Применение математического программирования для выбора оптимальных структур многомерных линейных регрессий

М. П. Базилевский

Иркутский государственный университет путей сообщения, Иркутск, Россия

Аннотация. В статье сформулирована задача одновременного отбора в многомерных линейных регрессиях как откликов, так и объясняющих переменных. Эта задача названа «отбор ключевых признаков и информативных регрессоров». Для оценивания регрессий применен метод наименьших квадратов. Сначала задача отбора заданного числа ключевых признаков и информативных регрессоров по критерию максимума суммы коэффициентов детерминации регрессий была сведена к задаче частично-булевого линейного программирования. Затем в нее были введены ограничения на знаки оценок, что позволило осуществлять отбор оптимальных структур многомерных регрессий. После чего добавлены ограничения на абсолютные вклады регрессоров в общие детерминации, что позволяет контролировать количество объясняющих переменных. При проведении вычислительных экспериментов на реальных данных при фиксированном числе ключевых признаков на построение многомерных моделей предложенным методом ушло примерно в 67,3 раза меньше времени, чем на построение их методом всех возможных регрессий. При этом ужесточение ограничений на абсолютные вклады регрессоров еще больше снизило время решения задач.

Ключевые слова: многомерная линейная регрессия, метод наименьших квадратов, коэффициент детерминации, отбор ключевых признаков и информативных регрессоров, задача частично-булевого линейного программирования, абсолютный вклад переменной в детерминацию, метод всех возможных регрессий.

DOI10.14357/20718632240404 **EDN** BBFOVP

Введение

На сегодняшний день модели машинного обучения представляют собой весьма мощный инструмент анализа данных, с помощью которого решаются сложные задачи в различных областях человеческой деятельности [1–3]. Существует множество типов таких моделей, среди которых искусственные нейронные сети, регрессионные модели, модели классификации, модели случайного леса, деревья решений и др. Данная статья посвящена проблеме построения регрессионных моделей – одним из главных представителей актуального сегодня интерпретируемого машинного обучения [4]. Идентифицированная регрессионная модель позволяет исследователю количественно оценить влияние факторов на результативный признак, получить прогноз отклика и принять обоснованные управленческие решения. Регрессионный анализ часто применяется для моделирования социально-экономических процессов и явлений [5], а построенная по экономическим данным модель называется эконометрической. Однако с помощью регрессионных моделей успешно решается множество задач и из других областей. Например,

в [6] с помощью регрессионного анализа получены прогнозы качества оросительной воды, в [7] построены регрессионные модели для выявления риска диабета, а в [8] – для прогнозирования энергопотребления зданий.

При построении регрессионной модели претендентов на роль факторов, влияющих на результативный признак, может быть так много, что приходится решать задачу выбора из них только наиболее значимых. Такая задача также известна в отечественной литературе как задача отбора наиболее существенных переменных [9] или информативных регрессоров (ОИР) в регрессии, а в зарубежной – subset selection in regression [10; 11]. Ее решение можно получить с помощью метода всех возможных регрессий [9]. Он состоит в оценивании моделей со всеми возможными комбинациями факторов и выборе такой комбинации объясняющих переменных, для которой, например, коэффициент детерминации с результирующим показателем максимален. Но, к сожалению, метод всех возможных регрессий требует больших объемов вычислений. Для преодоления вычислительных трудностей, как отмечено в [9], используется подхода. Первый из них связан с использованием полуэвристических методов оптимизации, например, метода «ветвей и границ», смысл которого состоит в отбрасывании большинства бесперспективных комбинаций объясняющих переменных на основе некоторого грубого правила. Второй связан с использованием пошаговых процедур отбора переменных, которые также не гарантируют получения оптимального решения с точки зрения коэффициента детерминации. На наш взгляд, эффективным и гарантирующим оптимальность набора переменных выходом из сложившейся ситуации будет использование аппарата частично-целочисленного линейного программирования, компьютерное оборудование и алгоритмы которого, как отмечено в [12], за последние десятилетия стали быстрее примерно в 1000 раз.

В настоящее время существует множество различных формулировок задачи ОИР в терминах математического программирования. Так, в [13] сформулирована задача отбора заданного числа регрессоров по критерию суммы квадратов ошибок, в [14] – задачи отбора оптимального числа регрессоров по скорректированному коэффициенту детерминации, критерию Акаике и Шварца. В [15] представлена задача ОИР с контролем мультиколлинеарности по коэффициентам вздутия дисперсии VIF. В [16] приведена формулировка задачи ОИР для линейной регрессии по критерию минимальной избыточности и максимальной релевантности, а в [17] – для регрессии Пуассона по критерию максимизации логарифмической функции правдоподобия. Объединяет введенные в работах [13–17] формулировки то, что они сделаны в рамках аппарата частично-целочисленного квадратичного программирования. Принципиально иной подход был изложен в статье [18], в которой задача ОИР для линейной регрессии, оцениваемой методом наименьших квадратов (МНК), была сведена к задаче частично-булевого линейного программирования (ЧБЛП). В [19] эта задача дополнилась линейными ограничениями на мультиколлинеарность. В [20] предложена ее модификация с использованием скорректированного коэффициента детерминации, а в [21] введены линейные ограничения на абсолютные вклады переменных в общую детерминацию. В работе [22] экспериментально доказана высокая эффективность решения задач ОИР в виде задач ЧБЛП по сравнению с методом всех возможных регрессий.

Приведенные в работах [13–22] задачи справедливы для отбора наиболее существенных факторов только во множественных линейных регрессиях (multiple linear regressions), состоящих из одного уравнения. Однако они представляют собой лишь частный случай так называемых многомерных линейных регрессий (multivariate linear regressions) [23; 24], в которых зависимых переменных не меньше двух. Научный интерес вызывает задача одновременного выбора в них и результативных признаков, и наиболее информативных регрессоров. Понятно, что такая задача гораздо более объемная в вычислительном плане, чем задача ОИР для множественных регрессий. Но, располагая предложенным в [18–22] эффективным методом, эта объемная задача не выглядит абсолютно недоступной. Цель данной работы – сформулировать в терминах ЧБЛП задачу одновременного отбора в многомерных регрессиях и результативных

признаков и соответствующих им наиболее информативных регрессоров, а также протестировать скорость ее решения на реальных данных.

1. Постановка задачи

Предположим, что выборочная совокупность объема n содержит значения l переменных x_1, x_2, …, x_l, причем, $l \geq 3$. Допустим, что каждая переменная может быть как зависимой (объясняемой, эндогенной), так и независимой (объясняющей, экзогенной). Тогда введем следующую систему независимых уравнений:

$$\begin{cases} x_{i1} = \alpha_{01} + \alpha_{21}x_{i2} + \alpha_{31}x_{i3} + ... + \alpha_{l,1}x_{il} + \varepsilon_{i1}, \\ x_{i2} = \alpha_{02} + \alpha_{12}x_{i1} + \alpha_{32}x_{i3} + ... + \alpha_{l,2}x_{il} + \varepsilon_{i2}, \\ ... \\ x_{il} = \alpha_{0l} + \alpha_{1l}x_{i1} + \alpha_{2l}x_{i2} + ... + \alpha_{l-1,l}x_{i,l-1} + \varepsilon_{il}, \\ i = \overline{1,n}, \end{cases} \quad (1)$$

где α_{jk}, $k = \overline{1,l}$, $j = \overline{0,l}$, $k \neq j$ – неизвестный параметр при j-й переменной в k-м уравнении системы; ε_{ik}, $i = \overline{1,n}$, $k = \overline{1,l}$ – i-я ошибка аппроксимации в k-м уравнении системы.

Сформулируем следующую задачу: из исходных l факторов требуется выделить p результативных признаков и m влияющих на каждый из них объясняющих переменных (регрессоров) так, чтобы максимизировать сумму коэффициентов детерминации включенных в систему (1) уравнений. В результате ее решения будет построена многомерная линейная регрессия. Назовем этот процесс отбором ключевых признаков и информативных регрессоров (ОКПиИР). Такая задача может быть решена методом всех возможных регрессий. Для этого необходимо оценить $p \cdot C_l^p \cdot C_{l-p}^m$ линейных регрессий, что гораздо больше, чем общее число моделей C_{l-1}^m при решении задачи ОИР.

Для оценивания уравнений в системе (1) будем использовать МНК. Для упрощения расчетов оценок воспользуемся известным приемом, описанным в [25]. Для этого проведем стандартизацию (нормирование) всех переменных по правилам:

$$x_{ij}^{\bullet} = \frac{x_{ij} - \overline{x_j}}{\sigma_{x_j}}, \quad i = \overline{1,n}, \; j = \overline{1,l},$$

где $x_{ij}^{\bullet}$, $i = \overline{1,n}$, $j = \overline{1,l}$ – значения стандартизованных переменных; $\overline{x_j}$, $j = \overline{1,l}$ – средние значения переменных; σ_{x_j}, $j = \overline{1,l}$ – стандартные отклонения переменных.

Тогда система независимых уравнений (1) в стандартизованном масштабе примет вид:

$$\begin{cases} x_{i1}^{\bullet} = \beta_{21}x_{i2}^{\bullet} + \beta_{31}x_{i3}^{\bullet} + ... + \beta_{l,1}x_{il}^{\bullet} + \varepsilon_{i1}^{\bullet}, \\ x_{i2}^{\bullet} = \beta_{12}x_{i1}^{\bullet} + \beta_{32}x_{i3}^{\bullet} + ... + \beta_{l,2}x_{il}^{\bullet} + \varepsilon_{i2}^{\bullet}, \\ ... \\ x_{il}^{\bullet} = \beta_{1l}x_{i1}^{\bullet} + \beta_{2l}x_{i2}^{\bullet} + ... + \beta_{l-1,l}x_{i,l-1}^{\bullet} + \varepsilon_{il}^{\bullet}, \\ i = \overline{1,n}, \end{cases} \quad (2)$$

где β_{jk}, $k = \overline{1,l}$, $j = \overline{1,l}$, $k \neq j$ – неизвестный стандартизованный коэффициент при j-й переменной в k-м уравнении системы; $\varepsilon_{ik}^{\bullet}$, $i = \overline{1,n}$, $k = \overline{1,l}$ – i-я ошибка аппроксимации в k-м уравнении системы.

Введем матрицу корреляций между всеми переменными

$$R_{xx} = \begin{pmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_l} \\ r_{x_1x_2} & 1 & \dots & r_{x_2x_l} \\ \dots & \dots & \dots & \dots \\ r_{x_1x_l} & r_{x_2x_l} & \dots & 1 \end{pmatrix}.$$

Тогда, как следует из [25], для оценивания первой регрессии в системе (2) нужно сформировать систему линейных алгебраических уравнений. Для этого нужно исключить из матрицы R_{xx} первую строку, затем взять первый столбец как вектор свободных членов, а остальные элементы – как основную матрицу системы. В матричном виде она будет иметь вид:

$$\begin{pmatrix} 1 & r_{x_2x_3} & \dots & r_{x_2x_l} \\ r_{x_2x_3} & 1 & \dots & r_{x_3x_l} \\ \dots & \dots & \dots & \dots \\ r_{x_2x_l} & r_{x_3x_l} & \dots & 1 \end{pmatrix} \begin{pmatrix} \beta_{21} \\ \beta_{31} \\ \dots \\ \beta_{l,1} \end{pmatrix} = \begin{pmatrix} r_{x_1x_2} \\ r_{x_1x_3} \\ \dots \\ r_{x_1x_l} \end{pmatrix}. \quad (3)$$

Альтернативная форма записи системы (3) такова:

$$\sum_{j=1}^{l-1} r_{x_{s_i}x_{s_j}} \cdot \beta_{s_j,1} = r_{x_{s_i}x_1}, \qquad i=\overline{1,l-1},$$

где s_i, $i=\overline{1,l-1}$ – элементы вектора $\begin{pmatrix}2 & 3 & \dots & l\end{pmatrix}$.

Для оценивания второй регрессии в системе (2), путем вычеркивания второй строки из матрицы R_{xx}, формируется система

$$\begin{pmatrix} 1 & r_{x_1x_3} & \dots & r_{x_1x_l} \\ r_{x_1x_3} & 1 & \dots & r_{x_3x_l} \\ \dots & \dots & \dots & \dots \\ r_{x_1x_l} & r_{x_3x_l} & \dots & 1 \end{pmatrix} \begin{pmatrix} \beta_{12} \\ \beta_{32} \\ \dots \\ \beta_{l,2} \end{pmatrix} = \begin{pmatrix} r_{x_1x_2} \\ r_{x_2x_3} \\ \dots \\ r_{x_2x_l} \end{pmatrix},$$

альтернативная форма записи которой

$$\sum_{j=1}^{l-1} r_{x_{s_i}x_{s_j}} \cdot \beta_{s_j,2} = r_{x_{s_i}x_2}, \qquad i=\overline{1,l-1},$$

где s_i, $i=\overline{1,l-1}$ – элементы вектора $\begin{pmatrix}1 & 3 & \dots & l\end{pmatrix}$.

Тогда для МНК-оценивания всех регрессий в системе (2) нужно решить следующую совокупность систем линейных алгебраических уравнений:

$$\sum_{j=1}^{l-1} r_{x_{q_{ki}}x_{q_{kj}}} \cdot \beta_{q_{kj},k} = r_{x_{q_{ki}}x_k}, \quad k=\overline{1,l},\ i=\overline{1,l-1}, \quad (4)$$

где k – номер регрессии, i – номер уравнения в системе, q_{ij}, $i=\overline{1,l}$, $j=\overline{1,l-1}$ – элементы матрицы Q, полученной путем вычеркивания главной диагонали из квадратной матрицы $\begin{pmatrix} 1 & 2 & \dots & l \\ 1 & 2 & \dots & l \\ \dots & \dots & \dots & \dots \\ 1 & 2 & \dots & l \end{pmatrix}$.

Коэффициент детерминации регрессии в стандартизованном масштабе, как следует из [25], равен сумме произведений стандартизованных коэффициентов на соответствующие значения коэффициентов корреляции входных переменных с выходной. Тогда формулы коэффициентов детерминации регрессий системы (2) можно записать следующим образом:

$$R_k^2 = \sum_{j=1}^{l-1} r_{x_{q_{kj}}x_k} \cdot \beta_{q_{kj},k}, \qquad k=\overline{1,l}. \quad (5)$$

Заметим, что от оценок стандартизованных коэффициентов регрессий системы (2) можно легко перейти к оценкам регрессий системы (1), основываясь на представленных в [25] формулах.

Перейдем к формализации задачи ОКПиИР в многомерной линейной регрессии в терминах аппарата математического программирования. Введем булевы переменные σ_k, $k=\overline{1,l}$, по правилу

$$\sigma_k = \begin{cases} 1, & \text{если в систему (2) включается} \\ & k\text{-я линейная регрессия,} \\ 0, & \text{в противном случае.} \end{cases}$$

Учитывая, что в систему должно входить ровно p регрессий, введем ограничение

$$\sum_{k=1}^{l} \sigma_k = p. \quad (6)$$

Если регрессия под номером k не входит в систему (2), то соответствующая ей система линейных алгебраических уравнений должна каким-то образом исключаться из совокупности (4), а соответствующие оценки обнуляться. Это можно получить с помощью следующих линейных ограничений:

$$-M \cdot \sigma_k \le \beta_{q_{kj},k} \le M \cdot \sigma_k,\ k=\overline{1,l},\ j=\overline{1,l-1}, \quad (7)$$

$$-(1-\sigma_k)M \le \sum_{j=1}^{l-1} r_{x_{q_{ki}}x_{q_{kj}}} \cdot \beta_{q_{kj},k} - r_{x_{q_{ki}}x_k} \le (1-\sigma_k)M,$$

$$k=\overline{1,l},\ i=\overline{1,l-1}, \quad (8)$$

где M – большое положительное число.

Если $\sigma_k = 1$, то $\beta_{q_{kj},k} \in [-M, M]$, т.е. никаких ограничений на МНК-оценки в k-й регрессии нет, а $\sum_{j=1}^{l-1} r_{x_{q_{ki}}x_{q_{kj}}} \cdot \beta_{q_{kj},k} - r_{x_{q_{ki}}x_k} = 0$, т.е. в совокупность (4) включается k-я система линейных алгебраических уравнений. Если же $\sigma_k = 0$, то $\beta_{q_{kj},k} = 0$, а $\sum_{j=1}^{l-1} r_{x_{q_{ki}}x_{q_{kj}}} \cdot \beta_{q_{kj},k} - r_{x_{q_{ki}}x_k} \in [-M, M]$, что означает неучастие k-й системы линейных алгебраических уравнений в совокупности (4).

Очевидно, что если $\sigma_k = 1$, то R_k^2 принимает значение коэффициента детерминации k-й регрессии в системе (2), а если $\sigma_k = 0$, то $R_k^2 = 0$. Тогда, используя формулу (5), введем целевую функцию:

$$\sum_{k=1}^{l} R_k^2 = \sum_{k=1}^{l} \sum_{j=1}^{l-1} r_{x_{q_{kj}} x_k} \cdot \beta_{q_{kj},k} \to \max . \quad (9)$$

Тем самым, решение задачи ЧБЛП с целевой функцией (9) и с линейными ограничениями (6) – (8) приводит к отбору в системе (2) ровно p ключевых признаков с максимальной суммой коэффициентов детерминации. При этом на каждый признак влияют абсолютно все переменные. Поэтому далее дополним эту задачу ограничениями для отбора ровно m одинаковых информативных регрессоров в каждой регрессии.

Введем бинарные переменные δ_j, $j = \overline{1,l}$, по правилу

$$\delta_j = \begin{cases} 1, \text{ если } j - \text{я переменная входит} \\ \qquad \text{в отобранные регрессии,} \\ 0, \text{ в противном случае.} \end{cases}$$

Количество регрессоров в регрессиях регулируется ограничением

$$\sum_{j=1}^{l} \delta_j = m . \quad (10)$$

Если состав регрессоров во всех регрессиях одинаковый, то каждая переменная может быть только либо зависимой, либо независимой. Для этого сформулируем ограничения вида

$$\sigma_k + \delta_k \le 1, \; k = \overline{1,l} . \quad (11)$$

Из (11) следует, что если $\sigma_k = 1$, т.е. k-я переменная является ключевым признаком, то $\delta_k = 0$, т.е. она никогда не может быть включена в состав регрессоров. Если же $\sigma_k = 0$, то $\delta_k \le 1$, т.е. k-я переменная может как входить, так и не входить в состав регрессоров.

Для контроля конфигураций систем линейных алгебраических уравнений в совокупности (4) введем следующие ограничения:

$$-M \cdot \delta_{q_{kj}} \le \beta_{q_{kj},k} \le M \cdot \delta_{q_{kj}}, \; k = \overline{1,l}, \; j = \overline{1,l-1}, \quad (12)$$

$$-\left(1-\delta_{q_{ki}}\right) M \le \sum_{j=1}^{l-1} r_{x_{q_{ki}} x_{q_{kj}}} \cdot \beta_{q_{kj},k} - r_{x_{q_{ki}} x_k} \le \left(1-\delta_{q_{ki}}\right) M ,$$

$$k = \overline{1,l}, \; i = \overline{1,l-1} . \quad (13)$$

Если $\delta_{q_{kj}} = 1$, то $\beta_{q_{kj},k} \in [-M, M]$, т.е. никаких ограничений на МНК-оценку при q_{kj}-й переменной в k-й регрессии нет, а $\sum_{j=1}^{l-1} r_{x_{q_{ki}} x_{q_{kj}}} \cdot \beta_{q_{kj},k} - r_{x_{q_{ki}} x_k} = 0$, т.е. в совокупность (4) в k-ю систему линейных алгебраических уравнений включается q_{kj}-е уравнение. Если же $\delta_{q_{kj}} = 0$, то $\beta_{q_{kj},k} = 0$, а $\sum_{j=1}^{l-1} r_{x_{q_{ki}} x_{q_{kj}}} \cdot \beta_{q_{kj},k} - r_{x_{q_{ki}} x_k} \in [-M, M]$, что означает неучастие q_{kj}-го уравнения в k-й системе линейных алгебраических уравнений в совокупности (4).

К сожалению, совместное применение ограничений (8) и (13) приводит к некорректному отбору переменных. Например, если $\sigma_k = 1$, то в совокупности (4) в k-й системе все уравнения срабатывают, т.е. обращаются в строгие равенства. Поэтому ограничения (13) уже не будут оказывать на их срабатывания никакого влияния. Для устранения сложившейся проблемной ситуации совместим ограничения (8) и (13) следующим образом:

$$-\left(1-\sigma_k\right) M - \left(1-\delta_{q_{ki}}\right) M \le \sum_{j=1}^{l-1} r_{x_{q_{ki}} x_{q_{kj}}} \cdot \beta_{q_{kj},k} - r_{x_{q_{ki}} x_k} \le$$

$$\le \left(1-\delta_{q_{ki}}\right) M + \left(1-\sigma_k\right) M, \; k = \overline{1,l}, \; i = \overline{1,l-1}. \quad (14)$$

Теперь в совокупности (4) в k-й системе будет срабатывать q_{ki}-е уравнение только тогда, когда $\sigma_k = 1$ и $\delta_{q_{ki}} = 1$, т.е. когда выбрана k-я регрессия и q_{ki}-й регрессор.

Таким образом, решение задачи ЧБЛП с целевой функцией (9) и с линейными ограничениями (6), (7), (10) – (12), (14) приводит к отбору в системе (2) ровно p ключевых признаков и m влияющих на них информативных регрессоров по критерию максимума суммы коэффициентов детерминации.

Заметим, что если в этой задаче ОКПиИР ключевой признак ровно один и он известен, то ее решение будет равносильно решению задачи ОИР, рассмотренной в [18].

Поскольку решение задачи ОИР, как следует из [22], становится эффективнее при введении ограничений на знаки МНК-оценок, то было принято решение ввести их и для задачи ОКПиИР, предложив новую формулировку. Идея состоит в том, чтобы согласовать знаки МНК-оценок со знаками коэффициентов корреляции между ключевыми признаками и информативными регрессорами. Такое согласование позитивно скажется на мультиколлинеарности в регрессиях, а также сделает доступными величины абсолютных вкладов регрессоров в общие детерминации.

Введем следующие ограничения:

$$0 \le \beta_{q_{kj},k} \le M \cdot \sigma_k,\ k=\overline{1,l},\ j=\overline{1,l-1},\ r_{x_{q_{kj}},x_k} > 0, \quad (15)$$

$$-M \cdot \sigma_k \le \beta_{q_{kj},k} \le 0,\ k=\overline{1,l},\ j=\overline{1,l-1},\ r_{x_{q_{kj}},x_k} < 0, \quad (16)$$

$$0 \le \beta_{q_{kj},k} \le M \cdot \delta_{q_{kj}},\ k=\overline{1,l},\ j=\overline{1,l-1},\ r_{x_{q_{kj}},x_k} > 0, \quad (17)$$

$$-M \cdot \delta_{q_{kj}} \le \beta_{q_{kj},k} \le 0,\ k=\overline{1,l},\ j=\overline{1,l-1},\ r_{x_{q_{kj}},x_k} < 0. \quad (18)$$

Тогда решение задачи ЧБЛП с целевой функцией (9) и с линейными ограничениями (6), (10), (11), (14), (15) – (18), если оно существует, приводит к отбору в многомерной линейной регрессии ровно p ключевых признаков и m влияющих на них информативных регрессоров с согласованными знаками МНК-оценок. Если из этой задачи исключить ограничения (6), то осуществляется отбор оптимального числа ключевых признаков для m информативных регрессоров. Если же из нее исключить ограничения (10), то осуществляется отбор оптимального числа информативных регрессоров для p ключевых признаков. А если исключить (6) и (10), то осуществляется отбор и оптимального числа признаков, и оптимального числа информативных регрессоров.

Из-за ограничений (15) – (18) становятся доступны величины:

$$C^{\text{абс}}_{x_k,x_{q_{kj}}} = r_{x_{q_{kj}}x_k} \cdot \beta_{q_{kj},k}, \qquad k=\overline{1,l},\ j=\overline{1,l-1},$$

где $C^{\text{абс}}_{x_k,x_{q_{kj}}} \ge 0$ – абсолютный вклад q_{kj}-го регрессора в общую детерминацию R_k^2 для k-го ключевого признака. Очевидно, что для этих величин справедливы следующие тождества:

$$\sum_{j=1}^{l-1} C^{\text{абс}}_{x_k,x_{q_{kj}}} = R_k^2, \quad k=\overline{1,l}.$$

Тогда введем следующие линейные ограничения на эти величины:

$$r_{x_{q_{kj}}x_k} \cdot \beta_{q_{kj},k} \ge \theta\left(\delta_{q_{kj}} + \sigma_k - 1\right), \quad k=\overline{1,l},\ j=\overline{1,l-1}, \quad (19)$$

где $\theta \in [0,1)$ – нижнее пороговое значение абсолютных вкладов регрессоров во входящих в систему (2) регрессий. Из (19) следует, что ограничение вида $r_{x_{q_{kj}}x_k} \cdot \beta_{q_{kj},k} \ge \theta$ на абсолютный вклад q_{kj}-го регрессора в детерминацию для k-го ключевого признака срабатывает только тогда, когда в систему (2) включены обе этих переменных, т.е. когда $\sigma_k = 1$ и $\delta_{q_{kj}} = 1$. В противном случае ограничения (19) принимают либо вид $r_{x_{q_{kj}}x_k} \cdot \beta_{q_{kj},k} \ge 0$, либо $r_{x_{q_{kj}}x_k} \cdot \beta_{q_{kj},k} \ge -\theta$, что справедливо в любом случае. Чем больше значение θ, тем жестче требование к степени влияния регрессоров на ключевые признаки, и тем больше шансов на снижение числа регрессоров в регрессиях. Ограничения (19) могут быть использованы для контроля абсолютных вкладов регрессоров в задаче ЧБЛП (9), (6), (10), (11), (14), (15) – (18) и в упомянутых выше различных ее вариациях.

2. Вычислительные эксперименты

Для проведения экспериментов были использованы статистические данные из базы Федеральной службы государственной статистики, характеризующие научную и инновационную деятельность в Иркутской области за период с 2000 по 2021 гг. Всего было задействовано 25 переменных:

x_1 – организации, выполнявшие научные исследования и разработки (единиц);

x_2 – численность исследователей, занятых научными исследованиями и разработками (человек);

x_3 – численность техников (человек);

x_4 – численность вспомогательного персонала (человек);

x_5 – численность прочего персонала (человек);

x_6 – численность исследователей с ученой степенью доктора наук (человек);

x_7 – с ученой степенью кандидата наук (человек);

x_8 – внутренние текущие затраты на оплату труда (млн руб.);

x_9 – внутренние текущие затраты на страховые взносы на ОПС, ОМС, ОСС (млн руб.);

x_{10} – внутренние текущие затраты на приобретение оборудования (млн руб.);

x_{11} – другие материальные затраты (млн руб.);

x_{12} – прочие текущие затраты (млн руб.);

x_{13} – капитальные затраты на научные исследования и разработки (млн руб.);

x_{14} – внутренние текущие затраты на фундаментальные исследования (млн руб.);

x_{15} – внутренние текущие затраты на прикладные исследования (млн руб.);

x_{16} – внутренние текущие затраты на разработки (млн руб.);

x_{17} – подано патентных заявок на изобретения (единиц);

x_{18} – подано патентных заявок на полезные модели (единиц);

x_{19} – выдано патентов на изобретения (единиц);

x_{20} – выдано патентов на полезные модели (единиц);

x_{21} – используемые передовые производственные технологии (единиц);

x_{22} – общие затраты на инновационную деятельность организаций (млн руб.);

x_{23} – затраты на инновационную деятельность в процентах от общего объема отгруженных товаров, выполненных работ, услуг (%);

x_{24} – общий объем инновационных товаров, работ, услуг (млн руб.);

x_{25} – объем инновационных товаров в процентах от общего объема отгруженных товаров, выполненных работ, услуг (%).

Главной целью вычислительных экспериментов было сравнение эффективности решения задач построения многомерных линейных регрессий методом всех возможных регрессий и предложенным нами методом. Эксперименты проводились на персональном компьютере с процессором AMD Ryzen 3 4300U (2,70 ГГц) и объемом оперативной памяти 16 Гб. Для решения задач ЧБЛП был использован бесплатный оптимизационный решатель LPSolve, имеющий встроенный счетчик, с помощью которого можно фиксировать точное время поиска решения. Для оценивания времени решения задач методом всех возможных регрессий была использована справедливая при $2 \le m \le 36$ следующая экспериментально полученная в [22] зависимость:

$$v(m) = e^{12,9507} \cdot m^{-1,45763},$$

где $v(m)$ – число моделей с m регрессорами, оцениваемых методом всех регрессий в пакете Gretl за 1 секунду на нашем персональном компьютере. Например, если $m = 5$, то скорость составляет 40325,33 моделей в секунду. Зная эту скорость и общее число моделей, можно без труда определить время.

Во всех задачах ЧБЛП большое число M было выбрано равным 100. И все они решались с использованием ограничений на знаки МНК-оценок (15) – (18).

Эксперимент №1. Двумя методами решались задачи ОКПиИР при фиксированном числе ключевых признаков p ($1 \le p \le 3$) и числе информативных регрессоров m ($1 \le m \le 3$). В задачах ЧБЛП с целевой функцией (9) использовались ограничения (6), (10), (11), (14), (15) – (18), т.е. ограничения на абсолютные вклады (19) не ставились.

Результаты эксперимента представлены в Табл. 1. В ней в первом столбце указан номер задачи, во втором – число ключевых признаков p, в третьем – число информативных регрессоров m, в четвертом, пятом и шестом – коэффициент детерминации каждой регрессии в системе, состав признаков и состав регрессоров, в седьмом и восьмом – время t_1 решения задачи методом всех регрессий в пакете Gretl и время t_2 решения нашим методом в пакете LPSolve. Время решения методом всех регрессий определялось по формуле

Табл. 1. Результаты эксперимента №1

№	p	m	R^2	Признаки	Регрессоры	t_1, сек	t_2, сек
1	1	1	0,96938	x_8	x_9	0,0014	3,366
2	2	1	0,96938	x_9	x_8	0,033	41,623
			0,93411	x_{14}			
3	3	1	0,96938	x_9	x_8	0,361	334,073
			0,93411	x_{14}			
			0,89782	x_{15}			
4	1	2	0,98959	x_8	x_9, x_{12}	0,045	14,87
5	2	2	0,97004	x_9	x_8, x_{10}	0,99	237,727
			0,98827	x_{14}			
6	3	2	0,93783	x_2	x_8, x_{21}	10,395	2195,324
			0,98462	x_9			
			0,9403	x_{14}			
7	1	3	0,99319	x_8	x_9, x_{16}, x_{20}	0,596	57,892
8	2	3	0,9862	x_9	x_3, x_8, x_{10}	12,515	895,723
			0,98887	x_{14}			
9	3	3	0,94635	x_2	x_7, x_{14}, x_{15}	125,146	8349,368
			0,97484	x_8			
			0,97152	x_9			

$$t_1 = pC_{25}^{p}\frac{C_{25-p}^{m}}{v(m)}.$$

Как видно из Табл. 1, наш метод оказался существенно менее эффективным, чем метод всех возможных регрессий. Схожий эффект проявлялся и ранее при решении задач ОИР [18] для фиксированного числа регрессоров. Выправить ситуацию, например, в [22], позволило исключение из задачи ЧБЛП ограничений на число регрессоров. Поэтому было решено провести следующий эксперимент.

Эксперимент №2. Двумя методами решались задачи ОКПиИР только при фиксированном числе ключевых признаков p $(1 \le p \le 3)$. В задачах ЧБЛП с целевой функцией (9) использовались ограничения (6), (11), (14), (15) – (18). Сначала ограничения на абсолютные вклады не ставились, а потом пороговое значение θ в (19) принималось равным 0,1.

Результаты эксперимента представлены в Табл. 2. В ней в первом столбце указан номер задачи, во втором – число ключевых признаков p, в третьем – число отобранных регрессоров, в четвертом – сумма коэффициентов детерминации регрессий в системе, в пятом – общее число моделей для оценивания методом всех регрессий, в шестом – время t_1 решения задачи методом всех регрессий, в седьмом – время t_2 решения нашим методом. Поскольку объем выборки $n = 22$, то максимальное число регрессоров в уравнении может быть 21, поэтому общее число

Табл. 2. Результаты эксперимента №2

№	p	Регрессоры	Сумма R^2	Число моделей для перебора	t_1, сек	t_2, сек
Нет ограничений на вклады						
1	1	15	0,99869	419 422 850	37784,9	601,7
2	2	12	1,99259	5 033 149 800	426419,1	7206,6
3	3	12	2,98018	28 940 683 800	2299615,4	46860,1
θ = 0,1						
4	1	5	0,99384	113 509 625	6644,6	125,8
5	2	4	1,97506	1 705 335 000	98393,9	1083,1
6	3	3	2,87356	12 036 601 500	682792,3	7659,9

моделей для метода всех регрессий определялось по формуле

$$N = pC_{25}^{p}\left(C_{25-p}^{1} + C_{25-p}^{2} + \ldots + C_{25-p}^{21}\right),$$

а время

$$t_1 = pC_{25}^{p}\left(\frac{C_{25-p}^{1}}{v(1)} + \frac{C_{25-p}^{2}}{v(2)} + \ldots + \frac{C_{25-p}^{21}}{v(21)}\right).$$

При $\theta = 0{,}1$ максимальное число регрессоров в уравнении может быть 10, поэтому в этих формулах в скобках нужно оставить только первые десять слагаемых.

Из Табл. 2 видно, во-первых, что все задачи требуют больших объемов вычислений методом всех возможных регрессий. Так, при отсутствии ограничений на вклады при p=1 на решение уйдет примерно 10,5 часов, при p=2 – 118,45 часов, при p=3 – 638,78 часов. При наличии ограничений на вклады это время, конечно же, сократится, но все равно останется довольно большим: 1,85 часа для p=1, 27,33 часов для p=2 и 189,66 часов для p=3. Во-вторых, наш метод во всех шести случаях оказался гораздо эффективнее метода всех возможных регрессий. Так, для задачи №1 время решения уменьшилось в 62,8 раза, для задачи №2 – в 59,17 раз, для задачи №3 – в 49,07 раз, для задачи №4 – в 52,82 раза, для задачи №5 – в 90,84 раз, для задачи №6 – в 89,14 раз. Тем самым, в среднем время уменьшилось примерно в 67,3 раз. В-третьих, использование ограничений на вклады во всех трех случаях привело к снижению времени решения задач ЧБЛП. Так, при p=1 время уменьшилось в 4,78 раза, при p=2 – в 6,65 раз, при p=3 – в 6,18 раз. Исходя из этого, в среднем время уменьшилось в 5,85 раз. Судя по всему, при p>3 наш метод также будет эффективнее метода всех регрессий, но на доказательство этого факта потребовалось бы довольно много времени. Поэтому было принято решение провести другой более простой в вычислительном плане эксперимент.

Эксперимент №3. Двумя методами решались задачи ОИР для конкретного набора ключевых признаков. В задачах ЧБЛП с целевой функцией (9) использовались ограничения (11), (14), (15) – (18). Вместо ограничений (6) булевым переменным придавались конкретные значения. Сначала ограничения на абсолютные вклады не ставились, а потом пороговое значение θ в (19) назначалось равным 0,1.

Общий состав ключевых признаков был выбран на основе пятого столбца Табл. 1, содержащего переменные x_2, x_8, x_9, x_{14}, x_{15}. Комбинируя сочетания переменных из этого множества, в общей сложности решалась $2^5 - 1 = 31$ задача.

Результаты эксперимента представлены в Табл. 3. В ней в первом столбце указан номер задачи, во втором – состав признаков, в третьем, четвертом, пятом и шестом – полученные при отсутствии ограничений на вклады показатели: число отобранных регрессоров, сумма коэффициентов детерминации регрессий, время t_1 решения задачи методом всех регрессий и время t_2 решения задачи нашим методом. Аналогичные показатели, полученные при ограничениях на вклады, приведены в седьмом, восьмом, девятом и десятом столбцах. При отсутствии ограничений на вклады время решения задачи методом всех регрессий определялось по формуле

$$t_1 = p\left(\frac{C_{25-p}^{1}}{v(1)} + \frac{C_{25-p}^{2}}{v(2)} + \ldots + \frac{C_{25-p}^{21}}{v(21)}\right),$$

а при наличии ограничений на вклады в этой формуле в скобках брались только первые десять слагаемых.

Из Табл. 3 видно, что, во-первых, во всех шестидесяти двух случаях наш метод снова оказался эффективнее метода всех возможных регрессий. Так, при отсутствии ограничений на вклады при p=1 время решения уменьшилось в 7,8 – 162,52 раз, при p=2 – в 9,48 – 78,1 раз, при p=3 – в 8,25 – 25,97 раз, при p=4 – в 6,08 – 12,13 раз, при p=5 – в 5,61 раз. При наличии ограничений при p=1 время уменьшилось в 42,19 – 156,35 раз, при p=2 – в 45,54 – 149,04 раз, при p=3 – в 36,65 – 106,04 раз, при p=4 – в 38,23 – 70,67 раз, при p=5 – в 45,24 раз. Во-вторых, использование ограничений на вклады во всех тридцати двух случаях привело к снижению времени решения задач ЧБЛП. Так, без ограничений на вклады среднее время решения составило 68,358 сек., а с ограничениями – 4,367 сек., что в 15,65 раз меньше. В-третьих, применение ограничений на вклады приводит к построению более компактных моделей с меньших числом регрессоров.

Табл. 3. Результаты эксперимента №3

№	Признаки	Нет ограничений на вклады				$\theta = 0{,}1$			
		Регрес-соры	Сумма R^2	t_1, сек	t_2, сек	Регрес-соры	Сумма R^2	t_1, сек	t_2, сек
1	x_2	13	0,99359	1511,4	26,5	5	0,98019	265,8	2
2	x_8	15	0,99869	1511,4	193,7	5	0,99384	265,8	4,8
3	x_9	11	0,99739	1511,4	32,7	3	0,9916	265,8	3,2
4	x_{14}	14	0,99431	1511,4	53,1	4	0,98989	265,8	6,3
5	x_{15}	10	0,9506	1511,4	9,3	5	0,93687	265,8	1,7
6	x_2, x_8	13	1,98599	1421,4	150	3	1,86696	327,9	3,2
7	x_2, x_9	11	1,98253	1421,4	92,5	4	1,92356	327,9	4,2
8	x_2, x_{14}	12	1,97896	1421,4	49,5	3	1,83493	327,9	3,9
9	x_2, x_{15}	9	1,92555	1421,4	18,2	2	1,76002	327,9	2,2
10	x_8, x_9	12	1,99259	1421,4	136,4	4	1,97506	327,9	5,9
11	x_8, x_{14}	10	1,99082	1421,4	119,5	4	1,93126	327,9	7,2
12	x_8, x_{15}	9	1,93626	1421,4	51,1	2	1,91111	327,9	4,1
13	x_9, x_{14}	12	1,98493	1421,4	59,2	2	1,92839	327,9	6,7
14	x_9, x_{15}	8	1,93328	1421,4	27,1	3	1,89089	327,9	3,2
15	x_{14}, x_{15}	7	1,9238	1421,4	39,4	2	1,84956	327,9	5,5
16	x_2, x_8, x_9	11	2,97625	999,8	121,1	3	2,82871	296,9	4,1
17	x_2, x_8, x_{14}	11	2,97011	999,8	85,5	3	2,76785	296,9	4
18	x_2, x_8, x_{15}	9	2,9059	999,8	39,6	1	2,6881	296,9	2,9
19	x_2, x_9, x_{14}	10	2,96059	999,8	80,7	2	2,80107	296,9	4,7
20	x_2, x_9, x_{15}	7	2,8975	999,8	38,5	2	2,66691	296,9	2,9
21	x_2, x_{14}, x_{15}	8	2,87953	999,8	44,7	2	2,65802	296,9	2,8
22	x_8, x_9, x_{14}	12	2,98018	999,8	80	3	2,87228	296,9	8,1
23	x_8, x_9, x_{15}	8	2,92482	999,8	48,9	3	2,87356	296,9	5,1
24	x_8, x_{14}, x_{15}	7	2,91858	999,8	58,3	2	2,83915	296,9	6,9
25	x_9, x_{14}, x_{15}	7	2,91039	999,8	53,9	1	2,80131	296,9	5,2
26	x_2, x_8, x_9, x_{14}	10	3,94928	623,3	102,5	3	3,71668	233,2	4,6
27	x_2, x_8, x_9, x_{15}	7	3,88227	623,3	51,4	2	3,59417	233,2	3,3
28	x_2, x_8, x_{14}, x_{15}	7	3,85504	623,3	62,4	1	3,60394	233,2	3,5
29	x_2, x_9, x_{14}, x_{15}	7	3,86088	623,3	60,9	1	3,56602	233,2	3,4
30	x_8, x_9, x_{14}, x_{15}	9	3,89844	623,3	67,8	3	3,7239	233,2	6,1
31	$x_2, x_8, x_9, x_{14}, x_{15}$	7	4,8439	363,1	64,7	2	4,49217	167,4	3,7

Для демонстрации корректности полученных в Табл. 3 результатов рассмотрим многомерную линейную регрессию с ключевыми признаками x_8, x_9, x_{14}, x_{15}, построенную при $\theta = 0{,}1$:

$$\begin{cases} x_8^* = -576{,}196 + \underset{(4{,}285)}{2{,}584}\, x_{11} + \underset{(5{,}157)}{1{,}538}\, x_{12} + \underset{(8{,}014)}{0{,}519}\, x_{21}, \\ x_9^* = -209{,}975 + \underset{(3{,}927)}{0{,}789}\, x_{11} + \underset{(3{,}022)}{0{,}3}\, x_{12} + \underset{(9{,}015)}{0{,}194}\, x_{21}, \\ x_{14}^* = -848{,}568 + \underset{(0{,}981)}{1{,}199}\, x_{11} + \underset{(4{,}055)}{2{,}45}\, x_{12} + \underset{(5{,}494)}{0{,}721}\, x_{21}, \\ x_{15}^* = -205{,}271 + \underset{(3{,}055)}{1{,}143}\, x_{11} + \underset{(3{,}34)}{0{,}618}\, x_{12} + \underset{(3{,}033)}{0{,}122}\, x_{21}, \end{cases} \quad (20)$$

для которой $R_8^2 = 0{,}96399$, $R_9^2 = 0{,}94985$, $R_{14}^2 = 0{,}90318$, $R_{15}^2 = 0{,}90688$.

В модели (20) в скобках над коэффициентами указаны значения абсолютных вкладов переменных в общую детерминацию. Как видно, все они, как и ожидалось, превосходят величину 0,1. А в скобках под коэффициентами указаны значения t-критерия Стьюдента. Оказалось, что все коэффициенты значимы для уровня значимости $\alpha = 0{,}01$, кроме коэффициента в третьем уравнении при переменной x_{11}. Это легко исправить, исключив незначимую переменную x_{11} и переоценив регрессию с помощью МНК:

$$x_{14}^{*} = \underset{(8,783)}{-798,074} + \underset{(8,783)}{2,944}\, x_{12} + \underset{(5,429)}{0,708}\, x_{21},$$

для которой $R_{14}^2 = 0,898$. Теперь в ней все оценки значимы. Также хочется обратить внимание, что знаки всех оценок в многомерной линейной регрессии (20) согласуются с содержательным смыслом факторов.

Многомерную линейную регрессию (20) можно воспринимать, как первый шаг двухшагового МНК, предназначенного для оценки систем одновременных уравнений.

Заключение

В статье сформулирована задача отбора ключевых признаков и информативных регрессоров в многомерных линейных регрессиях. С использованием элементов корреляционного анализа эта задача была сведена к задаче частично-булевого линейного программирования. В ней можно контролировать число ключевых признаков, число информативных регрессоров, знаки оценок регрессий, абсолютные вклады переменных в общие детерминации. При фиксированном числе ключевых признаков и информативных регрессоров наш метод построения многомерных моделей существенно уступил методу всех возможных регрессий. Но при фиксировании только числа ключевых признаков превзошел метод всех регрессий по времени в 67,3 раз. При этом, при использовании ограничений на абсолютные вклады переменных в детерминации время решения оказалось в 5,85 раз меньше, чем без них. При заданном конкретном составе ключевых признаков наш метод превзошел метод всех регрессий по времени в 16,6 раз без ограничений на вклады, и в 65,8 раз с ограничениями на вклады. При этом снова подтвердилось снижение времени решения задач ЧБЛП при использовании ограничений на вклады.

Стоит подчеркнуть, что в данной работе все вычислительные эксперименты были проведены на конкретном персональном компьютере. При использовании более мощных вычислительных систем время решения рассмотренных задач ОКПиИР как методом всех возможных регрессий, так и с помощью аппарата ЧБЛП, естественным образом, уменьшится. Но, вероятнее всего, значения ускорений вычислений останутся практически без изменений. Для проверки достоверности этой гипотезы в дальнейшем планируется провести дополнительные исследования.

Предложенный математический аппарат может успешно применяться для решения реальных задач анализа данных. Однако особенно аккуратно его стоит использовать при построении многомерных линейных регрессий с неизвестными ключевыми признаками, чтобы в результате отбора в состав откликов не вошли переменные, не связанные с регрессорами причинно-следственными отношениями.

Литература

1. Joshi A., Raman B., Mohan C.K., Cenkeramaddi L.R. Application of a new machine learning model to improve earthquake ground motion predictions // Natural Hazards. 2024. Vol. 120. No. 1. P. 729–753. doi: 10.1007/s11069-023-06230-4.
2. Talukder M.A., Hasan K.F., Islam M.M., Uddin M.A., Akhter A., Yousuf M.A., Alharbi F., Moni M.A. A dependable hybrid machine learning model for network intrusion detection // Journal of Information Security and Applications. 2023. Vol. 72. P. 103405. doi: 10.1016/j.jisa.2022.103405.
3. Amini M., Sharifani K., Rahmani A. Machine learning model towards evaluating data gathering methods in manufacturing and mechanical engineering // International Journal of Applied Science and Engineering Research. 2023. Vol. 15. No. 2023. P. 349–362.
4. Molnar C. Interpretable machine learning. Lulu. com, 2020.
5. Тарасова Ю.А., Февралева Е.С. Прогнозирование банкротства: эконометрическая модель для российских страховщиков // Финансовый журнал. 2021. Т. 13. № 4. С. 75–90.
6. Mokhtar A., Elbeltagi A., Gyasi-Agyei Y., Al-Ansari N., Abdel-Fattah M.K. Prediction of irrigation water quality indices based on machine learning and regression models // Applied Water Science. 2022. Vol. 12. No. 4. P. 76. doi: 10.1007/s13201-022-01590-x.
7. Wang S., Chen Y., Cui Z., Lin L., Zong Y. Diabetes Risk Analysis Based on Machine Learning LASSO Regression Model // Journal of Theory and Practice of Engineering Science. 2024. Vol. 4. No. 01. P. 58–64. doi: 10.53469/jtpes.2024.04(01).08.
8. Cai W., Wen X., Li C., Shao J., Xu J. Predicting the energy consumption in buildings using the optimized support vector regression model // Energy. 2023. Vol. 273. P. 127188. doi: 10.1016/j.energy.2023.127188.
9. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998. 1005 с.
10. Miller A. Subset selection in regression. Chapman and hall/CRC, 2002.
11. Das A., Kempe D. Algorithms for subset selection in linear regression // Proceedings of the fortieth annual ACM

symposium on Theory of computing. 2008. P. 45–54. doi: 10.1145/1374376.1374384.
12. Koch T., Berthold T., Pedersen J., Vanaret C. Progress in mathematical programming solvers from 2001 to 2020 // EURO Journal on Computational Optimization. 2022. Vol. 10. P. 100031. doi: 10.1016/j.ejco.2022.100031.
13. Konno H., Yamamoto R. Choosing the best set of variables in regression analysis using integer programming // Journal of Global Optimization. 2009. Vol. 44. P. 273–282. doi: 10.1007/s10898-008-9323-9.
14. Miyashiro R., Takano Y. Mixed integer second-order cone programming formulations for variable selection in linear regression // European Journal of Operational Research. 2015. Vol. 247. No. 3. P. 721–731. doi: 10.1016/j.ejor.2015.06.081.
15. Tamura R., Kobayashi K., Takano Y., Miyashiro R., Nakata K., Matsui T. Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor // Journal of Global Optimization. 2019. Vol. 73. P. 431–446. doi: 10.1007/s10898-018-0713-3.
16. Park Y.W., Klabjan D. Subset selection for multiple linear regression via optimization // Journal of Global Optimization. 2020. Vol. 77. No. 3. P. 543–574. doi: 10.1007/s10898-020-00876-1.
17. Saishu H., Kudo K., Takano Y. Sparse Poisson regression via mixed-integer optimization // Plos one. 2021. Vol. 16. No. 4. P. e0249916. doi: 10.1371/journal.pone.0249916.
18. Базилевский М.П. Сведение задачи отбора информативных регрессоров при оценивании линейной регрессионной модели по методу наименьших квадратов к задаче частично-булевого линейного программирования // Моделирование, оптимизация и информационные технологии. 2018. Т. 6. № 1 (20). С. 108–117.
19. Базилевский М.П. Отбор информативных регрессоров с учетом мультиколлинеарности между ними в регрессионных моделях как задача частично-булевого линейного программирования // Моделирование, оптимизация и информационные технологии. 2018. Т. 6. № 2 (21). С. 104–118.
20. Базилевский М.П. Отбор оптимального числа информативных регрессоров по скорректированному коэффициенту детерминации в регрессионных моделях как задача частично целочисленного линейного программирования // Прикладная математика и вопросы управления. 2020. № 2. С. 41–54.
21. Базилевский М.П. Построение вполне интерпретируемых линейных регрессионных моделей с помощью метода последовательного повышения абсолютных вкладов переменных в общую детерминацию // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2022. № 2. С. 5–16. doi: 10.17308/sait/1995-5499/2022/2/5-16.
22. Базилевский М.П. Сравнительный анализ эффективности методов построения вполне интерпретируемых линейных регрессионных моделей // Моделирование и анализ данных. 2023. Т. 13. № 4. С. 59–83. doi: 10.17759/mda.2023130404.
23. Shukla S., Jain P.K., Babu C.R., Pamula R. A multivariate regression model for identifying, analyzing and predicting crimes // Wireless Personal Communications. 2020. Vol. 113. No. 4. P. 2447–2461. doi: 10.1007/s11277-020-07335-w.
24. Langenbucher A., Szentmáry N., Cayless A., Weisensee J., Wendelstein J., Hoffmann P. Prediction of corneal back surface power–deep learning algorithm versus multivariate regression // Ophthalmic and Physiological Optics. 2022. Vol. 42. No. 1. P. 185–194. doi: 10.1111/opo.12909.
25. Ферстер Э., Ренц Б. Методы корреляционного и регрессионного анализа. М.: Финансы и статистика, 1983. 303 с.

Базилевский Михаил Павлович. Иркутский государственный университет путей сообщения, Иркутск, Россия. Доцент, кандидат технических наук. Область научных интересов: информационные технологии, машинное обучение, искусственный интеллект. E-mail: mik2178@yandex.ru

Application of Mathematical Programming for Selection the Optimal Structures of Multivariate Linear Regressions

M. P. Bazilevskiy

Irkutsk State Transport University, Irkutsk, Russia

Abstract. In this article formulates the problem of simultaneous selection of both responses and explanatory variables in multivariate linear regressions. This problem is called «key responses and relevant features selection». The ordinary least squares method is used to estimate regressions. First, the problem of selecting a given number of key responses and relevant features by the criterion of the maximum sum of the regression determination coefficients was reduced to a mixed 0-1 integer linear programming problem. Then, restrictions on the signs of the estimates were introduced into it, which made it possible

to select optimal structures of multivariate regressions. After that, restrictions on the absolute contributions of regressors to the overall determinations were added, which allows controlling the number of explanatory variables. When conducting computational experiments on real data with a fixed number of key responses, the time required to construct multivariate models using the proposed method was approximately 67.3 times less than the time required to construct them using the generating all subsets method. At the same time, tightening the restrictions on the absolute contributions of regressors further reduced the time required to solve problems.

Keywords: multivariate linear regression, ordinary least squares, coefficient of determination, key responses and relevant features selection, mixed 0-1 integer linear programming problem, absolute contribution of a variable to determination, generating all subsets method.

DOI 10.14357/20718632240404 **EDN** BBFOVP

References

1. Joshi A., Raman B., Mohan C.K., Cenkeramaddi L.R. Application of a new machine learning model to improve earthquake ground motion predictions. Natural Hazards. 2024;120(1):729–753. doi: 10.1007/s11069-023-06230-4.
2. Talukder M.A., Hasan K.F., Islam M.M., Uddin M.A., Akhter A., Yousuf M.A., Alharbi F., Moni M.A. A dependable hybrid machine learning model for network intrusion detection. Journal of Information Security and Applications. 2023;72:103405. doi: 10.1016/j.jisa.2022.103405.
3. Amini M., Sharifani K., Rahmani A. Machine learning model towards evaluating data gathering methods in manufacturing and mechanical engineering. International Journal of Applied Science and Engineering Research. 2023;15(2023):349–362.
4. Molnar C. Interpretable machine learning. Lulu. Com; 2020.
5. Tarasova Ju.A., Fevraleva E.S. Forecasting of bankruptcy: Evidence from insurance companies in Russia. Financial Journal. 2021;13(4):75–90 (In Russ.).
6. Mokhtar A., Elbeltagi A., Gyasi-Agyei Y., Al-Ansari N., Abdel-Fattah M.K. Prediction of irrigation water quality indices based on machine learning and regression models. Applied Water Science. 2022;12(4):76. doi: 10.1007/s13201-022-01590-x.
7. Wang S., Chen Y., Cui Z., Lin L., Zong Y. Diabetes Risk Analysis Based on Machine Learning LASSO Regression Model. Journal of Theory and Practice of Engineering Science. 2024;4(01):58–64. doi: 10.53469/jtpes.2024.04(01).08.
8. Cai W., Wen X., Li C., Shao J., Xu J. Predicting the energy consumption in buildings using the optimized support vector regression model. Energy. 2023;273:127188. doi: 10.1016/j.energy.2023.127188.
9. Aivazjan S.A., Mhitarjan V.S. Applied statistics and basics of econometrics. Moscow: YUNITI; 1998. 1005 p. (In Russ.).
10. Miller A. Subset selection in regression. Chapman and hall/CRC; 2002.
11. Das A., Kempe D. Algorithms for subset selection in linear regression. Proceedings of the fortieth annual ACM symposium on Theory of computing. 2008:45–54. doi: 10.1145/1374376.1374384.
12. Koch T., Berthold T., Pedersen J., Vanaret C. Progress in mathematical programming solvers from 2001 to 2020. EURO Journal on Computational Optimization. 2022;10:100031. doi: 10.1016/j.ejco.2022.100031.
13. Konno H., Yamamoto R. Choosing the best set of variables in regression analysis using integer programming. Journal of Global Optimization. 2009;44:273–282. doi: 10.1007/s10898-008-9323-9.
14. Miyashiro R., Takano Y. Mixed integer second-order cone programming formulations for variable selection in linear regression. European Journal of Operational Research. 2015;247(3):721–731. doi: 10.1016/j.ejor.2015.06.081.
15. Tamura R., Kobayashi K., Takano Y., Miyashiro R., Nakata K., Matsui T. Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor. Journal of Global Optimization. 2019;73:431–446. doi: 10.1007/s10898-018-0713-3.
16. Park Y.W., Klabjan D. Subset selection for multiple linear regression via optimization. Journal of Global Optimization. 2020;77(3):543–574. doi: 10.1007/s10898-020-00876-1.
17. Saishu H., Kudo K., Takano Y. Sparse Poisson regression via mixed-integer optimization. Plos one. 2021;16(4):e0249916. doi: 10.1371/journal.pone.0249916.
18. Bazilevskiy M.P. 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. Modeling, Optimization and Information Technology. 2018;6(1):108–117. (In Russ.).
19. Bazilevskiy M.P. Subset selection in regression models with considering multicollinearity as a task of mixed 0-1 integer linear programming. Modeling, Optimization and Information Technology. 2018;6(2):104–118. (In Russ.).
20. Bazilevskiy M.P. Selection an optimal number of variables in regression models using adjusted coefficient of determination as a mixed integer linear programming problem. Applied Mathematics and Control Sciences. 2020;(2):41–54. (In Russ.).
21. Bazilevskiy M.P. Construction of quite interpretable linear regression models using the method of successive increase the absolute contributions of variables to the general determination. Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies. 2022;(2):5–16. (In Russ.). doi: 10.17308/sait/1995-5499/2022/2/5-16.
22. Bazilevskiy M.P. Comparative analysis of the effectiveness of methods for constructing quite interpretable linear regression

models. Modelling and Data Analysis. 2023;13(4):59–83. (In Russ.). doi: 10.17759/mda.2023130404.
23. Shukla S., Jain P.K., Babu C.R., Pamula R. A multivariate regression model for identifying, analyzing and predicting crimes. Wireless Personal Communications. 2020;113(4):2447–2461. doi: 10.1007/s11277-020-07335-w.
24. Langenbucher A., Szentmáry N., Cayless A., Weisensee J., Wendelstein J., Hoffmann P. Prediction of corneal back surface power–deep learning algorithm versus multivariate regression. Ophthalmic and Physiological Optics. 2022;42(1):185–194. doi: 10.1111/opo.12909.
25. Ferster E., Rents B. Methods of correlation and regression analysis. Moscow: Finance and Statistics; 1983. 303 p. (In Russ.).

Bazilevskiy Mikhail P. Associate Professor, Candidate of technical sciences, Irkutsk State Transport University, 15 Chernyshevskogo str., Irkutsk, 664074, Russia. E-mail: mik2178@yandex.ru