СТАТЬИ

УДК 519.862.6:004
DOI 10.17513/snt.39723

ФОРМАЛИЗАЦИЯ ПРОЦЕССА ОТБОРА ИНФОРМАТИВНЫХ РЕГРЕССОРОВ В ЛИНЕЙНОЙ РЕГРЕССИИ В ВИДЕ ЗАДАЧИ ЧАСТИЧНО-БУЛЕВОГО ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ С ОГРАНИЧЕНИЯМИ НА КОЭФФИЦИЕНТЫ ИНТЕРКОРРЕЛЯЦИЙ

Базилевский М.П.

ФГБОУ ВО «Иркутский государственный университет путей сообщения», Иркутск, e-mail: mik2178@yandex.ru

Статья посвящена исследованию задачи отбора наиболее информативных регрессоров в модели множественной линейной регрессии, оцениваемой с помощью метода наименьших квадратов. Ранее эта задача была формализована в виде задачи частично-булевого линейного программирования. В полученной в результате ее решения линейной регрессии интеркорреляции объясняющих переменных могли быть статистически значимыми, что затрудняло интерпретацию ее оценок из-за мультиколлинеарности. В данной статье для контроля в процессе отбора наиболее информативных регрессоров абсолютных величин интеркорреляций в задачу частично-булевого линейного программирования введены специальные линейные ограничения. В результате сформулирована задача, решение которой приводит к построению линейной регрессии с оптимальным по коэффициенту детерминации числом объясняющих переменных, в которой абсолютные величины интеркорреляций не превосходят заданного числа. Помимо этого знаки оценок линейной регрессии согласуются со знаками соответствующих коэффициентов корреляции, а абсолютные вклады переменных в общую детерминацию не меньше заданного числа. С помощью предложенных линейных ограничений на интеркорреляции сформулированы задачи булевого линейного программирования, позволяющие выделять в корреляционной матрице кластеры слабо и высоко коррелирующих переменных. Проведено успешное тестирование предложенного математического аппарата.

Ключевые слова: линейная регрессия, отбор информативных регрессоров, задача частично-булевого линейного программирования, метод наименьших квадратов, интеркорреляция, кластеризация, мультиколлинеарность

FORMALIZATION THE SUBSET SELECTION PROCESS IN LINEAR REGRESSION AS A MIXED INTEGER 0-1 LINEAR PROGRAMMING PROBLEM WITH CONSTRAINTS ON INTERCORRELATION COEFFICIENTS

Bazilevskiy M.P.

Irkutsk State Transport University, Irkutsk, e-mail: mik2178@yandex.ru

This article is devoted to the study of subset selection problem in a multiple linear regression model estimated using the ordinary least squares. Previously, this problem was formalized as a mixed integer 0-1 linear programming problem. In the linear regression obtained as a result of its solution, the explanatory variables intercorrelations could be statistically significant, which made it difficult to interpret her estimates due to multicollinearity. In this article, to control the absolute values of intercorrelations in subset selection process, special linear constraints are introduced into the mixed integer 0-1 linear programming problem. As a result, a problem is formulated, the solution of which leads to the construction of a linear regression with an optimal number of explanatory variables in terms of the coefficient of determination, in which the absolute values of intercorrelations do not exceed a given number. In addition, the signs of the linear regression estimates agree with the signs of the corresponding correlation coefficients, and the absolute contributions of the variables to the overall determination are not less than a given number. With the help of the proposed linear constraints on intercorrelations, integer 0-1 linear programming problems are formulated that make it possible to single out clusters of weakly and highly correlated variables in the correlation matrix. The proposed mathematical apparatus was successfully tested.

Keywords: linear regression, subset selection in regression, mixed integer 0-1 linear programming, ordinary least squares, intercorrelation, multicollinearity, clustering

Аппарат математического программирования на сегодняшний день успешно применяется для отбора наиболее информативных регрессоров (ОИР) в регрессионных моделях. В зарубежной литературе такая процедура известна как «feature selection», «subset selection» и т.д. При этом в иностранных источниках задача ОИР в основном сводится к задаче частично-булевого квадратичного программирования (ЧБКП). Одна из первых таких формализаций для линейных регрессий, оцениваемых с помощью метода наименьших квадратов (МНК), была представлена в [1]. В [2] сформулирована задача ОИР на основе скорректированного коэффициента детерминации и информационных критериев Акаике и Шварца, в [3] – на основе критерия Мэллоуза, в [4] – на основе критериев среднеквадратичных и абсолютных ошибок, в [5] – на основе критерия

кросс-валидации. В [6] исследуется целостная линейная регрессия, для которой сформулирована задача ОИР с ограничениями на значимость коэффициентов и степень мультиколлинеарности, а в [7] осуществляется так называемая регрессионная диагностика с использованием линейных ограничений на наблюдаемые значения t-критерия Стьюдента. В [8] сформулирована задача ОИР для устранения мультиколлинеарности с помощью факторов «вздутия» дисперсии, в [9] разработан алгоритм для решения задачи максимизации канонической корреляции, а в [10] сформулирована задача для выявления уравнений сложных динамических систем.

В [11–14] автором предложены различные формализации задачи ОИР в линейной регрессии, оцениваемой с помощью МНК, в виде задач частично-булевого линейного программирования (ЧБЛП). Для контроля мультиколлинеарности в [11] использованы ограничения на факторы «вздутия» дисперсии. Однако этих ограничений недостаточно для того, чтобы гарантировать отсутствие значимой корреляции абсолютно между всеми парами объясняющих переменных, что необходимо для корректного объяснения полученных с помощью МНК оценок.

Цель исследования состоит в формализации задачи ОИР для линейной регрессии, оцениваемой с помощью МНК, в виде задачи ЧБЛП с линейными ограничениями на корреляции объясняющих переменных (интеркорреляции).

Материалы и методы исследования

Модель множественной линейной регрессии имеет вид

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \varepsilon_i,\ i = \overline{1,n}, \qquad (1)$$

где y_i, $i = \overline{1,n}$ – значения объясняемой (зависимой) переменной y; x_{ij}, $i = 1,n, j = 1,l$ – значения l объясняющих (независимых) переменных x_1, x_2,…, x_l; n – объем выборки; α_j, $j = 0,l$ – неизвестные параметры; ε_i, $i = 1,n$ – ошибки аппроксимации.

Задача ОИР для модели (1), оцениваемой с помощью МНК, формулируется следующим образом: необходимо из l объясняющих переменных выбрать m наиболее информативных так, чтобы либо сумма квадратов остатков модели была минимальна, либо коэффициент детерминации R^2 был максимален.

Проведем нормирование (стандартизацию) всех переменных по правилам:

$$y_i^{\bullet} = \frac{y_i - \overline{y}}{\sigma_y},\ x_{i1}^{\bullet} = \frac{x_{i1} - \overline{x_1}}{\sigma_{x_1}},\ \ldots,\ x_{il}^{\bullet} = \frac{x_{il} - \overline{x_l}}{\sigma_{x_l}},$$
$$i = \overline{1,n},$$

где $\overline{y}$, $\overline{x_1}$, …, $\overline{x_l}$ – средние значения переменных, σ_y, σ_{x_1}, …, σ_{x_l} – среднеквадратические отклонения переменных.

Составим модель стандартизованной линейной регрессии:

$$y_i^{\bullet} = \beta_1 x_{i1}^{\bullet} + \beta_2 x_{i2}^{\bullet} + \ldots + \beta_l x_{il}^{\bullet} + \varepsilon_i^{\bullet},\ i = \overline{1,n}, \qquad (2)$$

где β_1, …, β_l – неизвестные параметры; $\varepsilon_i^{\bullet}$, $i = 1,n$ – ошибки аппроксимации.

МНК-оценки стандартизованной регрессии (2) находятся по формуле

$$\tilde{\beta}_{\text{МНК}} = R_{xx}^{-1} \cdot R_{yx},$$

где $R_{xx} = \begin{pmatrix} 1 & r_{x_1x_2} & \ldots & r_{x_1x_l} \\ r_{x_1x_2} & 1 & \ldots & r_{x_2x_l} \\ \ldots & \ldots & \ldots & \ldots \\ r_{x_1x_l} & r_{x_2x_l} & \ldots & 1 \end{pmatrix}$ – матрица интеркорреляций;

$R_{yx} = \begin{pmatrix} r_{yx_1} & r_{yx_2} & \ldots & r_{yx_l} \end{pmatrix}^T$ – вектор корреляций объясняющих переменных с y.

В [14] сформулирована следующая задача ЧБЛП для ОИР в модели (2):

$$R^2 = \sum_{j=1}^{l} r_{yx_j} \cdot \beta_j \to \max \qquad (3)$$

$$-(1-\delta_j) \cdot M \le \sum_{k=1}^{l} r_{x_jx_k} \cdot \beta_k - r_{yx_j} \le (1-\delta_j) \cdot M,$$
$$j = \overline{1,l}, \qquad (4)$$

$$0 \le \beta_j \le \delta_j \cdot M,\ j \in J^+, \qquad (5)$$

$$-\delta_j \cdot M \le \beta_j \le 0,\ j \in J^-, \qquad (6)$$

$$\delta_j \in \{0,1\},\ j = \overline{1,l}, \qquad (7)$$

$$\sum_{j=1}^{l} \delta_j = m, \qquad (8)$$

в формуле (8) $\delta_j = \begin{cases} 1, & \text{если j-я объясняющая переменная входит в модель,} \\ 0, & \text{в противном случае,} \end{cases}$

M – большое положительное число, J^+ и J^- – сформированные из множества $\{1,2,...,l\}$ индексные подмножества, элементы которых удовлетворяют условиям $r_{yx_j} > 0$ и $r_{yx_j} < 0$.

Решение задачи ЧБЛП (3)–(8) приводит к построению оптимальной по критерию R^2 линейной регрессии с m объясняющими переменными, в которой знаки коэффициентов согласованы со знаками соответствующих корреляций вектора R_{xy}. Для того чтобы можно было объяснить полученные оценки, еще до решения задачи (3)–(8) необходимо исключать все объясняющие переменные, противоречиво коррелирующие с y. Для объяснения необходимо использовать МНК-оценки модели (1), связанные с МНК-оценками регрессии (2) формулами

$$\tilde{\alpha}_j = \tilde{\beta}_j \cdot \sigma_y \cdot \sigma_{x_j}^{-1}, j = \overline{1,l},$$

$$\tilde{\alpha}_0 = \overline{y} - \tilde{\alpha}_1 \overline{x_1} - \tilde{\alpha}_2 \overline{x_2} - ... - \tilde{\alpha}_l \overline{x_l}.$$

Пусть множество Ф содержит номера отобранных переменных. Поскольку для оцененной линейной регрессии выполняются условия $\tilde{\beta}_j \cdot r_{yx_j} > 0$, $j \in$ Ф, то становятся справедливыми следующие формулы для абсолютных вкладов переменных в общую детерминацию R^2:

$$C_{x_j}^{\text{абс}} = r_{yx_j} \cdot \tilde{\beta}_j, \ \ j \in \Phi.$$

С помощью этих коэффициентов можно контролировать степень влияния любой объясняющей переменной на y. Поэтому в [14] было предложено в задаче (3)–(8) заменить линейное ограничение (8) на следующее:

$$r_{yx_j} \cdot \beta_j \geq \theta \cdot \delta_j, j = \overline{1,l} \qquad (9)$$

где $\theta \geq 0$ – назначенный исследователем наименьший вклад входящих в модель объясняющих переменных в общую детерминацию R^2. Чем выше значение θ, тем больше переменных «выдавливается» из модели, следовательно, регрессия становится проще и снижается эффект мультиколлинеарности.

Таким образом, решение задачи ЧБЛП (3)–(7), (9), приводит к построению линейной регрессии с оптимальным по критерию R^2 количеством объясняющих переменных, в которой $\tilde{\beta}_j \cdot r_{yx_j} > 0$, $j \in$ Ф, и вклады $C_{x_j}^{\text{абс}} \geq \theta$, $j \in$ Ф.

Для контроля в оптимизационной задаче ОИР абсолютных величин интеркорреляций введем следующие линейные ограничения:

$$\left|r_{x_i x_j}\right|\left(d_i + d_j - 1\right) \leq r, \ i = \overline{1,l-1}, \ j = \overline{i+1,l}, (10)$$

где $0 \leq r \leq 1$ – назначенная исследователем наибольшая величина абсолютных интеркорреляций входящих в модель объясняющих переменных. Если $r = 0$, то в построенной регрессии все интеркорреляции должны быть равны 0, а если $r = 1$, то нет ограничений на величины интеркорреляций в оцененной модели.

Если в ограничении (10) $d_{i^*} = d_{j^*} = 0$, что означает, что ни i^*-я, ни j^*-я переменная не входят в регрессию, то оно принимает вид $-\left|r_{x_{i^*} x_{j^*}}\right| \leq r$, поэтому справедливо всегда. Если в ограничении (10) либо $d_{i^*} = 0$, $d_{j^*} = 1$, либо $d_{i^*} = 1$, $d_{j^*} = 0$ (в модель входит либо i^*-я, либо j^*-я переменная), то оно принимает вид $0 \leq r$, поэтому справедливо всегда. Если же в ограничении (10) $d_{i^*} = d_{j^*} = 1$ (в модель входит и i^*-я, и j^*-я переменная), то оно принимает вид $\left|r_{x_{i^*} x_{j^*}}\right| \leq r$. В последнем случае, если $\left|r_{x_{i^*} x_{j^*}}\right| \leq r$ выполняется, то i^*-я и j^*-я переменные могут входить в модель одновременно, а если не выполняется, то нет.

Заметим, что общее число ограничений (10) может быть сокращено. Действительно, если $r_{x_{i^*} x_{j^*}} = 0$, то ограничение (10) выполняется для любых значений бинарных переменных. Аналогично, если $\left|r_{x_{i^*} x_{j^*}}\right| \geq r$. Тогда ограничения (10) следует переписать в виде

$$\left|r_{x_i x_j}\right|\left(d_i + d_j - 1\right) \leq r,$$

$$(i,j) \in \left\{(s_1, s_2) \,|\, \left|r_{x_{s_1} x_{s_2}}\right| \geq r\right\} \qquad (11)$$

Таким образом, решение задачи ЧБЛП (3)–(7), (9), (11) приводит к построению линейной регрессии с оптимальным по критерию R^2 количеством объясняющих переменных, в которой $\tilde{\beta}_j \cdot r_{yx_j} > 0$, $j \in$ Ф, вклады $C_{x_j}^{\text{абс}} \geq \theta$, $j \in$ Ф, и интеркорреляции $\left|r_{x_i x_j}\right| \leq r$, $i, j \in$ Ф, $i < j$.

Предложенные ограничения (11) на величины интеркорреляций могут быть использованы для формализации задачи кластеризации корреляционной матрицы. Эта задача может быть сформулирована следующим образом: необходимо из l объясняющих переменных выбрать как можно больше переменных так, чтобы все их интеркорреляции не превосходили числа r.

Эта задача может быть формализована в виде задачи булевого линейного программирования (БЛП) с целевой функцией

$$\sum_{j=1}^{l} \delta_j \to \max \qquad (12)$$

и линейными ограничениями (7), (11).

Ее решение позволяет сформировать кластер слабо коррелирующих объясняющих переменных (КСКП).

Аналогично можно сформулировать задачу отбора из l объясняющих переменных наибольшего числа переменных так, чтобы все их интеркорреляции были не меньше числа r. Она формализуется в виде задачи БЛП с целевой функцией (12), линейными ограничениями (7) и

$$\left|r_{x_i x_j}\right| \geq r\left(d_i + d_j - 1\right),$$

$$(i, j) \in \left\{(s_1, s_2) \,|\, \left|r_{x_{s_1} x_{s_2}}\right| \leq r\right\} \qquad (13)$$

Решение задачи (12), (7), (13) позволяет сформировать кластер высоко коррелирующих объясняющих переменных (КВКП).

Заметим, что сформулированные задачи БЛП могут иметь несколько оптимальных решений.

Результаты исследования и их обсуждение

Для тестирования предложенного математического аппарата были использованы встроенные в эконометрический пакет Gretl статистические данные о зарплатах игроков НБА (data7-20.gdt). Для удобства были исключены все фиктивные переменные и составлена выборка из первых 30 наблюдений для зависимой переменной SALARY и оставшихся 17 объясняющих переменных. Для решения оптимизационных задач использовался решатель LPSolve IDE на персональном компьютере с 4-ядерным процессором (3100 МГц) и оперативной памятью 4 Гб. Для заданного числа r решалась задача ЧБЛП (3)–(7), (11) (без ограничений на вклады), задача БЛП (12), (7), (11) и задача БЛП (12), (7), (13). Для удобства назовем их задача А, задача В и задача С соответственно. В результате решения задачи А фиксировались номера отобранных переменных, время решения t в секундах и коэффициент детерминации R^2. А в результате решения задач В и С фиксировалось число отобранных переменных m и время решения t. Результаты тестирования представлены в таблице. Большое число M для решения задачи А выбиралось так, как это предложено делать в [14].

По таблице видно, что LPSolve IDE справился со всеми задачами практически мгновенно. Как и ожидалось, при решении задачи А с уменьшением числа r, т.е. с ужесточением требования на объясняющие переменные с высокими интеркорреляциями, число отобранных переменных в линейной регрессии снижается. При этом также снижается время решения задачи и значение коэффициента детерминации. А время решения задач В и С практически не менялось в зависимости от выбранных значений r. При этом в задаче В с уменьшением r объем кластера слабо коррелирующих переменных уменьшался, а в задаче С объем кластера высоко коррелирующих переменных увеличивался. Полученные результаты подтверждают корректность предложенного математического аппарата.

Заключение

В результате проведенных исследований сформулирована задача ЧБЛП, решение которой приводит к построению линейной регрессии с оптимальным по критерию R^2 числом объясняющих переменных, в которой знаки оценок согласуются со знаками соответствующих коэффициентов корреляции r_{yx_j}, абсолютные вклады переменных не меньше числа θ, а интеркорреляции не превосходят числа r. Построенная регрессионная модель гарантированно может быть интерпретирована, если на начальном этапе были исключены все противоречиво коррелирующие с y объясняющие переменные.

Результаты решения задач

r	Задача А			Задача В		Задача С	
	Номера переменных	t	R^2	m	t	m	t
0,9	2, 3, 6, 8, 9, 11, 12, 13, 14, 15, 17	1,481	0,5854	14	0,013	2	0,025
0,8	2, 3, 5, 9, 11, 12, 13, 14, 15, 17	1,217	0,5556	12	0,015	3	0,051
0,7	3, 5, 9, 11, 12, 13, 14, 15, 17	0,791	0,5413	10	0,021	4	0,031
0,6	4, 6, 9, 12, 13, 15, 17	0,557	0,4626	9	0,021	5	0,032
0,5	4, 5, 9, 12, 13, 15, 17	0,402	0,4438	8	0,025	7	0,027
0,4	4, 5, 9, 13, 15, 17	0,235	0,4233	6	0,032	7	0,023
0,3	4, 5, 15, 17	0,206	0,2841	5	0,036	8	0,021
0,2	4, 15, 17	0,155	0,2050	3	0,042	9	0,023
0,1	4, 15, 17	0,095	0,2050	3	0,027	10	0,016

Помимо этого разработанные линейные ограничения на интеркорреляции позволили сформировать задачи БЛП для построения кластеров слабо и высоко коррелирующих переменных. Первый из них способствует построению традиционных моделей множественной линейной регрессии, а второй – моделей полносвязной линейной регрессии, в которых все объясняющие переменные связаны между собой. В дальнейшем планируется провести тестирование предложенного математического аппарата для построения регрессионных моделей по выборкам большого объема.

Список литературы

1. 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. DOI: 10.1007/s10898-008-9323-9.

2. 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.

3. Miyashiro R., Takano Y. Subset selection by Mallows' Cp: A mixed integer programming approach // Expert Systems with Applications. 2015. Vol. 42. P. 325–331. DOI: 10.1016/j.eswa.2014.07.056.

4. Park Y.W., Klabjan D. Subset selection for multiple linear regression via optimization // Journal of Global Optimization. 2020. Vol. 77. P. 543–574. DOI: 10.1007/s10898-020-00876-1.

5. Takano Y., Miyashiro R. Best subset selection via cross-validation criterion. Top. 2020. Vol. 28, Is. 2. P. 475–488. DOI: 10.1007/s11750-020-00538-1.

6. Bertsimas D., Li M.L. Scalable holistic linear regression // Operations Research Letters. 2020. Vol. 48, Is. 3. P. 203–208. DOI: 10.1016/j.orl.2020.02.008.

7. 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. DOI: 10.1016/j.patcog.2020.107565.

8. Tamura R., Kobayashi K., Takano Y., Miyashiro R., Nakata K., Matsui T. Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor // Journal of Global Optimization. 2019. Vol. 73. P. 431–446.

9. Watanabe A., Tamura R., Takano Y., Miyashiro R. Branch-and-bound algorithm for optimal sparse canonical correlation analysis // Expert Systems with Applications. Vol. 217. P. 119530. DOI: 10.1016/j.eswa.2023.119530.

10. Bertsimas D., Gurnee W. Learning sparse nonlinear dynamics via mixed-integer optimization. Nonlinear Dynamics. 2023. Vol. 111. No. 7. P. 6585–6604. DOI: 10.1007/s11071-022-08178-9.

11. 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.

12. 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.

13. Базилевский М.П. Отбор значимых по критерию Стьюдента информативных регрессоров в оцениваемых с помощью МНК регрессионных моделях как задача частично-булевого линейного программирования // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2021. № 3. С. 5–16.

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