DOI: 10.15593/2499-9873/2020.2.03
УДК 519.862.6

М.П. Базилевский

Иркутский государственный университет путей сообщения, Иркутск, Россия

ОТБОР ОПТИМАЛЬНОГО ЧИСЛА ИНФОРМАТИВНЫХ РЕГРЕССОРОВ ПО СКОРРЕКТИРОВАННОМУ КОЭФФИЦИЕНТУ ДЕТЕРМИНАЦИИ В РЕГРЕССИОННЫХ МОДЕЛЯХ КАК ЗАДАЧА ЧАСТИЧНО ЦЕЛОЧИСЛЕННОГО ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

При построении регрессионной модели первоочередной проблемой, с которой сталкивается исследователь, является то, что непонятно, каким именно должно быть уравнение связи между объясняемой и объясняющими переменными. Этот начальный этап построения называется выбором структурной спецификации модели. При выборе спецификации регрессии параллельно возникает вопрос о том, какие именно объясняющие переменные должны быть включены в уравнение. Эта проблема называется задачей отбора информативных регрессоров. Ее суть состоит в том, чтобы выделить из множества «кандидатов» на включение подмножества наиболее информативных из них на основе некоторого критерия качества.

Посвящена проблеме отбора информативных регрессоров в регрессионных моделях, оцениваемых с помощью метода наименьших квадратов. Рассмотрен предложенный ранее подход к отбору заданного числа информативных регрессоров, основанный на задаче частично булевого линейного программирования. Неизвестными параметрами в этой задаче выступают бета-коэффициенты стандартизованной регрессии, а также булевы переменные, отвечающие за вхождение факторов в модель. Оптимальные значения неизвестных параметров находятся на основе максимизации значения коэффициента детерминации регрессии. К сожалению, для решения рассматриваемой задачи требуется вручную задавать количество отбираемых факторов, которое часто бывает невозможно определить заранее. Исходя из этого была поставлена цель формализовать задачу так, чтобы в результате ее решения определялось еще и оптимальное количество отбираемых регрессоров. Для этого в качестве целевой функции был использован скорректированный коэффициент детерминации, зависящий от количества факторов модели. В результате была сформулирована задача частично целочисленного линейного программирования. Неизвестными параметрами в ней по-прежнему выступают бета-коэффициенты и булевы переменные, а также целочисленная переменная – количество регрессоров.

На основе данных о ценах и характеристиках седанов и хэтчбеков американской автомобильной промышленности проведен вычислительный эксперимент, подтверждающий корректность разработанного математического аппарата. Формализованная в работе проблема в виде задачи частично целочисленного линейного программирования выглядит предпочтительнее с вычислительной точки зрения, чем та же проблема, формализованная в настоящее время в современной научной литературе в виде задачи частично квадратичного линейного программирования.

Ключевые слова: регрессионная модель, структурная спецификация, отбор информативных регрессоров, метод наименьших квадратов, стандартизованная регрессия, задача частично булевого линейного программирования, задача частично целочисленного линейного программирования, скорректированный коэффициент детерминации, LPSolve IDE, Gretl.

M.P. Bazilevskii

Irkutsk State Transport University, Irkutsk, Russian Federation

SELECTION AN OPTIMAL NUMBER OF VARIABLES IN REGRESSION MODELS USING ADJUSTED COEFFICIENT OF DETERMINATION AS A MIXED INTEGER LINEAR PROGRAMMING PROBLEM

When constructing a regression model, the primary problem faced by the researcher is that it is not clear what the equation of connection between the explained and explanatory variables should be. This initial stage of construction the selection of the model structural specification is called. When choosing a regression specification in parallel, the question arises of which explanatory variables should be included in the equation. This is the problem of variables selection in regression models. Its essence is to single out from the set of "candidates" for inclusion a subset of the most informative of them based on some quality criterion.

The article is devoted to the problem of variables selection in regression models estimated using the ordinary least squares. The previously proposed approach to selection a given number of variables based on mixed 0–1 linear programming is considered. The unknown parameters in this problem are the beta coefficients of standardized regression and Boolean variables that are responsible for the occurrence of factors in the model. The optimal values of unknown parameters are found on the basis of maximizing the value of the coefficient of determination of regression. Unfortunately, to solve the problem under consideration, it is required to manually set the number of selected factors, which is often impossible to determine in advance. Therefore, the goal was to formalize the problem so that as a result of its solution the optimal number of selected regressors was also determined. For this purpose, the adjusted determination coefficient, depending on the number of model factors, was used as the objective function. As a result, the problem of mixed integer linear programming was formulated. The unknown parameters in it are still beta coefficients and Boolean variables, as well as an integer variable – the number of regressors.

Based on data on prices and characteristics of sedans and hatchbacks of the American automobile industry, a computational experiment was carried out confirming the correctness of the developed mathematical apparatus. The problem formalized in this work in the form of a mixed integer linear programming looks more preferable from a computational point of view than the same problem formalized in modern scientific literature as a mixed quadratic linear programming.

Keywords: regression model, structural specification, variables selection in regression, ordinary least squares, standardized regression, mixed 0–1 linear programming, mixed integer linear programming, adjusted coefficient of determination, LPSolve IDE, Gretl.

Введение

Регрессионный анализ [1, 2] в настоящее время является распространенным статистическим методом исследования влияния одной или нескольких объясняющих переменных на объясняемую переменную. За последние годы появилось множество научных работ, посвященных использованию в регрессионном моделировании аппарата математического программирования [3–6]. Полученная в результате оценивания регрессионная модель применяется для установления степени влияния

факторов на выходную переменную, прогнозирования неизвестных значений объясняемой переменной, принятия широкого круга управленческих решений и т.д. При этом одной из основных проблем, возникающих в процессе построения регрессии, является выбор спецификации модели. Для этого в первую очередь необходимо определиться с составом входящих в модель факторов, т.е. выделить из множества «кандидатов» на включение подмножества наиболее информативных из них на основе некоторого критерия качества. Эта проблема называется задачей отбора информативных регрессоров (ОИР) [7]. В условиях роста объемов информации эта проблема весьма актуальна в области интеллектуального анализа данных и машинного обучения.

Проведенный в работе [7] анализ методов ОИР позволил сделать вывод, что единственным из них, который гарантирует точное решение задачи, является метод перебора всех возможных регрессий. Остальные алгоритмы, такие как шаговая регрессия, ступенчатая регрессия, алгоритм последовательной замены, лассо Тибширани, метод наименьших углов, носят, по сути, эвристический характер. Точное решение задачи ОИР также может быть получено, если формализовать ее в виде задачи математического программирования. Так, в работе [8] задача ОИР при оценивании регрессионной модели с помощью метода наименьших модулей (МНМ) сведена к задаче частично булевого линейного программирования, а при оценивании с помощью метода наименьших квадратов (МНК) – к задаче частично булевого квадратичного программирования. При этом число регрессоров должно быть зафиксировано исследователем. Но в связи с тем, что оптимальное число отбираемых регрессоров априори неизвестно, появилась работа [9], в которой выбор оптимального числа регрессоров формализован в виде задачи частично целочисленного линейного программирования для критерия средней абсолютной ошибки и в виде задачи частично целочисленного квадратичного программирования для критерия средней квадратичной ошибки. В статье [10] задача выбора оптимального числа информативных регрессоров сведена к задаче частично целочисленного квадратичного программирования для скорректированного коэффициента детерминации, критерия Акаике и Шварца, а в работе [11] еще и для критерия Мэллоуза. В работах [12, 13] рассмотрены вычислительные аспекты проблемы ОИР как задачи частично целочисленного программирования.

Таким образом, в современных литературных источниках проблема ОИР при оценивании регрессионной модели с помощью МНК формализована только в виде задачи частично булевого квадратичного программирования при фиксированном числе регрессоров и в виде задачи частично целочисленного квадратичного программирования при неизвестном числе регрессоров. Однако автору в работах [14, 15] удалось свести задачу ОИР при оценивании регрессионной модели с помощью МНК к задаче частично булевого линейного программирования для заданного числа регрессоров. Данная статья является логическим продолжением работ [14, 15]. Ее целью является формализация проблемы выбора оптимального числа информативных регрессоров по скорректированному коэффициенту детерминации при МНК-оценивании регрессионных моделей в виде задачи частично целочисленного линейного программирования.

1. Проблема ОИР для заданного числа регрессоров как задача частично булевого линейного программирования

Рассмотрим модель множественной линейной регрессии:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + ... + \alpha_m x_{im} + \varepsilon_i, \quad i = \overline{1,n}, \tag{1}$$

где y_i, $i = \overline{1,n}$, – значения зависимой (объясняемой) переменной y; x_{i1}, x_{i2}, …, x_{im}, $i = \overline{1,n}$, – значения m независимых (объясняющих) переменных (регрессоров) x_1, x_2, …, x_m; ε_i, $i = \overline{1,n}$, – ошибки аппроксимации; α_0, α_1, …, α_m – неизвестные параметры; n – объем выборки.

Пусть неизвестные параметры модели (1) оцениваются с помощью МНК, суть которого состоит в минимизации суммы квадратов ошибок аппроксимации:

$$J(\alpha_0, \alpha_1, ..., \alpha_m) = \sum_{i=1}^{n} \varepsilon_i^2 \to \min. \tag{2}$$

Приведем строгую постановку задачи отбора информативных регрессоров (ОИР) [7]. Пусть задана выборка из наблюдений для зависимой переменной y_i, $i = \overline{1,n}$, и для l возможных независимых переменных x_{ij}, $i = \overline{1,n}$, $j = \overline{1,l}$. Необходимо выделить из l возможных регрессоров m переменных, минимизируя функцию потерь (2) для регрессии (1).

Для того чтобы свести эту задачу к задаче частично булевого программирования так, как это сделано в работе [14], проведем нормирование (стандартизацию) всех переменных по формулам

$$v_i = \frac{y_i - \overline{y}}{\sigma_y},\ z_{i1} = \frac{x_{i1} - \overline{x}_1}{\sigma_{x_1}},\ \ldots,\ z_{im} = \frac{x_{im} - \overline{x}_m}{\sigma_{x_m}},$$

где $\overline{y}$, $\overline{x}_1$, …, $\overline{x}_m$ – средние значения переменных; σ_y, σ_{x_1}, …, σ_{x_m} – среднеквадратические отклонения переменных; v, z_1, …, z_m – стандартизованные переменные, для которых среднее значение равно 0, а среднеквадратическое отклонение равно 1.

Тогда регрессии (1) ставится в соответствие ее стандартизованная модель:

$$v_i = \beta_1 z_{i1} + \beta_2 z_{i2} + \ldots + \beta_m z_{im} + u_i,\ i = \overline{1,n}, \qquad (3)$$

где β_1, …, β_m – неизвестные параметры (бета-коэффициенты); u_i, $i = \overline{1,n}$, – ошибки аппроксимации.

В работах [14, 15] показано, что МНК-оценки стандартизованной модели (3) являются решением системы линейных алгебраических уравнений, представленной в матричном виде:

$$K\beta = h,$$

где $K = \begin{bmatrix} 1 & r_{x_1x_2} & \cdots & r_{x_1x_m} \\ r_{x_1x_2} & 1 & \cdots & r_{x_2x_m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{x_1x_m} & r_{x_2x_m} & \cdots & 1 \end{bmatrix}$ – матрица коэффициентов интеркорреляции;

$h = \begin{bmatrix} r_{yx_1} \\ r_{yx_2} \\ \cdots \\ r_{yx_m} \end{bmatrix}$ – вектор-столбец коэффициентов корреляции между объясняемой переменной и объясняющими переменными;

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_m \end{bmatrix}$$ – вектор-столбец бета-коэффициентов.

Коэффициент детерминации R^2 регрессии (1) связан с бета-коэффициентами соотношением

$$R^2 = \sum_{i=1}^{m} r_{yx_i} \beta_i. \tag{4}$$

Тогда проблема ОИР может быть сформулирована в виде следующей задачи частично булевого линейного программирования:

$$\sum_{i=1}^{l} r_{yx_i} \beta_i \to \max, \tag{5}$$

$$-(1-\delta_i)M \le K_i \beta - h_i \le (1-\delta_i)M, \quad i = \overline{1,l}, \tag{6}$$

$$-\delta_i M \le \beta_i \le \delta_i M, \quad i = \overline{1,l}, \tag{7}$$

$$\delta_i \in \{0,1\}, \quad i = \overline{1,l}. \tag{8}$$

$$\sum_{i=1}^{l} \delta_i = m, \tag{9}$$

где K_i – i-я строка матрицы коэффициентов интеркорреляции K; h_i – i-й элемент вектора h; M – заранее выбранное большое положительное число;

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

Решение задачи (5)–(9) позволяет оценить лишь бета-коэффициенты β_i, $i = \overline{1,l}$. Для перехода к оценкам параметров регрессии (1) необходимо воспользоваться следующими формулами:

$$\alpha_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}}, \quad i = \overline{1,l},$$

$$\alpha_0 = \overline{y} - \alpha_1 \overline{x}_1 - \alpha_2 \overline{x}_2 - \dots - \alpha_l \overline{x}_l.$$

Для решения задачи частично булевого линейного программирования (5)–(9) требуется задавать число информативных регрессоров m, которое априори практически всегда неизвестно. Исходя из этого переформулируем эту задачу так, чтобы ее решение выдавало оптимальное число информативных регрессоров.

2. Проблема выбора оптимального числа регрессоров как задача частично целочисленного линейного программирования

Рассмотрим скорректированный коэффициент детерминации модели (1):

$$R^2_{\text{adj}} = 1 - \left(1 - R^2\right)\frac{n-1}{n-m-1}, \tag{10}$$

где R^2 – коэффициент детерминации; n – количество наблюдений; m – количество объясняющих переменных.

Коэффициент (10) применяется для того, чтобы можно было сравнивать модели с разным числом объясняющих переменных. Он «штрафует» регрессию за дополнительно включенные факторы. Чем больше значение скорректированного коэффициента детерминации, тем адекватнее модель.

С учетом формул (4) и (10) можно ввести функционал

$$1 - \left(1 - \sum_{i=1}^{l} r_{yx_i}\beta_i\right)\frac{n-1}{n-m-1} \to \max, \tag{11}$$

где переменная m удовлетворяет ограничениям

$$1 \le m \le l, \quad m \in Z. \tag{12}$$

Тогда решение задачи частично целочисленного нелинейного программирования с целевой функцией (11) и с ограничениями (6)–(9), (12) гарантирует выбор оптимального числа m информативных регрессоров.

Проведем линеаризацию задачи (11), (6)–(9), (12).

Из выражения (10) следует, что

$$nR^2_{\text{adj}} - mR^2_{\text{adj}} - R^2_{\text{adj}} = nR^2 - m - R^2. \tag{13}$$

В равенстве (13) нелинейным является только слагаемое mR^2_{adj}. Перепишем выражение (13) в виде

$$nR^2_{\text{adj}} - \left(R_1^2 + R_2^2 + ... + R_l^2\right) - R^2_{\text{adj}} = nR^2 - m - R^2, \quad (14)$$

где переменные R_i^2, $i = \overline{1,l}$, подчинены условию

$$R_i^2 = \begin{cases} 0, \text{ если } \delta_i = 0, \\ R^2_{\text{adj}}, \text{ если } \delta_i = 1, \end{cases} \quad i = \overline{1,l}. \quad (15)$$

Условия (15) можно заменить следующими линейными ограничениями:

$$-M\delta_i \le R_i^2 \le M\delta_i, \quad i = \overline{1,l}, \quad (16)$$

$$-M\left(1-\delta_i\right) \le R_i^2 - R^2_{\text{adj}} \le M\left(1-\delta_i\right), \quad i = \overline{1,l}. \quad (17)$$

Так, если $\delta_i = 0$, то из выражения (16) следует, что $R_i^2 = 0$; а если $\delta_i = 1$, то из формулы (17) следует, что $R_i^2 = R^2_{\text{adj}}$.

Введем целевую функцию

$$R^2_{\text{adj}} \to \max. \quad (18)$$

Тогда задача частично целочисленного линейного программирования с целевой функцией (18) и с линейными ограничениями (4), (6)–(9), (14), (12), (16), (17) равносильна задаче нелинейного программирования (11), (6)–(9), (12) и дает точное решение проблемы выбора оптимального числа информативных регрессоров по скорректированному коэффициенту детерминации в оцениваемой с помощью МНК регрессионной модели.

3. Вычислительный эксперимент

Для проведения вычислительного эксперимента были использованы статистические данные эконометрического пакета Gretl (встроенный файл data7-12.gdt) о ценах и характеристиках седанов и хэтчбеков американской автомобильной промышленности за 1995 г. Объем выборки $n = 82$. Объясняемая переменная:

price – цена, тыс. долл.;

объясняющие переменные:

hatch – тип автомобиля (1 – хэтчбек, 0 – седан);

wbase – колесная база (расстояние между передней и задней осями), дюйм;

length – длина автомобиля, дюйм;

width – ширина автомобиля, дюйм;

height – высота автомобиля, дюйм;

weight – вес автомобиля, сотни фунтов;

cyl – количество цилиндров двигателя;

liters – объем двигателя, литры;

gasmpg – экономичность расхода топлива, миль на галлон;

trans – трансмиссия (1 – автомат, 0 – в противном случае).

Была поставлена следующая задача: из 10 объясняющих переменных выбрать оптимальное число информативных регрессоров по скорректированному коэффициенту детерминации в оцениваемой с помощью МНК регрессионной модели. Заметим, что эта задача может быть решена полным перебором всех возможных регрессий, общее количество которых $2^{10} = 1024$.

Поставленная задача с использованием пакета LPSolve IDE была сформулирована в виде задачи частично целочисленного линейного программирования с целевой функцией (18) и с линейными ограничениями (4), (6)–(9), (14), (12), (16), (17). Подробные результаты итерационного решения этой задачи в LPSolve IDE представлены в таблице.

Подробные результаты решения задачи

Итерация	δ_1	δ_2	δ_3	δ_4	δ_5	δ_6	δ_7	δ_8	δ_9	δ_{10}	m	R^2	R^2_{adj}
1	1	1	1	1	1	1	1	1	1	1	10	0,636366	0,585150
2	1	1	1	1	1	1	1	1	1	0	9	0,635668	0,590127
3	1	1	1	1	1	1	1	0	1	1	9	0,636122	0,590638
4	1	1	1	1	1	1	1	0	1	0	8	0,635276	0,595306
5	1	1	1	0	1	1	1	1	1	0	8	0,635525	0,595582
6	1	1	1	0	1	1	1	0	1	1	8	0,636051	0,596166
7	1	1	1	0	1	1	1	0	1	0	7	0,635182	0,600672

Полученная в результате МНК-оценивания семифакторная регрессионная модель имеет вид

$$\text{price} = 107,753 - 5,159\text{hatch} - 0,289\text{wbase} - 0,453\text{length} -$$

$$-1,236\text{height} + 2,586\text{weight} + 0,975\text{cyl} + 0,364\text{gasmpg}. \qquad (19)$$

Ее коэффициент детерминации $R^2 = 0,635182$, а скорректированный коэффициент детерминации $R^2_{\text{adj}} = 0,600672$. Отметим, что тот же самый результат показал метод полного перебора всех возможных регрессий, что подтверждает корректность предложенного в данной работе математического аппарата.

Заключение

В данной работе проблема выбора оптимального числа информативных регрессоров по скорректированному коэффициенту детерминации в регрессионных моделях, оцениваемых с помощью МНК, сведена к задаче частично целочисленного линейного программирования. Отметим, что в работе Р. Мияширо и Ю. Такано [10] эта проблема представлена в виде задачи частично целочисленного квадратичного программирования. А поскольку методы решения задач линейного программирования гораздо более эффективны, чем методы решения задач квадратичного программирования, применение описанного в данной работе математического аппарата на практике должно привести к снижению времени решения задачи.

Помимо этого, представленный в настоящей работе новый подход, так же как и метод полного перебора регрессий, гарантирует точное решение поставленной задачи. Однако в первом случае это решение находится с помощью метода ветвей и границ, отсекающего подмножества решений, заведомо не содержащих оптимальных решений. А это означает, что теоретически метод полного перебора всех регрессий должен уступать по скорости предложенному подходу. Исследование этих двух вопросов будет представлено в дальнейших работах автора.

Список литературы

1. Harrell Jr., Frank E. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. – Springer Series in Statistics, 2015. – 582 p.

2. Kuhn M., Johnson K. Applied predictive modeling. – Springer, 2018. – 600 p.

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

4. 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 Иркутск. гос. техн. ун-та. – 2017. – Т. 21, № 9 (128). – С. 37–44.

5. 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.

6. 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.

7. Носков С.И., Базилевский М.П. Построение регрессионных моделей с использованием аппарата линейно-булевого программирования / ИрГУПС. – Иркутск, 2018. – 176 с.

8. Konno H., Yamamoto R. Choosing the best set of variables in regression analysis using integer programming // Journal of Global Optimization. – 2009. – Vol. 44, no. 2. – P. 272–282.

9. Park Y.W., Klabjan D. Subset selection for multiple linear regression via optimization. – 2020. – URL: https://arxiv.org/pdf/1701.07920.pdf 081 (accessed 13 March 2020).

10. 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. DOI: 10.1016/j.ejor.2015.06.081

11. 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. DOI: 10.1016/j.eswa.2014.07.056

12. Bertsimas D., King A., Mazumder R. Best subset selection via a modern optimization lens // The Annals of Statistics. – 2016. – Vol. 44, no. 2. – P. 813–852.

13. Bertsimas D., King A. OR Forum – An algorithmic approach to linear regression // Operations Research. – 2016. – Vol. 64, no. 1. – P. 2–16.

14. 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.

15. 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.

References

1. Harrell Jr., Frank E. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. *Springer Series in Statistics*. 2015, 582 p.

2. Kuhn M., Johnson K. Applied predictive modeling. Springer. 2018, 600 p.

3. Bazilevskiy M.P., Noskov S.I. Otsenivanie indeksnykh modeley regressii s pomoshch'yu metoda naimen'shikh moduley [Estimation of index regression models using the least absolute deviations]. *Vestnik Rossiyskogo novogo universiteta. Seriya: Slozhnye sistemy: modeli, analiz i upravlenie*, 2020, no. 1, pp. 17-23.

4. Bazilevskiy M.P., Noskov S.I. Programmnyy kompleks postroeniya lineynoy regressionnoy modeli s uchetom kriteriya soglasovannosti povedeniya faktycheskoy i raschetnoy traektoriy izmeneniya znacheniy ob"yasnyaemoy peremennoy [Program complex for linear regression model construction considering behavior consistency criterion of actual and calculated trajectories of explained variable value change]. *Vestnik Irkutskogo gosudarstvennogo tekhnicheskogo universiteta*, 2017, vol. 128, no. 9, pp. 37-44.

5. Bazilevskiy M.P., Noskov S.I. Formalizatsiya zadachi postroeniya lineyno-mul'tiplikativnoy regressii v vide zadachi chastichno-bulevogo lineynogo programmirovaniya [Formalization of the problem of construction of linear multiplicative regressions in the form of a partial-Boolean linear programming problem]. *Sovremennye tekhnologii. Sistemnyy analiz. Modelirovanie*, 2017, vol. 55, no. 3, pp. 101-105.

6. Noskov S.I. O metode smeshannogo otsenivaniya parametrov lineynoy regressii [On the method of mixed estimation of linear regression parameters]. *Informatsionnye tekhnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami*, 2019, vol. 2, no. 1, pp. 41-45.

7. Noskov S.I., Bazilevskiy M.P. Postroenie regressionnykh modelei s ispol'zovaniem apparata lineino-bulevogo programmirovaniia [Constructing regression models using a linear Boolean programming apparatus]. Irkutsk, Irkutsk State Transport University, 2018, 176 p.

8. Konno H., Yamamoto R. Choosing the best set of variables in regression analysis using integer programming. *Journal of Global Optimization*, 2009, vol. 44, no. 2, pp. 272-282.

9. Park Y.W., Klabjan D. Subset selection for multiple linear regression via optimization. 2020, available at https://arxiv.org/pdf/1701.07920.pdf 081 (accessed 13 March 2020).

10. 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. DOI: 10.1016/j.ejor.2015.06.081 081

11. 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. DOI: 10.1016/j.eswa.2014.07.056 081

12. Bertsimas D., King A., Mazumder R. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 2016, vol. 44, no. 2, pp. 813-852.

13. Bertsimas D., King A. OR Forum – An algorithmic approach to linear regression. *Operations Research*, 2016, vol. 64, no. 1, pp. 2-16.

14. Bazilevskiy M.P. Svedenie zadachi otbora informativnykh regressorov pri otsenivanii lineinoi regressionnoi modeli po metodu naimen'shikh kvadratov k zadache chastichno-bulevogo lineinogo programmirovaniia [Reduction the problem of selecting informative regressors when estimating a linear regression model by the method of least squares to the problem of partial-Boolean linear programming]. *Modelirovanie, optimizatsiia i informatsionnye tekhnologii*, 2018, vol. 20, no. 1, pp. 108-117.

15. Bazilevskiy M.P. Otbor informativnykh regressorov s uchetom mul'tikollinearnosti mezhdu nimi v regressionnykh modeliakh kak zadacha chastichno-bulevogo lineinogo programmirovaniia [Subset selection in regression models with considering multicollinearity as a task of mixed 0-1 integer linear programming]. *Modelirovanie, optimizatsiia i informatsionnye tekhnologii*, 2018, vol. 21, no. 2, pp. 104-118.

Получено 13.03.2020

Принято 16.05.2020

Сведения об авторе

Базилевский Михаил Павлович (Иркутск, Россия) – кандидат технических наук, доцент кафедры «Математика», Иркутский государственный университет путей сообщения (664074, Иркутск, ул. Чернышевского, 15, e-mail: mik2178@yandex.ru).

About the author

Mikhail P. Bazilevskii (Irkutsk, Russian Federation) – Ph.D. in Engineering, Associate Professor, Department of Mathematics, Irkutsk State Transport University (664074, Irkutsk, Chernyshevskogo st., 15, e-mail: mik2178@yandex.ru).

Библиографическое описание статьи согласно ГОСТ Р 7.0.100–2018:

Базилевский, М.П. Отбор оптимального числа информативных регрессоров по скорректированному коэффициенту детерминации в регрессионных моделях как задача частично целочисленного линейного программирования / П. М. Базилевский. –

DOI 10.15593/2499-9873/2020.2.03. – Текст : непосредственный // Прикладная математика и вопросы управления = Applied Mathematics and Control Sciences. – 2020. – № 2. – С. 41–54.

Цитирование статьи в изданиях РИНЦ:

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. DOI: 10.15593/2499-9873/2020.2.03

Цитирование статьи в references и международных изданиях:

Cite this article as:

Bazilevskii M.P. Selection an optimal number of variables in regression models using adjusted coefficient of determination as a mixed integer linear programming problem. *Applied Mathematics and Control Sciences*, 2020, no. 2, pp. 41–54. DOI: 10.15593/2499-9873/2020.2.03 (*in Russian*)