УДК 519.862.6 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait/1995-5499/2022/2/5-16
Поступила в редакцию 04.04.2022
Подписана в печать 29.06.2022

ПОСТРОЕНИЕ ВПОЛНЕ ИНТЕРПРЕТИРУЕМЫХ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ С ПОМОЩЬЮ МЕТОДА ПОСЛЕДОВАТЕЛЬНОГО ПОВЫШЕНИЯ АБСОЛЮТНЫХ ВКЛАДОВ ПЕРЕМЕННЫХ В ОБЩУЮ ДЕТЕРМИНАЦИЮ

© 2022 М. П. Базилевский✉

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

Аннотация. Статья посвящена проблеме построения вполне интерпретируемых линейных регрессионных моделей, оцениваемых с помощью метода наименьших квадратов. Линейная регрессия называется вполне интерпретируемой, если знаки её коэффициентов соответствуют физическому смыслу входящих в уравнение факторов, а эффект мультиколлинеарности незначителен. При этом желательно, чтобы модель обладала высоким качеством аппроксимации, а все её коэффициенты были значимы. В статье впервые сформулирована задача частично-булевого линейного программирования для выбора в линейной регрессии оптимального числа информативных регрессоров, знаки коэффициентов при которых согласуются со знаками соответствующих коэффициентов их корреляции с зависимой переменной, а абсолютные вклады переменных в общую детерминацию не меньше заданного числа. Эффективность решения этой задачи обусловлена наличием ограничений на согласованность знаков коэффициентов модели, а ограничения на абсолютные вклады переменных позволяют контролировать эффект мультиколлинеарности. Разработан метод последовательного повышения абсолютных вкладов переменных в общую детерминацию, гарантирующий построение вполне интерпретируемой линейной регрессии. Для решения сформулированных задач разработана программа ВИнтер-1. Сначала с помощью неё на обычном персональном компьютере решалась довольно сложная вычислительная задача, решение которой методом полного перебора требует оценки примерно 16,5 квадриллионов моделей. Программа Винтер-1 справилась с этой задачей примерно за 293 секунды, что подтверждает её эффективность. Помимо этого с помощью ВИнтер-1 была построена вполне интерпретируемая модель грузоперевозок железнодорожного транспорта в Иркутской области.

Ключевые слова: вполне интерпретируемая линейная регрессия, метод наименьших квадратов, мультиколлинеарность, абсолютные вклады переменных в общую детерминацию, задача частично-булевого линейного программирования, железнодорожные грузоперевозки.

ВВЕДЕНИЕ

В настоящее время регрессионный анализ [1–3] является весьма распространенным методом анализа статистических данных из самых разных предметных областей. Можно выделить две главных цели проведения регрессионного анализа — прогнозирование и интерпретация. При построении регрессионной модели с целью прогнозирования зависимой переменной достаточно ориентироваться только лишь на получение регрессии с высоким аппроксимационным качеством, т. е.

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

чтобы такая модель очень точно описывала исходную выборку данных. Если целью исследователя является интерпретация модели, то к регрессии, как и к процедуре её построения, должно предъявляться гораздо больше требований.

К сожалению, на сегодняшний день в принципе нет четкого определения интерпретируемой математической модели. Так, в монографии [4], посвященной интерпретируемому машинному обучению, одно из определений формулируется так: «интерпретируемость — это степень, с которой человек может понять причину решения». Таким образом, интерпретируемость модели это трудно формализуемое понятие. Тем не менее, в последнее время проблеме построения интерпретируемых математических моделей в научной литературе уделяется всё больше и больше внимания. Так, в [5] рассматривается проблема оценки интерпретационной пригодности математических моделей, основанных на нечеткой логике, в [6] — проблема интерпретируемости моделей количественного отношения «структура–активность» (QSAR), в [7] — интерпретация матричных моделей текстов и формируемых на их основе моделей текстовых коллекций и др. В обзоре [8] некоторых современных тенденций в технологии машинного обучения также подчеркивается актуальность проблемы интерпретируемости моделей.

Что же касается регрессионного анализа, то научных работ, посвященных проблеме интерпретируемости регрессионных моделей, пока гораздо меньше. В основном они посвящены борьбе с мультиколлинеарностью в линейных регрессиях (см., например, [9]). Стоит выделить работу [10], в которой предложена операторная блок-схема алгоритма построения регрессионных моделей применительно к объекту маркетинговых исследований. При этом авторы формулируют достаточное условие интерпретируемости функции регрессии, состоящее в «диагональности корреляционной матрицы базисных функций». Если же базисные функции являются парными произведениями факторов, то достаточное условие в [10] дополняется требованием низкой корреляции факторов, включенных в парные произведения. В зависимости от степени корреляции факторов авторы работы [10] выделяют следующие 5 уровней интерпретируемости регрессий: точная, приближенная, грубая, неверная и невозможная.

В работе [11] дано следующее более общее (не связанное только с мультиколлинеарностью) определение интерпретируемой регрессионной модели.

Регрессионная модель называется вполне интерпретируемой, если она удовлетворяет трем условиям:

1) её спецификация изначально выбрана так, что после оценивания модели можно объяснить любой её коэффициент или некоторый его аналог, за исключением, быть может, свободного члена;

2) знаки коэффициентов оцененной модели соответствуют физическому смыслу входящих в уравнение факторов;

3) эффект мультиколлинеарности незначителен.

Первое условие зачастую не выполняется, когда спецификация модели представляет собой сложную математическую конструкцию. Например, даже если удастся оценить нелинейную модель $Y = \alpha_0 + \alpha_1 (\sin x)^{\alpha_2} + \varepsilon$, то сказать что-либо содержательное про её оценки $\tilde{\alpha}_1$ и $\tilde{\alpha}_2$ вряд ли получится, поэтому такая спецификация изначально не может считаться вполне интерпретируемой. Зато для любой линейной или степенной модели первое условие выполняется всегда.

Второе и третье условия тесно связаны между собой. Как известно, мультиколлинеарность [12, 13] это наличие линейной зависимости между объясняющими переменными регрессии. Сильная мультиколлинеарность искажает оценки параметров регрессионной модели, поэтому становится затруднительно интерпретировать влияние объясняющих переменных на зависимую переменную. Пусть, например, зависимая переменная y — валовой региональный продукт (ВРП) некоторого региона, а x_2 — объем производства сельскохозяйственной продукции того же региона. С экономической точки зрения очевидно, что чем больше значение переменной

x_2, тем выше должно быть значение y. Однако при построении даже самой простой модели множественной линейной регрессии коэффициент при переменной x_2 может получиться отрицательным, т. е. будет нарушено второе условие и интерпретация такой модели уже не будет иметь никакого смысла. На том же примере с переменными y и x_2 рассмотрим ситуации, при которых могут нарушаться второе и третье условия.

1. Коэффициент корреляции r_{yx_2} между переменными y (ВРП) и x_2 (объем с/х продукции) отрицательный (что может быть связано с явлением ложной корреляции), а объясняющие переменные тесно коррелируют между собой. В такой ситуации в любом случае нарушается третье условие. При этом из-за мультиколлинеарности коэффициент при переменной x_2 может как остаться со знаком «минус», что приведет к нарушению второго условия, так и исказится и стать положительным. В последнем случае второе условие для переменной x_2 будет выполнено, но относить такую модель к интерпретируемым в условиях мультиколлинеарности будет самообманом, поскольку коэффициенты будут значительно искажены. Важно понимать, что нельзя использовать мультиколлинеарность как инструмент для подгонки коэффициентов регрессии под содержательный смысл задачи.

2. Коэффициент $r_{yx_2} < 0$, а мультиколлинеарность слабая. В такой ситуации в любом случае не будет выполнено второе условие. Это связано с тем, что при отсутствии мультиколлинеарности знаки коэффициентов регрессии полностью согласуются со знаками соответствующих коэффициентов корреляции r_{yx_j}. Если $r_{yx_2} < 0$, то и соответствующий коэффициент будет отрицательным, а это противоречит смыслу задачи. Отсюда следует, что на начальном этапе построения регрессионной модели обязательно нужно анализировать согласованность коэффициентов корреляции r_{yx_j} содержательному смыслу факторов. Для этого можно привлекать экспертов из требуемых предметных областей. Если окажется, что коэффициент r_{yx_j} противоречит смыслу, то такую переменную следует или исключить, или каким-либо образом преобразовать.

3. Коэффициент $r_{yx_2} > 0$, а мультиколлинеарность сильная. В такой ситуации в любом случае нарушается третье условие. Второе условие в некоторых случаях тоже нарушается из-за мультиколлинеарности.

4. Коэффициент $r_{yx_2} > 0$, а мультиколлинеарность слабая. В такой ситуации в любом случае оба условия выполняются.

С одной стороны, если хотя бы одно из перечисленных трех условий не выполняется, то регрессионную модель нельзя считать вполне интерпретируемой. С другой стороны, помимо выполнения трех условий желательно, чтобы модель обладала высоким качеством аппроксимации, а все её коэффициенты были значимы по некоторому критерию.

В [11] помимо формулировки приводится описание алгоритма и программы построения вполне интерпретируемых и RTF-адекватных линейных регрессионных моделей. Этот алгоритм основан на формировании всех возможных вариантов регрессий и последовательной проверке каждой из них на мультиколлинеарность, соответствие знаков оценок физическому смыслу факторов и значимость коэффициентов по t-критерию Стьюдента. Не прошедшие проверку хотя бы по одному критерию модели исключаются из рассмотрения. Из оставшихся регрессий выбирается лучшая по величине коэффициента детерминации. К сожалению, для обработки больших массивов данных лежащая в основе этого алгоритма переборная процедура справедливо требует наличия у исследователя значительных вычислительных мощностей.

В настоящей статье предлагается иной подход к построению вполне интерпретируемых регрессионных моделей, основанный на применении аппарата математического программирования. Началом формирования предлагаемого подхода можно считать работу [14], в которой автор формализовал задачу отбора заданного числа информативных регрессоров при оценивании линейной регрессии с помощью метода наименьших квадратов (МНК) к задаче частично-булевого

линейного программирования (ЧБЛП). Дальнейшая модернизация привела к появлению статьи [15], в которой задача ЧБЛП расширена ограничениями для контроля эффекта мультиколлинеарности и значимости коэффициентов регрессии по t-критерию Стьюдента. Фактически, если добавить в задачу из [15] ограничения на знаки коэффициентов модели, то получим формализацию задачи построения вполне интерпретируемых линейных регрессий. Но такая постановка имеет следующие недостатки:

1) на начальном этапе требуется ввод числа отбираемых регрессоров;

2) задача содержит слишком много ограничений, что увеличивает время поиска оптимального решения;

3) нерешенным остается вопрос выбора больших чисел M в ограничениях задачи.

Последняя проблема, связанная с выбором больших чисел M, была решена автором в работе [16]. В ней же подтвердилась гипотеза из [17], состоящая в том, что если в задачу ЧБЛП построения регрессии интегрировать ограничения на согласованность знаков коэффициентов модели со знаками соответствующих коэффициентов корреляции r_{yx_j}, то такая задача будет решаться на порядок быстрее, чем при их отсутствии. К тому же в этом случае становятся справедливыми формулы для абсолютных вкладов объясняющих переменных в общую детерминацию, которые можно использовать для оценки значимости факторов вместо наблюдаемых значений t-критерия Стьюдента.

Целью данной работы является разработка нового эффективного метода построения вполне интерпретируемых линейных регрессионных моделей, реализация его в виде специализированного программного обеспечения и демонстрация работы последнего на примере моделирования грузоперевозок железнодорожного транспорта в Иркутской области.

1. МАТЕРИАЛЫ И МЕТОДЫ

1.1. Метод последовательного повышения вкладов переменных

Модель множественной линейной регрессии имеет вид:

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

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

Для модели (1) стандартизованная регрессия записывается в виде:

$$w_i = \sum_{j=1}^{l} \beta_j z_{ij} + \xi_i, \quad i = \overline{1,n}, \qquad (2)$$

где $w_i = \dfrac{y_i - \overline{y}}{\sigma_y}$, $z_{i1} = \dfrac{x_{i1} - \overline{x_1}}{\sigma_{x_1}}$, …, $z_{il} = \dfrac{x_{il} - \overline{x_l}}{\sigma_{x_l}}$, $i = \overline{1,n}$ — значения нормированных переменных; β_j, $j = \overline{1,l}$ — неизвестные стандартизованные коэффициенты; ξ_i, $i = \overline{1,n}$ — новые ошибки аппроксимации.

МНК-оценки модели (2) находятся по формуле

$$\tilde{\beta} = R_{xx}^{-1} \cdot R_{yx}, \qquad (3)$$

где $R_{xx} = \begin{pmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_l} \\ r_{x_1x_2} & 1 & \dots & r_{x_2x_l} \\ \dots & \dots & \dots & \dots \\ r_{x_1x_l} & r_{x_2x_l} & \dots & 1 \end{pmatrix}$ — матрица коэффициентов корреляции объясняющих переменных; $R_{yx} = \begin{pmatrix} r_{yx_1} & r_{yx_2} & \dots & r_{yx_l} \end{pmatrix}^T$ — вектор коэффициентов корреляции переменной y с каждой из переменных x_1, x_2, …, x_l.

Коэффициент детерминации модели (2) находится по формуле

$$R^2 = \sum_{j=1}^{l} r_{yx_j} \cdot \tilde{\beta}_j. \qquad (4)$$

На основе (3) и (4) в [14] была сформулирована следующая задача отбора m информативных регрессоров из общего числа l:

$$R^2(\beta_1,...,\beta_l,\delta_1,...,\delta_l)=\sum_{j=1}^{l} r_{yx_j}\cdot\beta_j\to\max, \quad (5)$$

$$-\left(1-\delta_j\right)M\le\sum_{k=1}^{l} r_{x_jx_k}\cdot\beta_k-$$

$$-r_{yx_j}\le(1-\delta_j)M,\quad j=\overline{1,l}, \quad (6)$$

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

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

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

где δ_j, $j=\overline{1,l}$ — булевы переменные, заданные по правилу

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

M — большое положительное число.

Достоинством задачи ЧБЛП (5)–(9) является то, что число её ограничений не зависит от объема выборки n.

В [16] ограничения (6) и (7) были заменены на следующие:

$$\left(1-\delta_j\right)M_{u_j}^{-}\le\sum_{k=1}^{l} r_{x_jx_k}\cdot\beta_k-$$

$$-r_{yx_j}\le(1-\delta_j)M_{u_j}^{+},\quad j=\overline{1,l}, \quad (10)$$

$$0\le\beta_j\le M_{\beta_j}\cdot\delta_j,\quad j\in J^{+}, \quad (11)$$

$$M_{\beta_j}\cdot\delta_j\le\beta_j\le 0,\quad j\in J^{-}, \quad (12)$$

где J^{+} и J^{-} — индексные множества, построенные из множества $\{1,2,...,l\}$, элементы которых удовлетворяют условиям $r_{yx_j}>0$ и $r_{yx_j}<0$; M_{β_j}, $M_{u_j}^{-}$, $M_{u_j}^{+}$, $j=\overline{1,l}$ — числа, которые определяются следующим образом.

Числа M_{β_j} в (11) и (12) находятся по формулам $M_{\beta_j}=\dfrac{R^2_{\max}}{r_{yx_j}}$, $j=\overline{1,l}$, где $R^2_{\max}$ — значение коэффициента детерминации регрессии, построенной со всеми l объясняющими переменными.

Для нахождения чисел $M_{u_j}^{-}$ в ограничениях (10) нужно решить серию из l задач линейного программирования с целевыми функциями $u_j\to\min$ при ограничениях

$$0\le\beta_j\le M_{\beta_j},\quad j\in J^{+}; \quad (13)$$

$$M_{\beta_j}\le\beta_j\le 0,\quad j\in J^{-}, \quad (14)$$

$$\sum_{k=1}^{l} r_{x_jx_k}\cdot\beta_k-r_{yx_j}=u_j,\quad j=\overline{1,l}, \quad (15)$$

$$\sum_{j=1}^{l} r_{yx_j}\cdot\beta_j\le R^2_{\max}, \quad (16)$$

где u_j — ошибка j-го уравнения линейной системы $\sum_{k=1}^{l} r_{x_jx_k}\cdot\beta_k-r_{yx_j}=0,\ j=\overline{1,l}.$

Для нахождения чисел $M_{u_j}^{+}$ в ограничениях (10) нужно решить серию из l задач линейного программирования с целевыми функциями $u_j\to\max$ при ограничениях (13)–(16).

Таким образом, решение задачи ЧБЛП с целевой функцией (5) и ограничениями (8) — (12) приводит к построению линейной регрессии с m объясняющими переменными, в которой знаки МНК-оценок β_j согласованы с соответствующими знаками коэффициентов корреляции r_{yx_j}, т. е. $\beta_j\cdot r_{yx_j}>0$.

Можно выделить 3 достоинства сформулированной задачи (5), (8)–(12):

1) наличие способа предварительной идентификации чисел M, в отличие от задачи (5)–(9);

2) как отмечено в [17], из-за ограничений на знаки (11) и (12) она решается на порядок быстрее, чем задача (5)–(9);

3) из-за согласованности знаков β_j и r_{yx_j} становятся справедливыми формулы для абсолютных вкладов переменных в общую детерминацию R^2:

$$C_{x_j}^{\text{абс}}=r_{yx_j}\cdot\tilde{\beta}_j,\quad j=\overline{1,l}. \quad (17)$$

По критериям (17) можно делать выводы о степени влияния каждой объясняющей переменной на объясняемую переменную y.

К недостаткам задачи (5), (8)–(12) можно отнести то, что:

1) иногда она может не иметь решений;

2) необходимо предварительно задавать число регрессоров m.

Перечисленные недостатки легко устраняются исключением из задачи (5), (8)–(12) ограничения (9) на число входящих в уравнение регрессоров. Тогда решение задачи с целевой функцией (5) и ограничениями (8), (10)–(12) приводит к построению линейной

регрессии с оптимальным числом объясняющих переменных, в которой $\beta_j \cdot r_{yx_j} > 0$.

К сожалению, в задаче (5), (8), (10)–(12) нет ограничений ни на значимость коэффициентов модели, ни на эффект мультиколлинеарности, поэтому полученная в результате решения линейная регрессия может оказаться как вполне интерпретируемой, так и нет. Эту проблему можно решить следующим образом.

Введем ограничения на абсолютные вклады переменных $C_{x_j}^{\text{абс}}$ в общую детерминацию R^2:

$$r_{yx_j} \cdot \beta_j \geq \theta \cdot \delta_j, \quad j = \overline{1,l}, \qquad (18)$$

где $\theta \geq 0$ — заданное минимальное значение вклада каждой входящей в регрессию переменной в общую детерминацию. Тогда, если j-я переменная не входит в модель, то $\delta_j = 0$ и соответствующее ограничение (18) выполняется всегда, а если j-я переменная входит в модель, то $\delta_j = 1$ и соответствующее ограничение (18) принимает вид $r_{yx_j} \cdot \beta_j \geq \theta$.

Таким образом, решение задачи (5), (8), (10)–(12), (18) приводит к построению линейной регрессии с оптимальным числом объясняющих переменных, абсолютные вклады которых в общую детерминацию будут не меньше, чем число θ. Понятно, что при $\theta = 0$ решение этой задачи совпадает с решением задачи (5), (8), (10)–(12). А с ростом числа θ будет происходить снижение числа входящий в регрессию объясняющих переменных, а значит, и снижение эффекта мультиколлинеарности. Последнее обстоятельство по механизму функционирования напоминает известный метод LASSO [18].

Для того чтобы можно было регулировать эффект мультиколлинеарности нужно поступать следующим образом. Решить задачу (5), (8), (10)–(12). Для полученной регрессии вычислить абсолютные вклады переменных и оценить степень мультиколлинеарности любым известным методом, например, с помощью коэффициентов вздутия дисперсии. Если вклады достаточно высоки, а мультиколлинеарность слабая, то модель получена. В противной случае назначается величина θ, чуть большая, чем минимальный из текущих абсолютных вкладов переменных, и решается задача (5), (8), (10)–(12), (18). И так до тех пор, пока не будет получена модель со слабой мультиколлинеарностью и необходимыми абсолютными вкладами переменных в детерминацию. Назовем предложенный подход методом последовательного повышения вкладов переменных (МППВП). МППВП гарантирует построение вполне интерпретируемой линейной регрессии.

Следует подчеркнуть, что перед применением МППВП нужно исключить или преобразовать объясняющие переменные, знаки коэффициентов корреляции r_{yx_j} которых не удовлетворяют физическому смыслу решаемой задачи.

1.2. Программа ВИнтер-1

Для решения предложенных задач ЧБЛП в среде программирования Delphi была разработана программа построения вполне интерпретируемых линейных регрессионных моделей (ВИнтер-1). Её интерфейс представлен на рис. 1.

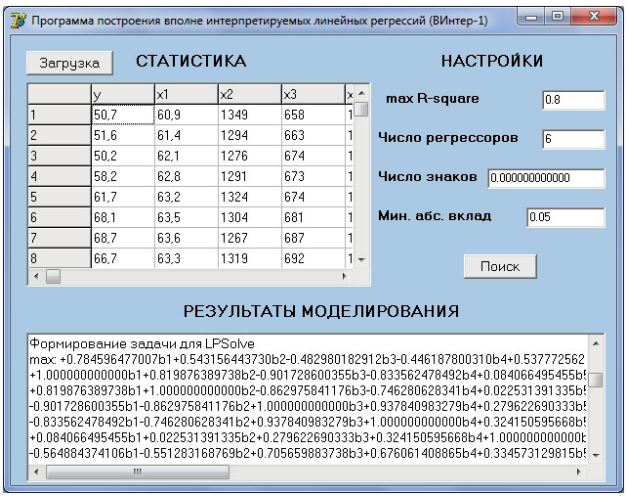

Рис. 1. Интерфейс программы ВИнтер-1
[Fig. 1. Interface of program ВИнтер-1]

В ВИнтер-1 решателем задач линейного программирования выступает пакет LPSolve.

Для работы в программе сначала нужно загрузить исходные статистические данные, нажав кнопку «Загрузка». Напомним, что предварительно с этими данными должны поработать эксперты, исключив противоре-

чивые по смыслу объясняющие переменные. Данные должны храниться в текстовом файле с расширением .txt, в котором наблюдения расположены по строкам, переменные — по столбцам, и отделяются друг от друга символом табуляции. Разделителем целых и дробных частей вещественных чисел является «,».

Завершив загрузку данных, нужно задать следующие параметры.

1. max R-square. Это значение коэффициента детерминации $R^2_{\max}$ регрессии, построенной со всеми l объясняющими переменными. Если $l \geq n$, то $R^2_{\max}$ нужно принять равным 1.

2. Число регрессоров. Это количество отбираемых объясняющих переменных m. Если выбрано натуральное число, то в задачу включается соответствующее ограничение на число регрессоров, а если 0, то такое ограничение игнорируется.

3. Число знаков. Этот количество знаков после запятой при округлении вещественных чисел. По умолчанию задано 12 знаков.

4. Мин. абс. вклад. Это минимальное значение вклада каждой переменной в общую детерминацию θ. Если $\theta = 0$, то в задаче игнорируются ограничения (18).

После ввода всех необходимых параметров нужно нажать кнопку «Поиск». Сначала в программе ВИнтер-1 автоматически будут определены числа M_{β_j}, $j = \overline{1,l}$ для ограничений (11) и (12), а затем на основе решения серии из $2l$ задач линейного программирования будут найдены числа $M^-_{u_j}$, $M^+_{u_j}$, $j = \overline{1,l}$ для ограничений (10). После чего с использованием заданных параметров будет сформирована задача ЧБЛП для решателя LPSolve. Все полученные в ходе работы программы результаты фиксируются в поле «РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ».

Отметим, что на момент написания данной статьи программа ВИнтер-1 представляет собой лишь первую версию разрабатываемой крупной системы, имеющей гораздо больше функциональных возможностей. Поэтому пока для решения сформулированной задачи ЧБЛП нужно копировать её математическую модель из ВИнтер-1 в LPSolve, а уже там запускать решатель. В дальнейшем планируется полностью автоматизировать этот процесс, наделив программу возможностями построения вполне интерпретируемых нелинейных моделей.

2. РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Моделирование грузоперевозок на железнодорожном транспорте является актуальной научной задачей (см., например, [19, 20]). Для демонстрации работы разработанной программы ВИнтер-1 были собраны годовые статистические данные (источник https://rosstat.gov.ru/) по Иркутской области за период с 2000 г. по 2020 г. для зависимой переменной y (отправление грузов железнодорожным транспортом общего пользования, млн руб.) и 62 объясняющих переменных — x_1, x_2, …, x_{62}. Описание некоторых из этих переменных будет приведено далее.

Моделирование проводилось на персональном компьютере с процессором Intel Core i5 (3.40 ГГц, 4 ядра) и объемом оперативной памяти 8 ГБ.

Эксперимент № 1. Целью эксперимента было не построение вполне интерпретируемой линейной регрессии, а демонстрация эффективности применения ВИнтер-1 для решения задач большой размерности. Для этого по исходным данным со всеми переменными с помощью ВИнтер-1 решалась задача ЧБЛП с целевой функцией (5) и ограничениями (8), (10)–(12), т. е. не учитывалось ограничение (9) на число входящих в уравнение регрессоров. На решение этой задачи LPSolve потребовалось всего 292,7 секунд. В результате была построена регрессионная модель с восьмью переменными:

$$\tilde{y} = 105,207 - 0,092x_8 - 0,119x_{10} - \\ -4,654 \cdot 10^{-6} x_{20} + 0,066x_{22} + 0,131x_{40} - \\ -0,001x_{42} - 0,608x_{50} + 0,004x_{60}, \qquad (19)$$

где x_8 — число собственных легковых автомобилей на 1000 человек населения (штук); x_{10} — численность обучающихся по программам начального, основного и среднего общего образования (тыс. человек); x_{20} — кредиторская задолженность организаций (млн руб.); x_{22} — производство электроэнергии

(млрд киловатт-часов); x_{40} — процент организаций, использующих сеть Интернет; x_{42} — число подключенных абонентских устройств мобильной связи на 1000 человек населения; x_{50} — средние потребительские цены на проезд в трамвае за декабрь (руб.); x_{60} — импорт из стран дальнего зарубежья (млн долларов США).

Коэффициент детерминации модели (19) $R^2 = 0,92657$. Очевидно, что построенная регрессия не является вполне интерпретируемой.

Для решения задачи построения регрессии (19) методом полного перебора потребовалось бы оценить $\sum_{j=1}^{20} C_{62}^{j} = 16483857933928472$ (примерно 16,5 квадриллионов) моделей, а ВИнтер-1 справился с такой задачей менее чем за 5 минут, что подтверждает его высокую эффективность.

Эксперимент № 2. Целью эксперименты было построение вполне интерпретируемой линейной регрессии. Для этого специальной группой экспертов из 62 объясняющих переменных сначала были исключены те, которые по смыслу не влияют на y. Затем из оставшихся 44 переменных исключили те, у которых знаки коэффициентов корреляции r_{yx_j} не согласуются с экономическим смыслом задачи, а также те, для которых $\left|r_{yx_j}\right| < 0,2$. В результате осталось 11 переменных: x_8; x_{20}; x_{22}; x_2 — процент трудоспособного населения от общей численности; x_3 — численность рабочей силы (тыс. человек); x_5 — численность пенсионеров (тыс. человек); x_{18} — число предприятий и организаций; x_{35} — удельный вес автодорог с твердым покрытием в общей протяженности автодорог общего пользования (%); x_{36} — удельный вес автодорог с усовершенствованным покрытием в протяженности автодорог с твердым покрытием общего пользования (%); x_{45} — средние потребительские цены на бензин автомобильной марки АИ-92 (литр) за декабрь (рублей); x_{58} — тарифы на грузовые перевозки (железнодорожный транспорт).

После чего по исходным данным для отобранных 11 переменных с помощью ВИнтер-1 был реализован МППВП.

Шаг 1. Решается задача (5), (8), (10)–(12). В ВИнтер-1 были заданы следующие параметры: max R-square — 0,93207, число регрессоров — 0, число знаков — 12, минимальный абсолютный вклад — 0. В результате практически мгновенно была построена регрессия с пятью переменными:

$$\tilde{y} = \underset{}{-67,3} + \underset{(1,519)}{\overset{(0,408)}{1,222}}\, x_2 - \underset{(-0,524)}{\overset{(0,071)}{0,023}}\, x_8 + \underset{(3,209)}{\overset{(0,285)}{+0,00057}}\, x_{18} - \underset{(-0,037)}{\overset{(0,0011)}{4,396 \cdot 10^{-7}}}\, x_{20} + \underset{(1,492)}{\overset{(0,102)}{0,363}}\, x_{22}, \quad (20)$$

где в скобках под коэффициентами указаны значения t-критерия Стьюдента, а над коэффициентами — абсолютные вклады переменных.

Коэффициент детерминации регрессии (20) $R^2 = 0,867218$.

Как видно, в уравнении (20) самой незначимой по критерию Стьюдента переменной оказалась x_{20}. А её абсолютный вклад в общую детерминацию оказался практически равен нулю. Отсюда можно сделать о слабом влиянии переменной x_{20} на y. При этом по критерию Стьюдента самое сильное влияние на y оказывает переменная x_{18}, а по абсолютному вкладу — x_2.

Найденные для модели (20) коэффициенты вздутия дисперсии для переменных x_2 и x_8 превышают пороговое значение 10, что говорит о присутствии в модели эффекта мультиколлинеарности. Таким образом, регрессию (20) нельзя считать вполне интерпретируемой.

Шаг 2. Минимальный абсолютный вклад в (20) равен 0,0011. На основе этого была назначена величина $\theta = 0,002$. В ВИнтер-1 решается задача (5), (8), (10)–(12), (18). В результате была построена регрессия с четырьмя переменными:

$$\tilde{y} = \underset{}{-67,79} + \underset{(1,761)}{\overset{(0,412)}{1,235}}\, x_2 - \underset{(-0,541)}{\overset{(0,071)}{0,023}}\, x_8 + \underset{(3,638)}{\overset{(0,284)}{+0,00057}}\, x_{18} + \underset{(1,721)}{\overset{(0,101)}{0,359}}\, x_{22}. \quad (21)$$

Коэффициент детерминации регрессии (21) $R^2 = 0,867206$.

Как и следовало ожидать, произошло исключение переменной x_{20}. При этом на вели-

чину коэффициента детерминации модели (21) это практически никак не повлияло.

В уравнении (21) самой незначимой как по критерию Стьюдента, так и по величине абсолютного вклада переменной оказалась x_8. Это значит, что переменная x_8 слабо влияет на y. В то же время по критерию Стьюдента самое сильное влияние на y оказывает переменная x_{18}, а по абсолютному вкладу — x_2.

Найденные для модели (21) коэффициенты вздутия дисперсии для переменных x_2 и x_8 по-прежнему превышают пороговое значение 10, что говорит о присутствии в модели эффекта мультиколлинеарности. Следовательно, регрессию (21) нельзя считать вполне интерпретируемой.

Шаг 3. Минимальный абсолютный вклад в (21) равен 0,071. На основе этого была назначена величина $\theta = 0,072$. В ВИнтер-1 решается задача (5), (8), (10)–(12), (18). В результате была построена регрессия с тремя переменными:

$$\tilde{y} = \underset{}{-86,106} + \underset{(6,978)}{\overset{(0,532)}{1,593}}\, x_2 + \underset{(5,209)}{\overset{(0,251)}{+0,0005}}\, x_{18} + \underset{(1,773)}{\overset{(0,082)}{0,293}}\, x_{22}. \qquad (22)$$

Коэффициент детерминации регрессии (22) $R^2 = 0,864777$.

Как и ожидалось, произошло исключение переменной x_8. На качество регрессии (22) это практически не повлияло.

В модели (22) все коэффициенты значимы по критерию Стьюдента для уровня значимости 0,1. Минимальный абсолютный вклад равен 0,082 для переменной x_{22}, что в относительном выражении составляет примерно 9,5 %, поэтому нельзя утверждать о слабом влиянии x_{22} на y. Стоит обратить внимание, что переменные ранжируется одинаково как по критерию Стьюдента, так и по абсолютному вкладу.

Все найденные для модели (22) коэффициенты вздутия дисперсии оказались меньше, чем 1,2, что говорит об отсутствии в модели эффекта мультиколлинеарности. Следовательно, регрессию (22) можно считать вполне интерпретируемой.

ЗАКЛЮЧЕНИЕ

В работе сформулирована задача ЧБЛП для выбора в линейной регрессии оптимального числа информативных регрессоров, знаки коэффициентов при которых согласуются со знаками соответствующих коэффициентов их корреляции с зависимой переменной, а абсолютные вклады переменных в общую детерминацию не меньше заданного числа. На основе этой задачи разработан метод последовательного повышения абсолютных вкладов переменных в общую детерминацию, гарантирующий построение вполне интерпретируемой линейной регрессии. Для решения предложенных задач разработана программа ВИнтер-1. В ходе экспериментов была продемонстрирована эффективность ВИнтер-1 и была построена вполне интерпретируемая модель грузоперевозок железнодорожного транспорта в Иркутской области.

КОНФЛИКТ ИНТЕРЕСОВ

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

СПИСОК ЛИТЕРАТУРЫ

1. *Arkes, J.* Regression analysis: a practical introduction / J. Arkes. – Routledge, 2019. – 362 p. https://doi.org/10.4324/9781351011099

2. *Westfall, P. H.* Understanding regression analysis: a conditional distribution approach / P. H. Westfall, A. L. Arias. – Chapman and Hall/CRC, 2020. – 514 p. https://doi.org/10.1201/9781003025764

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

4. *Molnar, C.* Interpretable machine learning / C. Molnar. – Lulu. com, 2020.

5. *Долгий, А. И.* Интерпретируемость нечетких темпоральных моделей / А. И. Долгий, С. М. Ковалев // Известия Южного федераль-

ного университета. Технические науки. – 2018. – № 5 (199). – С. 131–142.

6. *Адылова, Ф. Т.* Практика и потенциал развития интерпретируемости моделей количественного отношения структура-активность (QSAR) / Ф. Т. Адылова, Р. Р. Давронов, У. У. Жамилов, О. А. Каюмов // Проблемы вычислительной и прикладной математики. – 2018. – № 5. – С. 7–26.

7. *Крейнес, М. Г.* Матричные модели текстов. Интерпретация моделей и экспериментальная верификация / М. Г. Крейнес, Е. М. Крейнес // Математическое моделирование. – 2020. – Т. 32, № 7. – С. 24-46. https://doi.org/10.20948/mm-2020-07-02

8. *Коротеев, М. В.* Обзор некоторых современных тенденций в технологии машинного обучения / М. В. Коротеев // E-Management. – 2018. – Т. 1, № 1. – С. 26–35.

9. *Мокшина, С. И.* Метод построения содержательно интерпретируемых регрессионных моделей в условиях мультиколлинеарности / С. И. Мокшина, Г. В. Шуршикова, С. С. Щекунских // Современная экономика: проблемы и решения. – 2017. – Т. 5.

10. *Горбач, А. Н.* Покупательское поведение: анализ спонтанных последовательностей и регрессионных моделей в маркетинговых исследованиях / А. Н. Горбач, Н. А. Цейтлин. – Київ : Освіта України, 2011.

11. *Базилевский, М. П.* Программа построения вполне интерпретируемых и RTF-адекватных линейных регрессионных моделей / М. П. Базилевский // Системы и средства информатики. – 2021. – Т. 31, № 4. – С. 18–26 https://doi.org/10.14357/08696527210402

12. *Giacalone, M.* Multicollinearity in regression: an efficiency comparison between Lp-norm and least squares estimators / M. Giacalone, D. Panarello, R. Mattera // Quality & Quantity. – 2018. – Vol. 52, No 4. – P. 1831–1859. https://doi.org/10.1007/s11135-017-0571-y

13. *Tamura, R.* Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor / R. Tamura, K. Kobayashi, Y. Takano, R. Miyashiro, K. Nakata, T. Matsui // Journal of Global Optimization. – 2019. – Vol. 73, No. 2. – P. 431–446. https://doi.org/10.1007/s10898-018-0713-3

14. *Базилевский, М. П.* Сведение задачи отбора информативных регрессоров при оценивании линейной регрессионной модели по методу наименьших квадратов к задаче частично-булевого линейного программирования / М. П. Базилевский // Моделирование, оптимизация и информационные технологии. – 2018. – Т. 6, № 1 (20). – С. 108–117.

15. *Базилевский, М. П.* Отбор значимых по критерию Стьюдента информативных регрессоров в оцениваемых с помощью МНК регрессионных моделях как задача частично-булевого линейного программирования / М. П. Базилевский // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. – 2021. – № 3. – С. 5–16. https://doi.org/10.17308/sait.2021.3/3731

16. *Базилевский, М. П.* Способ определения параметра M в задаче частично-булевого линейного программирования для отбора регрессоров в линейной регрессии / М. П. Базилевский // Вестник Технологического университета. – 2022. – Т. 25, № 2. – С. 62–66.

17. *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. https://doi.org/10.1007/s10898-008-9323-9

18. *Emmert-Streib, F.* High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection / F. Emmert-Streib, M. Dehmer // Machine Learning and Knowledge Extraction. – 2019. – Vol. 1, No 1. – P. 359–383. https://doi.org/10.3390/make1010021

19. *Носков, С. И.* Моделирование объема погрузки на железнодорожном транспорте методом смешанного оценивания / С. И. Носков, К. С. Перфильева // Известия Тульского государственного университета. Технические науки. – 2021. – № 2. – С. 148–153.

20. *Носков, С. И.* Анализ регрессионной модели грузооборота железнодорожного транспорта / С. И. Носков, И. П. Врублевский // Вестник транспорта Поволжья. – 2020. – №1 (79). – С. 86–90.

Базилевский Михаил Павлович — канд. техн. наук, доц., 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/1995-5499/2022/2/5-16 ISSN 1995-5499
Received 04.04.2022
Accepted 29.06.2022

CONSTRUCTION OF QUITE INTERPRETABLE LINEAR REGRESSION MODELS USING THE METHOD OF SUCCESSIVE INCREASE THE ABSOLUTE CONTRIBUTIONS OF VARIABLES TO THE GENERAL DETERMINATION

© 2022 M. P. Bazilevskiy✉

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

Annotation. This article is devoted to the problem of constructing quite interpretable linear regression models estimated using the ordinary least squares. Linear regression is called quite interpretable if the signs of its coefficients correspond to the physical meaning of the factors included in the equation, and the effect of multicollinearity is insignificant. At the same time, it is desirable that the model has a high quality of approximation, and all its coefficients are significant. In this article, for the first time, the problem of mixed 0-1 integer linear programming was formulated to select the optimal number of variables in linear regression, the signs of the coefficients for which are consistent with the signs of the corresponding coefficients of their correlation with the dependent variable, and the absolute contributions of the variables to the general determination are not less than a given number. The efficiency of solving this problem is due to the presence of restrictions on the consistency of the model coefficients signs, and restrictions on the absolute contributions of the variables make it possible to control the multicollinearity. A method has been developed for successive increase the absolute contributions of variables to the general determination, which guarantees the construction of quite interpretable linear regression. To solve the formulated tasks, the program ВИнтер-1 was developed. At first, using it on an ordinary personal computer, a rather complex computational problem was solved, the solution of which by the exhaustive search method requires the estimation of approximately 16.5 quadrillion models. The ВИнтер-1 program completed this task in about 293 seconds, which confirms its effectiveness. In addition, with the help of ВИнтер-1, a quite interpretable model of rail freight transportation in the Irkutsk region was construct.
Keywords: quite interpretable linear regression, ordinary least squares, multicollinearity, absolute contributions of variables to the general determination, mixed 0-1 integer linear programming, rail freight transportation.

CONFLICT OF INTEREST

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

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

REFERENCES

1. *Arkes J.* (2019) Regression analysis: a practical introduction. Routledge. 362 p. https://doi.org/10.4324/9781351011099

2. *Westfall P. H. and Arias A. L.* (2020) Understanding regression analysis: a conditional distribution approach. Chapman and Hall/CRC. 514 p. https://doi.org/10.1201/9781003025764

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

4. *Molnar C.* (2020) Interpretable machine learning. Lulu. com.

5. *Dolgy A. I. and Kovalev S. M.* (2018) Interpretability of fuzzy temporal models. *Izvestiya SFedU. Engineering Sciences.* No 5 (199). P. 131–142.

6. *Adilova F. T., Davronov R. R., Jamilov U. U. and Kayumov O. A.* (2018) Practice and potential for the development of the interpretability of quantitative «structure-activity» models (QSAR). *Problems of Computational and Applied Mathematics.* No 5. P. 7–26.

7. *Kreines M. G. and Kreines E. M.* (2020) Matrix text models. Interpretation and experimental verification of models. *Matematicheskoe modelirovanie*. V. 32, No 7. P. 24–46. https://doi.org/10.20948/mm-2020-07-02

8. *Koroteev M. V.* (2018) Review of some contemporary trends in machine learning technology. *E-Management.* V. 1, No 1. P. 26–35.

9. *Mokshina S. I., Shurshikova G. V. and Shchekunskih S. S.* (2017) The construction method of meaningful interpreted regression models in conditions of multicollinearity. *Sovremennaya ekonomika: problemy i resheniya*. V. 5.

10. *Gorbach A. N. and Tseytlin N. A.* (2011) Buying behavior: an analysis of spontaneous sequences and regression models in marketing research. Kyiv: Education of Ukraine.

11. *Bazilevskiy M. P.* (2021) A program for constructing of quite interpretable and RTF-adequate linear regression models. *Systems and Means of Informatics*. V. 31, No 4. P. 18–26. https://doi.org/10.14357/08696527210402

12. *Giacalone M., Panarello D. and Mattera R.* (2018) Multicollinearity in regression: an efficiency comparison between Lp-norm and least squares estimators. *Quality & Quantity*. V. 52, No 4. P. 1831–1859. https://doi.org/10.1007/s11135-017-0571-y

13. *Tamura R., Kobayashi K., Takano Y., Miyashiro R., Nakata K. and Matsui T.* (2019) Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor. *Journal of Global Optimization*. V. 73, No 2. P. 431–446. https://doi.org/10.1007/s10898-018-0713-3

14. *Bazilevskiy M. P.* (2018) 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*. V. 6, No 1 (20). P. 108–117.

15. *Bazilevskiy M. P.* (2021) Selection of informative regressors significant by Student's t-test in regression models estimated using OLS as a partial Boolean linear programming problem. *Proceedings of VSU, series: System analysis and information technologies*. No 3. P. 5–16. https://doi.org/10.17308/sait.2021.3/3731

16. *Bazilevskiy M. P.* (2022) Method foe the M parameter determination in 0-1 mixed-integer linear programming problem for subset selection in linear regression. *Bulletin of the Technological University*. V. 25, No 2. P. 62–66.

17. *Konno H. and 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

18. *Emmert-Streib F. and Dehmer M.* (2019) High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection. *Machine Learning and Knowledge Extraction*. V. 1, No 1. P. 359–383. https://doi.org/10.3390/make1010021

19. *Noskov S. I. and Perfilieva K. S.* (2021) Application of the mixed estimation method in modeling the loading volume in railway transport. *Proceedings of TSU. Technical sciences*. No 2. P. 148–153.

20. *Noskov S. I. and Vrublevskiy I. P.* (2020) Analysis of the regression model of railway freight turnover. *Vestnik transporta Povolzhya*. No 1 (79). P. 86–90.

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