МАТЕМАТИЧЕСКИЕ МЕТОДЫ СИСТЕМНОГО АНАЛИЗА, УПРАВЛЕНИЯ И МОДЕЛИРОВАНИЯ

УДК 519.862.6 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait.2021.3/3731
Поступила в редакцию 19.02.2021
Подписана в печать 20.11.2021

ОТБОР ЗНАЧИМЫХ ПО КРИТЕРИЮ СТЬЮДЕНТА ИНФОРМАТИВНЫХ РЕГРЕССОРОВ В ОЦЕНИВАЕМЫХ С ПОМОЩЬЮ МНК РЕГРЕССИОННЫХ МОДЕЛЯХ КАК ЗАДАЧА ЧАСТИЧНО-БУЛЕВОГО ЛИНЕЙНОГО ПРОГРАММИРОВАНИЯ

© 2021 М. П. Базилевский✉

Иркутский государственный университет путей сообщения
ул. Чернышевского, 15, 664074 Иркутск, Российская Федерация

Аннотация. Настоящая статья посвящена проблеме отбора фиксированного числа информативных регрессоров в оцениваемых с помощью метода наименьших квадратов линейных регрессионных моделях. В современных научных работах для решения этой задачи применяется хорошо развитый за последние годы аппарат целочисленного математического программирования. В большинстве этих работ задача отбора регрессоров формализована в виде задач частично-квадратичного линейного программирования. Относительно недавно начали появляться статьи, в которых авторы стремятся сформулировать единую задачу математического программирования, которая параллельно с отбором факторов гарантирует построение регрессии, удовлетворяющей различным статистических тестам. Данная работа является логическим продолжением предыдущих статей автора, в которых задача отбора информативных регрессоров формализована в виде задачи частично-булевого линейного, а не квадратичного, программирования. Ранее уже были рассмотрены способы контроля в этой задаче степени мультиколлинеарности. В данной статье с помощью известного подхода к определению наблюдаемых значений t-критерия Стьюдента, основанного на вычислении частных F-критериев, в упомянутую задачу частично-булевого линейного программирования были интегрированы линейные ограничения на степень значимости коэффициентов регрессии. Сформулирована двухкритериальная задача, позволяющая строить модель с позиции соотношения «качество — значимость», и трехкритериальная задача, осуществляющая построение регрессии с позиции соотношения «качество — мультиколлинеарность — значимость». Успешно проведены вычислительные эксперименты, подтверждающие корректность предложенного математического аппарата.

Ключевые слова: регрессионная модель, стандартизованная регрессия, отбор информативных регрессоров, мультиколлинеарность, t-критерий Стьюдента, коэффициент детерминации, задача частично-булевого линейного программирования.

ВВЕДЕНИЕ

При построении регрессионной модели зачастую приходится решать задачу выделения из заданного множества объясняющих переменных подмножества наиболее информативных из них. Процедура отбора информативных регрессоров (ОИР) известна в зарубежной литературе, как «feature selection», «variable selection», «attribute selection» или «subset selection». Для решения этой проблемы к настоящему времени разработано большое количество разнообразных математиче-

✉ Базилевский Михаил Павлович
e-mail: mik2178@yandex.ru

ских методов, современный обзор которых можно найти в работе [1].

За последние пару десятков лет была существенно развита технология решения задач частично-целочисленного программирования. Вместе с увеличением вычислительных мощностей программ-решателей целочисленных задач были исследованы и разработаны новые эффективные алгоритмы их решения. Поэтому на сегодняшний день повышенное внимание в литературе уделяется решению задач ОИР с использованием аппарата частично-целочисленного программирования.

Трудно ответить на вопрос, кто первый сформулировал задачу ОИР в регрессионном анализе как задачу математического программирования. Так, например, в отечественной литературе сведение задачи ОИР для линейной регрессии, оцениваемой с помощью метода наименьших модулей (МНМ), к задаче частично-булевого линейного программирования (ЧБЛП) для фиксированного числа регрессоров можно найти в монографии [2]. В ней же формализована задача построения с помощью МНМ линейной регрессии с минимальным количеством регрессоров и со средней относительной ошибкой аппроксимации, не превышающей заданное значение. Позднее была сформулирована задача ОИР с одновременной корректировкой МНМ-оценок линейной регрессии по критерию «согласованности» поведения [3].

В зарубежной литературе первой статьей, на которую ссылается большинство авторов при обзоре методов решения задач ОИР, является работа [4]. В ней задача ОИР для линейной регрессии, оцениваемой с помощью метода наименьших квадратов (МНК), для фиксированного числа регрессоров сформулирована в виде задачи частично-булевого квадратичного программирования (ЧБКП). На основе [4] в [5] разработан алгоритм, существенно сокращающий время вычислений. Работа [6] посвящена задачам ОИР на основе скорректированного критерия детерминации, критерия Акаике и Шварца. Эти задачи не требуют фиксации числа отбираемых переменных. В работе [7] представлена задача ОИР на основе критерия Мэллоуза. В [8] рассмотрен ОИР для случая, когда число переменных больше, чем объем выборки.

В 2020 году одновременно вышли 2 независимых зарубежных работы [9] и [10], посвященные ОИР с использованием аппарата математического программирования. В [9] авторы используют понятие «regression diagnostics» (регрессионная диагностика). Как следует из содержания статьи, регрессионная диагностика означает, что задача математического программирования должна не просто осуществлять отбор переменных, но и по возможности гарантировать статистическую значимость оценок регрессии и выполнение таких предпосылок МНК, как гомоскедастичность остатков, отсутствие автокорреляции и мультиколлинеарности. Поэтому в [9] задача ЧБКП дополнена так называемыми «relaxed» (ослабленными) линейными ограничениями на наблюдаемые значения t-критерия Стьюдента. Поскольку из-за ослабленных ограничений при решении этой задачи коэффициенты регрессии могут получиться незначимыми, авторы предлагают использовать «lazy» (ленивый) алгоритм, который на каждом узле проверяет выполнение необходимых условий и в случае из несрабатывания вносит коррективы в процесс вычисления. Авторы утверждают, что этот алгоритм можно применять и для диагностики других проблем, но никакого конкретного руководства к действию не приводят.

В работе [10] рассматривается так называемая «holistic» (целостная) линейная регрессия. Её идея очень похожа на идею регрессионной диагностики [9]. Суть в том, чтобы сформулировать задачу ЧБКП, решение которой давало бы модель с желаемыми свойствами. В [10] так же, как и в [9], использованы ленивые ограничения, но проверка значимости коэффициентов осуществляется не по t-критерию Стьюдента, а с помощью теста асимптотической нормальности. Помимо этого, в [10] задача ОИР дополнена ограничениями на степень мультиколлинеарности.

Как видно, в зарубежных источниках [4–10] сформировался «квадратичный» подход к задачам ОИР. Однако в 2018 году автору настоящей статьи удалось свести задачу ОИР

в оцениваемой с помощью МНК линейной регрессии для фиксированного числа регрессоров к задаче ЧБЛП [11]. Затем последовала работа [12], в которой задача ЧБЛП была дополнена линейными ограничениями на степень мультиколлинеарности. В [13] на основе скорректированного коэффициента детерминации задача ОИР, не требующая фиксации числа отбираемых переменных, была сформулирована в виде задачи частично-целочисленного линейного программирования (ЧЦЛП). Таким образом, в отечественной литературе формируется «линейный» подход к задачам ОИР.

Данная статья является логическим продолжением работ [11–13]. Её цель – формулировка единой задачи ЧБЛП, осуществляющей ОИР в оцениваемой с помощью МНК линейной регрессии с заданной степенью мультиколлинеарности и значимыми по t-критерию Стьюдента оценками.

1. МАТЕРИАЛЫ И МЕТОДЫ

Рассмотрим модель множественной линейной регрессии:

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

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

Задача ОИР состоит в том, чтобы выбрать для включения в линейную модель (1) из l объясняющих переменных m наиболее информативных по некоторому критерию качества. Пусть в качестве такого критерия используется сумма квадратов ошибок, т. е. регрессия оценивается с помощью МНК.

Проведем нормирование (стандартизацию) всех переменных по формулам:

$$v_i = \frac{y_i - \overline{y}}{\sigma_y},\ z_{i1} = \frac{x_{i1} - \overline{x_1}}{\sigma_{x_1}},\ ...,\ z_{il} = \frac{x_{il} - \overline{x_l}}{\sigma_{x_l}},$$

где $\overline{y}$, $\overline{x_1}$, …, $\overline{x_l}$ — средние значения переменных; σ_y, σ_{x_1}, …, σ_{x_l} — среднеквадратические отклонения переменных; v, z_1, …, z_l — стандартизованные переменные, для которых среднее значение равно 0, а среднеквадратическое отклонение равно 1.

Введем стандартизованное уравнение регрессии:

$$v_i = \beta_1 z_{i1} + \beta_2 z_{i2} + ... + \beta_l z_{il} + u_i,\ i = \overline{1,n}, \quad (2)$$

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

Известно (см., например, [14]), что МНК-оценки стандартизованной регрессии (2) находятся на основе решения системы нормальных уравнений

$$\sum_{k=1}^{l} R_{xx}^{(j,k)} \cdot \beta_k = R_{yx}^{(j,1)},\ \ j = \overline{1,l}, \quad (3)$$

где $R_{yx}^{(j,1)}$ — элементы вектора коэффициентов парной корреляции между объясняемой переменной y и объясняющими переменными x_1, x_2, ..., x_l; $R_{xx}^{(j,k)}$ — элементы матрицы коэффициентов парной корреляции между объясняющими переменными.

Коэффициент детерминации R^2 для линейной (1) и стандартизованной (2) регрессий находится по формуле [14]

$$R^2 = \sum_{j=1}^{l} R_{yx}^{(j,1)} \cdot \beta_j\ . \quad (4)$$

С использованием формул (3) и (4) в работе [11] задача ОИР в оцениваемой с помощью МНК линейной регрессии была сведена к следующей задаче ЧБЛП:

$$R^2\left(\beta_1, \beta_2, ..., \beta_l\right) = \sum_{j=1}^{l} R_{yx}^{(j,1)} \cdot \beta_j \to \max, \quad (5)$$

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

$$-R_{yx}^{(j,1)} \le (1-\delta_j)M,\quad j = \overline{1,l}, \quad (6)$$

$$-\delta_j M \le \beta_j \le \delta_j M,\quad j = \overline{1,l}, \quad (7)$$

$$\delta_j \in \{0,1\},\quad j = \overline{1,l}, \quad (8)$$

$$\sum_{j=1}^{l} \delta_j = m, \quad (9)$$

где M — большое положительное число; δ_j, $j = \overline{1,l}$ — булевы переменные, заданные по правилу:

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

Как видно, если в задаче (5)–(9) $\delta_j = 0$, то $\beta_j = 0$ и одновременно снимается ограничение-равенство с j-го уравнения системы (3), т. е. происходит МНК-оценивание линейной регрессии без j-й объясняющей переменной. В противном случае, когда $\delta_j = 1$, переменная не исключается.

Стоит отметить, что для перехода от бета-коэффициентов стандартизованной регрессии (2) к МНК-оценкам линейной регрессии (1) необходимо воспользоваться формулами:

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

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

Одним из способов выявления мультиколлинеарности [15] в модели (1) является расчет значений коэффициентов детерминации вспомогательных регрессий, представляющих собой линейные зависимости j-й объясняющей переменной от всех остальных. Высокие значения этих коэффициентов (обычно больше 0,6) свидетельствуют о наличии мультиколлинеарности.

Для выявления мультиколлинеарности в стандартизованной регрессии (2) введем вспомогательные регрессии

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

где β_{kj}, $k = \overline{1,l}$, $j = \overline{1,l-1}$ — бета-коэффициенты вспомогательных регрессий; u_k, $k = \overline{1,l}$ — векторы ошибок аппроксимации.

В соответствии с (3), для оценивания этих вспомогательных регрессий требуется решить l линейных систем

$$\sum_{j=1}^{l-1} R_{xx}^{(q_{ki}, q_{kj})} \beta_{kj} = R_{xx}^{(q_{ki}, k)}, \; k = \overline{1,l}, \; i = \overline{1,l-1}, \quad (10)$$

где q_{ij} — элементы матрицы $Q_{l\times(l-1)}$, полученной путем вычеркивания главной диагонали из матрицы $\begin{pmatrix} 1 & 2 & ... & l \\ 1 & 2 & ... & l \\ ... & ... & ... & ... \\ 1 & 2 & ... & l \end{pmatrix}_{l\times l}$.

Коэффициенты детерминации R_k^2, $k = \overline{1,l}$ вспомогательных регрессий, согласно (4), находятся по формулам:

$$R_k^2 = \sum_{j=1}^{l-1} R_{xx}^{(q_{kj}, k)} \cdot \beta_{kj}, \quad k = \overline{1,l}. \quad (11)$$

В работе [12] на основе формул (10) и (11) рассмотрены различные способы контроля эффекта мультиколлинеарности. Так, например, можно дополнить задачу (5)–(9) линейными ограничениями:

$$-\left(1 - \delta_{q_{ki}}\right) M \le \sum_{j=1}^{l-1} R_{xx}^{(q_{ki}, q_{kj})} \beta_{kj} - R_{xx}^{(q_{ki}, k)} \le$$
$$\le \left(1 - \delta_{q_{ki}}\right) M, \quad k = \overline{1,l}, \quad i = \overline{1,l-1}, \quad (12)$$

$$-\delta_{q_{ki}} M \le \beta_{ki} \le \delta_{q_{ki}} M, \quad k = \overline{1,l}, \quad i = \overline{1,l-1}, \quad (13)$$

$$\sum_{j=1}^{l-1} R_{xx}^{(q_{kj}, k)} \cdot \beta_{kj} - \left(1 - \delta_k\right) M \le r, \quad k = \overline{1,l}, \quad (14)$$

где r — ограничение на значения коэффициентов детерминации вспомогательных регрессий из интервала $(0,1]$. Если $r = 1$, то у исследователя нет никаких требований к степени мультиколлинеарности, а если $r \to 0$, то он желает полностью устранить этот негативный эффект.

К сожалению, задача (5)–(9), (12)–(14) для заданного числа r может вообще не иметь решений. Поэтому в работе [12] сформулирована двухкритериальная задача ОИР. Для этого нужно дополнить задачу (5)–(9), (12)–(14) еще одно целевой функцией

$$r \to \min. \quad (15)$$

Решение этой двухкритериальной задачи позволяет идентифицировать спецификацию модели с наилучшим качеством аппроксимации и одновременно с минимальным эффектом мультиколлинеарности. Поскольку такая задача является еще и частично-целочисленной, то возникает проблема с её решением. В этом случае можно использовать стандартный прием и свести двухкритериальную задачу к однокритериальной с помощью линейной свертки:

$$\left(1 - \lambda\right) R^2 - \lambda r \to \max, \quad (16)$$

где λ — заданное число из интервала $[0;1]$. Так, если $\lambda = 0$, то будет построена модель только с наилучшим качеством аппроксима-

ции, а если $\lambda = 1$, то только с минимальным эффектом мультиколлинеарности.

На основе (16) можно сформировать множество Парето в пространстве критериев $\left(R^2, r\right)$, которое исследователь может использовать для выбора оптимального на его взгляд соотношения пары «качество – мультиколлинеарность» в линейной регрессии.

Известно, что для линейной регрессии (1) t-критерии Стьюдента находятся по формулам:

$$t_j = \frac{\alpha_j}{s\sqrt{\left(X^T X\right)^{-1}_{jj}}}, \quad j = \overline{1,l},$$

где s^2 — величина остаточной дисперсии; X — матрица значений объясняющих переменных; $\left(X^T X\right)^{-1}_{jj}$ — j-й диагональный элемент матрицы $\left(X^T X\right)^{-1}$.

Тогда для стандартизованной регрессии (2) наблюдаемые значения t-критерия Стьюдента находятся по формулам:

$$t_j = \frac{\beta_j}{s\sqrt{\left(R_{xx}\right)^{-1}_{jj}}}, \quad j = \overline{1,l},$$

где $s^2 = \dfrac{1-R^2}{n-m-1}$.

Как видно, получить линейные ограничения с помощью этих формул не представляется возможным. Поэтому используем другой известный подход [14] к определению значений t-критериев, основанный на вычислении частных F-критериев.

Частный F-критерий F_j используется для оценки значимости влияния объясняющей переменной x_j как дополнительного включенного в линейную модель (1) фактора и находится по формуле:

$$F_{x_j} = \frac{R^2 - R^2_{y|x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m}}{1-R^2} \cdot \frac{n-m-1}{1}, \quad (17)$$

где $R^2_{y|x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m}$ — коэффициент детерминации модели (1) без включения в неё фактора x_j.

Зная величину частного F-критерия, можно найти для коэффициента при j-м факторе величину t-критерия Стьюдента:

$$t_j = \sqrt{F_{x_j}}. \quad (18)$$

С вероятностно-статистической точки зрения, коэффициент регрессии считается значимым по t-критерию Стьюдента, если его наблюдаемое по абсолютной величине значение $\left|t_j\right|$ больше, чем табличное критическое значение $t_{\text{крит}}\left(\alpha, n-m-1\right)$, т. е.

$$\left|t_j\right| > t_{\text{крит}}\left(\alpha, n-m-1\right), \quad (19)$$

где α — заданный уровень значимости.

Возводя неравенство (19) в квадрат, с учетом (17) и (18) получим

$$R^2 - R^2_{y|x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m} > \left(1-R^2\right)T, \quad (20)$$

где $T = \dfrac{t^2_{\text{крит}}\left(\alpha, n-m-1\right)}{n-m-1}$ — найденное на основе заданного критического значения $t_{\text{крит}}\left(\alpha, n-m-1\right)$ число, указывающее на степень значимости коэффициентов регрессии. Если $T = 0$, то у исследователя нет никаких требований к значимости коэффициентов регрессии по t-критерию Стьюдента. Чем выше значение T, тем выше эти требования.

Линейное неравенство (20) для каждой объясняющей переменной x_j нетрудно интегрировать в задачу ЧБЛП (5)–(9). Для вычисления коэффициентов детерминации $R^2_{y|x_1,x_2,\ldots,x_{j-1},x_{j+1},\ldots,x_m}$ составим для стандартизованной регрессии (2) вспомогательные модели, не содержащие одной объясняющей переменной x_k, $k = \overline{1,l}$:

$$v_i = \beta^*_{11} z_{i2} + \beta^*_{12} z_{i3} + \ldots + \beta^*_{1,l-1} z_{il} + u^*_{i1},$$
$$v_i = \beta^*_{21} z_{i1} + \beta^*_{22} z_{i3} + \ldots + \beta^*_{2,l-1} z_{il} + u^*_{i2},$$
$$\ldots$$
$$v_i = \beta^*_{l,1} z_{i1} + \beta^*_{l,2} z_{i2} + \ldots + \beta^*_{l,l-1} z_{i,l-1} + u^*_{il},$$

где β^*_{kj}, $k = \overline{1,l}$, $j = \overline{1,l-1}$ — бета-коэффициенты вспомогательных регрессий; u^*_k, $k = \overline{1,l}$ – векторы ошибок аппроксимации.

Тогда, в соответствии с (3), для нахождения оценок бета-коэффициентов β^*_{kj}, $k = \overline{1,l}$, $j = \overline{1,l-1}$ вспомогательных регрессий необходимо решить l систем линейных алгебраических уравнений

$$\sum_{j=1}^{l-1} R^{(q_{ki},q_{kj})}_{xx} \beta^*_{kj} = R^{(q_{ki},1)}_{yx}, \quad k = \overline{1,l}, \quad i = \overline{1,l-1}. \quad (21)$$

Коэффициенты детерминации $\overline{R}^2_k$, $k = \overline{1,l}$ таких вспомогательных регрессий, согласно (4), находятся по формулам:

$$\overline{R}_k^2 = \sum_{j=1}^{l-1} R_{yx}^{(q_{kj},1)} \cdot \beta_{kj}^*, \quad k = \overline{1,l}. \qquad (22)$$

Формулы (21) и (22) справедливы при оценивании стандартизованной регрессии (2) со всеми l объясняющими переменными. Для того чтобы организовать процедуру отбора из них m регрессоров введем следующие линейные ограничения:

$$-\left(1-\delta_{q_{ki}}\right)M \le \sum_{j=1}^{l-1} R_{xx}^{(q_{ki},q_{kj})}\beta_{kj}^* - R_{yx}^{(q_{ki},1)} \le$$

$$\le \left(1-\delta_{q_{ki}}\right)M, \quad k = \overline{1,l}, \quad i = \overline{1,l-1}, \qquad (23)$$

$$-\delta_{q_{ki}} M \le \beta_{ki}^* \le \delta_{q_{ki}} M, \ k = \overline{1,l}, \ i = \overline{1,l-1}. \qquad (24)$$

На основе неравенства (20) для каждой объясняющей переменной введем линейные ограничения:

$$\sum_{i=1}^{l} R_{yx}^{(i,1)} \cdot \beta_i - \sum_{j=1}^{l-1} R_{yx}^{(q_{kj},1)} \cdot \beta_{kj}^* \ge$$

$$\ge \left(1 - \sum_{i=1}^{l} R_{yx}^{(i,1)} \cdot \beta_i\right) T - \left(1-\delta_k\right)M, \ k = \overline{1,l}. \qquad (25)$$

Так, если $\delta_k = 0$, т. е. k-я переменная не входит в модель, то соответствующее ограничение (25) на значимость коэффициента по t-критерию снимается, иначе — нет.

Интеграция ограничений (23)–(25) в задачу ЧБЛП (5)–(9) позволяет контролировать в процессе ОИР степень значимости коэффициентов регрессии по t-критерию Стьюдента.

К сожалению, задача (5)–(9), (23)–(25) для заданного числа T может вообще не иметь решений. Если попробовать ввести в неё еще одну целевую функцию $T \to \max$, то ограничения (25) станут нелинейными.

Поэтому обозначим в ограничениях (25) величину $\left(1-R^2\right)T = \left(1 - \sum_{i=1}^{l} R_{yx}^{(i,1)} \cdot \beta_i\right)T$ переменной Δ. Из (20) следует, что $0 \le \Delta \le 1$. Если $\Delta = 0$, то нет никаких требований к значимости коэффициентов регрессии по t-критерию Стьюдента. На основании (20) можно сделать вывод, что чем больше величина Δ, тем выше разница между коэффициентом детерминации R^2 линейной регрессии и коэффициентами детерминации $\overline{R}_k^2$, $k = \overline{1,l}$ вспомогательных регрессий, не содержащих одной объясняющей переменной. Это значит, что с увеличением Δ растет степень значимости коэффициентов по t-критерию Стьюдента. Но, к сожалению, одновременно происходит уменьшение величины R^2.

Дополним задачу (5)–(9), (23)–(25) целевой функцией

$$\Delta \to \max. \qquad (26)$$

По аналогии с (16), сведем полученную двухкритериальную задачу ОИР к однокритериальной с помощью введения линейной сверки:

$$\left(1-\lambda\right)R^2 + \lambda\Delta \to \max. \qquad (27)$$

На основе (27) можно сформировать множество Парето в пространстве критериев $\left(R^2, \Delta\right)$ и использовать его для выбора оптимального соотношения пары «качество – значимость» в линейной регрессии.

Наконец, можно сформулировать задачу ЧБЛП (5)–(9), (12)–(14), (23)–(25), предназначенную для построения с помощью МНК линейной m-факторной регрессии с заданными ограничениями на степень мультиколлинеарности и значимости коэффициентов по t-критерию Стьюдента.

Если исследователь затрудняется с назначением величин r и T, то можно сформулировать трехкритериальную задачу ОИР с целевыми функциями (5), (15), (26) и с линейными ограничениями (6)–(9), (12)–(14), (23)–(25). Линейная свертка для такой задачи выглядит следующим образом:

$$w_1 R^2 - w_2 r + w_3 \Delta \to \max, \qquad (28)$$

где w_1, w_2, w_3 – некоторые весовые коэффициенты. С помощью (20) можно сформировать множество Парето в пространстве критериев $\left(R^2, r, \Delta\right)$ для выбора оптимального соотношения тройки «качество – мультиколлинеарность – значимость».

2. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Для демонстрация корректности предложенного математического аппарата на персональном компьютере с 4-х ядерным процессором Intel Core i5-4670 с тактовой частотой 3400 МГц и объемом оперативной памяти 8 Гб проводились вычислительные эксперименты. Для этого использовались встроенные

в эконометрический пакет Gretl статистические данные (data7-10.gdt) о качестве воздуха в Калифорнии и его детерминантов за 1970–1972 гг. Объем выборки — 30. В качестве зависимой переменной выбрана переменная *airqual*, а в качестве независимых выступают *popln*, *valadd*, *rain*, *coast*, *density*, *medincm*, *poverty*, *electr*, *fueloil* и *indestab*. Для удобства будем обозначать их далее, как y, x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9 и x_{10}. С помощью элементарного преобразования x^2 были сформированы дополнительные переменные: $x_{11} = x_1^2$, $x_{12} = x_2^2$, $x_{13} = x_3^2$, $x_{14} = x_5^2$, $x_{15} = x_6^2$, $x_{16} = x_7^2$, $x_{17} = x_8^2$, $x_{18} = x_9^2$, $x_{19} = x_{10}^2$. На основе этих данных с использованием пакета LPSolve были проведены следующие эксперименты.

Эксперимент № 1. Ставилась задача построения четырехфакторной линейной регрессии с заданной степенью значимости её коэффициентов по t-критерию Стьюдента, т. е. решалась задача (5)–(9), (23)–(25). В ней $M = 1000$, $m = 4$, а величина T назначалась для уровней значимости 1 и 0,05. Так, для критических точек $t_{\text{крит}}(1,25) = 0$ и $t_{\text{крит}}(0.05,25) = 2,05954$ величина T составляет 0 и 0,16967 соответственно. Для решения этих задач в LPSolve была разработана программа, содержащая 1464 основных ограничений и 380 переменных. Для каждого T фиксировалось время и ход решения задачи. Результаты тестирования представлены в табл. 1.

После чего эти задачи были решены методом простого перебора. Для каждого значения T потребовалось перебрать $C_{19}^4 = 3876$ моделей.

Ограничению $T = 0$ удовлетворяют все регрессии, а лучшей из них по величине R^2 является

$$\tilde{y} = \underset{}{106,44} + \underset{(3,237)}{0,0535}\,x_1 - \underset{(-3,23)}{28,118}\,x_4 - \underset{(-2,955)}{0,0095}\,x_6 - \underset{(-1,664)}{5,05\cdot 10^{-6}}\,x_{16},$$

где в скобках указаны наблюдаемые значения t-критериев Стьюдента. Таким образом, коэффициенты при переменных x_1, x_4 и x_6 значимы для уровня $\alpha = 0,01$. А коэффициент при x_{16} незначим даже для уровня $\alpha = 0,1$.

Ограничению $T = 0,16967$ удовлетворяют 18 регрессий из 3876, а лучшей из них является

$$\tilde{y} = \underset{}{129,292} - \underset{(-3,556)}{30,105}\,x_4 - \underset{(-2,154)}{0,0236}\,x_{10} + \underset{(2,967)}{9,83\cdot 10^{-6}}\,x_{11} - \underset{(-2,922)}{2,98\cdot 10^{-7}}\,x_{15}.$$

В этой модели все коэффициенты значимы для уровня $\alpha = 0,05$.

Как видно, полученные решения полностью совпадают с результатами, приведенными в табл. 1.

Таблица 1. Эксперимент № 1
[Table 1. Experiment no. 1]

Узел	Переменные	R^2	Время, с
$T = 0$			
1	x_1, x_4, x_5, x_8	0,41868	24,117
2	x_1, x_4, x_5, x_9	0,443849	
3	x_1, x_4, x_5, x_{10}	0,460335	
4	x_1, x_4, x_5, x_6	0,516784	
5	x_1, x_4, x_6, x_8	0,517173	
6	x_1, x_3, x_4, x_6	0,538648	
7	x_1, x_4, x_6, x_7	0,551907	
8	x_1, x_4, x_6, x_{16}	0,564986	
$T = 0,16967$			
1	x_1, x_5, x_{13}, x_{16}	0,311285	31,39
2	x_1, x_3, x_7, x_{16}	0,499439	
3	$x_4, x_{10}, x_{11}, x_{15}$	0,514623	

Эксперимент № 2. Ставилась двухкритериальная задача построения четырехфакторной линейной регрессии (27), (6)–(9), (23)–(25). В ней $M = 1000$, $m = 4$, а величина λ задавалась из интервала $[0,1]$ с шагом 0,1. Результаты решений представлены в табл. 2.

Стоит отметить, что в этом эксперименте задача ЧБЛП (27), (6)–(9), (23)–(25) была дополнена линейным ограничением $\Delta \le 1$, что существенно повысило эффективность её решения.

На основе табл. 2 можно сформировать множество Парето в пространстве критериев (R^2, Δ): $(0,564986; 0,048205)$, $(0,499439; 0,134064)$, $(0,377346; 0,197692)$. С учетом того, что $t_{\text{крит}}(\alpha, n-m-1) = \sqrt{(n-m-1)\dfrac{\Delta}{1-R^2}}$, можно представить сформированное множество Парето в пространстве критериев (R^2, α): $(0,564986; 0,1085)$, $(0,499439; 0,01587)$, $(0,377346; 0,009322)$, в котором, например, альтернатива $(0,564986; 0,1085)$ означает, что коэффициент детерминации модели равен 0,564986, а все коэффициенты при объясняющих переменных значимы для уровня 0,1085. Выбор оптимального соотношения пары «качество – значимость» остается за исследователем.

Эксперимент № 3. Ставилась задача построения четырехфакторной линейной регрессии с заданной степенью мультиколлинеарности и значимости, т. е. решалась задача (5)–(9), (12)–(14), (23)–(25). В ней $M = 1000$, $m = 4$, T задавалось 0 и 0,05548 (для уровня $\alpha = 0,25$), а параметр r назначался 1 и 0,9. Для решения этих задач в LPSolve была разработана программа, содержащая 2851 основных ограничений и 722 переменных. Результаты тестирования представлены в табл. 3.

Полученные в табл. 3 решения полностью совпадают с результатами, полученными методом полного перебора моделей.

Эксперимент № 4. Ставилась трехкритериальная задача построения четырехфакторной линейной регрессии (28), (6)–(9), (12)–(14), (23)–(25). В ней $M = 1000$, $m = 4$, а веса в (28) задавалась одинаковыми. Результаты решения представлены в табл. 4.

Заметим, что в этом эксперименте в задачу (28), (6)–(9), (12)–(14), (23)–(25) вновь было введено ограничение $\Delta \le 1$, повысившее эффективность решения.

Как следует из табл. 4, лучшей по критерию $R^2 - r + \Delta \to \max$ стала модель с объясняющими переменными $x_4, x_{15}, x_{17}, x_{18}$, коэффициент детерминации R^2 которой равен 0,4118, что говорит о её невысоком качестве. Наибольшее значение коэффициента детерминации r вспомогательных регрессий, представляющих собой линейные зависимости j-й объясняющей переменной от всех остальных, составляет всего 0,0396, что указывает практически на полное отсутствие эф-

Таблица 2. Эксперимент № 2

[Table 2. Experiment no. 2]

λ	Переменные	R^2	Δ	Время, с
0	x_1, x_4, x_6, x_{16}	0,564986	0,048205	28,62
0,1	x_1, x_4, x_6, x_{16}	0,564986	0,048205	35,939
0,2	x_1, x_4, x_6, x_{16}	0,564986	0,048205	39,565
0,3	x_1, x_4, x_6, x_{16}	0,564986	0,048205	40,666
0,4	x_1, x_4, x_6, x_{16}	0,564986	0,048205	42,109
0,5	x_1, x_3, x_7, x_{16}	0,499439	0,134064	42,734
0,6	x_1, x_3, x_7, x_{16}	0,499439	0,134064	44,358
0,7	x_2, x_6, x_{11}, x_{19}	0,377346	0,197692	41,402
0,8	x_2, x_6, x_{11}, x_{19}	0,377346	0,197692	42,62
0,9	x_2, x_6, x_{11}, x_{19}	0,377346	0,197692	41,687
1	x_2, x_6, x_{11}, x_{19}	0,377346	0,197692	46,824

Таблица 3. Эксперимент № 3
[Table 3. Experiment no. 3]

Узел	Переменные	R^2	Время, с
$T=0,\ r=1$			
1	x_1, x_4, x_5, x_8	0,41868	87,116
2	x_1, x_4, x_5, x_9	0,443849	
3	x_1, x_4, x_5, x_{10}	0,460335	
4	x_1, x_4, x_5, x_6	0,516784	
5	x_1, x_4, x_6, x_8	0,517173	
6	x_1, x_3, x_4, x_6	0,538648	
7	x_1, x_4, x_6, x_7	0,551907	
8	x_1, x_4, x_6, x_{16}	0,564986	
$T=0,\ r=0{,}9$			
1	x_1, x_4, x_5, x_8	0,41868	96,011
2	x_1, x_4, x_5, x_9	0,443849	
3	x_1, x_4, x_8, x_9	0,448129	
4	x_1, x_4, x_8, x_{16}	0,452619	
5	x_4, x_8, x_{16}, x_{17}	0,458315	
$T=0{,}05548,\ r=1$			
1	x_1, x_4, x_8, x_{16}	0,452619	94,174
2	x_1, x_4, x_7, x_{16}	0,464511	
3	x_1, x_4, x_6, x_7	0,551907	
4	x_1, x_4, x_6, x_{16}	0,564986	
$T=0{,}05548,\ r=0{,}9$			
1	x_1, x_4, x_8, x_{16}	0,452619	103,456

Таблица 4. Эксперимент № 4
[Table 4. Experiment no. 4]

Узел	Переменные	R^2	r	Δ	$R^2 - r + \Delta$	Время, с
1	x_1, x_4, x_5, x_8	0,4187	0,2077	0,0056	0,2165	128,848
2	x_1, x_4, x_5, x_9	0,4438	0,2161	0,0119	0,2396	
3	x_1, x_3, x_4, x_5	0,4059	0,0746	0,0112	0,3424	
4	x_1, x_3, x_4, x_{14}	0,4043	0,0704	0,0096	0,3435	
5	x_4, x_8, x_{11}, x_{18}	0,4352	0,1212	0,0310	0,3450	
6	x_4, x_8, x_{14}, x_{15}	0,4041	0,0616	0,0071	0,3496	
7	x_3, x_4, x_8, x_{15}	0,4003	0,0540	0,0033	0,3497	
8	x_4, x_8, x_{15}, x_{18}	0,4330	0,1096	0,0360	0,3594	
9	x_4, x_9, x_{14}, x_{15}	0,4364	0,0815	0,0133	0,3683	
10	$x_4, x_{11}, x_{14}, x_{18}$	0,4105	0,0433	0,0100	0,3773	
11	$x_4, x_{11}, x_{17}, x_{18}$	0,4160	0,0518	0,0155	0,3797	
12	$x_4, x_{14}, x_{15}, x_{18}$	0,4047	0,0290	0,0105	0,3861	
13	$x_4, x_{15}, x_{17}, x_{18}$	0,4118	0,0396	0,0176	0,3899	

фекта мультиколлинеарности. Вычисленный для $\Delta = 0,0176$ уровень значимости α составил 0,3949, что говорит о присутствии в модели слабо значимых по t-критерию Стьюдента коэффициентов.

ЗАКЛЮЧЕНИЕ

В работе рассмотрена предложенная ранее формализация задачи ОИР в оцениваемых в помощью МНК регрессионных моделях с регулируемым эффектом мультиколлинеарности в виде задачи частично-булевого линейного программирования. Для обеспечения значимости коэффициентов по t-критерию Стьюдента эта задача была расширена соответствующими линейными ограничениями. Поскольку при заданных начальных параметрах такая задача может вовсе не иметь решений, то сформулирована двухкритериальная задача, позволяющая строить модели с позиции соотношения «качество – значимость». Для построения модели с позиции соотношения «качество — мультиколлинеарность — значимость» сформулирована соответствующая трехкритериальная задача. Проведено 4 вычислительных эксперимента, подтверждающих корректность и демонстрирующих эффективность разработанного математического аппарата.

КОНФЛИКТ ИНТЕРЕСОВ

Автор декларирует отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Desboulets, L.D.D.* A review on variable selection in regression analysis / L.D.D. Desboulets // Econometrics. – 2018. – Vol. 6. – P. 1–27. https://doi.org/10.3390/econometrics6040045.

2. *Носков, С. И.* Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных / С. И. Носков. – Иркутск : Облинформпечать, 1996. – 321 с.

3. *Базилевский, М. П.* Программный комплекс построения линейной регрессионной модели с учетом критерия согласованности поведения фактической и расчетной траекторий изменения значений объясняемой переменной / М. П. Базилевский, С. И. Носков // Вестник Иркутского государственного технического университета. – 2017. – Т. 21, № 9 (128). – С. 37–44.

4. *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. – P. 273–282. Available at: https://doi.org/10.1007/s10898-008-9323-9

5. *Bertsimas, D.* Best subset selection via a modern optimizations lens / D. Bertsimas, A. King, R. Mazumder // The Annals of Statistics. – 2016. – Vol. 44. – P. 813–852. Available at: https://doi.org/10.1214/15-AOS1388

6. *Miyashiro, R.* Mixed integer second-order cone programming formulations for variable selection in linear regression / R. Miyashiro, Y. Takano // European Journal of Operational Research. – 2015. – Vol. 247. – P. 721–731. Available at: https://doi.org/10.1016/j.ejor.2015.06.081

7. *Miyashiro, R.* Subset selection by Mallows' Cp: a mixed integer programming approach / R. Miyashiro, Y. Takano // Expert Systems with Applications. – 2015. – Vol. 42. – P. 325–331. Available at: https://doi.org/10.1016/j.eswa.2014.07.056

8. *Park, Y. W.* Subset selection for multiple linear regression via optimization / Y. W. Park, D. Klabjan // Journal of Global Optimization. – 2020. – Vol. 77. – P. 543–574. Available at: https://doi.org/10.1007/s10898-020-00876-1

9. *Chung, S.* A mathematical programming approach for integrated multiple linear regression subset selection and validation / S. Chung, Y. W. Park, T. Cheong // Pattern Recognition. – 2020. – Vol. 108. – P. 107565. Available at: https://doi.org/10.1016/j.patcog.2020.107565

10. *Bertsimas, D.* Scalable holistic linear regression / D. Bertsimas, M.L. Li // Operations Research Letters. – 2020. – Vol. 48, No. 3. – P. 203–208. Available at: https://doi.org/10.1016/j.orl.2020.02.008

11. *Базилевский, М. П.* Сведение задачи отбора информативных регрессоров при оценивании линейной регрессионной модели по методу наименьших квадратов к зада-

че частично-булевого линейного программирования / М. П. Базилевский // Моделирование, оптимизация и информационные технологии. – 2018. – Т. 6, № 1 (20). – С. 108–117.

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

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

14. Эконометрика : учебник / И. И. Елисеева [и др.]; под ред. И. И. Елисеевой. – 2-е изд., перераб. и доп. – М. : Финансы и статистика, 2007. – 576 с.

15. *Кремер, Н. Ш.* Эконометрика: учебник / Н. Ш. Кремер, Б. А. Путко. – 3-е изд., перераб. и доп. – М. : ЮНИТИ-ДАНА, 2010. – 328 с.

Базилевский Михаил Павлович — канд. техн. наук, доцент, доцент кафедры математики Иркутского государственного университета путей сообщения.
E-mail: mik2178@yandex.ru
ORCID iD: https://orcid.org/0000-0002-3253-5697

DOI: https://doi.org/10.17308/sait.2021.3/3731 ISSN 1995-5499
Received 19.02.2021
Accepted 20.11.2021

SELECTION OF INFORMATIVE REGRESSORS SIGNIFICANT BY STUDENT'S T-TEST IN REGRESSION MODELS ESTIMATED USING OLS AS A PARTIAL BOOLEAN LINEAR PROGRAMMING PROBLEM

© 2021 M. P. Bazilevskiy✉

Irkutsk State Transport University
15, Chernyshevskogo Street, 664074 Irkutsk, Russian Federation

Annotation. This article is devoted to the problem of selecting a fixed number of informative regressors in linear regression models estimated using ordinary least squares method. In modern scientific works, the apparatus of integer mathematical programming, which has been well developed in recent years, is used to solve this problem. In most of these works, the problem of subset selection is formalized in the form of partial-quadratic linear programming problems. Relatively recently, articles began to appear in which the authors strive to formulate a unified problem of mathematical programming, which, in parallel with the selection of factors, guarantees the construction of a regression that satisfies various statistical tests. This work is a logical continuation of the previous articles of the author, in which the problem of subset selection is formalized as a partial Boolean linear, not quadratic, programming problem. Methods of control in this problem of the degree of multicollinearity have already been considered. In this article, using the well-known approach to determining the values of Student's t-tests, based on the calculation of partial F-tests, linear constraints on the degree of significance of the regression coefficients were integrated into the mentioned problem of partial-Boolean linear programming. A two-criteria problem is formulated, which allows to build a model from the perspective of the ratio «quality —significance», and a three-criteria problem, which constructs a

✉ Bazilevskiy Mikhail P.
e-mail: mik2178@yandex.ru

regression from the position of the relationship «quality — multicollinearity — significance». Computational experiments have been successfully carried out to confirm the correctness of the proposed mathematical apparatus.
Keywords: regression model, standardized regression, subset selection, multicollinearity, Student's t-test, coefficient of determination, partial-Boolean linear programming problem.

CONFLICT OF INTEREST

The authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.

REFERENCES

1. *Desboulets L.D.D.* (2018) A review on variable selection in regression analysis. Econometrics. V. 6. P. 1–27. https://doi.org/10.3390/econometrics6040045

2. *Noskov S. I.* (1996) Technology for modeling objects with unstable functioning and uncertainty in data. Irkutsk: Oblinformpechat' . 320 p.

3. *Bazilevskiy M. P., Noskov S. I.* (2017) A software package for constructing a linear regression model taking into account the criterion of consistency of the behavior of the actual and calculated trajectories of change in the values of the explained variable. Proceedings of Irkutsk State Technical University. V. 128. No. 9. P. 37–44.

4. *Konno H., Yamamoto R.* (2009) Choosing the best set of variables in regression analysis using integer programming. Journal of global optimization. V. 44. P. 273–282. https://doi.org/10.1007/s10898-008-9323-9

5. *Bertsimas D., King A., Mazumder R.* (2016) Best subset selection via a modern optimizations lens. The Annals of Statistics. V. 44. P. 813–852. https://doi.org/10.1214/15-AOS1388

6. *Miyashiro R., Takano Y.* (2015) Mixed integer second-order cone programming formulations for variable selection in linear regression. European Journal of Operational Research. V. 247. P. 721–731. https://doi.org/10.1016/j.ejor.2015.06.081

7. *Miyashiro R., Takano Y.* (2015) Subset selection by Mallows' Cp: a mixed integer programming approach. Expert Systems with Applications. V. 42. P. 325–331. https://doi.org/10.1016/j.eswa.2014.07.056

8. *Park Y.W., Klabjan D.* (2020) Subset selection for multiple linear regression via optimization. Journal of Global Optimization. V. 77. P. 543–574. https://doi.org/10.1007/s10898-020-00876-1

9. *Chung S., Park Y.W., Cheong T.* (2020) A mathematical programming approach for integrated multiple linear regression subset selection and validation. Pattern Recognition. V. 108. P. 107565. https://doi.org/10.1016/j.patcog.2020.107565

10. *Bertsimas D., Li M. L.* (2020) Scalable holistic linear regression. Operations Research Letters. V. 48, No. 3. P. 203-208. https://doi.org/10.1016/j.orl.2020.02.008

11. *Bazilevskiy M. P.* (2018) Reduction of the informative regressor selection problem in estimating a linear regression model using the least squares method to a partial Boolean linear programming problem. Modeling, optimization and information technology. V. 20, No.1. P. 108-117.

12. *Bazilevskiy M. P.* (2018) Selection of informative regressors taking into account the multicollinearity between them in regression models as a partial Boolean linear programming problem. Modeling, optimization and information technology. V. 21, No. 2. P. 104–118.

13. *Bazilevskiy M. P.* (2020) Selection of the optimal number of informative regressors by the adjusted coefficient of determination in regression models as a problem of partially integer linear programming. Applied Mathematics and Control Sciences. No. 2. P. 41–54.

14. *Eliseeva I. I. et al.* (2007) Econometrics. Moscow: Finance and Statistics. 576 p.

15. *Kremer N. Sh., Putko B. A.* (2010) Econometrics. Moscow: UNITY-DANA. 328 p.

Bazilevskiy Mikhail P. — PhD in Technical Sciences, Associate Professor, Department of Mathematics, Irkutsk State Transport University.
E-mail: mik2178@yandex.ru
ORCID iD: https://orcid.org/0000-0002-3253-5697