УДК 519.862.6

М. П. Базилевский

ОТБОР ИНФОРМАТИВНЫХ РЕГРЕССОРОВ С УЧЕТОМ МУЛЬТИКОЛЛИНЕАРНОСТИ МЕЖДУ НИМИ В РЕГРЕССИОННЫХ МОДЕЛЯХ КАК ЗАДАЧА ЧАСТИЧНО-БУЛЕВОГО ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

Иркутский государственный университет путей сообщения, Иркутск, Россия

Статья посвящена проблеме отбора информативных регрессоров в линейной регрессионной модели, точное решение которой может быть гарантировано либо полным перебором всех возможных вариантов регрессий, либо решением специальным образом сформулированной задачи математического программирования с булевыми переменными. Часто задача отбора информативных регрессоров решается с использованием лишь одного критерия адекватности, например, минимизируются только ошибки модели. Но в случае оценивания регрессии с помощью метода наименьших квадратов необходимо стремиться не только к увеличению качества аппроксимации, но и к соблюдению условий теоремы Гаусса – Маркова, одним из которых является отсутствие линейной зависимости между объясняющими переменными. Если это условие не выполняется, то говорят, что имеет место мультиколлинеарность. Таким образом, при отборе информативных регрессоров целесообразно решать двухкритериальную задачу – стремиться максимизировать качество аппроксимации и одновременно минимизировать мультиколлинеарность между объясняющими переменными. Поскольку точных количественных критериев для определения наличия / отсутствия мультиколлинеарности не существует, в данной работе на основе известной рекомендации сформулирован критерий верхней границы мультиколлинеарности. С использованием этого критерия предложены четыре возможные постановки задачи отбора информативных регрессоров, каждая из которых сведена к задаче частично-булевого линейного программирования. Для демонстрации предложенного математического аппарата разработана пробная версия специализированного программного комплекса, с помощью которого решена задача моделирования грузооборота Красноярской железной дороги.

Ключевые слова: регрессионная модель, метод наименьших квадратов, мультиколлинеарность, отбор информативных регрессоров, задача частично-булевого линейного программирования.

Введение. Одной из основных проблем в регрессионном анализе является выбор спецификации [1] модели, т.е. выбор состава входящих в неё объясняющих переменных (регрессоров) и математической формы связи между ними и зависимой переменной. Проблема отбора информативных регрессоров (ОИР) в регрессионной модели упоминается в зарубежной литературе как «subset selection» или «feature selection» in regression [2]. Точное решение задачи ОИР возможно только с помощью метода полного перебора всех возможных вариантов регрессий. При этом

осуществлять выбор наилучшего уравнения регрессии желательно не по какому-то одному критерию адекватности, например, минимизируя только суммарные ошибки модели, а, по возможности, по как можно большему числу критериев, отвечающих за самые разные качественные стороны модельного описания исследуемого объекта или процесса. Решение многокритериальной задачи ОИР можно получить с помощью организации «конкурса» моделей [3]. Безусловно, построенная таким образом регрессия будет представлять большую практическую ценность, чем модель, построенная только по одному критерию адекватности.

При оценивании регрессионной модели с помощью метода наименьших квадратов (МНК), одной из его предпосылок является линейная независимость объясняющих переменных. Если это условие не выполняется, то говорят, что имеет место мультиколлинеарность [1, 4, 5]. При этом мультиколлинеарность бывает полной и частичной. В случае полной мультиколлинеарности одна из объясняющих переменных является линейной комбинацией остальных, что приводит к невозможности определения оценок регрессии. На практике чаще приходится сталкиваться с частичной мультиколлинеарностью, когда объясняющие переменные тесно коррелируют между собой. В этом случае МНК-оценки найти можно, но по ним практически невозможно оценивать отдельное влияние каждой объясняющей переменной на зависимую переменную. В связи с этим, при решении конкретных многокритериальных задач ОИР множество известных критериев адекватности моделей [3] целесообразно расширять неким критерием наличия / отсутствия мультиколлинеарности, который будет рассмотрен ниже.

В работе [3] впервые показано, что задача ОИР в линейной регрессии, оцениваемой с помощью метода наименьших модулей, может быть сведена к задаче частично-булевого линейного программирования (ЧБЛП), которая упоминается в зарубежной литературе как «mixed 0-1 integer linear programming» (MILP). В последнее время решению задач ОИР в регрессионном анализе методами математического программирования уделяется особое внимание (см., например, [6-8]), что связано с развитием вычислительной техники и с совершенствованием алгоритмов решения этих задач. Вопросам мультиколлинеарности с использованием методов математического программирования посвящены работы [9, 10], в которых задача ОИР сводится к задаче частично-булевого квадратичного программирования «mixed 0-1 integer quadratic programming» (MIQP), решение которой вызывает определенные трудности. Поскольку в работе [11] автору удалось свести задачу ОИР в линейной регрессии, оцениваемой с помощью МНК, к задаче ЧБЛП, то целью данной статьи является интеграция в эту задачу критерия наличия / отсутствия мультиколлинеарности.

Выявление мультиколлинеарности и критерий её верхней границы. Рассмотрим модель множественной линейной регрессии:

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

где n – объем выборки;

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

x_{i1}, x_{i2}, …, x_{im}, $i = \overline{1,n}$ – значения m объясняющих (независимых) переменных x_1, x_2, …, x_m;

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

α_0, α_1, …, α_m – неизвестные параметры, оценки которых будем находить с помощью МНК.

Как отмечается в работе [5], точных количественных критериев для определения наличия или отсутствия мультиколлинеарности не существует. Тем не менее, имеются некоторые эвристические рекомендации по её выявлению, которые можно найти в работах [1, 4, 5]. Одним из таких способов является анализ матрицы K парных коэффициентов корреляции между объясняющими переменными. Считается [4], что если коэффициент корреляции по абсолютной величине превосходит значение 0,75-0,8, то это свидетельствует о присутствии мультиколлинеарности. Но этот метод позволяет лишь в первом приближении и достаточно поверхностно судить о наличии / отсутствии мультиколлинеарности в данных [4]. Более глубокое изучение вопроса достигается с помощью расчета значений коэффициентов детерминации каждой из m объясняющих переменных модели (1) по всем оставшимся $(m-1)$ факторам. Высокое значение коэффициента детерминации [5] (обычно больше 0,6) свидетельствует о наличии мультиколлинеарности.

Иногда для обнаружения мультиколлинеарности вычисляются факторы «вздутия» (инфляции) дисперсии (variance inflation factor, VIF) [12] по формулам:

$$VIF_j = \frac{1}{1 - R_j^2}, \qquad j = \overline{1,m},$$

где R_j^2 – коэффициент детерминации в регрессии, представляющей собой зависимость j-й объясняющей переменной от всех остальных объясняющих переменных. Если в данных присутствует слабая мультиколлинеарность, то $R_j^2 \to 0$, следовательно, $VIF_j \to 1$, а в случае сильной мультиколлинеарности $R_j^2 \to 1$ и, следовательно, $VIF_j \to \infty$. Считается, что если значение фактора «вздутия» дисперсии больше 10, что

соответствует значениям коэффициента детерминации $R_j^2 \geq 0{,}9$, то можно утверждать, что в модели присутствует мультиколлинеарность.

В дальнейшем будем называть линейную модель зависимости j-й объясняющей переменной от остальных $(m-1)$ объясняющих переменных j-й вспомогательной регрессией. На основе вышеперечисленных рекомендаций по выявлению мультиколлинеарности, введем критерий наличия / отсутствия мультиколлинеарности:

$$\Psi = \max\left\{R^2_{1|2,3,\ldots,\mathrm{m}}, R^2_{2|1,3,\ldots,m}, \ldots, R^2_{m|1,2,\ldots,m-1}\right\}, \quad (2)$$

где $R^2_{1|2,3,\ldots,m}$, …, $R^2_{m|1,2,\ldots,m-1}$ – коэффициенты детерминации вспомогательных регрессий. Таким образом, смысл критерия (2) заключается в том, что он показывает верхнюю границу мультиколлинеарности между объясняющими переменными в линейной регрессии. Область возможных значений критерия (2) $\Psi \in [0;1]$. Чем меньше его значение, тем слабее эффект мультиколлинеарности. Четкой границы между областями присутствия и отсутствия мультиколлинеарности для этого критерия не существует, но, согласно [5], значение $\Psi \geq 0{,}6$ может свидетельствовать о её наличии. Будем называть критерий (2) верхней границей мультиколлинеарности.

Отметим, что критерий (2) справедлив для количества объясняющих переменных $m \geq 2$, т. е. только для множественных регрессионных моделей. При этом, если $m = 2$, то коэффициенты детерминации вспомогательных регрессий равны, т. е. $R^2_{1|2} = R^2_{2|1}$, следовательно, $\Psi = R^2_{1|2} = R^2_{2|1}$.

Аналогично можно ввести следующий критерий верхней границы мультиколлинеарности:

$$\Psi^* = \max\left\{VIF_1, VIF_2, \ldots, VIF_m\right\},$$

где VIF_j, $j = \overline{1,m}$ – факторы «вздутия» дисперсии. Область возможных значений этого критерия $\Psi^* \in [1;\infty]$. Чем меньше его значение, тем слабее эффект мультиколлинеарности. Значение $\Psi^* \geq 10$ может свидетельствовать о наличии мультиколлинеарности.

Очевидно, что одновременное использование критериев верхней границы мультиколлинеарности Ψ и Ψ^* в силу их определения избыточно, поэтому в дальнейшем будем пользоваться критерием (2).

Сведение задачи ОИР с учетом критерия верхней границы мультиколлинеарности к задаче ЧБЛП. Строгая постановка задачи ОИР может быть сформулирована следующим образом. Пусть задана выборка из n наблюдений для объясняемой переменной y_i, $i = \overline{1,n}$, и для

объясняющих переменных x_{ij}, $i=\overline{1,n}$, $j=\overline{1,l}$. Необходимо выделить из l возможных регрессоров m переменных на основе некоторого критерия качества.

В работе [11] задача ОИР в регрессионной модели, оцениваемой с помощью МНК, была сведена к задаче ЧБЛП. Кратко рассмотрим процедуру такого сведения.

Предварительно проведем нормирование зависимой переменной y и независимых переменных x_1, x_2, …, x_l по формулам:

$$y_i^* = \frac{y_i - \overline{y}}{\sigma_y}, \qquad i=\overline{1,n},$$

$$x_{ij}^* = \frac{x_{ij} - \overline{x_j}}{\sigma_{x_j}}, \qquad i=\overline{1,n}, \qquad j=\overline{1,l},$$

где y^*, x_j^*, $j=\overline{1,l}$ – стандартизованные переменные, для которых средние значения равны 0, а суммы квадратов равны 1.

Тогда регрессионной модели зависимости y от x_1, x_2, …, x_l можно поставить в соответствие стандартизованное уравнение регрессии:

$$y_i^* = \beta_1 x_{i1}^* + \beta_2 x_{i2}^* + \ldots + \beta_l x_{il}^* + \varepsilon_i^*, \qquad i=\overline{1,n}, \tag{3}$$

где β_j, $j=\overline{1,l}$ – стандартизованные коэффициенты регрессии (бета-коэффициенты); ε_i^*, $i=\overline{1,n}$ – ошибки аппроксимации.

МНК-оценки $\hat{\beta}_j$, $j=\overline{1,l}$, стандартизованной регрессии (3) находятся с помощью решения системы линейных алгебраических уравнений [1, 11]:

$$K\hat{\beta} = h, \tag{4}$$

где K – матрица коэффициентов корреляции между объясняющими переменными (интеркорреляционная матрица):

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

h – вектор коэффициентов корреляции объясняющих переменных с объясняемой переменной:

$$h = \begin{bmatrix} r_{yx_1} & r_{yx_2} & \cdots & r_{yx_m} \end{bmatrix}^T.$$

Тогда задача ОИР может быть сформулирована следующим образом [11]:

$$R^2 = \sum_{i=1}^{l} r_{yx_i}\beta_i \to \max, \tag{5}$$

$$-(1-\delta_i)M \le \sum_{j=1}^{l} r_{x_i x_j}\beta_j - r_{yx_i} \le (1-\delta_i)M, \qquad i=\overline{1,l}, \tag{6}$$

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

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

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

где R^2 – критерий детерминации; M – заранее выбранное большое положительное число; δ_i, $i=\overline{1,l}$ – булевы переменные, заданные по правилу:

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

Таким образом, решение задачи (5) с $(4l+1)$ линейными ограничениями (6), (7), (9) и условиями целочисленности переменных (8) гарантирует точное решение задачи ОИР на основе критерия детерминации в регрессионной модели, оцениваемой с помощью МНК.

Перейдем к сведению задачи ОИР в линейной регрессии по коэффициенту детерминации R^2 и критерию верхней границы мультиколлинеарности Ψ к задаче ЧБЛП. Для этого предложены 4 возможные постановки такой задачи.

Задача 1. Необходимо выделить из l возможных регрессоров m переменных на основе критерия детерминации R^2 так, чтобы критерий верхней границы мультиколлинеарности Ψ удовлетворял условию:

$$\Psi = \max\left\{R^2_{1|2,3,\dots,l}, R^2_{2|1,3,\dots,l}, \dots, R^2_{l|1,2,\dots,l-1}\right\} \le r, \tag{10}$$

где r – число из интервала $[0;1]$. Если $r=1$, то имеем стандартную задачу ОИР по критерию детерминации, а если $r=0$, то имеем задачу ОИР по критерию детерминации с полным отсутствием мультиколлинеарности, которая в редких случаях может иметь решение.

Для стандартизованной регрессии (3) вспомогательные стандартизованные модели зависимости объясняющей переменной x_k, $k=\overline{1,l}$, от остальных регрессоров имеют вид:

$$x^*_{i1} = \beta_{11}x^*_{i2} + \beta_{12}x^*_{i3} + \dots + \beta_{1,l-1}x^*_{il} + u_{i1},$$
$$x^*_{i2} = \beta_{21}x^*_{i1} + \beta_{22}x^*_{i3} + \dots + \beta_{2,l-1}x^*_{il} + u_{i2},$$
$$\dots$$
$$x^*_{il} = \beta_{l,1}x^*_{i1} + \beta_{l,2}x^*_{i2} + \dots + \beta_{l,l-1}x^*_{i,l-1} + u_{il},$$

где β_{kj}, $k=\overline{1,l}$, $j=\overline{1,l-1}$ – бета-коэффициенты вспомогательных стандартизованных регрессий; u_k, $k=\overline{1,l}$ – ошибки аппроксимации.

Коэффициенты детерминации R_k^2, $k=\overline{1,l}$, вспомогательных регрессий по аналогии с выражением (5), можно найти по формулам:

$$R_k^2 = \sum_{j=1}^{l-1} r_{x_k x_{s_j^*}} \beta_{kj}, \qquad k=\overline{1,l}, \tag{11}$$

где $S^* = \left\{ s^* \mid s^* \in S \setminus \{k\} \right\}$, $S = \{1,2,...,l\}$.

Бета-коэффициенты β_{kj}, $k=\overline{1,l}$, $j=\overline{1,l-1}$, вспомогательных стандартизованных регрессий находятся подобно выражению (4) с помощью решения системы линейных алгебраических уравнений:

$$\sum_{j=1}^{l-1} r_{x_i x_{s_j^*}} \beta_{kj} - r_{x_k x_i} = 0, \qquad k=\overline{1,l}, \qquad i \in S^*. \tag{12}$$

Формулы (11) и (12) справедливы при оценивании стандартизованной регрессии (3) со всеми l объясняющими переменными. Для того чтобы в задаче ЧБЛП из этих l регрессоров выделялось m переменных необходимо представить выражения (11) и (12) в виде линейных ограничений с учётом булевых переменных δ_i, $i=\overline{1,l}$.

Зададим ограничения на коэффициенты детерминации R_k^2, $k=\overline{1,l}$, вспомогательных регрессий с учетом условия (10):

$$R_k^2 = \sum_{j=1}^{l-1} r_{x_k x_{s_j^*}} \beta_{kj} - (1-\delta_k) M \le r, \qquad k=\overline{1,l}, \tag{13}$$

где M – заранее выбранное большое положительное число. Если $\delta_k = 0$, т. е. k-я объясняющая переменная не входит в модель, то неравенство (13) всегда выполняется, а если $\delta_k = 1$, то имеем ограничение на k-й коэффициент детерминации.

Наложим на бета-коэффициенты β_{kj}, $k=\overline{1,l}$, $j=\overline{1,l-1}$, вспомогательных стандартизованных регрессий линейные ограничения так, как это сделано в работе [11]:

$$-(1-\delta_i) M \le \sum_{j=1}^{l-1} r_{x_i x_{s_j^*}} \beta_{kj} - r_{x_k x_i} \le (1-\delta_i) M, \qquad k=\overline{1,l}, \qquad i \in S^*, \tag{14}$$

$$-\delta_{s_j^*} M \le \beta_{kj} \le \delta_{s_j^*} M, \qquad k=\overline{1,l}, \qquad j=\overline{1,l-1}. \tag{15}$$

Если $\delta_{s_j^*} = 0$, то в ограничениях (15) бета-коэффициенты $\beta_{kj} = 0$, т. е. во все вспомогательные стандартизованные регрессии не будут входить объясняющие переменные $x_{s_j^*}$, $j=\overline{1,l-1}$. В противном случае $\delta_k = 1$ и бета-коэффициенты β_{kj} принимают любые значения. При этом, если $\delta_i = 0$, то в ограничениях (14) для k-й объясняющей переменной исключается i-е уравнение, в противном случае – не исключается.

Таким образом, решение задачи ЧБЛП (5) с $\left(4l^2+l+1\right)$ линейными ограничениями (6), (7), (9), (13) – (15) и условиями целочисленности переменных (8) позволяет осуществить ОИР в модели с наибольшим значением коэффициента детерминации R^2 и величиной верхней границы мультиколлинеарности Ψ, не превосходящей заданного значения r. Недостатком такой постановки задачи является то, что она может и вовсе не иметь решений.

Задача 2. Необходимо выделить из l возможных регрессоров m переменных только на основе критерия верхней границы мультиколлинеарности Ψ.

В этом случае целевую функцию (5) в предыдущей задаче необходимо заменить на функционал:

$$r \to \min, \tag{16}$$

где r – новая переменная, представляющая собой величину, превышающую коэффициенты детерминации R_k^2 вспомогательных стандартизованных регрессий, и равная максимальному коэффициенту детерминации для такого номера k, для которого неравенство (13) обращается в равенство.

Задача (16) с ограничениями (6) – (9), (13) – (15) всегда совместна и гарантирует ОИР в модели с минимальным эффектом мультиколлинеарности между объясняющими переменными. Недостатком такой постановки задачи является рассмотрение критерия верхней границы мультиколлинеарности Ψ в качестве альтернативного по отношению к коэффициенту детерминации R^2, поскольку, как справедливо отмечено в [3], наиболее важной интегрирующей характеристикой адекватности модели исследуемому объекту или процессу является все-таки точность аппроксимации.

Задача 3. Пусть R^{2*} – найденное максимальное значение коэффициента детерминации при решении задачи ОИР (5) с линейными ограничениями (6) – (9). Предположим, что исследователь может назначить некоторую величину ΔR^2, на которую допустимо уменьшение значение R^{2*} без существенного ухудшения качества аппроксимации. Необходимо выделить из l возможных регрессоров m переменных на основе критерия верхней границы мультиколлинеарности Ψ, при условии, что $R^2 \geq R^{2*} - \Delta R^{2*}$.

Отметим, что подобная задача для коррекции критерия согласованности поведения была сформулирована гораздо ранее в работе [3].

С учетом условия задачи введем ограничение на коэффициент детерминации R^2 стандартизованной регрессии (3):

$$R^2 = \sum_{i=1}^{l} r_{yx_i} \beta_i \geq R^{2*} - \Delta R^{2*}. \quad (17)$$

Задача (16) с ограничениями (6) – (9), (13) – (15), (17) всегда совместна и гарантирует ОИР в линейной регрессии по критерию верхней границы мультиколлинеарности Ψ, при условии, что коэффициент детерминации R^2 будет не меньше величины $R^{2*} - \Delta R^{2*}$.

*Задача 4***.** Необходимо выделить из l возможных регрессоров m переменных на основе линейной свёртки критериев детерминации R^2 и верхней границы мультиколлинеарности Ψ:

$$D = (1-\lambda)R^2 - \lambda r,$$

где λ – заданное число из интервала $[0;1]$.

Для такой задачи целевая функция примет вид:

$$D = (1-\lambda)\sum_{i=1}^{l} r_{yx_i} \beta_i - \lambda r \to \max. \quad (18)$$

Таким образом, решение задачи ЧБЛП (18) с линейными ограничениями (6) – (9), (13) – (15) позволяет осуществить ОИР в модели с наибольшим значением критерия D. Недостатком данной постановки задачи является то, что не ясно, какое именно выбирать значение параметра λ. Для этого можно разбить интервал $[0;1]$ точками на заданное число одинаковых отрезков, и для каждой точки решить задачу (18) с ограничениями (6) – (9), (13) – (15). Это позволит сформировать множество Парето в пространстве критериев (R^2, Ψ), которое исследователь может использовать для выбора оптимального на его взгляд соотношения пары «детерминация – мультиколлинеарность» в регрессии. Такой прием представляет собой достаточно трудоёмкую вычислительную задачу.

Пример. Для демонстрации возможностей и корректности предложенного в данной работе математического аппарата решалась задача моделирования грузооборота Красноярской железной дороги (КЖД). Для этого была использована статистическая информация из работы [13] за период с 2000 г. по 2015 г. по следующим показателям КЖД: y – грузооборот, млн. т. км; x_1 – прием порожних вагонов, штук; x_2 – динамическая нагрузка, т. км / км; x_3 – среднесуточный пробег локомотива, км; x_4 – эксплуатируемый парк локомотивов, штук; x_5 – техническая скорость локомотивов, км / час.

По исходным статистическим данным были построены все возможные двухфакторные и трехфакторные линейные регрессионные модели, общее количество которых равно 20. Для каждой регрессии были найдены значения критериев детерминации R^2 и нижней границы

мультиколлинеарности Ψ. Эти значения приведены в таблице 1. Они были использованы в качестве эталонных для проверки корректности решенных задач ЧБЛП.

Таблица 1 – Критерии адекватности регрессионных моделей

Модель	Регрессоры	R^2	Ψ	Модель	Регрессоры	R^2	Ψ
1	1, 2	0,8804	0,7756	11	1, 2, 3	0,8886	0,8954
2	1, 3	0,8878	0,1308	12	1, 2, 4	0,9632	0,8331
3	1, 4	0,9362	0,7856	13	1, 2, 5	0,8967	0,9008
4	1, 5	0,8958	0,4977	14	1, 3, 4	0,9798	0,8389
5	2, 3	0,8184	0,0003	15	1, 3, 5	0,8958	0,7566
6	2, 4	0,8891	0,7643	16	1, 4, 5	0,9633	0,8396
7	2, 5	0,8451	0,1938	17	2, 3, 4	0,9730	0,7826
8	3, 4	0,9722	0,0227	18	2, 3, 5	0,8542	0,7149
9	3, 5	0,6009	0,5322	19	2, 4, 5	0,9515	0,8129
10	4, 5	0,9515	0,3397	20	3, 4, 5	0,9734	0,7611

Так как предложенные в данной работе задачи ЧБЛП содержат достаточно большое количество переменных и ограничений, то для их решения была разработана пробная версия специализированного программного комплекса, вычислительным ядром которого является пакет LPSolve. С помощью данного программного обеспечения осуществлялось решение рассмотренных выше четырех задач. При этом большое положительное число M во всех задачах задавалось равным 100.

Задача 1. Результаты решения этой задачи в зависимости от значений параметра r для двухфакторных регрессий представлены в таблице 2, для трехфакторных – в таблице 3.

Таблица 2 – Результаты решения задачи 1 для двухфакторных регрессий

r	R^2	Ψ	Регрессоры
1	0,9722	0,0227	3, 4
0,9	0,9722	0,0227	3, 4
0,1	0,9722	0,0227	3, 4
0,01	0,8184	0,0003	2, 3
0,0005	0,8184	0,0003	2, 3
0,0001	–	–	–

Таблица 3 – Результаты решения задачи 1 для трехфакторных регрессий

r	R^2	Ψ	Регрессоры
1	0,9798	0,8389	1, 3, 4
0,9	0,9798	0,8389	1, 3, 4
0,8	0,9734	0,7611	3, 4, 5

0,75	0,8542	0,7149	2, 3, 5
0,7	–	–	–

Как видно по таблицам 2 и 3, при $r=1$ имеем стандартный ОИР по коэффициенту детерминации R^2. Наилучшая с его точки зрения двухфакторная регрессия с коэффициентом $R^2=0{,}9722$ содержит регрессоры x_3 и x_4, а трехфакторная с коэффициентом $R^2=0{,}9798$ – регрессоры x_1, x_3 и x_4. Результаты полностью согласуются с данными таблицы 1.

С постепенным уменьшением значений параметра r для двухфакторной регрессии (см. таблицу 2) решение задачи изменилось только при $r=0{,}01$: критерий детерминации $R^2=0{,}8184$, регрессоры – x_2 и x_3. При этом верхняя граница мультиколлинеарности достигла своего наименьшего значения $\Psi=0{,}0003$. Дальнейшее уменьшение значений параметра r до порогового значения $\Psi=0{,}0003$ решение задачи не меняло, а при задании $r=0{,}0001$, т. е. переступив через этот порог и потребовав построение двухфакторной модели практически без мультиколлинеарности, решение задачи, естественно, получено не было.

Для трехфакторных моделей (см. таблицу 3) с уменьшением значений параметра r решение задачи изменялось 2 раза: при $r=0{,}8$ и при $r=0{,}75$. Таким образом, сравнивая все полученные для задачи 1 результаты с данными таблицы 1, можно сделать вывод о корректности предложенного математического аппарата.

Задача 2. В этом случае двухфакторной моделью с наименьшим эффектом мультиколлинеарности оказалась модель с регрессорами x_2, x_3, коэффициентом детерминации $R^2=0{,}8184$, верхней границей мультиколлинеарности $\Psi=0{,}0003$. Трехфакторной моделью с наименьшим эффектом мультиколлинеарности оказалась модель с регрессорами x_2, x_3, x_5, коэффициентом детерминации $R^2=0{,}8542$, верхней границей мультиколлинеарности $\Psi=0{,}7149$. Те же самые результаты можно видеть в таблице 1.

Задача 3. Максимальное значение коэффициента детерминации для двухфакторных регрессий $R^{2*}=0{,}9722$. При этом верхняя граница мультиколлинеарности $\Psi=0{,}0227$. В этом случае пытаться снижать мультиколлинеарность нет смысла, поскольку она практически отсутствует.

Максимальное значение коэффициента детерминации для трехфакторных регрессий $R^{2*}=0{,}9798$. При этом верхняя граница мультиколлинеарности $\Psi=0{,}8389$. Допустим, что ради снижения эффекта

мультиколлинеарности исследователь готов пойти на уменьшение коэффициента детерминации R^{2*} на величину $\Delta R^{2*} = 0{,}2$. Решением такой задачи ОИР будет модель с регрессорами x_2, x_3, x_5, с коэффициентом детерминации $R^2 = 0{,}8542$ и с верхней границей мультиколлинеарности $\Psi = 0{,}7149$.

Задача 4. Варьируя значения параметра λ в функционале (18) от 0 до 1 с шагом 0,01, для двухфакторной регрессии в пространстве критериев (R^2, Ψ) было сформировано множество Парето:

$$(0{,}9722;\ 0{,}0227),$$

$$(0{,}8184;\ 0{,}0003).$$

Множество Парето для трехфакторной регрессии:

$$(0{,}9798;\ 0{,}8389),$$

$$(0{,}9734;\ 0{,}7611),$$

$$(0{,}8542;\ 0{,}7149).$$

Окончательный выбор наилучшей регрессионной модели с позиции «качество – мультиколлинеарность» остается за исследователем.

Заключение. Таким образом, в данной работе для обнаружения мультиколлинеарности между объясняющими переменными в регрессионной модели был сформулирован критерий верхней границы мультиколлинеарности. С использованием этого критерия и коэффициента детерминации предложены четыре возможные постановки, как однокритериальных, так и двухкритериальных, задач отбора информативных регрессоров. При этом каждая такая задача сведена к соответствующей задаче частично-булевого линейного программирования. Для демонстрации предложенного математического аппарата решена задача моделирования грузооборота Красноярской железной дороги.

ЛИТЕРАТУРА

1. Эконометрика / И.И. Елисеева, С.В. Курышева, Т.В. Костеева и др. – М.: Финансы и статистика, 2007. – 576 с.
2. Miller A.J. Subset selection in regression / A.J. Miller. – Chapman & Hall/CRC, 2002. – p. 247.

3. Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных / С.И. Носков. – Иркутск: РИЦ ГП «Облинформпечать», 1996. – 321 с.
4. Айвазян С.А. Методы эконометрики / С.А. Айвазян. – М. : Магистр : ИНФРА-М, 2010. – 512 с.
5. Кремер Н.Ш. Эконометрика / Н.Ш. Кремер, Б.А. Путко. – М.: ЮНИТИ-ДАНА, 2002. – 311 с.
6. 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.
7. 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.
8. 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.
9. Best subset selection for eliminating multicollinearity / R. Tamura, K. Kobayashi, Y. Takano, R. Miyashiro, K. Nakata, T. Matsui // Journal of the Operations Research Society of Japan. Vol. 60, No. 3, 2017, pp. 321-336.
10. 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. 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. – Воронеж, 2018. – Т. 6. – № 1 – URL: https://moit.vivt.ru/wp-content/uploads/2018/01/Bazilevskiy_1_1_18.pdf (дата обращения 10.05.2018).
12. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных. – URL: http://www.machinelearning.ru/wiki/index.php?title=Фактор_инфляции_регрессии (дата обращения 10.05.2018).
13. Среднесрочное прогнозирование эксплуатационных показателей функционирования Красноярской железной дороги / М.П. Базилевский, И.П. Врублевский, С.И. Носков, И.С. Яковчук // Фундаментальные исследования. – 2016. – №10(3). – С.471-476.

M. P. Bazilevskiy

SUBSET SELECTION IN REGRESSION MODELS WITH CONSIDERING MULTICOLLINEARITY AS A TASK OF MIXED 0-1 INTEGER LINEAR PROGRAMMING

Irkutsk State Transport University,
Irkutsk, Russia

The article is devoted to the problem of subset selection in linear regression model, the exact solution of which guarantees either a full search of all possible regressions or a solution of a specially formulated mathematical programming problem with Boolean variables. Often the problem of subset selection is solved using only one criterion of adequacy, for example, only model errors are minimized. But in the case of estimating regression using ordinary least squares, it is necessary to strive not only to increase the quality of the approximation, but also to observe the conditions of the Gauss-Markov theorem, one of which is the absence of a linear dependence between the explanatory variables. If this condition is not satisfied, then it is said that multicollinearity takes place. Thus, when selecting informative regressors, it is expedient to solve the two-criteria problem - to strive to maximize the quality of approximation and at the same time minimize the multicollinearity between explanatory variables. Since there are no exact quantitative criteria for determining the presence / absence of multicollinearity, in this paper, based on the well-known recommendation, a criterion for the upper bound of multicollinearity is formulated. Using this criterion, four possible statements of the two-criteria problem of subset selection are proposed, each of which is reduced to task of mixed 0-1 integer linear programming. To demonstrate the proposed mathematical apparatus, a trial version of a specialized software package was developed, with the help of which the task of modeling the freight turnover of the Krasnoyarsk railroad was solved.

Keywords: regression model, ordinary least squares, multicollinearity, subset selection in regression, task of mixed 0-1 integer linear programming.

REFERENCES

1. Jekonometrika / Eliseeva I.I., Kurysheva S.V., Kosteeva T.V. Moscow, Finansy i statistika, 2007. 576 p. (in Russian)
2. Miller A.J. Subset selection in regression / A.J. Miller. Chapman & Hall/CRC, 2002. 247 p.
3. 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)
4. Ajvazjan S.A. Metody jekonometriki / S.A. Ajvazjan. Moscow : Magistr : INFRA-M, 2010. 512 p. (in Russian)
5. Kremer N.Sh. Jekonometrika / N.Sh. Kremer, B.A. Putko. Moscow : JuNITI-DANA, 2002. 311 p. (in Russian)

6. 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.
7. 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.
8. 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.
9. Best subset selection for eliminating multicollinearity / R. Tamura, K. Kobayashi, Y. Takano, R. Miyashiro, K. Nakata, T. Matsui // Journal of the Operations Research Society of Japan. Vol. 60, No. 3, 2017, pp. 321-336.
10. 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. Bazilevskij M.P. Svedenie zadachi otbora informativnyh regressorov pri ocenivanii linejnoj regressionnoj modeli po metodu naimen'shih kvadratov k zadache chastichno-bulevogo linejnogo programmirovanija // Modelirovanie, optimizacija i informacionnye tehnologii. Voronezh, 2018. Vol. 6, no. 1. Available from https://moit.vivt.ru/wp-content/uploads/2018/01/Bazilevskiy_1_1_18.pdf. (in Russian)
12. Professional'nyj informacionno-analiticheskij resurs, po-svjashhennyj mashinnomu obucheniju, raspoznavaniju obrazov i intellektual'nomu analizu dannyh. Available from http://www.machinelearning.ru/wiki/index.php?title=Фактор_инфляции_регрессии. (in Russian)
13. Srednesrochnoe prognozirovanie jekspluatacionnyh pokazatelej funkcionirovanija Krasnojarskoj zheleznoj dorogi / M.P. Bazilevskij, I.P. Vrublevskij, S.I. Noskov, I.S. Jakovchuk // Fundamental'nye issledovanija. 2016. Vol. 10, no. 3, pp. 471-476. (in Russian)