УДК 519.862.6

ОПТИМИЗАЦИОННЫЕ ЗАДАЧИ ОТБОРА ИНФОРМАТИВНЫХ РЕГРЕССОРОВ В ЛИНЕЙНОЙ РЕГРЕССИИ С КОНТРОЛЕМ ЕЁ ЗНАЧИМОСТИ ПО КРИТЕРИЮ ФИШЕРА

© 2024 М.П. Базилевский

Иркутский государственный университет путей сообщения, г. Иркутск, Россия

Статья поступила в редакцию 30.07.2024

Статья посвящена проблеме отбора информативных регрессоров в моделях множественной линейной регрессии. При реализации такого отбора с помощью коэффициента детерминации, полученная в результате модель может быть незначимой по критерию Фишера. Для решения этой проблемы предложено две задачи частично-булевого линейного программирования, алгоритмы решения которых улучшены в десятки раз за последние 20 лет. Решение первой из них дает оптимальную модель с назначенным числом регрессоров, при решении второй оптимальное число регрессоров определяется автоматически. Проведены вычислительные эксперименты. Для второй задачи на примере показано, что с ужесточением требований на значимость модели по критерию Фишера число регрессоров при отборе снижается. Предложенный в статье приём, связанный с вводом дополнительных бинарных переменных, может быть использован в дальнейшем для контроля в моделях мультиколлинеарности и значимости оценок по t-критерию Стьюдента.

Ключевые слова: регрессионный анализ, линейная регрессия, метод наименьших квадратов, отбор информативных регрессоров, задача частично-булевого линейного программирования, коэффициент детерминации, критерий Фишера.

DOI: 10.37313/1990-5378-2024-26-6-200-207
EDN : MEZQQZ

ВВЕДЕНИЕ

Задачи поиска математической формы связи между выходной переменной и одной или несколькими входными переменными в научных исследованиях и не только возникают довольно часто. Одним из инструментов решения таких задач считается регрессионный анализ [1,2], тесно связанный с актуальной на сегодняшний день технологией машинного обучения [3]. С помощью регрессионных моделей решаются задачи интерпретации и прогнозирования в различных областях деятельности человека (см., например, [4–6]). При этом входных переменных, предположительно влияющих на исследуемый фактор, может быть так много, что приходиться осуществлять из них выбор некоторого количества только наиболее значимых, иными словами, решать задачу отбора информативных регрессоров (ОИР) [7–9].

Методов и алгоритмов ОИР, как отмечено в работах [7–9], существует много. К ним относятся эвристические процедуры включения-исключения, метод последовательной замены переменных, алгоритм Лассо, метод группового учёта аргументов и пр. Однако оптимальность модели с точки зрения некоторого критерия гарантирует только полный перебор (метод всех регрессий), что представляется довольно трудоёмкой вычислительной задачей, поскольку её сложность возрастает экспоненциально в зависимости от числа входных переменных. На помощь приходит хорошо развитый за последние годы аппарат математического программирования. Так, например, в [10] авторы делают вывод, что в среднем с 2001 по 2020 гг. для задач частично-целочисленного линейного программирования (ЧЦЛП) компьютерное оборудование стало примерно в 20 раз быстрее, а алгоритмы улучшились примерно в 50 раз, что дает общее ускорение в 1000 раз. В [11] автор утверждает, что за последние 10 лет коммерческий решатель задач ЧЦЛП Gurobi продемонстрировал почти 60-кратное аппаратно-независимое ускорение. И процесс совершенствования алгоритмов решения задач ЧЦЛП продолжается (см., например, [12]). В связи с достигнутым прогрессом в настоящее время публикуется довольно много научных работ, в которых различные задачи ОИР в регрессионных моделях формализуются в терминах математического программирования.

Известно, что при идентификации линейной регрессии с помощью метода наименьших модулей задача ОИР может быть сведена к задаче частично-булевого линейного программирования (ЧБЛП) [13,14], а с помощью метода наименьших квадратов (МНК) – к задаче частично-булевого квадратичного программирования (ЧБКП) [14,15]. В [16] предложена задача ОИР по критерию минимальной избыточности и максимальной релевантности, формализованная в терминах ЧЦЛП. Но всё же в зарубеж-

Базилевский Михаил Павлович, кандидат технических наук, доцент кафедры математики.
E-mail: mik2178@yandex.ru

ной литературе на данный момент большинство работ посвящено МНК, когда при ОИР в линейной регрессии решается задача ЧБКП. Например, в [17] сформулирована задача ОИР с одновременным контролем значимости оценок по t-критерию Стьюдента. А в [11] исследуется так называемая процедура «robust subset selection», при которой в линейной регрессии осуществляется как выбор регрессоров, так и выбор наблюдений.

Иной подход к построению линейной регрессии с помощью МНК, требующий решения задачи ЧБЛП вместо ЧБКП, был предложен в [18]. Его особенность в том, чтобы формировать ограничения не по выборке данных, а по корреляционной матрице переменных, а вместо минимизации суммы квадратов ошибок максимизировать коэффициент детерминации [19,20]. Тем самым, число ограничений в такой задаче не зависит от числа наблюдений, а зависит только от числа входных переменных. Предложенная в [18] задача ЧБЛП за последние годы расширилась, дополнившись новыми линейными ограничениями. Так, в [21] для контроля мультиколлинеарности появились ограничения на коэффициенты VIF вздутия дисперсий вспомогательных регрессий, в [22] – ограничения на значимость МНК-оценок по t-критерию Стьюдента, в [23] – ограничения для контроля автокорреляции остатков. В [24] вместо коэффициента детерминации в целевой функции использован его скорректированный аналог. Однако никогда прежде в эту задачу ЧБЛП не вводились ограничения на имеющий статистический характер критерий Фишера (F-тест) [25], поэтому полученная регрессия могла признаваться статистически незначимой. Цель настоящей работы и заключается в решении этой проблемы.

1. ПОСТАНОВКА ЗАДАЧИ

Прежде чем переходить к описанию математического аппарата, хотелось бы продемонстрировать результаты следующего эксперимента, подтверждающие необходимость контроля критерия Фишера при решении задачи ОИР. Для проведения эксперимента случайно были сгенерированы статистические данные (табл. 1) для выходной переменной y и входных переменных x_1, x_2, x_3 и x_4. В практике регрессионного анализа рекомендуется, чтобы объем выборки превышал число входных факторов минимум в 4 раза, поэтому объем был взят 16.

Далее по этим данным с помощью эконометрического пакета Gretl был организован ОИР методом всех регрессий. Всего с помощью МНК было оценено $2^4-1=15$ моделей. Результаты моделирования представлены в табл. 2. В ней в первом столбце указан номер регрессии, во втором – состав входящих в неё факторов, в третьем – значение коэффициента детерминации R^2, в четвертом – наблюдаемое значение критерия Фишера $F_{набл}$, в пятом – критическое значение критерия Фишера $F_{крит}$ для уровня значимости $\alpha=0,01$, в шестом, на основе сравнения $F_{набл}$ и $F_{крит}$, – значима или незначима регрессия.

По табл. 2 можно сделать следующие выводы.

Наибольшее значение коэффициент детерминации R^2 принимает для модели № 15. Но по критерию Фишеру она признана незначимой, поэтому нельзя считать её до конца адекватной.

Наибольшее значение критерия Фишера имеет значимая модель № 9. Но её коэффициент детерминации составляет всего 0,5345. Поэтому регрессию № 9 тоже нельзя считать самой адекватной, поскольку есть значимая модель № 12, для которой R^2 выше и составляет 0,6057.

Таблица 1. Статистические данные для эксперимента

№	y	x_1	x_2	x_3	x_4
1	348,65	18,02	8,48	12,03	9,30
2	287,01	7,26	7,81	13,83	14,52
3	298,32	11,30	2,96	12,27	8,76
4	189,83	16,25	2,07	9,66	11,12
5	64,74	4,23	5,89	12,46	13,05
6	386,03	1,15	4,35	14,74	0,36
7	284,58	15,40	7,84	13,60	16,88
8	525,97	18,89	7,07	14,85	6,40
9	122,16	7,67	2,29	3,00	9,19
10	163,71	14,79	1,32	13,33	8,17
11	189,98	15,36	3,42	13,42	10,13
12	281,19	15,75	6,48	2,14	4,82
13	134,24	0,73	0,84	2,98	9,90
14	272,90	6,86	2,45	9,03	4,16
15	160,98	7,85	6,10	4,56	17,83
16	225,32	8,14	6,12	11,41	6,11

Таблица 2. Результаты моделирования

№	Состав	R^2	$F_{набл}$	$F_{крит}$	Значима?
1	x_1	0,1478	2,4280	8,86	нет
2	x_2	0,2141	3,8138	8,86	нет
3	x_3	0,2079	3,6751	8,86	нет
4	x_4	0,1739	2,9476	8,86	нет
5	x_1, x_2	0,2806	2,5351	6,7	нет
6	x_1, x_3	0,2892	2,6447	6,7	нет
7	x_1, x_4	0,3560	3,5933	6,7	нет
8	x_2, x_3	0,3247	3,1249	6,7	нет
9	x_2, x_4	0,5345	7,4642	6,7	да
10	x_3, x_4	0,3620	3,6879	6,7	нет
11	x_1, x_2, x_3	0,3676	2,3255	5,95	нет
12	x_1, x_2, x_4	0,6057	6,1441	5,95	да
13	x_1, x_3, x_4	0,4719	3,5744	5,95	нет
14	x_2, x_3, x_4	0,5973	5,9320	5,95	нет
15	x_1, x_2, x_3, x_4	0,6496	5,0974	5,67	нет

Таким образом, при моделировании нельзя ориентироваться на максимизацию только коэффициента детерминации, или только критерия Фишера. В идеале сначала нужно исключать незначимые по F-критерию регрессии, а потом из оставшихся выбирать модель с наибольшей величиной R^2. В такой последовательности из табл. 2 будет выбрана регрессия № 12.

Перейдем к формализации задачи ОИР с контролем критерия Фишера в терминах математического программирования. Сначала рассмотрим задачу отбора из l объясняющих переменных x_1, x_2, ..., x_l ровно m штук по выборке объема n для линейной регрессии вида $y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \varepsilon_i$, $i = \overline{1,n}$, где α_0, α_1, ..., α_l – неизвестные параметры; ε_1, ε_2, ..., ε_n – ошибки аппроксимации. Для оценивания регрессии будем использовать МНК.

В работе [18] такая задача ОИР сведена к следующей задаче ЧБЛП с целевой функцией

$$R^2 = \sum_{j=1}^{l} r_{yx_j} \cdot \beta_j \to \max, \qquad (1)$$

и с линейными ограничениями

$$-(1-\delta_j)\cdot M \le \sum_{k=1}^{l} r_{x_j x_k} \cdot \beta_k - r_{yx_j} \le (1-\delta_j)\cdot M, \ j = \overline{1,l}, \qquad (2)$$

$$-\delta_j \cdot M \le \beta_j \le \delta_j \cdot M, \ j = \overline{1,l}, \qquad (3)$$

$$\delta_j \in \{0,1\}, \ j = \overline{1,l}, \qquad (4)$$

$$\sum_{j=1}^{l} \delta_j = m. \qquad (5)$$

В задаче (1) – (5) β_j, $j = \overline{1,l}$ – неизвестные коэффициенты стандартизованной линейной регрессии вида $y_i^\bullet = \beta_1 x_{i1}^\bullet + \beta_2 x_{i2}^\bullet + \ldots + \beta_l x_{il}^\bullet + \varepsilon_i^\bullet$, $i = \overline{1,n}$, в которой переменные преобразованы по правилам $y_i^\bullet = \frac{y_i - \overline{y}}{\sigma_y}$, $i = \overline{1,n}$, $x_{ij}^\bullet = \frac{x_{ij} - \overline{x_j}}{\sigma_{x_j}}$, $i = \overline{1,n}$, $j = \overline{1,l}$; символом r обозначены парные коэффициенты корреляции; δ_j, $j = \overline{1,l}$ – булевы переменные, отвечающие за включение объясняющих переменных в модель; M – большое положительное число. Заметим, что коэффициенты детерминации обычной и стандартизованной регрессий равны.

Пусть задан уровень значимости α. Тогда проверка значимости линейной регрессии по критерию Фишера осуществляется по следующей схеме.

1. Определяется наблюдаемое значение критерия Фишера по формуле

$$F_{набл} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}. \qquad (6)$$

2. По таблице распределения находится критическое значение критерия Фишера $F_{крит}(\alpha, m, n-m-1)$.

3. Найденные значения сравниваются между собой. Если $F_{набл} < F_{крит}$, то модель признается незначимой, а если $F_{набл} > F_{крит}$, то значимой.

Условие $F_{набл} > F_{крит}$ на значимость линейной регрессии, с учётом (6), можно записать в виде

$$\frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} > F_{крит}. \qquad (7)$$

Решив неравенство (7) относительно переменной R^2, получим ограничение:

$$R^2 \ge \frac{F_{крит}}{\frac{n-m-1}{m} + F_{крит}}. \qquad (8)$$

Неравенство (8) означает, что если значение коэффициента детерминации R^2 превосходит величину $\dfrac{F_{\text{крит}}}{\dfrac{n-m-1}{m}+F_{\text{крит}}}$, то модель является значимой в целом по F-критерию Фишера, а если не превосходит, то нет. Поэтому, дополнив задачу ОИР (1) – (5) ограничением (8), получим возможность контролировать значимость линейной регрессии.

Однако зачастую число отбираемых регрессоров m неизвестно. В таком случае поступим следующим образом. Используя соответствующую таблицу F-распределения, для заданного уровня значимости α и для каждого из чисел $m = 1, 2, ..., \mu$ (μ – назначенное исследователем наибольшее число регрессоров, причем, $\mu \le n-1$) вычислим величины $\dfrac{F_{\text{крит}}}{\dfrac{n-m-1}{m}+F_{\text{крит}}}$, которые обозначим G_1, G_2 ..., G_μ. Тогда в зависимости от числа регрессоров m должно срабатывать ровно одно неравенство из следующего набора:

$$R^2 \ge G_j,\ j = \overline{1,\mu}.$$

Введём булевы переменные ρ_j, $j = \overline{1,\mu}$, по правилу:

$$\rho_j = \begin{cases} 1, & \text{если } j = m, \\ 0, & \text{если } j \ne m. \end{cases}$$

С учётом введённых переменных сформируем следующие линейные ограничения:

$$-M\left(1-\rho_j\right) \le \sum_{k=1}^{l} \delta_k - j \le M\left(1-\rho_j\right),\ j = \overline{1,\mu}, \quad (9)$$

$$\rho_j \in \{0,1\},\ j = \overline{1,\mu}, \quad (10)$$

$$\sum_{j=1}^{\mu} \rho_j = 1, \quad (11)$$

$$R^2 \ge G_j - M\left(1-\rho_j\right),\ j = \overline{1,\mu}. \quad (12)$$

Ограничения (9) – (12) означают, что если булева переменная $\rho_j = 1$, то $j = m$, поэтому из набора (12) строгим становится только одно ограничение под номером m.

Таким образом, решение задачи ЧБЛП (1) – (4), (9) – (12) приводит к построению значимой по критерию Фишера линейной регрессии с оптимальным по критерию R^2 количеством объясняющих переменных.

2. ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ

Стоит отметить, что вычислительные эксперименты проводились с целью доказать работоспособность предложенного математического аппарата, а именно то, что при варьировании уровня значимости α в результате решения задачи ЧБЛП (1) – (4), (9) – (12) состав входящих в модель регрессоров действительно меняется, а сама регрессия оказывается значимой по критерию Фишера. Тестирование эффективности решения предложенной задачи по сравнению с другими методами при обработке больших массивов данных пока не выполнялось.

Вычислительные эксперименты были проведены на основе встроенных в пакет Gretl статистических данных, содержащихся в файле data7-10.gdt, о качестве воздуха в Калифорнии. Объем выборки $n = 30$. Среди входных переменных – численность населения, количество осадков, потребление электроэнергии промышленными производителями и пр. Подробное описание этих переменных можно найти в Gretl. Выходную переменную, которая в Gretl названа *airqual*, обозначим *y*, а 9 входных переменных *popln*, *valadd*, *rain*, *density*, *medincm*, *poverty*, *electr*, *fueloil*, *indestab* – x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8 и x_9.

Сначала по этим данным с помощью МНК была оценена линейная регрессия, уравнение которой имеет вид

$$\tilde{y} = 105,699 + 0,0913x_1 - 0,00114x_2 - 0,145x_3 - 0,000529x_4 - \\ -0,0159x_5 - 0,0268x_6 + 0,00769x_7 - 0,00283x_8 - 0,00121x_9, \quad (13)$$

а её критерии адекватности $R^2 = 0,4206$, $F_{\text{набл}} = 1,6132$. Значение R^2 гораздо меньше 1, поэтому использовать модель (13) для прогнозирования категорически нельзя. Однако перед нами сейчас другая цель. Критическое значение $F_{\text{крит}}$ критерия Фишера для уровня значимости $\alpha = 0,1$ составляет 1,9648, для $\alpha = 0,05$ – 2,3928, для $\alpha = 0,01$ – 3,4567. В любом случае наблюдаемое значение меньше критического, поэтому регрессию (13) следует признать незначимой.

Затем решалась задача ЧБЛП (1) – (4), (9) – (12) для уровней значимости 0,1, 0,05 и 0,01. При этом большое число *M* было выбрано 1000, а наибольшее число регрессоров $\mu = 9$. В качестве решателя задачи ЧБЛП был использован оптимизационный пакет LPSolve. Предварительно в зависимости от уровня значимости α были найдены критические значения критерия Фишера и значения, фигурирующие в ограничениях (12) коэффициентов G_j, $j = \overline{1,\mu}$. Эти значения приведены в табл. 3.

Модели строились по мере снижения α.

При $\alpha = 0,1$ (низший уровень) была построена модель:

$$\tilde{y} = 105,864 + 0,0918x_1 - 0,00123x_2 - 0,145x_3 - \\ -0,000563x_4 - 0,0162x_5 - 0,0268x_6 - 0,00281x_8, \quad (14)$$

для которой $R^2 = 0,4204$, $F_{\text{набл}} = 2,2792$;

Таблица 3. Значения коэффициентов G_j для ограничений (12)

m	*n-m-1*	$F_{крит}$			G		
		α = 0,1	α = 0,05	α = 0,01	α = 0,1	α = 0,05	α = 0,01
1	28	2,8938	4,1959	7,6356	0,094	0,130	0,214
2	27	2,5106	3,3541	5,4881	0,157	0,199	0,289
3	26	2,3074	2,9751	4,6365	0,210	0,256	0,349
4	25	2,1842	2,7587	4,1774	0,259	0,306	0,401
5	24	2,1030	2,6206	3,8950	0,305	0,353	0,448
6	23	2,0472	2,5276	3,7102	0,348	0,397	0,492
7	22	2,0084	2,4637	3,5866	0,390	0,439	0,533
8	21	1,9818	2,4204	3,5056	0,430	0,480	0,572
9	20	1,9648	2,3928	3,4566	0,469	0,518	0,609

при $\alpha = 0,05$ (достаточный уровень):

$$\tilde{y} = 103,255 + 0,0947x_1 - 0,00129x_2 - 0,13x_3 - \\ -0,0168x_5 - 0,0262x_6 - 0,00291x_8, \quad (15)$$

для которой $R^2 = 0,4178$, $F_{набл} = 2,7509$;

при $\alpha = 0,01$ (высший уровень):

$$\tilde{y} = 97,3966 + 0,0888x_1 - 0,0161x_5 - \\ - 0,0245x_6 - 0,00297x_8, \quad (16)$$

для которой $R^2 = 0,4098$, $F_{набл} = 4,3403$.

Как видно, с уменьшением α, т.е. с ужесточением уровня значимости от низшего к высшему, происходит уменьшение числа входящих в модель регрессоров. Так, при $\alpha = 0,1$ построена регрессия (14) с 7-ю переменными, при $\alpha = 0,05$ регрессия (15) с 6-регрессорами, а при $\alpha = 0,01$ модель (16) с 4-мя факторами. При этом все построенные модели значимы по критерию Фишера. Полученные результаты полностью совпали с результатами, полученными методом всех регрессий, что подтверждает корректность предложенного математического аппарата.

ЗАКЛЮЧЕНИЕ

В статье продемонстрировано, что при ОИР в линейной регрессии наилучшая по коэффициенту детерминации модель может оказаться статистически незначимой по критерию Фишера. Поэтому сформулировано 2 задачи ЧБЛП для построения линейной регрессии с максимальным значением коэффициента R^2 и с ограничениями на её значимость. В первой из них число отбираемых регрессоров должно быть известно, во второй – нет. Проведены вычислительные эксперименты, доказывающие корректность предложенного математического аппарата.

В завершение хотелось бы отметить следующее.

На основе проведенного автором анализа отечественной и зарубежной литературы можно сделать вывод о том, что никогда ранее в задачу математического программирования для ОИР в линейной регрессии ограничения на значимость модели по критерию Фишера не вводились. Быть может это связано с тем, что такая задача раньше не формулировалась в терминах ЧБЛП. Теперь же в этой задаче можно контролировать сразу оба критерия – коэффициент детерминации и критерий Фишера, что повышает ценность полученной в результате её решения модели.

Ранее автору уже удалось расширить задачу ОИР в линейной регрессии ограничениями на t-критерий Стьюдента [22] и на коэффициенты вздутия дисперсии [21]. Но сделано это было только для ситуации с известным числом регрессоров в модели. Предложенный в данной работе приём для контроля значимости по критерию Фишера при неизвестном числе регрессоров может серьезно расширить функциональность рассмотренных в [21,22] задач. К тому же этот приём можно использовать и при построении нелинейных по факторам моделей.

Предложенные в статье задачи ЧБЛП при высоких требованиях на значимость модели по критерию Фишера, естественным образом, могут вовсе не иметь решения. В этом случае исследователю необходимо просто снизить свои требования, заново перестроив модель.

Рассмотренный математический аппарат, безусловно, требует разработки в будущем специализированного программного продукта. С помощью последнего можно будет исследовать эффективность решения предложенных задач ЧБЛП на реальных данных. Актуально это еще и потому, что в одной из задач при контроле значимости модели по критерию Фишера появляются дополнительные бинарные переменные, что может негативно влиять на скорость решения.

СПИСОК ЛИТЕРАТУРЫ

1. *Montgomery, D.C.* Introduction to linear regression analysis / D.C. Montgomery, E.A. Peck, G.G. Vining. – John Wiley & Sons, 2021.

2. *Chatterjee, S.* Regression analysis by example / S. Chatterjee, A.S. Hadi. – John Wiley & Sons, 2015.
3. *Mahesh, B*. Machine learning algorithms-a review / B. Mahesh // International Journal of Science and Research. – 2020. – Vol. 9. – No. 1. – P. 381–386.
4. *Abid, N*. A blessing or a burden? Assessing the impact of climate change mitigation efforts in Europe using quantile regression models / N. Abid, F. Ahmad, J. Aftab, A. Razzaq // Energy Policy. – 2023. – Vol. 178. – P. 113589.
5. *Pina-Sánchez, J*. The impact of measurement error in regression models using police recorded crime rates / J. Pina-Sánchez, D. Buil-Gil, I. Brunton-Smith, A. Cernat // Journal of Quantitative Criminology. – 2023. – Vol. 39. – No. 4. – P. 975–1002.
6. *Wang, S*. Diabetes Risk Analysis Based on Machine Learning LASSO Regression Model / S. Wang, Y. Chen, Z. Cui, L. Lin, Y. Zong // Journal of Theory and Practice of Engineering Science. – 2024. – Vol. 4. – No. 1. – P. 58–64.
7. *Miller, A*. Subset selection in regression / A. Miller. – Chapman and hall/CRC, 2002.
8. *Das, A*. Algorithms for subset selection in linear regression / A. Das, D. Kempe // In Proceedings of the fortieth annual ACM symposium on Theory of computing. – 2008. – P. 45–54.
9. *Стрижов, В.В*. Методы выбора регрессионных моделей // В.В. Стрижов, Е.А. Крымова. – М.: ВЦ РАН, 2010. – 60 с.
10. *Koch, T*. Progress in mathematical programming solvers from 2001 to 2020 / T. Koch, T. Berthold, J. Pedersen, C. Vanaret // EURO Journal on Computational Optimization. – 2022. – Vol. 10. – P. 100031.
11. *Thompson, R*. Robust subset selection / R. Thompson // Computational Statistics & Data Analysis. – 2022. – Vol. 169. – P. 107415.
12. *Turner, M*. Adaptive cut selection in mixed-integer linear programming / M. Turner, T. Koch, F. Serrano, M. Winkler // arXiv preprint arXiv:2202.10962. – 2022.
13. *Носков, С.И*. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных / С.И. Носков. – Иркутск: РИЦ ГП «Облинформпечать», 1996. – 321 с.
14. *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.
15. *Bertsimas, D*. Best subset selection via a modern optimization lens / D. Bertsimas, A. King, R. Mazumder // The Annals of Statistics. – 2016. – Vol. 44. – No. 2. – P. 813–852.
16. *Park, Y.W*. Subset selection for multiple linear regression via optimization / Y.W. Park, D. Klabjan // Journal of Global Optimization. – 2020. – Vol. 77. – No. 3. – P. 543–574.
17. *Chung, S*. A mathematical programming approach for integrated multiple linear regression subset selection and validation / S. Chung, Y.W. Park, T. Cheong // Pattern Recognition. – 2020. – Vol. 108. – P. 107565.
18. *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.
19. *Chicco, D*. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation / D. Chicco, M.J. Warrens, G. Jurman // Peerj computer science. – 2021. – Vol. 7. – P. e623.
20. *Piepho, H.P*. An adjusted coefficient of determination (R2) for generalized linear mixed models in one go / H.P. Piepho // Biometrical Journal. – 2023. – Vol. 65. – No. 7. – P. 2200290.
21. *Базилевский, М.П.* Отбор информативных регрессоров с учетом мультиколлинеарности между ними в регрессионных моделях как задача частично-булевого линейного программирования / М.П. Базилевский // Моделирование, оптимизация и информационные технологии. – 2018. – Т. 6. – № 2 (21). – С. 104–118.
22. *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.
23. *Базилевский, М.П.* Контроль автокорреляции остатков с помощью коэффициента Фехнера в задаче математического программирования для отбора информативных регрессоров в линейной регрессии / М.П. Базилевский // System Analysis and Mathematical Modeling. – 2024. – Т. 6. – № 2. – С. 146–158.
24. *Базилевский, М.П*. Отбор оптимального числа информативных регрессоров по скорректированному коэффициенту детерминации в регрессионных моделях как задача частично целочисленного линейного программирования / М.П. Базилевский // Прикладная математика и вопросы управления. – 2020. – № 2 . – С. 41–54.
25. *Garcia-Queiruga, J*. A cross-sectional study of non-modifiable and modifiable risk factors of dry eye disease states / J. Garcia-Queiruga, H. Pena-Verdeal, B. Sabucedo-Villamarin, M.J. Giraldez, C. Garcia-Resua, E. Yebra-Pimentel // Contact Lens and Anterior Eye. – 2023. – Vol. 46. – No. 3. – P. 101800.

OPTIMIZATION PROBLEMS OF SUBSET SELECTION IN LINEAR REGRESSION WITH CONTROL OF ITS SIGNIFICANCE USING F-TEST

Irkutsk State Transport University, Irkutsk, Russia

This article is devoted to the problem of subset selection in multiple linear regression models. When implementing such a selection using the determination coefficient, the resulting model may be insignificant according to the F-test. To solve this problem, two problems of mixed 0-1 integer linear programming are proposed, the solution algorithms for which have been improved dozens of times over the past 20 years. The solution to the first of them gives an optimal model with an assigned number of factors, while the optimal number of factors is determined automatically when solving the second one. Computational experiments were carried out. For the second problem, an example shows that with tightening the requirements for the significance of the model according to the F-test, the number of factors in the selection decreases. The technique proposed in the article, associated with the introduction of additional binary variables, can be used in the future to control multicollinearity in models and the significance of estimates according to the Student's t-test.

Keywords: regression analysis, linear regression, ordinary least squares, subset selection, mixed 0-1 integer linear programming, coefficient of determination, F-test.

DOI: 10.37313/1990-5378-2024-26-6-200-207
EDN : MEZQQZ

REFERENCES

1. *Montgomery, D.C.* Introduction to linear regression analysis / D.C. Montgomery, E.A. Peck, G.G. Vining. – John Wiley & Sons, 2021.
2. *Chatterjee, S.* Regression analysis by example / S. Chatterjee, A.S. Hadi. – John Wiley & Sons, 2015.
3. *Mahesh, B*. Machine learning algorithms-a review / B. Mahesh // International Journal of Science and Research. – 2020. – Vol. 9. – No. 1. – P. 381–386.
4. *Abid, N*. A blessing or a burden? Assessing the impact of climate change mitigation efforts in Europe using quantile regression models / N. Abid, F. Ahmad, J. Aftab, A. Razzaq // Energy Policy. – 2023. – Vol. 178. – P. 113589.
5. *Pina-Sánchez, J*. The impact of measurement error in regression models using police recorded crime rates / J. Pina-Sánchez, D. Buil-Gil, I. Brunton-Smith, A. Cernat // Journal of Quantitative Criminology. – 2023. – Vol. 39. – No. 4. – P. 975–1002.
6. *Wang, S*. Diabetes Risk Analysis Based on Machine Learning LASSO Regression Model / S. Wang, Y. Chen, Z. Cui, L. Lin, Y. Zong // Journal of Theory and Practice of Engineering Science. – 2024. – Vol. 4. – No. 1. – P. 58–64.
7. *Miller, A*. Subset selection in regression / A. Miller. – Chapman and hall/CRC, 2002.
8. *Das, A*. Algorithms for subset selection in linear regression / A. Das, D. Kempe // In Proceedings of the fortieth annual ACM symposium on Theory of computing. – 2008. – P. 45–54.
9. *Strizhov, V.V.* Metody vybora regressionnykh modeley // V.V. Strizhov, E.A. Krymova. – M.: VTs RAN, 2010. – 60 s.
10. *Koch, T*. Progress in mathematical programming solvers from 2001 to 2020 / T. Koch, T. Berthold, J. Pedersen, C. Vanaret // EURO Journal on Computational Optimization. – 2022. – Vol. 10. – P. 100031.
11. *Thompson, R*. Robust subset selection / R. Thompson // Computational Statistics & Data Analysis. – 2022. – Vol. 169. – P. 107415.
12. *Turner, M*. Adaptive cut selection in mixed-integer linear programming / M. Turner, T. Koch, F. Serrano, M. Winkler // arXiv preprint arXiv:2202.10962. – 2022.
13. *Noskov, S.I.* Tekhnologiya modelirovaniya ob"ektov s nestabil'nym funktsionirovaniem i neopredelennost'yu v dannykh / S.I. Noskov. – Irkutsk: RITs GP «Oblinformpechat'», 1996. – 321 s.
14. *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.
15. *Bertsimas, D*. Best subset selection via a modern optimization lens / D. Bertsimas, A. King, R. Mazumder // The Annals of Statistics. – 2016. – Vol. 44. – No. 2. – P. 813–852.
16. *Park, Y.W*. Subset selection for multiple linear regression via optimization / Y.W. Park, D. Klabjan // Journal of Global Optimization. – 2020. – Vol. 77. – No. 3. – P. 543–574.
17. *Chung, S*. A mathematical programming approach for integrated multiple linear regression subset selection and validation / S. Chung, Y.W. Park, T. Cheong // Pattern Recognition. – 2020. – Vol. 108. – P. 107565.
18. *Bazilevskiy, M.P.* Svedenie zadachi otbora informativnykh regressorov pri otsenivanii lineynoy regressionnoy modeli po metodu naimen'shikh kvadratov k zadache chastichno-bulevogo lineynogo programmirovaniya / M.P. Bazilevskiy // Modelirovanie, optimizatsiya i informatsionnye tekhnologii. – 2018. – T. 6. – № 1 (20). – S. 108–117.
19. *Chicco, D*. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation / D. Chicco, M.J. Warrens, G. Jurman // Peerj computer science. – 2021. – Vol. 7. – P. e623.
20. *Piepho, H.P*. An adjusted coefficient of determination (R2) for generalized linear mixed models in one go / H.P. Piepho // Biometrical Journal. – 2023. – Vol. 65. –

No. 7. – P. 2200290.
21. *Bazilevskiy, M.P.* Otbor informativnykh regressorov s uchetom mul'tikollinearnosti mezhdu nimi v regressionnykh modelyakh kak zadacha chastichno-bulevogo lineynogo programmirovaniya / M.P. Bazilevskiy // Modelirovanie, optimizatsiya i informatsionnye tekhnologii. – 2018. – T. 6. – № 2 (21). – S. 104–118.
22. *Bazilevskiy, M.P.* Otbor znachimykh po kriteriyu St'yudenta informativnykh regressorov v otsenivaemykh s pomoshch'yu MNK regressionnykh modelyakh kak zadacha chastichno-bulevogo lineynogo programmirovaniya / M.P. Bazilevskiy // Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: Sistemnyy analiz i informatsionnye tekhnologii. – 2021. – № 3. – S. 5–16.
23. *Bazilevskiy, M.P.* Kontrol' avtokorrelyatsii ostatkov s pomoshch'yu koeffitsienta Fekhnera v zadache matematicheskogo programmirovaniya dlya otbora informativnykh regressorov v lineynoy regressii / M.P. Bazilevskiy // System Analysis and Mathematical Modeling. – 2024. – T. 6. – № 2. – S. 146–158.
24. *Bazilevskiy, M.P.* Otbor optimal'nogo chisla informativnykh regressorov po skorrektirovannomu koeffitsientu determinatsii v regressionnykh modelyakh kak zadacha chastichno tselochislennogo lineynogo programmirovaniya / M.P. Bazilevskiy // Prikladnaya matematika i voprosy upravleniya. – 2020. – № 2 . – S. 41–54.
25. *Garcia-Queiruga, J.* A cross-sectional study of non-modifiable and modifiable risk factors of dry eye disease states / J. Garcia-Queiruga, H. Pena-Verdeal, B. Sabucedo-Villamarin, M.J. Giraldez, C. Garcia-Resua, E. Yebra-Pimentel // Contact Lens and Anterior Eye. – 2023. – Vol. 46. – No. 3. – P. 101800.

Mikhail Bazilevskiy, Candidate of Technics, Associate Professor at the Mathematics Department.
E-mail: mik2178@yandex.ru