# Отбор информативных операций при построении линейно-неэлементарных регрессионных моделей

М. П. Базилевский

***Аннотация*—Настоящая статья посвящена одной из главных проблем регрессионного анализа – выбору структурной спецификации регрессионной модели. Работа основана на предложенных ранее автором линейно-неэлементарных регрессиях, в которые помимо объясняющих переменных входят бинарные операции всех возможных их пар. В таких моделях с ростом числа объясняющих переменных существенно возрастает число бинарных операций. Целью данной работы является разработка алгоритмов отбора в линейно-неэлементарных регрессиях наиболее информативных переменных и операций. Рассмотрен алгоритм приближенного оценивания линейно-неэлементарных регрессий с помощью метода наименьших квадратов. Сформулирована задача отбора информативных операций. Предложено две стратегии построения линейно-неэлементарных регрессий. В первой из них нет ограничений на число вхождений объясняющих переменных в модель и на число бинарных операций. Во второй – модель содержит наибольшее число бинарных операций, а каждая объясняющая переменная входит в неё только один раз. С использованием комбинаторики была определена вычислительная сложность каждой из этих стратегий. Оказалось, что задача построения линейно-неэлементарной модели на основе второй стратегии на практике решается значительно быстрее, чем аналогичная задача на основе первой стратегии. Предложенные алгоритмы с помощью пакета Gretl были реализованы в виде специальной программы. С помощью неё были построены высококачественные линейно-неэлементарные регрессионные модели грузовых железнодорожных перевозок в Иркутской области.**

***Ключевые слова*—линейно-неэлементарная регрессия, метод наименьших квадратов, отбор информативных операций, вычислительная сложность, грузовые железнодорожные перевозки в Иркутской области.**

## I. Введение

В регрессионном анализе [1,2] одной из главных является проблема выбора при построении модели состава объясняющих переменных и математической формы связи между ними. Таких форм в настоящее время уже существует значительное количество (см., например, [3,4]). И этот арсенал продолжает расширяться. Так, например, в [5] разработаны степенно-показательные регрессионные модели, в [6] – модели полносвязной линейной регрессии, в [7] – линейно-мультипликативные регрессии, в [8] – индексные регрессии, в [9, 10] – регрессии, основанные на производственной функции Леонтьева. Все эти формы позволяют выявлять самые различные скрытые механизмы функционирования исследуемых объектов или процессов. Для выбора лучшей из этих форм целесообразно реализовывать технологию организации «конкурса» моделей [11].

В работе [12] автором предложена новая форма связи между переменными – линейно-неэлементарная регрессия (ЛНР). К сожалению, при построении ЛНР с ростом числа объясняющих переменных существенно возрастает количество входящих в неё бинарных операций. Поэтому целью данной работы является разработка стратегий построения ЛНР, предназначенных для отбора из всего множества объясняющих переменных и бинарных операций наиболее информативных с точки зрения заданных критериев качества.

## II. Линейно-неэлементарные регрессии

ЛНР [12] представляет собой модель, в которую помимо регрессоров $x_1$, $x_2$, ..., $x_l$ входят бинарные операции всех возможные комбинаций их пар:

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \sum_{j=1}^{p} \alpha_{j+l} \min\left\{x_{i\mu_{j1}}, \lambda_j x_{i\mu_{j2}}\right\} + \varepsilon_i, \quad i = \overline{1,n}, \qquad (1)$$

где $n$ – объем выборки; $l$ – количество объясняющих переменных; $y_i$ и $x_{ij}$, $i = \overline{1,n}$, $j = \overline{1,l}$ – известные значения объясняемой и объясняющих переменных; $\varepsilon_i$, $i = \overline{1,n}$ – ошибки аппроксимации; $p = C_l^2$ – количество пар переменных; $\mu_{j1}$ и $\mu_{j2}$, $j = \overline{1,p}$ – элементы первого и второго столбца матрицы пар индексов переменных размера $p \times 2$; $\alpha_j$, $j = \overline{0,l+p}$, $\lambda_j$, $j = \overline{1,p}$ – подлежащие оцениванию параметры.

Будем считать, что значения всех объясняющих переменных, входящих в регрессионную модель (1), являются положительными, т.е. $x_{ij} > 0$, $i = \overline{1,n}$, $j = \overline{1,l}$.

Статья получена 11 марта 2021.

Базилевский Михаил Павлович, Иркутский государственный университет путей сообщения, Иркутск, Российская Федерация (e-mail: mik2178@yandex.ru).

Если в регрессии (1) параметры $\lambda_j$, $j=\overline{1,p}$ неизвестны, то она является в значительной степени нелинейной. Однако если коэффициенты $\lambda_j$, $j=\overline{1,p}$ заданы, то модель (1) представляет собой квазилинейную регрессию только с неизвестными параметрами $\alpha_0$, $\alpha_1$, ..., $\alpha_{l+C_l^2}$, для оценивания которых можно применить обычный метод наименьших квадратов (МНК).

Алгоритм приближенного МНК-оценивания ЛНР (1) [12] при неизвестных параметрах $\lambda_j$, $j=\overline{1,p}$ представлен на рис. 1.

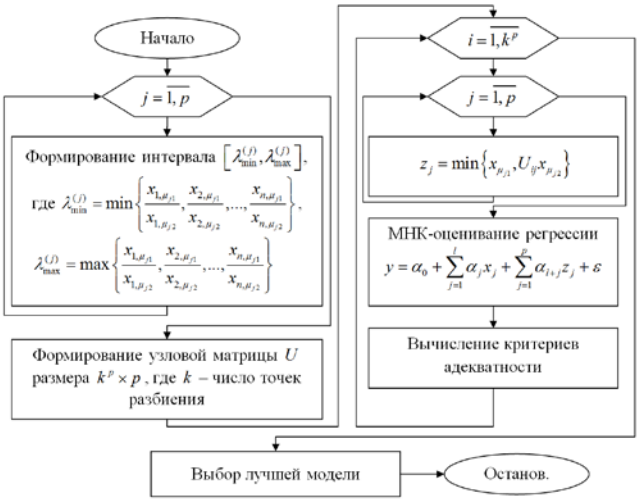


**Рис. 1.** Алгоритм приближенного оценивания ЛНР

Суть алгоритма, изображенного на рис. 1, состоит в том, чтобы с помощью перебора значений параметров $\lambda_j$, $j=\overline{1,p}$ найти наилучшие с точки зрения минимума суммы квадратов ошибок оценки параметров $\alpha_0$, $\alpha_1$, ..., $\alpha_{l+C_l^2}$ ЛНР (1). В [12] показано, что для перебора достаточно ограничиться промежутками

$$\lambda_j\in\left(\lambda_{\min}^{(j)},\lambda_{\max}^{(j)}\right), \qquad (2)$$

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 факторов.

Пусть для каждого промежутка (2) задано одинаковое число точек его разбиения – $k$. Тогда для приближенного оценивания регрессионной модели (1) необходимо с помощью МНК идентифицировать $k^p$ штук ЛНР и выбрать из них лучшую на основании одного или нескольких критериев адекватности.

## III. Отбор информативных операций

Перейдем к рассмотрению задачи отбора информативных регрессоров [11] для ЛНР (1). При этом сразу стоит подчеркнуть, что в состав ЛНР могут входить не только регрессоры, но и операции минимум. Поэтому правильнее в данном случае будет говорить не об отборе регрессоров, а об отборе информативных операций (ОИО).

Приведем постановку задачи ОИО. Пусть задана выборка из наблюдений для объясняемой переменной $y$ и для возможных независимых переменных $x_j$, $j=\overline{1,l}$ ($l\geq 2$). Из этих регрессоров составим все возможные комбинации их пар, общее число которых $p=C_l^2$, и применим для этих пар бинарные операции минимум так, как это сделано в спецификации (1). Из общего числа $l$ регрессоров и $C_l^2$ операций требуется выбрать $m$ штук на основе некоторого критерия качества.

Рассмотрим следующие стратегии построения регрессионной модели (1).

*Стратегия 1*. Нет ограничений на число вхождений объясняющих переменных в ЛНР и на число бинарных операций.

Пусть в модели (1) значения коэффициентов $\lambda_j$, $j=\overline{1,p}$ известны. В этом случае число перебираемых альтернатив $r$ находится по формуле:

$$r=C_{l+C_l^2}^m. \qquad (3)$$

Для наглядного представления вычислительной сложности этой стратегии, по формуле (3) было рассчитано число альтернатив $r$ для заданного количества операций из интервала $1\leq m\leq 10$ и общего числа переменных из интервала $2\leq l\leq 6$. Результаты вычислений приведены в таблице 1.

**Таблица 1.** Количество альтернатив для первой стратегии (при известных параметрах $\lambda_j$)

| m \ l | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 1 | 3 | 6 | 10 | 15 | 21 |
| 2 | 3 | 15 | 45 | 105 | 210 |
| 3 | 1 | 20 | 120 | 455 | 1330 |
| 4 | 0 | 15 | 210 | 1365 | 5985 |
| 5 | 0 | 6 | 252 | 3003 | 20349 |
| 6 | 0 | 1 | 210 | 5005 | 54264 |
| 7 | 0 | 0 | 120 | 6435 | 116280 |
| 8 | 0 | 0 | 45 | 6435 | 203490 |
| 9 | 0 | 0 | 10 | 5005 | 293930 |
| 10 | 0 | 0 | 1 | 3003 | 352716 |

По таблице 1 можно сделать вывод, что вычислительная сложность представленной стратегии построения ЛНР гораздо ниже, чем для первой стратегии построения линейно-мультипликативной регрессии [7]. Но, к сожалению, для данной стратегии остается нерешенным вопрос с выбором значений коэффициентов $\lambda_j$, $j=\overline{1,p}$. Поэтому гораздо больший интерес представляет оценивание регрессии (1) при неизвестных значениях $\lambda_j$, $j=\overline{1,p}$. Алгоритм решения задачи ОИО для этого случая представлен на рис. 2.

Как видно по рис. 2, для реализации данной стратегии каждая альтернатива из общего числа (3) должна пройти ещё дополнительную процедуру приближенной идентификации неизвестных параметров $\lambda_j$, $j=\overline{1,p}$ по отображенному на рис. 1 алгоритму. Тогда общее количество перебираемых альтернатив значительно

увеличится. Подсчитаем их количество. Модель (1) состоит из $l$ регрессоров и $C_l^2$ операций. Выбирая из неё $m$ информативных факторов можно получить $C_{C_l^2}^m$ спецификаций, полностью состоящих из операций, $C_{C_l^2}^{m-1}C_l^1$ спецификаций, состоящих из $(m-1)$-й операции и одного регрессора, $C_{C_l^2}^{m-2}C_l^2$ спецификаций, состоящих из $(m-2)$-х операций и двух регрессоров и т.д. Причем, общее число таких спецификаций находится по формуле (3), т.е. справедливо соотношение

$$C_{l+C_l^2}^m = \sum_{q=0}^{m} C_{C_l^2}^{m-q} C_l^q .$$

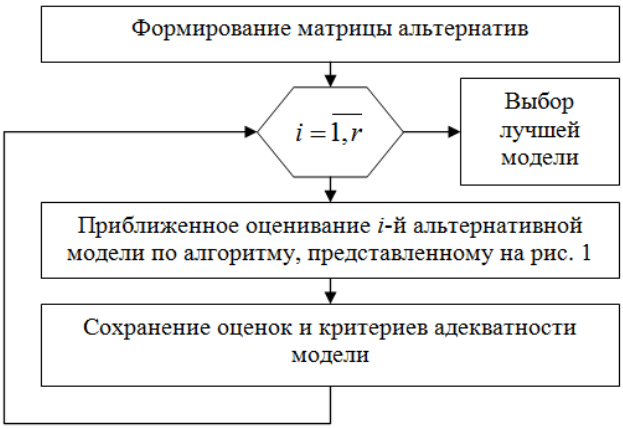


**Рис. 2.** Алгоритм ОИО

Если при выборе из модели (1) $m$ информативных факторов её спецификация полностью состоит из операций, то общее количество перебираемых альтернатив увеличится в $k^m$ раз, если спецификация состоит из $(m-1)$-й операции, то в $k^{m-1}$ раз и т.д. Тогда общее количество альтернатив для первой стратегии построения модели (1) при неизвестных значениях $\lambda_j$, $j=\overline{1,p}$ находится по формуле:

$$r' = \sum_{q=0}^{m} C_{C_l^2}^{m-q} C_l^q k^{m-q} . \qquad (4)$$

Пусть число точек разбиения $k=10$. Используя формулу (4) было вычислено количество альтернатив при $1 \le m \le 10$ и $2 \le l \le 6$. Результаты представлены в таблице 2.

**Таблица 2.** Количество альтернатив для первой стратегии (при неизвестных параметрах $\lambda_j$)

| l / m | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| 1 | 12 | 33 | 64 | 105 | 156 |
| 2 | 21 | 393 | 1746 | 5010 | 11415 |
| 3 | 10 | 1991 | 26364 | 143510 | 520270 |
| 4 | 0 | 3930 | 239241 | 2746005 | 16540515 |
| 5 | 0 | 3300 | 1326060 | 36945501 | 389237256 |
| 6 | 0 | 1000 | 4381500 | 358222600 | 7020808401 |
| 7 | 0 | 0 | 8220000 | 2523604500 | 99164388150 |
| 8 | 0 | 0 | 8550000 | 12862620000 | 1110888490500 |
| 9 | 0 | 0 | 4600000 | 46728100000 | 9935936855000 |
| 10 | 0 | 0 | 1000000 | 118075200000 | 71076390450000 |

Таким образом, по вычислительной сложности данная стратегия построения модели (1) при неизвестных коэффициентах не только превосходит стратегию её построения при известных коэффициентах (табл. 1), но и первую стратегию построения линейно-мультипликативных регрессий [7]. Поэтому было принято решение разработать менее сложную в вычислительном плане стратегию построения регрессии (1), вводя некоторые ограничения на спецификацию её альтернативных вариантов.

*Стратегия 2*. ЛНР содержит ровно $s$ переменных, наибольшее число бинарных операций, а каждая независимая переменная входит в неё только 1 раз.

Отметим, что увеличение количества бинарных операций в ЛНР является хорошим инструментом для борьбы с мультиколлинеарностью. Например, сильная корреляция объясняющих переменных $x_1$, $x_2$, $x_3$ приведет в линейной регрессии к возникновению эффекта мультиколлинеарности. А в ЛНР вида $y = \alpha_0 + \alpha_1 \min\{x_1, \lambda x_2\} + \alpha_2 x_3 + \varepsilon$ это явление будет значительно слабее за счет меньшего числа неизвестных параметров. Причем, контролируя величину параметра $\lambda$, в некоторых случаях можно и вовсе свести эффект мультиколлинеарности к нулю.

Если задано 3 объясняющих переменных $x_1$, $x_2$, $x_3$, то данная стратегия должна привести к выбору лучшей регрессии из следующего множества альтернатив:

$$y = \alpha_0 + \alpha_1 \min\{x_1, \lambda x_2\} + \alpha_2 x_3 ,$$
$$y = \alpha_0 + \alpha_1 \min\{x_1, \lambda x_3\} + \alpha_2 x_2 ,$$
$$y = \alpha_0 + \alpha_1 \min\{x_2, \lambda x_3\} + \alpha_2 x_1 .$$

Обозначим наибольшее число бинарных операций – bin . Тогда это число определяется по формуле:

$$\text{bin} = \left[\frac{s}{2}\right],$$

где $[\cdot]$ – целая часть числа.

Если остаток от деления $\frac{s}{2}$ равен нулю, то спецификация ЛНР состоит только из bin бинарных операций, а если единице, то из bin бинарных операций и одной переменной.

Сначала предположим, что в модели (1) параметры $\lambda_j$, $j=\overline{1,p}$ известны. Формирование альтернативных вариантов моделей для этой стратегии осуществляется по следующему алгоритму.

1. Из общего набора $l$ переменных выбирается $s$ штук. Этот выбор можно сделать $C_l^s$ способами.

2. Для каждого такого выбора формируется ЛНР (1), состоящая из bin , либо из bin+1 факторов. При этом первую бинарную операцию можно сформировать $C_s^2$ способами, вторую – $C_{s-2}^2$ и т.д. Если остаток от деления $\frac{s}{2}$ равен нулю, то последнюю бинарную операцию можно сформировать $C_2^2 = 1$ способом, а если остаток равен 1, то последнюю операцию – $C_3^2$ способами, и оставшийся регрессор – $C_1^1 = 1$ способом.

Учитывая повторения в комбинациях, запишем формулу для вычисления количества альтернатив для данной стратегии:

$$r = C_l^s \frac{\prod_{q=0}^{\text{bin}-1} C_{s-2q}^2}{\text{bin}!}. \quad (5)$$

С помощью формулы (5) была оценена вычислительная сложность предложенной стратегии. Количества формируемых альтернатив для заданных параметров $1 \le s \le 10$ и $1 \le l \le 10$ можно найти в таблице 3.

**Таблица 3.** Количество альтернатив для второй стратегии (при известных параметрах $\lambda_j$)

| l \ s | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 2 | 0 | 1 | 3 | 6 | 10 | 15 | 21 | 28 | 36 | 45 |
| 3 | 0 | 0 | 3 | 12 | 30 | 60 | 105 | 168 | 252 | 360 |
| 4 | 0 | 0 | 0 | 3 | 15 | 45 | 105 | 210 | 378 | 630 |
| 5 | 0 | 0 | 0 | 0 | 15 | 90 | 315 | 840 | 1890 | 3780 |
| 6 | 0 | 0 | 0 | 0 | 0 | 15 | 105 | 420 | 1260 | 3150 |
| 7 | 0 | 0 | 0 | 0 | 0 | 0 | 105 | 840 | 3780 | 12600 |
| 8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 105 | 945 | 4725 |
| 9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 945 | 9450 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 945 |

Как видно по таблице 3, вычислительная сложность данной стратегии существенно ниже, чем для первой стратегии.

Предположим, что в модели (1) коэффициенты $\lambda_j$, $j = \overline{1, p}$ неизвестны. Тогда для каждой альтернативы из таблицы 3 нужно провести процедуру идентификации этих коэффициентов, представленную на рис. 1. Поэтому формула (5) примет вид:

$$r' = k^{\text{bin}} C_l^s \frac{\prod_{q=0}^{\text{bin}-1} C_{s-2q}^2}{\text{bin}!}. \quad (6)$$

В таблице 4 приведены количества альтернативных вариантов данной стратегии, вычисленные по формуле (6) при $k = 10$, $1 \le s \le 10$ и $1 \le l \le 10$.

**Таблица 4.** Количество альтернатив для второй стратегии (при неизвестных параметрах $\lambda_j$)

| l \ s | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 1 | 1 | 2 | 3 | 4 | 5 |
| 2 | 0 | 10 | 30 | 60 | 100 |
| 3 | 0 | 0 | 30 | 120 | 300 |
| 4 | 0 | 0 | 0 | 300 | 1500 |
| 5 | 0 | 0 | 0 | 0 | 1500 |
| 6 | 0 | 0 | 0 | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 |
| 9 | 0 | 0 | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 |

Продолжение таблицы 4

| l \ s | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|
| 1 | 6 | 7 | 8 | 9 | 10 |
| 2 | 150 | 210 | 280 | 360 | 450 |
| 3 | 600 | 1050 | 1680 | 2520 | 3600 |
| 4 | 4500 | 10500 | 21000 | 37800 | 63000 |
| 5 | 9000 | 31500 | 84000 | 189000 | 378000 |
| 6 | 15000 | 105000 | 420000 | 1260000 | 3150000 |
| 7 | 0 | 105000 | 840000 | 3780000 | 12600000 |
| 8 | 0 | 0 | 1050000 | 9450000 | 47250000 |
| 9 | 0 | 0 | 0 | 9450000 | 94500000 |
| 10 | 0 | 0 | 0 | 0 | 94500000 |

По таблице 4 видно, что задача построение модели (1) на основе данной стратегии при неизвестных коэффициентах на практике будет решаться значительно быстрее, чем аналогичная задача для первой стратегии (табл. 2).

## IV. Моделирование перевозок

Актуальной на сегодняшний день задачей является моделирование перевозок железнодорожного транспорта (см., например, [13-15]). Для построения с помощью предложенного в этой статье математического аппарата моделей грузовых железнодорожных перевозок в Иркутской области были использованы данные за период 2000-2018 гг. по следующим переменным:

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

$x_1$ – численность населения (тыс. человек);

$x_2$ – население в трудоспособном возрасте (в процентах от общей численности населения);

$x_3$ – численность рабочей силы (тыс. человек);

$x_4$ – численность безработных (тыс. человек);

$x_{14}$ – валовой региональный продукт (млн рублей);

$x_{15}$ – инвестиции в основной капитал (млн рублей);

$x_{16}$ – доходы консолидированного бюджета (млн рублей);

$x_{17}$ – расходы консолидированного бюджета (млн рублей);

$x_{21}$ – объем промышленной продукции (млн рублей);

$x_{22}$ – производство электроэнергии (млрд киловатт-часов);

$x_{23}$ – среднегодовая номинальная начисленная заработная плата работников в области добычи полезных ископаемых (рублей);

$x_{24}$ – среднегодовая номинальная начисленная заработная плата работников в области обрабатывающих производств (рублей);

$x_{25}$ – продукция сельского хозяйства (млн рублей);

$x_{26}$ – среднегодовая номинальная начисленная заработная плата работников сельского хозяйства, охоты и лесного хозяйства (рублей);

$x_{35}$ – удельный вес автомобильных дорог с твердым покрытием в общей протяженности автомобильных дорог общего пользования (в процентах);

$x_{36}$ – удельный вес автомобильных дорог с усовершенствованным покрытием в протяженности автомобильных дорог с твердым покрытием общего пользования (в процентах);

$x_{37}$ – плотность автомобильных дорог общего

пользования с твердым покрытием (км путей на 1000 км$^2$ территории);

$x_{39}$ – среднегодовая номинальная начисленная заработная плата работников транспорта (рублей);

$x_{43}$ – затраты на технологические инновации (млн рублей);

$x_{58}$ – индексы тарифов на грузовые перевозки (железнодорожный транспорт) (процент).

Для организации ОИО по предложенным выше стратегиям при построении ЛНР с использованием эконометрического пакета Gretl была разработана специальная программа. Сначала с помощью неё решалась задача ОИО по второй стратегии. При этом были заданы следующие параметры поиска: число переменных $s=5$, число точек разбиения $k=10$, бинарная операция $\min$. В результате из 23256000 альтернативных вариантов регрессий на основе коэффициента детерминации $R^2$ выбрана лучшая модель:

$$\tilde{y} = \underset{(32,87)}{42,9648} + \underset{(24,54)}{0,0001403}\min\{x_{14}, 4382.8x_{22}\} - \underset{(-7,186)}{0,001292}\min\{x_{25}, 2.4967x_{39}\} + \underset{(5,378)}{0,001306}x_{24}, \quad (7)$$

для которой $R^2 = 0,97909$, что подтверждает её высокое качество. Под коэффициентами в скобках указаны значения t-критериев Стьюдента, указывающие на значимость всех переменных при уровне $\alpha = 0,01$.

Затем решалась задача ОИО с теми же параметрами поиска, но вместо бинарной операции $\min$ задавалась бинарная операция $\max$. Лучшей оказалась модель

$$\tilde{y} = \underset{(29,31)}{108,576} - \underset{(-12,98)}{0,713}\max\{x_{22}, 0.0026358x_{25}\} - \underset{(-6,784)}{0,3313}\max\{x_{35}, 0.029352x_{58}\} + \underset{(14)}{0,0023}x_{24}, \quad (8)$$

для которой $R^2 = 0,95827$. Как видно, эта модель оказалась незначительно хуже по величине коэффициента $R^2$, чем регрессия (7).

## V. Заключение

В работе рассмотрены линейно-неэлементарные регрессионные модели и алгоритм их приближенного МНК-оценивания. Предложены 2 стратегии построения ЛНР. В первой из них нет ограничений на число вхождений объясняющих переменных в ЛНР и на число бинарных операций. Во второй – ЛНР содержит заданное число переменных, наибольшее число бинарных операций, а каждая независимая переменная входит в неё только 1 раз. Установлено, что вычислительная сложность второй стратегии значительно ниже, чем первой. Построены высококачественные ЛНР грузовых железнодорожных перевозок в Иркутской области.

Предложенные в работе алгоритмы могут успешно применяться для моделирования объектов или процессов самой различной природы. Дальнейшие работы автора будут посвящены исследованию интерпретационных возможностей ЛНР.

## Библиография


[1] Westfall P.H., Arias A.L. Understanding regression analysis: a conditional distribution approach. Chapman and Hall/CRC, 2020. 514 p.

[2] Pardoe I. Applied regression modeling. Wiley, 2020. 336 p.

[3] Клейнер Г.Б. Производственные функции: теория, методы, применение. М.: Финансы и статистика, 1986. 239 с.

[4] Хацкевич Г.А., Проневич А.Ф., Чайковский М.В. Двухфакторные производственные функции с заданной предельной нормой замещения // Экономическая наука сегодня. 2019. № 10. С. 169-181.

[5] Базилевский М.П. Построение степенно-показательных регрессионных моделей и их интерпретация // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2020. № 4. С. 19-28.

[6] Базилевский М.П. Многофакторные модели полносвязной линейной регрессии без ограничений на соотношения дисперсий ошибок переменных // Информатика и ее применения. 2020. Т. 14. № 2. С. 92-97.

[7] Базилевский М.П., Носков С.И. Формализация задачи построения линейно-мультипликативной регрессии в виде задачи частично-булевого линейного программирования // Современные технологии. Системный анализ. Моделирование. 2017. № 3 (55). С. 101-105.

[8] Базилевский М.П., Носков С.И. Оценивание индексных моделей регрессии с помощью метода наименьших модулей // Вестник Российского нового университета. Серия: Сложные системы: модели, анализ и управление. 2020. № 1. С. 17-23.

[9] 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.

[10] 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.

[11] Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. Иркутск: РИЦ ГП «Облинформпечать», 1996. 321 с.

[12] Базилевский М.П. Оценивание линейно-неэлементарных регрессионных моделей с помощью метода наименьших квадратов // Моделирование, оптимизация и информационные технологии. 2020. № 8 (4). Доступно по: https://moitvivt.ru/ru/journal/pdf?id=872

[13] Носков С.И., Врублевский И.П. Анализ регрессионной модели грузооборота железнодорожного транспорта // Вестник транспорта Поволжья. 2020. № 1 (79). С. 86-90.

[14] Носков С.И., Врублевский И.П. Регрессионная модель динамики эксплуатационных показателей функционирования железнодорожного транспорта // Современные технологии. Системный анализ. Моделирование. 2016. № 2 (50). С. 192-197.

[15] Базилевский М.П., Врублевский И.П., Носков С.И., Яковчук И.С. Среднесрочное прогнозирование эксплуатационных показателей функционирования Красноярской железной дороги // Фундаментальные исследования. 2016. № 10-3. С. 471-476.


**Базилевский Михаил Павлович**, к.т.н., доцент кафедры математики Иркутского государственного университета путей сообщения, Иркутск, Россия; ORCID 0000-0002-3253-5697 (e-mail: mik2178@yandex.ru)

# Selection of Informative Operations in the Construction of Linear Non-elementary Regression Models

M. P. Bazilevskiy

*Abstract*— **This article is devoted to one of the main problems of regression analysis – the choice of regression model structural specification. The work is based on the linear non-elementary regressions proposed earlier by the author, which, in addition to explanatory variables, include binary operations of all their possible pairs. In such models, with an increase in the number of explanatory variables, the number of binary operations increases significantly. The aim of this work is to develop selection algorithms in linear non-elementary regressions of the most informative variables and operations. An algorithm for approximate estimation of linear non-elementary regressions using the ordinary least squares is considered. The problem of selection of informative operations is formulated. Two strategies for constructing linear non-elementary regressions are proposed. In the first of them there are no restrictions on the number of occurrences of explanatory variables in the model and on the number of binary operations. In the second, the model contains the largest number of binary operations, and each explanatory variable is included in it only once. Using combinatorics, the computational complexity of each of these strategies was determined. It turned out that the problem of constructing a linear non-elementary model based on the second strategy is solved in practice much faster than a similar problem based on the first strategy. The proposed algorithms were implemented using the Gretl package as a special program. With the help of it, high-quality linear non-elementary regression models of freight rail transportation in the Irkutsk region were built.**

***Keywords*— linear non-elementary regression, ordinary least squares, selection of informative operations, computational complexity, freight rail transportation in the Irkutsk region.**

REFERENCES

[1]Westfall P.H., Arias A.L. Understanding regression analysis: a conditional distribution approach. Chapman and Hall/CRC, 2020. 514p.
[2]Pardoe I. Applied regression modeling. Wiley, 2020. 336 p.
[3]Kleyner G.B. Proizvodstvennye funktsii: teoriya, metody, primenenie. Moscow: Finance and Statistics, 1986. 239 p.
[4]Khatskevich G.A., Pronevich A.F., Chaykovskiy M.V. Dvukhfaktornye proizvodstvennye funktsii s zadannoy predel'noy normoy zameshcheniya // Ekonomicheskaya nauka segodnya. 2019. No. 10. P. 169-181.
[5]Bazilevskiy M.P. Postroenie stepenno-pokazatel'nykh regressionnykh modeley i ikh interpretatsiya // Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: Sistemnyy analiz i informatsionnye tekhnologii. 2020. No. 4. P. 19-28.
[6]Bazilevskiy M.P. Mnogofaktornye modeli polnosvyaznoy lineynoy regressii bez ogranicheniy na sootnosheniya dispersiy oshibok peremennykh // Informatika i ee primeneniya. 2020. Vol. 14. No 2. P. 92-97.
[7]Bazilevskiy M.P., Noskov S.I. Formalizatsiya zadachi postroeniya lineyno-mul'tiplikativnoy regressii v vide zadachi chastichno-bulevogo lineynogo programmirovaniya // Sovremennye tekhnologii. Sistemnyy analiz. Modelirovanie. 2017. Vol. 55. No. 3. P. 101-105.
[8]Bazilevskiy M.P., Noskov S.I. Otsenivanie indeksnykh modeley regressii s pomoshch'yu metoda naimen'shikh moduley // Vestnik Rossiyskogo novogo universiteta. Seriya: Slozhnye sistemy: modeli, analiz i upravlenie. 2020. No. 1. P. 17-23.
[9]Ivanova N.K., Lebedeva S.A., Noskov S.I. Identifikatsiya parametrov nekotorykh negladkikh regressiy // Informatsionnye tekhnologii i problemy matematicheskogo modelirovaniya slozhnykh sistem. 2016. No. 17. P. 107-110.
[10] Noskov S.I., Khonyakov A.A. Programmnyy kompleks postroeniya nekotorykh tipov kusochno-lineynykh regressiy // Informatsionnye tekhnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami. 2019. Vol. 4. No. 3. P. 47-55.
[11] Noskov S.I. Tekhnologiya modelirovaniya ob''ektov s nestabil'nym funktsionirovaniem i neopredelennost'yu v dannykh. Irkutsk: RITs GP «Oblinformpechat'», 1996. 321 p.
[12] Bazilevskiy M.P. Otsenivanie lineyno-neelementarnykh regressionnykh modeley s pomoshch'yu metoda naimen'shikh kvadratov // Modelirovanie, optimizatsiya i informatsionnye tekhnologii. 2020. Vol. 4. No. 8. Available at: https://moitvivt.ru/ru/journal/pdf?id=872
[13] Noskov S.I., Vrublevskiy I.P. Analiz regressionnoy modeli gruzooborota zheleznodorozhnogo transporta // Vestnik transporta Povolzh'ya. 2020. Vol. 79. No. 1. P. 86-90.
[14] Noskov S.I., Vrublevskiy I.P. Regressionnaya model' dinamiki ekspluatatsionnykh pokazateley funktsionirovaniya zheleznodorozhnogo transporta // Sovremennye tekhnologii. Sistemnyy analiz. Modelirovanie. 2016. Vol. 50. No. 2. P. 192-197.
[15] Bazilevskiy M.P., Vrublevskiy I.P., Noskov S.I., Yakovchuk I.S. Srednesrochnoe prognozirovanie ekspluatatsionnykh pokazateley funktsionirovaniya Krasnoyarskoy zheleznoy dorogi // Fundamental'nye issledovaniya. 2016. No. 10-3. P. 471-476.

Bazilevskiy Mikhail Pavlovich, Ph.D., Associate Professor of the Department of Mathematics, Irkutsk State Transport University, Irkutsk, Russia; ORCID 0000-0002-3253-5697 (e-mail: mik2178@yandex.ru)