1 УДК 519.862.6

М. П. Базилевский [1]

1 Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация

ОТБОР ИНФОРМАТИВНЫХ РЕГРЕССОРОВ В ОЦЕНИВАЕМЫХ С ПОМОЩЬЮ МНК РЕГРЕССИОННЫХ МОДЕЛЯХ КАК ЗАДАЧА ЧАСТИЧНО-БУЛЕВОГО ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ: ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ

Аннотация. Статья посвящена проблеме отбора информативных регрессоров в регрессионных моделях, оцениваемых с помощью метода наименьших квадратов. Ранее эта проблема была сформулирована в виде задачи частично-булевого линейного программирования. С помощью пакета LPSolve проведено тестирование сформулированной задачи. Помимо этого тестировалась задача отбора информативных регрессоров с ограничением на степень мультиколлинеарности. Результаты тестирования полностью совпали с решениями, полученными методом полного перебора регрессий, что подтверждает корректность разработанного математического аппарата.

Ключевые слова: регрессионная модель, отбор информативных регрессоров, частично-булево линейное программирование, метод наименьших квадратов, мультиколлинеарность.

M.P. Bazilevskiy [1]

[1] *Irkutsk State Transport University, Irkutsk, Russia*

FEATURE SELECTION IN REGRESSION MODELS ESTIMATED BY OLS AS A PARTIAL BOOLEAN LINEAR PROGRAMMING PROBLEM: COMPUTATIONAL EXPERIMENTS

Abstract. This article is devoted to the problem of feature selection in regression models estimated using the least squares method. Earlier this problem was formulated as a partial Boolean linear programming problem. The LPSolve package was used to test the formulated problem. In addition, the problem of feature selection with a constraint on the degree of multicollinearity was tested. The test results completely coincided with the solutions obtained by the method of full enumeration of regressions, which confirms the correctness of the developed mathematical apparatus.

Keywords: regression model, feature selection, partial Boolean linear programming, ordinary least squares, multicollinearity.

Введение. Аппарат математического программирования находит широкое применение в регрессионном анализе. Например, известными методами оценивания параметров регрессионной модели являются метод наименьших модулей (МНМ) и антиробастного оценивания (МАО), основанные на решении задач линейного программирования (ЛП) [1]. Существуют и другие методы. Так, работы [2,3] посвящены методу множественного оценивания регрессионных моделей, заключающемуся в одновременной минимизации ошибок сразу по двум критериям – МНМ и МАО. В [4,5] можно найти описание метода смешанного оценивания (МСО), состоящего в отыскании вектора оценок регрессии минимизацией суммы функций потерь для МНМ и МАО на разных участках выборки. В [6] описан метод построения регрессионной модели по интервальной информации. Все эти методы представляют собой задачи ЛП.

Часто в регрессионном анализе применяется аппарат частично булевого линейного программирования (ЧБЛП) [7]. Например, в работе [8] описан критерий "согласованности" поведения, который применен для корректировки МНМ-оценок регрессионной модели. В [9] предложно обобщение этого критерия. В [10] критерий "согласованности" поведения использован для выделения из множественных оценок единственного вектора параметров.

Статья [11] посвящена оцениванию индексных моделей регрессии с помощью МНМ. Все эти задачи представляют собой задачи ЧБЛП.

Аппарат ЧБЛП в регрессионном анализе в основном применяется для решения задачи отбора информативных регрессоров (ОИР) [1]. Так, в [1,7] можно найти формализацию задачи ОИР для МНМ в виде ЧБЛП. В [12,13] задача построения линейно-мультипликативных регрессий с помощью МНМ сведена к задаче ЧБЛП. В [14] рассмотрен групповой ОИР. В [15] рассмотрена задача ОИР с одновременной корректировкой МНМ-оценок регрессии по критерию "согласованности" поведения.

В работе [16] рассмотрена задача ОИР с фиксированным числом регрессоров для метода наименьших квадратов (МНК). Она сформулирована в виде задачи частично булевого квадратичного программирования (ЧБКП). В [17] отбор оптимального числа регрессоров осуществляется по критерию остаточной дисперсии, в [18,19] – по скорректированному критерию детерминации, критерию Акаике, Шварца и Мэллоуза. Формулировку задач ЧБКП для линейной регрессии можно также найти в работах [20–22].

В статье [23] задача ОИР для известного числа регрессоров при оценивании линейной регрессии с помощью МНК сведена автором к задаче ЧБЛП. В [24] эта задача расширена линейными ограничениями на степень мультиколлинеарности. В [25] задача ОИР для неизвестного числа регрессоров при оценивании линейной регрессии по скорректированному критерию детерминации сведена к задаче частично целочисленного линейного программирования.

Целью данной работы является тестирование предложенного автором в работах [23] и [24] математического аппарата.

Задача ОИР в регрессионной модели, оцениваемой с помощью МНК, как задача ЧБЛП

Рассмотрим модель множественной линейной регрессии:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + ... + \alpha_l x_{il} + \varepsilon_i, \qquad i = \overline{1,n}, \tag{1}$$

где y_i, $i = \overline{1,n}$ – значения зависимой (объясняемой) переменной y;

x_{i1}, x_{i2}, …, x_{il}, $i = \overline{1,n}$ – значения l независимых (объясняющих) переменных (регрессоров) x_1, x_2, …, x_l;

ε_i, $i = \overline{1,n}$ – ошибки аппроксимации;

α_0, α_1, …, α_l – неизвестные параметры;

n – объем выборки.

Задача ОИР состоит в том, чтобы выбрать для включения в линейную модель из l объясняющих переменных m наиболее информативных по некоторому критерию качества. Пусть в качестве такого критерия используется сумма квадратов ошибок, т.е. регрессия оценивается с помощью МНК. Такая задача формализована в работе [23] в виде следующей задачи ЧБЛП:

$$R^2(\beta_1, \beta_2, ..., \beta_l) = \sum_{j=1}^{l} R_{yx}^{(j,1)} \cdot \beta_j \to \max, \tag{2}$$

$$-(1-\delta_j)M \le \sum_{k=1}^{l} R_{xx}^{(j,k)} \cdot \beta_k - R_{yx}^{(j,1)} \le (1-\delta_j)M, \qquad j = \overline{1,l}, \tag{3}$$

$$-\delta_j M \le \beta_j \le \delta_j M, \qquad j = \overline{1,l}, \tag{4}$$

$$\delta_j \in \{0,1\}, \qquad j = \overline{1,l}, \tag{5}$$

$$\sum_{j=1}^{l} \delta_j = m, \tag{6}$$

где β_1, β_2, ..., β_l – параметры (бета-коэффициенты) стандартизованной линейной регрессии

$$v_i = \beta_1 z_{i1} + \beta_2 z_{i2} + ... + \beta_l z_{il} + u_i, \qquad i = \overline{1,n},$$

которая строится на основе нормирования всех переменных; R^2 – коэффициент детерминации; $R_{yx}^{(j,1)}$ – элементы вектора коэффициентов парной корреляции между объясняемой переменной y и объясняющими переменными x_1, x_2, ..., x_l; $R_{xx}^{(j,k)}$ – элементы матрицы коэффициентов парной корреляции между объясняющими переменными; M – большое положительное число; δ_j, $j = \overline{1,l}$ – булевы переменные, заданные по правилу:

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

Для перехода от бета-коэффициентов к традиционным МНК-оценкам линейной регрессии (1) необходимо воспользоваться формулами:

$$\tilde{\alpha}_i = \tilde{\beta}_i \frac{\sigma_y}{\sigma_{x_i}}, \quad i = \overline{1,l}; \qquad \tilde{\alpha}_0 = \overline{y} - \tilde{\alpha}_1 \overline{x_1} - \tilde{\alpha}_2 \overline{x_2} - ... - \tilde{\alpha}_l \overline{x_l}.$$

В работе [24] показано, каким образом в задаче (2) – (6) можно контролировать степень мультиколлинеарности. Это можно сделать, в частности, введением в эту задачу следующих линейных ограничений:

$$-\left(1-\delta_{q_{ki}}\right)M \le \sum_{j=1}^{l-1} R_{xx}^{(q_{ki},q_{kj})}\beta_{kj} - R_{xx}^{(q_{ki},k)} \le \left(1-\delta_{q_{ki}}\right)M, \qquad k = \overline{1,l}, \ i = \overline{1,l-1}, \quad (7)$$

$$-\delta_{q_{ki}} M \le \beta_{ki} \le \delta_{q_{ki}} M, \ k = \overline{1,l}, \qquad i = \overline{1,l-1}, \quad (8)$$

$$\sum_{j=1}^{l-1} R_{xx}^{(q_{kj},k)} \cdot \beta_{kj} - \left(1-\delta_k\right)M \le r, \qquad k = \overline{1,l}, \quad (9)$$

где β_{kj}, $k = \overline{1,l}$, $j = \overline{1,l-1}$ – бета-коэффициенты вспомогательных регрессий

$$z_{i1} = \beta_{11} z_{i2} + \beta_{12} z_{i3} + ... + \beta_{1,l-1} z_{il} + u_{i1},$$
$$z_{i2} = \beta_{21} z_{i1} + \beta_{22} z_{i3} + ... + \beta_{2,l-1} z_{il} + u_{i2},$$
$$\ldots$$
$$z_{il} = \beta_{l,1} z_{i1} + \beta_{l,2} z_{i2} + ... + \beta_{l,l-1} z_{i,l-1} + u_{il};$$

q_{ij} – элементы матрицы $Q_{l\times(l-1)}$, полученной путем вычеркивания главной диагонали из матрицы $\begin{pmatrix} 1 & 2 & ... & l \\ 1 & 2 & ... & l \\ ... & ... & ... & ... \\ 1 & 2 & ... & l \end{pmatrix}_{l\times l}$; r – задаваемое исследователем ограничение на значения коэффициентов детерминации вспомогательных регрессий.

Вычислительные эксперименты

Вычислительные эксперименты проводились на персональном компьютере с 4-х ядерным процессором Intel Core i5-4670 с тактовой частотой 3400 МГц и объемом оперативной памяти 8 Гб. Для этого использовались встроенные в эконометрический пакет Gretl статистические данные (data7-12.gdt) о ценах и характеристиках двухдверных седанов и хетчбэков американской автомобильной промышленности за 1995 год. Объем выборки – 82. В качестве зависимой переменной выбрана переменная *price*, а в качестве независимых выступают *wbase*, *length*, *width*, *height*, *weight*, *liters* и *gasmpg*. Для удобства будем обозначать их далее, как y, x_1, x_2, x_3, x_4, x_5, x_6 и x_7. С помощью элементарных преобразований x^2, ln(x) и $\sqrt{x}$ были сформированы дополнительные переменные: $x_8 = x_1^2$, $x_9 = x_2^2$, $x_{10} = x_3^2$, $x_{11} = x_4^2$,

$x_{12} = x_5^2$, $x_{13} = x_6^2$, $x_{14} = x_7^2$, $x_{15} = \ln x_1$, $x_{16} = \ln x_2$, $x_{17} = \ln x_3$, $x_{18} = \ln x_4$, $x_{19} = \ln x_5$, $x_{20} = \ln x_6$, $x_{21} = \ln x_7$, $x_{22} = \sqrt{x_1}$, $x_{23} = \sqrt{x_2}$, $x_{24} = \sqrt{x_3}$, $x_{25} = \sqrt{x_4}$, $x_{26} = \sqrt{x_5}$, $x_{27} = \sqrt{x_6}$, $x_{28} = \sqrt{x_7}$.

Сначала по этим данным проводилось тестирование задачи ОИР (2) – (6). Из 28 переменных осуществлялся отбор при $m = 1$, $m = 2$, $m = 3$, $m = 4$ и $m = 5$. Для этого в пакете LPSolve была разработана программа, содержащая 113 основных ограничений и 56 переменных. Фрагменты этой программы представлены на рис. 1 (а), 1 (б), 1 (в) и 1 (г).

```
/* Objective function */
max: +0.185345009916 b1 +0.398314675136 b2 +0.477576919

/* Constraints */
+b1 +0.749546923377 b2 +0.39239054017 b3 +0.41230412508
+b1 +0.749546923377 b2 +0.39239054017 b3 +0.41230412508
+0.749546923377 b1 +b2 +0.578506065213 b3 +0.2743391520
+0.749546923377 b1 +b2 +0.578506065213 b3 +0.2743391520
+0.39239054017 b1 +0.578506065213 b2 +b3 -0.04076377199
+0.39239054017 b1 +0.578506065213 b2 +b3 -0.04076377199
+0.412304125089 b1 +0.274339152077 b2 -0.0407637719949
+0.412304125089 b1 +0.274339152077 b2 -0.0407637719949
```

(а)

```
+0.337905949496 b1 +0.707213613537 b2 +0.654074388033 b
+0.337905949496 b1 +0.707213613537 b2 +0.654074388033 b
-0.413415213946 b1 -0.776194880562 b2 -0.569242088974 b
-0.413415213946 b1 -0.776194880562 b2 -0.569242088974 b
+b1 -1000 d1 <= 0;
+b1 +1000 d1 >= 0;
+b2 -1000 d2 <= 0;
+b2 +1000 d2 >= 0;
+b3 -1000 d3 <= 0;
+b3 +1000 d3 >= 0;
+b4 -1000 d4 <= 0;
+b4 +1000 d4 >= 0;
```

(б)

```
+b26 -1000 d26 <= 0;
+b26 +1000 d26 >= 0;
+b27 -1000 d27 <= 0;
+b27 +1000 d27 >= 0;
+b28 -1000 d28 <= 0;
+b28 +1000 d28 >= 0;
+d1 +d2 +d3 +d4 +d5 +d6 +d7 +d8 +d9 +d10 +d11 +d12 +d13

/* Variable bounds */
b1 >= -Inf;
b2 >= -Inf;
b3 >= -Inf;
```

(в)

5

```
168 d21 <= 1;
169 d22 <= 1;
170 d23 <= 1;
171 d24 <= 1;
172 d25 <= 1;
173 d26 <= 1;
174 d27 <= 1;
175 d28 <= 1;
176
177 /* Integer definitions */
178 int d1,d2,d3,d4,d5,d6,d7,d8,d9,d10,d11,d12,d13,d14,d15,
179
```

(г)

Рис. 1. Фрагменты программы

Большое число M задавалось равным 1000. Для каждого m фиксировалось время и ход решения задачи. Результаты тестирования представлены в таблице 1.

Таблица 1

Вычислительный эксперимент № 1

Итерация	Переменные	R^2	Время, с
m=1			
1	x_5	0,430504	0,022
2	x_{12}	0,457473	
m=2			
1	x_3, x_{12}	0,457502	0,13
2	x_4, x_5	0,521219	
3	x_4, x_{12}	0,551162	
4	x_5, x_{16}	0,562752	
5	x_5, x_{23}	0,563087	
6	x_2, x_{12}	0,593074	
7	x_9, x_{12}	0,598261	
m=3			
1	x_3, x_4, x_5	0,521253	1,145
2	x_3, x_5, x_{18}	0,529720	
3	x_3, x_5, x_{16}	0,562823	
4	x_3, x_5, x_{23}	0,563204	
5	x_2, x_3, x_{12}	0,593538	
6	x_3, x_9, x_{12}	0,598657	
7	x_4, x_5, x_{18}	0,737746	
8	x_4, x_{12}, x_{18}	0,753629	
9	x_{12}, x_{18}, x_{25}	0,754003	
m=4			
1	x_3, x_4, x_5, x_{16}	0,586392	6,974
2	x_3, x_4, x_5, x_{12}	0,588594	
3	x_3, x_4, x_5, x_{18}	0,738374	
4	x_3, x_4, x_{12}, x_{18}	0,753629	
5	x_3, x_{12}, x_{18}, x_{25}	0,754003	
6	x_2, x_4, x_{12}, x_{18}	0,765441	
7	x_4, x_9, x_{12}, x_{18}	0,765629	
8	x_2, x_{12}, x_{18}, x_{25}	0,765960	
9	x_9, x_{12}, x_{18}, x_{25}	0,766139	

m=5			
1	x_2, x_3, x_4, x_5, x_6	0,593877	36,98
2	$x_2, x_3, x_4, x_5, x_{13}$	0,600862	
3	$x_2, x_3, x_4, x_5, x_{12}$	0,628993	
4	$x_2, x_3, x_4, x_5, x_{18}$	0,748241	
5	$x_2, x_3, x_4, x_{12}, x_{18}$	0,765471	
6	$x_2, x_3, x_{12}, x_{18}, x_{25}$	0,765986	
7	$x_2, x_4, x_5, x_{12}, x_{18}$	0,770847	
8	$x_2, x_4, x_{11}, x_{12}, x_{18}$	0,775012	
9	$x_2, x_4, x_{11}, x_{12}, x_{25}$	0,775035	
10	$x_4, x_5, x_7, x_{12}, x_{18}$	0,775717	
11	$x_4, x_5, x_{12}, x_{18}, x_{19}$	0,798992	
12	$x_4, x_5, x_{12}, x_{18}, x_{26}$	0,800438	
13	$x_5, x_{12}, x_{18}, x_{25}, x_{26}$	0,800963	

Затем эти задачи были решены для M=100. Результаты оказались теми же, а время решения при m=1 составило 0,024 с, при m=2 – 0,141 с, при m=3 – 1,124 с, при m=4 – 6,657 с, при m=5 – 33,886 с. Как видно, уменьшение величины M приводит к снижению времени решения задачи. Однако при малых M полученное решение может оказаться не оптимальным. Выбор числа M пока остается нерешенной проблемой.

После чего все эти задачи были решены методом простого перебора. Для m=1 потребовалось перебрать 28 моделей, для m=2 – 378 моделей, для m=3 – 3276 моделей, для m=4 – 20475 моделей, для m=5 – 98280 моделей. Полученные решения полностью совпали с результатами, приведенными в таблице 1, что подтверждает корректность математического аппарата (2) – (6).

Далее проводилось тестирование задачи ОИР (2) – (9). Из 28 переменных осуществлялся отбор $m = 3$. Для этого в пакете LPSolve была разработана программа, содержащая 3165 основных ограничений и 812 переменных. Большое число M задавалось равным 1000. Задача решалась при r равном 1, 0.9, 0.8, ..., 0.1. Для каждого r фиксировалось время и ход решения задачи. Результаты тестирования представлены в таблице 2.

Таблица 2

Вычислительный эксперимент № 2

Итерация	Переменные	R^2	Время, с
r=1			
Результат тот же, что в табл. 1 при m=3			77,086
r=0,9			
1	x_3, x_4, x_5	0,521253	93,8
2	x_3, x_5, x_{18}	0,52972	
3	x_3, x_5, x_{16}	0,562823	
4	x_3, x_5, x_{23}	0,563204	
5	x_2, x_3, x_{12}	0,593538	
6	x_3, x_9, x_{12}	0,598657	
7	x_4, x_{12}, x_{16}	0,615097	
8	x_2, x_4, x_{12}	0,617789	
9	x_4, x_9, x_{12}	0,619255	
10	x_2, x_{12}, x_{18}	0,621703	
11	x_9, x_{12}, x_{18}	0,622912	
r=0,8			
1	x_3, x_4, x_5	0,521253	94,3

2	x_3, x_5, x_{18}	0,52972	
3	x_3, x_4, x_{12}	0,551999	
4	x_3, x_{12}, x_{18}	0,560338	
5	x_3, x_{12}, x_{16}	0,586076	
6	x_4, x_{12}, x_{16}	0,615097	
7	x_2, x_4, x_{12}	0,617789	
8	x_2, x_{12}, x_{18}	0,621703	
r=0,7			
1	x_3, x_4, x_5	0,521253	111,356
2	x_3, x_5, x_{18}	0,52972	
3	x_3, x_4, x_{12}	0,551998	
4	x_3, x_{12}, x_{18}	0,560338	
5	x_4, x_8, x_{12}	0,568242	
6	x_4, x_{12}, x_{15}	0,570857	
7	x_1, x_{12}, x_{18}	0,575363	
8	x_{12}, x_{15}, x_{18}	0,576582	
r=0,6			
Результат тот же, что при r=0,7			112,241
r=0,5			
Результат тот же, что при r=0,7			109,99
r=0,4			
1	x_3, x_{18}, x_{21}	0,438976	180,78
2	x_4, x_6, x_8	0,448644	
3	x_4, x_7, x_{13}	0,470268	
4	x_7, x_{13}, x_{18}	0,4748	
r=0,3			
1	x_3, x_{14}, x_{18}	0,379830	171,735
2	x_4, x_6, x_8	0,448644	
3	x_4, x_8, x_{13}	0,449929	
4	x_5, x_7, x_{26}	0,488894	
r=0,2			
1	x_5, x_7, x_{26}	0,488894	207,39
r=0,1			
1	x_5, x_7, x_{26}	0,488894	198,59

При решении задач для r=0,3, r=0,2 и r=0,1 LPSolve выдал предупреждение "unacceptable accuracy found" (обнаружена неприемлемая точность).

Затем все эти задачи были решены методом простого перебора. Полученные решения для r=1, r=0,9, r=0,8, r=0,7, r=0,6, r=0,5, r=0,4 полностью совпали с результатами, приведенными в таблице 2. Но при r=0,3 был получен набор переменных x_8, x_{13}, x_{18} с коэффициентом детерминации 0,455839, а при r=0,2 и r=0,1 задача решений не имеет. Была выдвинута гипотеза, что это несоответствие связано с неверным выбором числа M. Увеличение числа M до 3000 позволило зафиксировать верные решения при r=0,3, r=0,2 и r=0,1 за 197,29 с, 399,22 с и 385,498 с. Таким образом, подтверждена корректность математического аппарата (2) – (9).

Заключение. В данной работе проведено тестирование сформулированных ранее задач ЧБЛП для ОИР в регрессионных моделях, оцениваемых с помощью МНК. Все найденные решения совпали с результатами, полученными методом полного перебора. Установлено, что время решения задачи ОИР (2) – (6) возрастает с увеличением числа регрессоров m, а время

решения задачи ОИР (2) – (9) при фиксированном m возрастает с уменьшением параметра r. Нерешенной проблемой остается неопределенность при назначении величины М.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. – Иркутск: Облинформпечать, 1996. – 321 с.
2. Носков С.И., Баенхаева А.В. Множественное оценивание параметров линейного регрессионного уравнения // Современные технологии. Системный анализ. Моделирование. – 2016 – № 3 (51). – С. 133-138.
3. Баенхаева А.В., Базилевский М.П., Носков С.И. Моделирование валового регионального продукта Иркутской области на основе применения методики множественного оценивания регрессионных параметров // Фундаментальные исследования. – 2016. – № 10-1. – С. 9–14.
4. Носков С.И., Перфильева К.С. Эмпирический анализ некоторых свойств метода смешанного оценивания параметров линейного регрессионного уравнения // Наука и бизнес: пути развития. – 2020. – № 6 (108). – С. 62-66.
5. Носков С.И. О методе смешанного оценивания параметров линейной регрессии // Информационные технологии и математическое моделирование в управлении сложными системами. – 2019. – № 1 (2). – С. 41-45.
6. Носков С.И., Врублевский И.П., Заянчуковская В.О. Применение интервального регрессионного анализа для моделирования объектов транспорта // Вестник Уральского государственного университета путей сообщения. – 2020. – № 3 (47). – С. 45-52.
7. Носков С.И., Базилевский М.П. Построение регрессионных моделей с использованием аппарата линейно-булевого программирования. – Иркутск: ИрГУПС, 2018. – 176 с.
8. Носков С.И. Критерий "согласованность поведения" в регрессионном анализе // Современные технологии. Системный анализ. Моделирование. – 2013 – № 1 (37). – С. 107-110.
9. Носков С.И. Обобщенный критерий согласованности поведения в регрессионном анализе // Информационные технологии и математическое моделирование в управлении сложными системами. – 2018. – № 1 (1). – С. 14-20.
10. Носков С.И., Базилевский М.П. Множественное оценивание параметров и критерий согласованности поведения в регрессионном анализе // Вестник Иркутского государственного технического университета. – 2018. – Т. 22. – № 4 (135). – С. 101-110.
11. 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.
12. 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.
13. Базилевский М.П. Программный комплекс построения линейно-мультипликативных регрессий // Прикладная информатика. – 2018. – Т. 13. – № 3 (75). – С. 110-123.
14. Базилевский М.П., Вергасов А.С., Носков С.И. Групповой отбор информативных переменных в регрессионных моделях // Южно-Сибирский научный вестник. – 2019. – № 4-1 (28). – С. 36-39.
15. Базилевский М.П., Носков С.И. Программный комплекс построения линейной регрессионной модели с учетом критерия согласованности поведения фактической и расчетной траекторий изменения значений объясняемой переменной // Вестник Иркутского государственного технического университета. – 2017. – Т. 21. – № 9 (128). – С. 37-44.

16. Konno H., Yamamoto R. Choosing the best set of variables in regression analysis using integer programming // Journal of Global Optimization. – 2009. – Vol. 44. – P. 272–282.

17. Park Y.W., Klabjan D. Subset selection for multiple linear regression via optimization // Journal of Global Optimization. – 2020. – Vol. 77. – P. 543–574.

18. Miyashiro R., Takano Y. Mixed integer second-order cone programming formulations for variable selection in linear regression // European Journal of Operational Research. – 2015. – Vol. 247. – P. 721–731.

19. Miyashiro R., Takano Y. Subset selection by Mallows' C_p: a mixed integer programming approach // Expert Systems with Applications. – 2015. – Vol. 42. – P. 325–331.

20. Bertsimas D., King A., Mazumder R. Best subset selection via a modern optimizations lens // The Annals of Statistics. – 2016. – Vol. 44. – P. 813–852.

21. Bertsimas D., King A. OR forum – An algorithmic approach to linear regression // Operations Research. – 2016. – Vol. 64. – P. 2–16.

22. Konno H., Takaya Y. Multi-step methods for choosing the best set of variables in regression analysis // Computational Optimization and Applications. – 2010. – Vol. 46. – P. 417–426.

23. 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.

24. Базилевский М.П. Отбор информативных регрессоров с учетом мультиколлинеарности между ними в регрессионных моделях как задача частично-булевого линейного программирования // Моделирование, оптимизация и информационные технологии. – 2018. – Т. 6. – № 2 (21). – С. 104-118.

25. Базилевский М.П. Отбор оптимального числа информативных регрессоров по скорректированному коэффициенту детерминации в регрессионных моделях как задача частично целочисленного линейного программирования // Прикладная математика и вопросы управления. – 2020. – № 2. – С. 41-54.

REFERENCES

1. Noskov S.I. *Tehnologija modelirovanija ob'ektov s nestabil'nym funkcionirovaniem i neopredelennost'ju v dannyh* [Modeling technology for objects with unstable operation and data uncertainty]. Irkutsk, RIC GP «Oblinformpechat'» Publ., 1996. 321 p.

2. Noskov S.I., Baenhaeva A.V. *Mnozhestvennoe ocenivanie parametrov linejnogo regressionnogo uravnenija* [Multiple Estimation of Linear Regression Equation Parameters]. *Sovremennye tehnologii. Sistemnyj analiz. Modelirovanie* [Modern technologies. System analysis. Modeling]. 2016, no. 3, vol. 51, pp. 133–138.

3. Baenhaeva A.V., Bazilevskiy M.P., Noskov S.I. *Modelirovanie valovogo regional'nogo produkta Irkutskoy oblasti na osnove primeneniya metodiki mnozhestvennogo otsenivaniya regressionnykh parametrov* [Modeling of gross regional product Irkutsk region of the basis of methods of multiple estimation of regression parameters]. *Fundamental'nye issledovaniya* [Fundamental research]. 2016, no. 10-1, pp. 9–14.

4. Noskov S.I., Perfil'eva K.S. *Jempiricheskij analiz nekotoryh svojstv metoda smeshannogo ocenivanija parametrov linejnogo regressionnogo uravnenija* [An empirical analysis of some properties of the method of mixed estimation of parameters of a linear regression equation]. *Nauka i biznes: puti razvitija* [Science and business: ways of development]. 2020, no. 6, vol. 108, pp. 62–66.

5. Noskov S.I. *O metode smeshannogo ocenivanija parametrov linejnoj regressii* [On the method of mixed estimation of linear regression parameters]. *Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami* [Information technology and mathematical modeling in the management of complex systems]. 2019, no. 1, vol. 2, pp. 41-45.

6. Noskov S.I., Vrublevskij I.P., Zajanchukovskaja V.O. *Primenenie interval'nogo regressionnogo analiza dlja modelirovanija ob'ektov transporta* [Using interval regression analysis for modeling transport objects]. *Vestnik Ural'skogo gosudarstvennogo universiteta putej soobshhenija* [Bulletin of the Ural State Transport University]. 2020, no. 3, vol. 47, pp. 45-52.

7. Noskov S.I., Bazilevskij M.P. *Postroenie regressionnyh modelej s ispol'zovaniem apparata linejno-bulevogo programmirovanija* [Building regression models using the linear-boolean programming apparatus]. Irkutsk, IrGUPS, 2018. 176 p.

8. Noskov S.I. *Kriterij "soglasovannost' povedenija" v regressionnom analize* [Criterion "consistency of behavior" in regression analysis]. *Sovremennye tehnologii. Sistemnyj analiz. Modelirovanie* [Modern technologies. System analysis. Modeling]. 2013, no. 1, vol. 37, pp. 107-110.

9. Noskov S.I. *Obobshhennyj kriterij soglasovannosti povedenija v regressionnom analize* [Generalized criterion for the consistency of behavior in regression analysis]. *Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami* [Information technology and mathematical modeling in the management of complex systems]. 2018, no. 1, vol. 1, pp. 14-20.

10. Noskov S.I., Bazilevskij M.P. *Mnozhestvennoe ocenivanie parametrov i kriterij soglasovannosti povedenija v regressionnom analize* [Multiple Parameter Estimation and Behavior Consistency Criterion in Regression Analysis]. *Vestnik Irkutskogo gosudarstvennogo tehnicheskogo universiteta* [Irkutsk State Technical University Bulletin]. 2018, no. 4, vol. 135, pp. 101-110.

11. Bazilevskij M.P., Noskov S.I. *Ocenivanie indeksnyh modelej regressii s pomoshh'ju metoda naimen'shih modulej* [Estimating Index Regression Models Using Least Modules]. *Vestnik Rossijskogo novogo universiteta. Serija: Slozhnye sistemy: modeli, analiz i upravlenie* [Bulletin of the Russian New University. Series: Complex Systems: Models, Analysis and Management]. 2020, no. 1, pp. 17-23.

12. Bazilevskij M.P., Noskov S.I. *Formalizacija zadachi postroenija linejno-mul'tiplikativnoj regressii v vide zadachi chastichno-bulevogo linejnogo programmirovanija* [Formalization of the problem of constructing linear multiplicative regression as a partial boolean linear programming problem]. *Sovremennye tehnologii. Sistemnyj analiz. Modelirovanie* [Modern technologies. System analysis. Modeling]. 2017, no. 3, vol. 55, pp. 101-105.

13. Bazilevskij M.P. *Programmnyj kompleks postroenija linejno-mul'tiplikativnyh regressij* [A software package for constructing linear multiplicative regressions]. *Prikladnaja informatika* [Applied Informatics]. 2018, no. 3, vol. 75, pp. 110-123.

14. Bazilevskij M.P., Vergasov A.S., Noskov S.I. *Gruppovoj otbor informativnyh peremennyh v regressionnyh modeljah* [Group selection of informative variables in regression models]. *Juzhno-Sibirskij nauchnyj vestnik* [South Siberian Scientific Bulletin]. 2019, no. 4-1, vol. 28, pp. 36-39.

15. Bazilevskij M.P., Noskov S.I. *Programmnyj kompleks postroenija linejnoj regressionnoj modeli s uchetom kriterija soglasovannosti povedenija faktycheskoj i raschetnoj traektorij izmenenija znachenij ob'jasnjaemoj peremennoj* [A software package for constructing a linear regression model taking into account the criterion of consistency of the behavior of the actual and calculated trajectories of change in the values of the explained variable]. *Vestnik Irkutskogo gosudarstvennogo tehnicheskogo universiteta* [Irkutsk State Technical University Bulletin]. 2017, no. 9, vol. 128, pp. 37-44.

16. Konno H., Yamamoto R. Choosing the best set of variables in regression analysis using integer programming. Journal of Global Optimization. 2009, vol. 44, pp. 272–282.

17. Park Y.W., Klabjan D. Subset selection for multiple linear regression via optimization. Journal of Global Optimization. 2020, vol. 77, pp. 543–574.

18. Miyashiro R., Takano Y. Mixed integer second-order cone programming formulations for variable selection in linear regression. European Journal of Operational Research. 2015, vol. 247, pp. 721–731.

19. Miyashiro R., Takano Y. Subset selection by Mallows' C_p: a mixed integer programming approach. Expert Systems with Applications. 2015, vol. 42, pp. 325–331.

20. Bertsimas D., King A., Mazumder R. Best subset selection via a modern optimizations lens. The Annals of Statistics. 2016, vol. 44, pp. 813–852.

21. Bertsimas D., King A. OR forum – An algorithmic approach to linear regression. Operations Research. 2016, vol. 64, pp. 2–16.

22. Konno H., Takaya Y. Multi-step methods for choosing the best set of variables in regression analysis. Computational Optimization and Applications. 2010, vol. 46, pp. 417–426.

23. Bazilevskij M.P. *Svedenie zadachi otbora informativnyh regressorov pri ocenivanii linejnoj regressionnoj modeli po metodu naimen'shih kvadratov k zadache chastichno-bulevogo linejnogo programmirovanija* [Reduction of the problem of selection of informative regressors when estimating a linear regression model using the least squares method to a partial boolean linear programming problem]. *Modelirovanie, optimizacija i informacionnye tehnologii* [Modeling, optimization and information technology]. 2018, no. 1, vol. 20, pp. 108-117.

24. Bazilevskij M.P. *Otbor informativnyh regressorov s uchetom mul'tikollinearnosti mezhdu nimi v regressionnyh modeljah kak zadacha chastichno-bulevogo linejnogo programmirovanija* [Selection of informative regressors taking into account the multicollinearity between them in regression models as a partial Boolean linear programming problem]. *Modelirovanie, optimizacija i informacionnye tehnologii* [Modeling, optimization and information technology]. 2018, no. 2, vol. 21, pp. 104-118.

25. Bazilevskij M.P. *Otbor optimal'nogo chisla informativnyh regressorov po skorrektirovannomu kojefficientu determinacii v regressionnyh modeljah kak zadacha chastichno celochislennogo linejnogo programmirovanija* [Selection of the optimal number of informative regressors by the corrected coefficient of determination in regression models as a partial integer linear programming problem]. *Prikladnaja matematika i voprosy upravlenija* [Applied Mathematics and Management Issues]. 2020, no. 2, pp. 41-54.

Информация об авторах

Михаил Павлович Базилевский – к. т. н., доцент, доцент кафедры «Математика», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: mik2178@yandex.ru

Authors

Mikhail Pavlovich Bazilevskiy – Ph. D. in Engineering Science, Associate Professor, the Subdepartment of Mathematics, Irkutsk State Transport University, Irkutsk, e-mail: mik2178@yandex.ru

Для цитирования

Базилевский М.П. Отбор информативных регрессоров в оцениваемых с помощью МНК регрессионных моделях как задача частично-булевого линейного программирования: вычислительные эксперименты // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2021. – №2(10). – С. 1-12 – DOI: 10.26731/2658-3704.2021.2(10).1-12 – Режим доступа: https://ismm.irgups.ru/toma/210-2021, свободный.. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 29.04.2021)

For citations

Bazilevskiy M.P. Feature selection in regression models estimated by OLS as a partial boolean linear programming problem: computational experiments // Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal [Information technology and mathematical modeling in the management of complex systems:

electronic scientific journal], 2021. No. 2. P.1-12. DOI: 10.26731/2658-3704.2021.2(10).1-12 [Accessed 29/04/21]