case, the model parameters are approximately estimated using the ordinary least squares method. The formulated problem has advantages: the number of constraints does not depend on the sample size, and the signs of the estimates for the explanatory variables are consistent with the signs of their correlation coefficients with the dependent variable. Regressors are eliminated at the initial stage to reduce the time for solving the problem and make the model quite interpretable. A nonelementary linear regression of rail freight in Irkutsk oblast is constructed, and its interpretation is given.

Keywords: nonelementary linear regression, ordinary least squares method, 0-1 mixed integer linear programming problem, subset selection, coefficient of determination, interpretation, rail freight.