УДК 519.862.6

М.П. Базилевский

СВЕДЕНИЕ ЗАДАЧИ ОТБОРА ИНФОРМАТИВНЫХ РЕГРЕССОРОВ ПРИ ОЦЕНИВАНИИ ЛИНЕЙНОЙ РЕГРЕССИОННОЙ МОДЕЛИ ПО МЕТОДУ НАИМЕНЬШИХ КВАДРАТОВ К ЗАДАЧЕ ЧАСТИЧНО-БУЛЕВОГО ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

Иркутский государственный университет путей сообщения, Иркутск, Россия

Одной из главных проблем в регрессионном анализе является проблема выбора структурной спецификации регрессионной модели, т.е. выбора состава переменных и математической формы связи между ними. В случае линейной регрессионной модели такая задача сводится только лишь к отбору наиболее информативных регрессоров. Точное решение задачи отбора информативных регрессоров при оценивании линейной регрессии с помощью метода наименьших квадратов может быть получено либо алгоритмом полного перебора, либо посредством введения в рассмотрение булевых переменных и последующем решении весьма непростой вычислительной задачи частично-булевого квадратичного программирования. В данной статье задача отбора информативных регрессоров в линейной регрессии, оцениваемой с помощью метода наименьших квадратов, сведена к задаче частично-булевого линейного программирования, решение которой не вызывает никаких затруднений при использовании соответствующих пакетов программ. Новая постановка задачи предполагает для оценивания неизвестных параметров линейной регрессионной модели производить предварительное нормирование всех переменных с целью нахождения бета-коэффициентов стандартизованной регрессии. Бета-коэффициенты определяются по известной интеркорреляционной матрице и вектору корреляций между зависимой переменной и независимыми факторами. Для оценки адекватности линейной регрессии применяется коэффициент детерминации.

Ключевые слова: регрессионная модель, метод наименьших квадратов, отбор информативных регрессоров, задача частично-булевого линейного программирования, стандартизованная регрессия, коэффициент корреляции, критерий детерминации.

Введение. В регрессионном анализе проблема выбора спецификации регрессионной модели, т.е. выбора состава и математической формы связи между переменными, является одной из ключевых. При построении линейной регрессии эта задача сводится только лишь к выделению из множества «кандидатов» на включение в состав независимых переменных уравнения подмножества наиболее информативных из них на основе некоторого критерия качества. Будем называть такую проблему задачей отбора информативных регрессоров (ОИР). В зарубежной литературе она более известна, как «subset selection», «variable selection» или «feature selection» in regression [1, 2]. В настоящее время, в результате роста объемов хранимых данных в соответствующих базах, эта проблема весьма актуальна в области интеллектуального анализа данных и машинного обучения.

Проблема ОИР достаточно полно освещена в работах Дж. Себера [3], А. Миллера [1], В.В. Стрижова и Е.А. Крымовой [4], Х. Лю и Х. Мотоды [5], И. Гион и А. Елиссеева [6], А.Г. Ивахненко [7] и др. При этом большинство методов ОИР предназначено главным образом для оценивания параметров линейной регрессии с помощью метода наименьших квадратов (МНК). Но, как отмечается в [1], решением этой проблемы занимались и в случае антиробастного оценивания (АО), и в случае оценивания по методу наименьших модулей (МНМ).

В случае оценивания линейной регрессии с помощью МНК применяются следующие методы ОИР: алгоритм полного перебора, техника «ветвей и границ» (поиск вдоль перспективных ветвей), шаговая регрессия, ступенчатая регрессия, алгоритм последовательной замены, лассо Тибширани, метод наименьших углов, метод группового учёта аргументов и т.д. Единственным из этих методов, который гарантирует точное решение задачи ОИР, является алгоритм полного перебора. Все остальные алгоритмы носят, по сути, эвристический характер и позволяют находить «хорошее» решение, в редких случаях даже совпадающее с оптимальным.

Точное решение задачи ОИР в линейной регрессии помимо алгоритма полного перебора может быть найдено еще одним способом, а именно, посредством введения в рассмотрение булевых переменных и последующем решении задачи целочисленного программирования. Стоит отметить, что в научном сообществе принято ошибочно полагать, что этот приём впервые описан в работе [8], на которую имеются множественные ссылки (см., например, [9-13]). Однако гораздо ранее такая задача была сформулирована для случая МНМ в работе С.И. Носкова [14].

При оценивании линейной регрессии по МНМ задача ОИР представляет собой задачу частично-булевого линейного программирования (ЧБЛП) [14], а при оценивании по МНК – задачу частично-булевого квадратичного программирования (ЧБКП). При этом если решение задач ЧБЛП с использованием соответствующих пакетов программ не представляет никаких проблем, то решение задач ЧБКП представляет собой сложную вычислительную проблему, требующую разработки эффективных численных методов её решения. Целью данной работы является сведение задачи ОИР в модели линейной регрессии, оцениваемой с помощью МНК, к задаче ЧБЛП.

Оценки стандартизованной регрессии с помощью МНК. Рассмотрим модель множественной линейной регрессии:

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

где y_i, $i=\overline{1,n}$ – значения зависимой (объясняемой) переменной y;
x_{i1}, x_{i2}, …, x_{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 для оценки неизвестных параметров модели (1) используется метод наименьших квадратов (МНК), суть которого заключается в минимизации суммы квадратов ошибок:

$$\sum_{i=1}^{n}\varepsilon_i^2 \to \min . \tag{2}$$

При этом МНК-оценки линейной регрессии (1) можно найти явно в матричной форме:

$$\hat{\alpha}_{\text{МНК}} = \left(X^T X\right)^{-1} X^T y ,$$

где $y=\begin{pmatrix} y_1 \\ y_2 \\ \cdots \\ y_n \end{pmatrix}$, $X=\begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{pmatrix}$, $\alpha=\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \cdots \\ \alpha_m \end{pmatrix}$,

T – операция транспонирование.

Рассмотрим другой известный подход к определению параметров регрессии (1) [15, 16]. Для этого проведем нормирование (стандартизацию) всех переменных по формулам:

$$v_i=\frac{y_i-\overline{y}}{\sigma_y},\; z_{i1}=\frac{x_{i1}-\overline{x_1}}{\sigma_{x_1}},\; \ldots,\; z_{im}=\frac{x_{im}-\overline{x_m}}{\sigma_{x_m}}, \tag{3}$$

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 стандартизованное уравнение регрессии (1) имеет вид:

$$v_i=\beta_1 z_{i1}+\beta_2 z_{i2}+\ldots+\beta_m z_{im}+u_i , \qquad i=\overline{1,n}, \tag{4}$$

где β_1, …, β_m – стандартизованные коэффициенты регрессии, называемые также бета-параметрами;
u_i, $i=\overline{1,n}$ – стандартизованные ошибки аппроксимации.

Основным достоинством стандартизованных коэффициентов β_1, …, β_m в отличие от коэффициентов α_0, α_1, …, α_m регрессии является то, что, сравнивая их между собой, можно ранжировать факторы по силе их воздействия на результат [15].

Уравнение стандартизованной регрессии в матричной форме имеет вид:

$$v = Z\beta + u, \tag{5}$$

где $v = \begin{bmatrix} v_1 \\ v_2 \\ \cdots \\ v_n \end{bmatrix}$, $Z = \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1m} \\ z_{21} & z_{22} & \cdots & z_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ z_{n1} & z_{n2} & \cdots & z_{nm} \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdots \\ \beta_m \end{bmatrix}$, $u = \begin{bmatrix} u_1 \\ u_2 \\ \cdots \\ u_n \end{bmatrix}$.

Для оценки бета-параметров по МНК воспользуемся формулой:

$$\beta = \left(Z^T Z\right)^{-1} Z^T v. \tag{6}$$

При этом с учётом формул (3) нетрудно показать, что $Z^T Z = K$, где K – матрица коэффициентов корреляции между объясняющими переменными (интеркорреляционная матрица) вида:

$$K = \begin{bmatrix} 1 & r_{x_1x_2} & \cdots & r_{x_1x_m} \\ r_{x_1x_2} & 1 & \cdots & r_{x_2x_m} \\ \cdots & \cdots & \cdots & \cdots \\ r_{x_1x_m} & r_{x_2x_m} & \cdots & 1 \end{bmatrix}, \tag{7}$$

а $Z^T v = h$, где h – вектор коэффициентов корреляции между объясняемой переменной и объясняющими переменными вида:

$$h = \begin{bmatrix} r_{yx_1} \\ r_{yx_2} \\ \cdots \\ r_{yx_m} \end{bmatrix}. \tag{8}$$

По определению коэффициента парной корреляции, все элементы матриц (7) и (8) по модулю не превосходят единицы, а матрица (7) является симметричной относительно главной диагонали.

Тогда формула (6) для расчета бета-коэффициентов примет вид:

$$\beta = K^{-1} h. \tag{9}$$

Умножив равенство (9) на матрицу K слева, получим соотношение:

$$K\beta = h, \tag{10}$$

которое представляет собой систему линейных алгебраических уравнений вида:

$$\begin{cases} \beta_1 + r_{x_1x_2}\beta_2 + r_{x_1x_3}\beta_3 + \ldots + r_{x_1x_m}\beta_m = r_{yx_1}, \\ r_{x_1x_2}\beta_1 + \beta_2 + r_{x_2x_3}\beta_3 + \ldots + r_{x_2x_m}\beta_m = r_{yx_2}, \\ \ldots\ldots\ldots \\ r_{x_1x_m}\beta_1 + r_{x_2x_m}\beta_2 + r_{x_3x_m}\beta_3 + \ldots + \beta_m = r_{yx_m}. \end{cases} \quad (11)$$

Согласно формуле (9), система (11) всегда имеет единственное решение, если определитель матрицы коэффициентов интеркорреляции $|K| \neq 0$, т.е. при отсутствии линейной зависимости (мультиколлинеарности) между объясняющими переменными.

Для оценки адекватности регрессии (1) воспользуемся коэффициентом детерминации:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \varepsilon_i^2}{\sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2}, \quad (12)$$

значения которого изменяются от 0 до 1. Чем ближе его значение к 1, тем теснее связь между зависимой переменной и регрессорами.

В случае линейной зависимости факторов формула коэффициента детерминации представима в виде [15]:

$$R^2 = \sum_{i=1}^{m} r_{yx_i}\beta_i, \quad (13)$$

где β_i, $i = \overline{1,m}$ – рассмотренные выше бета-коэффициенты.

В матричном виде равенство (13) имеет вид:

$$R^2 = h^T\beta. \quad (14)$$

Подставив в это выражение соотношение (9), получим:

$$R^2 = h^T K^{-1} h. \quad (15)$$

Достоинством формулы (15) является то, что значение коэффициента детерминации можно находить еще на начальном этапе моделирования, минуя этап оценивания неизвестных параметров модели. Для этого достаточно построить корреляционную матрицу для всех переменных регрессии.

Сведение задачи ОИР к задаче ЧБЛП. Приведем строгую постановку задачи ОИР. Пусть задана выборка из наблюдений для зависимой переменной y_i, $i = \overline{1,n}$, и для возможных независимых переменных x_{ij}, $i = \overline{1,n}$, $j = \overline{1,l}$. Необходимо, учитывая требование ограниченности числа степеней свободы уравнения, выделить из l возможных регрессоров m переменных на основе некоторого критерия качества, например, минимизируя выбранную функцию потерь (2) для регрессии (1).

Сведем задачу ОИР регрессии (1) к задаче ЧБЛП. При этом известными в ней будем считать коэффициенты парной корреляции для всех переменных, а неизвестными переменными – бета-коэффициенты β_i, $i=\overline{1,l}$. В качестве целевой функции используем линейную зависимость (13) для коэффициента детерминации:

$$\sum_{i=1}^{l} r_{yx_i}\beta_i \to \max. \qquad (16)$$

Следует отметить, что в силу равенства (12) задача (16) равносильна задаче $\sum_{i=1}^{n}\varepsilon_i^2 \to \min$.

Понятно, что если k-я переменная не входит в состав регрессоров стандартизованной регрессии (4), то в матрице K (7) необходимо соответственно вычеркнуть k-ю строку и k-й столбец, а из вектора h (8) требуется исключить k-й элемент. Иначе говоря, при исключении k-й переменной из регрессии в системе (11) необходимо принять коэффициент $\beta_i=0$, и одновременно исключить из неё i-е уравнение. Аналогично нужно поступать при исключении нескольких переменных. Введем булевы переменные δ_i, $i=\overline{1,l}$ по правилу:

$$\delta_i=\begin{cases}1, \text{если i - я стандартизованная переменная входит в регрессию (4);}\\ 0, \text{в противномслучае.}\end{cases}$$

Тогда наложим на бета-коэффициенты следующие линейные ограничения:

$$-(1-\delta_i)M \le K_i\beta - h_i \le (1-\delta_i)M, \quad i=\overline{1,l}, \qquad (17)$$

$$-\delta_i M \le \beta_i \le \delta_i M, \qquad i=\overline{1,l}, \qquad (18)$$

$$\delta_i \in \{0,1\}, \quad i=\overline{1,l}. \qquad (19)$$

В неравенствах (17) K_i это i –я строка интеркорреляционной матрицы K, h_i это i –й элемент вектора h, $K_i\beta$ – матричное произведение, M – заранее выбранное положительное число.

Легко видеть, что если $\delta_i=0$, то согласно (18), коэффициент $\beta_i=0$, т.е. i –я стандартизованная переменная не входит в состав регрессоров, при этом, согласно (17), величина $K_i\beta-h_i$ принимает любое значение, т.е. из системы (11) исключается i –е уравнение. Наоборот, если $\delta_i=1$, то согласно (18), коэффициент β_i может принимать любые значения, т.е. i –я стандартизованная переменная входит в состав регрессоров, при этом, согласно (17), $K_i\beta-h_i=0$, т.е. в систему (11) входит i –е уравнение.

Поскольку регрессоров в уравнении должно быть m, естественным является ограничение

$$\sum_{i=1}^{l} \delta_i = m. \qquad (20)$$

Таким образом, задача ОИР в линейной регрессии, оцениваемой по МНК, сведена к задаче ЧБЛП с целевой функцией (16) и с линейными ограничениями (17) – (20). Для её решения может быть использовано специализированное программное обеспечение.

Стоит отметить, что решение задачи (16) – (20) дает лишь стандартизованные коэффициенты регрессии β_i, $i = \overline{1,l}$. Для перехода к оценкам параметров исходной регрессии (1) α_0, α_1, …, α_l необходимо воспользоваться формулами [15]:

$$\alpha_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}}, \ i = \overline{1,l}, \qquad (21)$$

$$\alpha_0 = \overline{y} - \alpha_1 \overline{x_1} - \alpha_2 \overline{x_2} - ... - \alpha_l \overline{x_l}. \qquad (22)$$

Заключение. Таким образом, задача отбора информативных регрессоров в линейной регрессии, оцениваемой с помощью метода наименьших квадратов, сведена к задаче частично-булевого линейного программирования (16) – (20), решение которой не вызывает никаких затруднений при использовании соответствующих пакетов программ. Подчеркнем, что решение задачи (16) – (20) является оптимальным, т.е. из l факторов осуществляется выбор m переменных для вхождения в линейную регрессию с наибольшим значением критерия детерминации. При этом множество решений задачи (16) – (20) в многомерном пространстве представляет собой множество точек, каждая из которых соответствует конкретной модели, поэтому накладывая ограничения на коэффициенты регрессии можно всегда оставаться в рамках МНК-оценок, которые, при выполнении определенных условий, являются несмещенными, состоятельными и эффективными. Безусловно, сформулированная задача (16) – (20) требует своей реализации в виде разработки специализированного программного комплекса, который будет представлен в дальнейших работах автора.

ЛИТЕРАТУРА

1. Miller, A.J. Subset selection in regression / A.J. Miller. – Chapman & Hall/CRC, 2002. – p. 247.
2. Burnham, K.P. Model selection and multimodel inference: a practical information theoretic approach / K.P. Burnham, D.R. Anderson. – Springer, 2002. – P. 515.
3. Себер, Дж. Линейный регрессионный анализ / Дж. Себер. – М.: Издательство «Мир», 1980. – 456 с.

4. Стрижов, В.В. Методы выбора регрессионных моделей / В.В. Стрижов, Е.А. Крымова. – М.: Вычислительный центр РАН, 2010. – 60 с.
5. Liu, H. Computational methods of feature selection / H. Liu, H. Motoda. – Chapman and Hall/CRC, 2007. – 419 p.
6. Guyon, I. An introduction to variable and feature selection / I. Guyon, A. Elisseeff // Journal of machine learning research, 2003. – Vol. 3. – Pp. 1157-1182.
7. Ивахненко, А.Г. Индуктивный метод самоорганизации моделей сложных систем / А.Г. Ивахненко. – Киев: Наукова думка, 1981. – 296 с.
8. Konno, H. Choosing the best set of variables in regression analysis using integer programming / H. Konno, R. Yamamoto // Journal of Global Optimization, 2009. Vol. 44, no. 2, pp. 272-282.
9. Park, Y.W. Subset selection for multiple linear regression via optimization / Y.W. Park, D. Klabjan // Technical report, 2013. Available from http://www.klabjan.dynresmanagement.com.
10. Tamura, R. Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor / R. Tamura, K. Kobayashi, Y. Takano, R. Miyashiro, K. Nakata, T. Matsui // Optimization online, 2016. Available from http://www.optimization-online.org/DB_HTML/2016/09/5655.html.
11. Chung, S. A mathematical programming approach for integrated multiple linear regression subset selection and validation / S. Chung, Y.W. Park, T. Cheong. arXiv.org, 2017. Available from https://arxiv.org/abs/1712.04543.
12. Miyashiro, R. Mixed integer second-order cone programming formulations for variable selection / R. Miyashiro, Y. Takano // Technical Report, 2013. Available from http://www.me.titech.ac.jp/technicalreport/h25/2013-7.pdf.
13. Miyashiro, R. Subset selection by Mallows' Cp: a mixed integer programming approach / R. Miyashiro, Y. Takano. Technical report, 2014. Available from http://www.me.titech.ac.jp/technicalreport/h26/2014-1.pdf.
14. Носков, С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных / С.И. Носков. – Иркутск: РИЦ ГП «Облинформпечать», 1996. – 321 с.
15. Елисеева И.И. Эконометрика / И.И. Елисеева, С.В. Курышева, Т.В. Костеева и др. – М.: Финансы и статистика, 2007. – 576 с.
16. Фёрстер, Э. Методы корреляционного и регрессионного анализа / Э. Фёрстер, Б. Рёнц. – М.: Финансы и статистика, 1983. – 303 с.

M. P. Bazilevskiy

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

Irkutsk State Transport University,
Irkutsk, Russia

One of the main problems in regression analysis is the problem of choosing the structural specification of the regression model, i.e. the choice of the composition of variables and the mathematical form of the connection between them. In the case of a linear regression model, this task is reduced only to the selection of the most informative regressors. The exact solution of the problem of selecting informative regressors when evaluating linear regression using the least squares method can be obtained either by a complete search algorithm or by introducing Boolean variables into consideration and then solving a very complicated computational problem of partial Boolean quadratic programming. In this paper, the problem of selecting informative regressors in a linear regression estimated using the least squares method is reduced to the problem of partial-boolean linear programming, the solution of which does not cause any difficulties when using the corresponding software packages. The new formulation of the problem assumes the preliminary normalization of all variables for estimating the unknown parameters of the linear regression model in order to find the beta coefficients of the standardized regression. Beta coefficients are determined from the known intercorrelation matrix and the correlation vector between the dependent variable and the independent factors. To assess the adequacy of linear regression, the determination coefficient is applied.

Keywords: regression model, least squares method, selection of informative regressors, partial boolean linear programming problem, standardized regression, correlation coefficient, determination criterion.

REFERENCES

1. Miller A.J. Subset selection in regression. Chapman & Hall/CRC, 2002. 247 p.
2. Burnham K.P., D.R. Anderson. Model selection and multimodel inference: a practical information theoretic approach. Springer, 2002. 515 p.
3. Seber Dzh. Linejnyj regressionnyj analiz. Moscow, Izdatel'stvo «Mir», 1980. 456 p. (in Russian)
4. Strizhov V.V., Krymova E.A. Metody vybora regressionnyh modelej. Moscow, Vychislitel'nyj centr RAN, 2010. 60 p. (in Russian)
5. Liu H, Motoda H. Computational methods of feature selection. Chapman and Hall/CRC, 2007. 419 p.
6. Guyon I., Elisseeff I. An introduction to variable and feature selection // Journal of machine learning research, 2003, vol. 3, pp. 1157-1182.
7. Ivahnenko A.G. Induktivnyj metod samoorganizacii modelej slozhnyh system. Kiev, Naukova dumka, 1981. 296 p. (in Russian)

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., D. Klabjan. Subset selection for multiple linear regression via optimization // Technical report, 2013. Available from URL: http://www.klabjan.dynresmanagement.com.
10. 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 // Optimization online, 2016. Available from URL: http://www.optimization-online.org/DB_HTML/2016/09/5655.html.
11. Chung S., Park Y.W., Cheong T. A mathematical programming approach for integrated multiple linear regression subset selection and validation. arXiv.org, 2017. Available from URL: https://arxiv.org/abs/1712.04543.
12. Miyashiro R., Takano Y. Mixed integer second-order cone programming formulations for variable selection // Technical Report, 2013. Available from URL: http://www.me.titech.ac.jp/technicalreport/h25/2013-7.pdf.
13. Miyashiro R., Takano Y. Subset selection by Mallows' Cp: a mixed integer programming approach. Technical report, 2014. Available from http://www.me.titech.ac.jp/technicalreport/h26/2014-1.pdf.
14. Noskov S.I. Tehnologija modelirovanija ob#ektov s nestabil'nym funkcionirovaniem i neopredelennost'ju v dannyh. Irkutsk: RIC GP «Oblinformpechat'», 1996. 321 p. (in Russian)
15. Eliseeva I.I., Kurysheva S.V., Kosteeva T.V. Jekonometrika. Moscow, Finansy i statistika, 2007. 576 p. (in Russian)
16. Fjorster Je, Rjonc B. Metody korreljacionnogo i regressionnogo analiza. Moscow, Finansy i statistika, 1983. 303 p. (in Russian)