Моделирование и анализ данных
2024. Том 14. № 4. С. 91–103
DOI: https://doi.org/10.17759/mda.2024140406
ISSN: 2219-3758 (печатный)
ISSN: 2311-9454 (online)

Modelling and Data Analysis
2024. Vol. 14, no. 4, pp. 91–103
DOI: https://doi.org/10.17759/mda.2024140406
ISSN: 2219-3758 (print)
ISSN: 2311-9454 (online)

◇◇◇◇◇◇◇ МЕТОДЫ ОПТИМИЗАЦИИ ◇◇◇◇◇◇◇

УДК 519.862.6

Оптимизационная задача построения линейных регрессий с минимальной величиной средней абсолютной ошибки на тестовых выборках

Базилевский М.П.*
Иркутский государственный университет путей сообщения
(ФГБОУ ВО ИрГУПС), г. Иркутск, Российская Федерация
ORCID: https://orcid.org/0000-0002-3253-5697
e-mail: mik2178@yandex.ru

Статья посвящена проблеме отбора заданного числа наиболее информативных регрессоров в линейных регрессиях. При использовании метода наименьших квадратов точное решение этой задачи по критерию максимизации коэффициента детерминации при задействовании всей выборки данных может быть получено в результате решения особым образом сформулированной задачи частично-булевого линейного программирования. Однако в машинном обучении важным этапом при создании надежной и эффективной модели считается её построение по обучающей выборке и проверка точности её предсказания по тестовой выборке. Поэтому в статье сформулирована оптимизационная задача отбора информативных регрессоров в линейных регрессиях по критерию минимизации средней абсолютной ошибки на тестовой выборке. Формулировка основана на известном приёме, согласно которому абсолютные ошибки должны быть представлены в виде разности между двумя неотрицательными переменными. С использованием встроенных в пакет Gretl статистических данных о заработной плате спортсменов и решателя оптимизационных задач LPSolve проведены вычислительные эксперименты. Для этого обучающая выборка формировалась из 70%, 75% и 80% наблюдений. Во всех этих случаях среднее снижение значения коэффициента детерминации моделей составило 24,76%, 18,4% и 12,22%, но при этом средняя абсолютная ошибка уменьшилась на 24,8%, 26,3% и 21,05% соответственно. Эксперименты показали, что среднее время решения задач при минимизации средней абсолютной ошибки на тестовых выборках оказалось в 2,33–2,85 раза выше, чем время решения задач при максимизации коэффициента детерминации на обучающих выборках.

Ключевые слова: машинное обучение, регрессионный анализ, метод наименьших квадратов, отбор информативных регрессоров, коэффициент

детерминации, средняя абсолютная ошибка, обучающая выборка, тестовая выборка, задача частично-булевого линейного программирования.

Для цитаты:
Базилевский М.П. Оптимизационная задача построения линейных регрессий с минимальной величиной средней абсолютной ошибки на тестовых выборках // Моделирование и анализ данных. 2024. Том 14. № 4. С. 91–103. DOI: https://doi.org/10.17759/mda.2024140406

***Базилевский Михаил Павлович,** кандидат технических наук, доцент кафедры математики, Иркутский государственный университет путей сообщения (ФГБОУ ВО ИрГУПС), г. Иркутск, Российская Федерация, ORCID: https://orcid.org/0000-0002-3253-5697, e-mail: mik2178@yandex.ru

1. ВВЕДЕНИЕ

В современном мире модели машинного обучения [1, 2] активно применяются для решения самых разнообразных задач анализа данных в энергетике [3], экономике [4], медицине [5], промышленности [6] и других областях человеческой деятельности. Существует множество разновидностей моделей машинного обучения, однако наиболее высокими интерпретационными качествами из них обладают регрессионные модели [7]. Регрессионный анализ в настоящее время развивается не менее стремительно, чем искусственный интеллект. Например, в [8] предложен метод частичных наименьших квадратов расстояний, в [9] – двухуровневый метод регрессионного анализа, использующий ансамбли деревьев с оптимальной дивергенцией, в [10] исследовано двухкритериальное оценивание регрессий методами наименьших квадратов (МНК) и модулей, в [11] рассмотрена методика построения линейно-неэлементарных регрессий.

Очень часто при проведении регрессионного анализа на практике приходится решать задачу отбора наиболее информативных регрессоров (ОИР) [12], т.е. формировать регрессионную модель только со значимо влияющими на результативный признак объясняющими переменными. Для её решения разработано множество методов, многие из которых описаны в монографии [13]. Из них точное решение задачи ОИР дает метод всех возможных регрессий [14], суть которого заключается в оценивании моделей со всеми возможными комбинациями вариантов вхождений объясняющих переменных в регрессионное уравнение. Поэтому метод всех регрессий самый трудоёмкий из всех. Другой точный метод решения задачи ОИР был предложен в работе [15]. В ней сформулирована задача частично-булевого линейного программирования (ЧБЛП), решение которой дает оцененную с помощью МНК наилучшую по коэффициенту детерминации регрессию с заданным числом объясняющих переменных. В [16] эта задача была трансформирована в задачу частично-целочисленного линейного программирования, решение которой дает оптимальную по скорректированному коэффициенту детерминации регрессию. В [17] экспериментально доказано, что при построении вполне интерпретируемых линейных регрессий

предложенный в работах [15, 16] метод существенно выигрывает по эффективности у метода всех возможных регрессий.

В предложенной в работах [15–17] технологии ОИР осуществляется сразу по всей исходной выборке данных. Однако в машинном обучении важным этапом при создании надежной и эффективной модели считается предварительное разделение выборки на обучающую и тестовую [18–20]. По обучающей выборке строится модель, а тестовая выборка используется для проверки точности предсказания модели. Обычно на обучающую выборку приходится 70% наблюдений, а на тестовую – 30%. Разделение выборки делается, во-первых, для оценки работоспособности модели в реальных условиях, во-вторых, для предотвращения переобучения, в-третьих, для объективной оценки качества модели. Цель данной работы состоит в формализации задачи ОИР в линейной регрессии по критерию минимизации величины средней абсолютной ошибки на тестовой выборке.

2. ПОСТАНОВКА ОПТИМИЗАЦИОННОЙ ЗАДАЧИ

Предположим, что y – зависимая (объясняемая, выходная) переменная, а x_1, x_2, …, x_l – независимые (объясняющие, входные) переменные. Пусть общий объем выборки составляет n наблюдений, из них n_1 наблюдений приходится на обучающую выборку, а n_2 – на тестовую. По обучающей выборке с помощью МНК оцениваются неизвестные параметры α_0, α_1, …, α_l модели множественной линейной регрессии вида

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \varepsilon_i,\ i = \overline{1, n_1}, \tag{1}$$

где ε_i, $i = \overline{1, n_1}$ – ошибки аппроксимации.

Как отмечено в [21], эффективность МНК-оценивания регрессии (1) увеличится, если все исходные переменные нормировать по правилам:

$$y_i^{\bullet} = \frac{y_i - \overline{y}}{\sigma_y},\ i = \overline{1, n_1},$$

$$x_{ij}^{\bullet} = \frac{x_{ij} - \overline{x_j}}{\sigma_{x_j}},\ i = \overline{1, n_1},\ j = \overline{1, l},$$

где $\overline{y} = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$, $\overline{x_j} = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij}$, $j = \overline{1, l}$, $\sigma_y = \sqrt{\overline{y^2} - \left(\overline{y}\right)^2}$, $\sigma_{x_j} = \sqrt{\overline{x_j^2} - \left(\overline{x_j}\right)^2}$, $j = \overline{1, l}$.

Тогда модель множественной линейной регрессии (1) в стандартизованном масштабе принимает вид

$$y_i^{\bullet} = \sum_{j=1}^{l} \beta_j x_{ij}^{\bullet} + \varepsilon_i^{\bullet},\ i = \overline{1, n_1}, \tag{2}$$

где β_1, …, β_l – стандартизованные коэффициенты регрессии, $\varepsilon_i^{\bullet}$, $i=\overline{1,n_1}$ – ошибки аппроксимации.

Известно [21], что неизвестные коэффициенты регрессии (2) находятся в результате решения следующей системы линейных алгебраических уравнений:

$$R_{xx}\cdot\beta = R_{yx}, \tag{3}$$

где $R_{xx}=\begin{pmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_l} \\ r_{x_1x_2} & 1 & \dots & r_{x_2x_l} \\ \dots & \dots & \dots & \dots \\ r_{x_1x_l} & r_{x_2x_l} & \dots & 1 \end{pmatrix}$, $\beta=\begin{pmatrix} \beta_1 \\ \beta_2 \\ \dots \\ \beta_l \end{pmatrix}$, $R_{yx}=\begin{pmatrix} r_{yx_1} \\ r_{yx_2} \\ \dots \\ r_{yx_l} \end{pmatrix}$, т.е. R_{xx} – матрица коэффициентов корреляции между объясняющими переменными, β – вектор-столбец неизвестных коэффициентов, R_{yx} – вектор-столбец коэффициентов корреляции результативного признака с объясняющими переменными. Подчеркнем, что матрицы R_{xx} и R_{yx} находятся по обучающей выборке.

Альтернативная форма записи системы (3) имеет вид

$$\sum_{k=1}^{l} r_{x_jx_k}\cdot\beta_k = r_{yx_j},\ j=\overline{1,l}. \tag{4}$$

Обозначим найденные в результате решения системы (4) оценки $\tilde{\beta}_j$, $j=\overline{1,l}$. Тогда, как отмечено в [15], с помощью этих оценок можно найти оптимальные МНК-оценки линейной регрессии (1) по формулам:

$$\tilde{\alpha}_j = \tilde{\beta}_j\frac{\sigma_y}{\sigma_{x_j}},\ j=\overline{1,l}, \tag{5}$$

$$\tilde{\alpha}_0 = \overline{y} - \sum_{j=1}^{l}\tilde{\alpha}_j\overline{x_j}. \tag{6}$$

Коэффициент детерминации R^2 стандартизованной линейной регрессии (2) вычисляется по формуле

$$R^2 = \sum_{j=1}^{l} r_{yx_j}\cdot\tilde{\beta}_j. \tag{7}$$

Заметим, что коэффициенты детерминации регрессий (1) и (2) одинаковы.

Формализовать задачу ОИР в линейной регрессии в терминах математического программирования можно следующим образом. Введем бинарные переменные δ_j, $j=\overline{1,l}$ по правилу:

$$\delta_j = \begin{cases} 1, & \text{если } j\text{-я объясняющая переменная входит в модель,} \\ 0, & \text{в противном случае.} \end{cases}$$

С использованием этих переменных поставим на стандартизованные коэффициенты линейной регрессии следующие ограничения:

$$-\delta_j \cdot M \le \beta_j \le \delta_j \cdot M,\ j = \overline{1,l}\,, \tag{8}$$

где M – большое положительное число. Если $\delta_j = 1$, то $\beta_j \in [-M, M]$, а если $\delta_j = 0$, то $\beta_j = 0$.

Если в линейную регрессию должно входить ровно m регрессоров, то задачу математического программирования следует дополнить ограничением

$$\sum_{j=1}^{l} \delta_j = m\,. \tag{9}$$

Как уже было отмечено, если $\delta_j = 0$, то $\beta_j = 0$, поэтому в зависимости от бинарных переменных δ_j, $j = \overline{1,l}$ должна меняться конфигурация системы линейных алгебраических уравнений (4), а именно, из неё должны исключаться уравнения с номерами, совпадающими с номерами нулевых бинарных переменных δ_j. Это может быть реализовано с помощью следующих линейных ограничений

$$-\left(1-\delta_j\right)M \le \sum_{k=1}^{l} r_{x_j x_k} \cdot \beta_k - r_{yx_j} \le \left(1-\delta_j\right)M\,,\ j = \overline{1,l}\,. \tag{10}$$

Если $\delta_j = 1$, то $\sum_{k=1}^{l} r_{x_j x_k} \cdot \beta_k - r_{yx_j} = 0$, т.е. соответствующее уравнение включается в систему, а если $\delta_j = 0$, то $\sum_{k=1}^{l} r_{x_j x_k} \cdot \beta_k - r_{yx_j} \in [-M, M]$, т.е. соответствующее уравнение исключается из системы.

Оптимальной регрессией считается та, у которой значение коэффициента детерминации наибольшее. Учитывая, что коэффициент детерминации R^2 находится по формуле (7), введем целевую функцию

$$\sum_{j=1}^{l} r_{yx_j} \cdot \beta_j \to \max\,. \tag{11}$$

Решение задачи ЧБЛП с целевой функцией (11) и с линейными ограничениями (8) – (10) приводит к построению по обучающей выборке объема n_1 линейной регрессии с m регрессорами и с наибольшим значением R^2.

Предположим теперь, что оптимальной линейной регрессией считается не та, у которой на обучающей выборке значение R^2 наибольшее, а та, у которой на тестовой выборке значение средней абсолютной ошибки MAE наименьшее. Средняя абсолютная ошибка находится по формуле:

$$MAE = \frac{1}{n_2} \sum_{i=n_1+1}^{n_2} \left| y_i - \tilde{y}_i \right|, \tag{12}$$

где $\tilde{y}_i$, $i = \overline{n_1+1, n_2}$ – прогнозные по линейной регрессии (1) значения зависимой переменной y. С учетом (5) и (6) эти прогнозные значения находятся по формулам:

$$\tilde{y}_i = \overline{y} + \sum_{j=1}^{l} \tilde{\beta}_j \frac{\sigma_y}{\sigma_{x_j}} \left(x_{ij} - \overline{x_j} \right), \; i = \overline{n_1+1, n_2} \; . \tag{13}$$

Подставляя (13) в (12), получим, что

$$MAE = \frac{1}{n_2} \sum_{i=n_1+1}^{n_2} \left| y_i - \overline{y} - \sum_{j=1}^{l} \tilde{\beta}_j \frac{\sigma_y}{\sigma_{x_j}} \left(x_{ij} - \overline{x_j} \right) \right| . \tag{14}$$

Учитывая (14), введем целевую функцию

$$\sum_{i=n_1+1}^{n_2} \left| y_i - \overline{y} - \sum_{j=1}^{l} \beta_j \frac{\sigma_y}{\sigma_{x_j}} \left(x_{ij} - \overline{x_j} \right) \right| \to \min. \tag{15}$$

Линеаризовать функционал (15) можно с использованием известного приёма, предложенного в [22]. Для этого введем неотрицательные переменные u_i, v_i, $i = \overline{n_1+1, n_2}$ следующим образом:

$$u_i = \begin{cases} y_i - \overline{y} - \sum_{j=1}^{l} \beta_j \frac{\sigma_y}{\sigma_{x_j}} \left(x_{ij} - \overline{x_j} \right), \text{ если } y_i - \overline{y} - \sum_{j=1}^{l} \beta_j \frac{\sigma_y}{\sigma_{x_j}} \left(x_{ij} - \overline{x_j} \right) > 0, \\ 0, \text{ в противном случае,} \end{cases}$$

$$v_i = \begin{cases} -y_i + \overline{y} + \sum_{j=1}^{l} \beta_j \frac{\sigma_y}{\sigma_{x_j}} \left(x_{ij} - \overline{x_j} \right), \text{ если } y_i - \overline{y} - \sum_{j=1}^{l} \beta_j \frac{\sigma_y}{\sigma_{x_j}} \left(x_{ij} - \overline{x_j} \right) < 0, \\ 0, \text{ в противном случае.} \end{cases}$$

Тогда имеют место равенства

$$\sum_{j=1}^{l} \beta_j \frac{\sigma_y}{\sigma_{x_j}} \left(x_{ij} - \overline{x_j} \right) + u_i - v_i = y_i - \overline{y} \; , \; i = \overline{n_1+1, n_2} . \tag{16}$$

При этом целевая функция (15) принимает вид

$$\sum_{i=n_1+1}^{n_2} \left(u_i + v_i \right) \to \min . \tag{17}$$

Решение задачи ЧБЛП с целевой функцией (17) и с линейными ограничениями (8) – (10), (16) приводит к построению по обучающей выборке объема n_1 линейной регрессии с m регрессорами и с наименьшим на тестовой выборке объема n_2 значением MAE.

3. ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ

Вычислительные эксперименты, прежде всего, проводились с целью подтвердить работоспособность предложенного математического аппарата. К тому же ставилась задача выяснить экспериментально, будет ли время решения задач ЧБЛП при максимизации R^2 на обучающей выборке существенно отличаться от времени решения задач при минимизации MAE на тестовой выборке.

Для проведения вычислительных экспериментов использовались статистические данные, хранящиеся в файле data7–20.gdt эконометрического пакета Gretl. В этом файле содержится информация о зарплате 56-ти игроков национальной баскетбольной ассоциации и влияющих на неё 25-ти показателей. Среди этих показателей антропометрические характеристики спортсменов, их позиции на площадке, количество набранных очков и пр. В результате предварительного анализа выборки было установлено, что первые 55 значений фиктивной переменной XPAN, обозначающей расширение команды за последние 2 сезона, равны 0. И только последнее 56-е значение равно 1. Поскольку изменчивость переменной XPAN практически нулевая, то было принято решение исключить её из рассмотрения. Итого осталось 24 показателя, которым были присвоены имена x_1, x_2, …, x_{24} в соответствие с порядком их следования в файле Gretl.

Вычислительные эксперименты проводились на персональном компьютере с процессором AMD Ryzen 3 4300U с тактовой частотой 2,7 ГГц и объемом оперативной памяти 16 Гб. Для решения задач ЧБЛП использовался бесплатный оптимизационный решатель LPSolve. Большое число M в задачах ЧБЛП задавалось равным 1000.

Эксперименты проводились при трёх различных делениях исходной выборки на обучающую и тестовую. В первом случае на обучающую выборку приходилось 70% наблюдений, во втором – 75%, в третьем – 80%. В каждом случае решалась задача максимизации коэффициента детерминации R^2 на обучающей выборке, т.е. задача ЧБЛП (8) – (11), и задача минимизации средней абсолютной ошибки MAE на тестовой выборке, т.е. задача ЧБЛП (8) – (10), (16), (17). При этом число регрессоров m менялось в диапазоне от 1 до 5. В каждом эксперименте фиксировался состав входящих в модель регрессоров, значение R^2, MAE и время решения задачи в LPSolve. Результаты вычислительных экспериментов приведены в табл. 1. В ней в первом столбце указан номер эксперимента, во втором – число назначенных регрессоров, в третьем, четвертом, пятом и шестом столбцах – состав регрессоров, значение R^2, MAE и время решения задачи при максимизации R^2 на обучающей выборке, в седьмом, восьмом, девятом и десятом столбцах аналогичные показатели при минимизации MAE на тестовой выборке.

Таблица 1

Результаты вычислительных экспериментов

№	m	Максимизация R^2				Минимизация MAE			
		Регрессоры	R^2	MAE	t, c	Регрессоры	R^2	MAE	t, c
Обучающая – 70% (n_1 = 39), тестовая – 30% (n_2 = 17)									
1	1	x_{24}	0,19039	11797,4	0,03	x_{24}	0,19039	11797,4	0,055
2	2	x_{10}, x_{19}	0,28829	13328,5	0,091	x_3, x_{24}	0,20567	10938,0	0,224

№	m	Максимизация R^2				Минимизация MAE			
		Регрессоры	R^2	MAE	t, c	Регрессоры	R^2	MAE	t, c
Обучающая – 70% (n_1 = 39), тестовая – 30% (n_2 = 17)									
3	3	x_{10}, x_{16}, x_{19}	0,38982	16451,4	0,466	x_3, x_{14}, x_{24}	0,25024	10480,4	1,422
4	4	x_6, x_{15}, x_{17}, x_{19}	0,44877	15699,6	2,675	x_3, x_{10}, x_{14}, x_{24}	0,28966	9979,9	5,904
5	5	x_6, x_8, x_{15}, x_{17}, x_{19}	0,50891	14889,6	9,751	x_7, x_8, x_{18}, x_{19}, x_{24}	0,38726	9927,5	20,579
Обучающая – 75% (n_1 = 42), тестовая – 25% (n_2 = 14)									
6	1	x_{24}	0,22619	9323,8	0,032	x_{24}	0,22619	9323,8	0,057
7	2	x_{10}, x_{19}	0,29207	10668,6	0,099	x_3, x_{24}	0,24270	8500,0	0,274
8	3	x_{10}, x_{16}, x_{19}	0,36893	13300,5	0,525	x_3, x_{14}, x_{24}	0,27157	7897,9	1,710
9	4	x_3, x_{14}, x_{16}, x_{24}	0,42466	11555,5	2,650	x_3, x_{14}, x_{15}, x_{24}	0,33933	7355,3	7,195
10	5	x_3, x_{14}, x_{16}, x_{19}, x_{24}	0,47636	11051,6	10,345	x_3, x_{14}, x_{15}, x_{23}, x_{24}	0,33993	7263,1	24,417
Обучающая – 80% (n_1 = 45), тестовая – 20% (n_2 = 11)									
11	1	x_{24}	0,25612	7248,1	0,03	x_{24}	0,25612	7248,1	0,054
12	2	x_{19}, x_{24}	0,30327	7395,4	0,092	x_3, x_{24}	0,27518	6509,3	0,261
13	3	x_{10}, x_{19}, x_{24}	0,37019	6985,8	0,480	x_3, x_{14}, x_{24}	0,31325	6204,2	1,831
14	4	x_3, x_{14}, x_{16}, x_{24}	0,45369	10028,8	2,571	x_3, x_{14}, x_{15}, x_{24}	0,37268	5916,6	7,941
15	5	x_3, x_{14}, x_{16}, x_{19}, x_{24}	0,49648	9474,7	9,668	x_7, x_8, x_{18}, x_{19}, x_{24}	0,40417	5582,6	26,345

Прокомментируем полученные в табл. 1 результаты.

1. При решении задач минимизации MAE по тестовым выборкам, естественным образом, получены регрессии, выигравшие по этому критерию у построенных в результате максимизации R^2 по обучающим выборкам моделей, но проигравшие им по критерию R^2. Так, при включении в обучающую выборку 70% наблюдений значение коэффициента детерминации R^2 снижалось в диапазоне от 0 до 0,159, а значение MAE – в диапазоне от 0 до 5971. При этом в среднем снижение значения R^2 составило 24,76%, а MAE – 24,8%. При включении в обучающую выборку 75% наблюдений значение R^2 снижалось в диапазоне от 0 до 0,136, а значение MAE – в диапазоне от 0 до 5402,6. В этом случае в среднем снижение значения R^2 составило 18,4%, а MAE – 26,3%. При включении в обучающую выборку 80% наблюдений значение R^2 снижалось в диапазоне от 0 до 0,092, а значение MAE – в диапазоне от 0 до 4112,2. В такой ситуации в среднем снижение значения R^2 составило 12,22%, а MAE – 21,05%.
2. Во всех случаях с ростом числа регрессоров *m* возрастало время решения задач ЧБЛП. При включении в обучающую выборку 70%, 75% и 80% наблюдений среднее время решения задач при минимизации MAE оказалось соответственно в 2,33, 2,58 и 2,85 раза выше, чем время решения задач при максимизации R^2. При этом замечено, что с ростом числа *m* отношение времени решения при минимизации MAE ко времени решения при максимизации R^2 сначала возрастает, достигая наибольшего значения при *m*=3, а затем убывает. Так, при включении в обучающую выборку 70% наблюдений эти отношения составляют 1,83, 2,46, 3,05, 2,21,

2,11, при включении 75% – 1,78, 2,76, 3,25, 2,71, 2,36, при включении 80% – 1,8, 2,83, 3,81, 3,08, 2,72.

4. ЗАКЛЮЧЕНИЕ

В статье сформулирована оптимизационная задача ОИР в линейных регрессиях по критерию минимизации средней абсолютной ошибки на тестовой выборке. Проведены вычислительные эксперименты, подтверждающие корректность математических выкладок. Во всех экспериментах значения средних абсолютных ошибок на тестовых выборках были снижены в большей степени, чем значения коэффициентов детерминации моделей. Время решения задач ЧБЛП при минимизации средних абсолютных ошибок на тестовых выборках оказалось в 2,33–2,85 раза больше, чем время решения задач при максимизации коэффициентов детерминации на обучающих выборках. Научный интерес вызывает внедрение предложенного метода в процедуру построения вполне интерпретируемых регрессионных моделей, которая работает значительно эффективнее метода всех возможных регрессий. К тому же открытым остается вопрос идентификации в сформулированной задаче ЧБЛП больших чисел M, влияющих на её скорость решения.

Литература

1. *Рашка С.* Python и машинное обучение. М.: ДМК Пресс, 2017. 418 с.
2. *Janiesch C., Zschech P., Heinrich K.* Machine learning and deep learning // Electronic Markets. 2021. Vol. 31. No. 3. P. 685–695. DOI:10.1007/s12525-021-00475-2.
3. *Mhlanga D.* Artificial intelligence and machine learning for energy consumption and production in emerging markets: a review // Energies. 2023. Vol. 16. No. 2. P. 745. DOI:10.3390/en16020745.
4. *Xu Z., Mohsin M., Ullah K., Ma X.* Using econometric and machine learning models to forecast crude oil prices: Insights from economic history // Resources Policy. 2023. Vol. 83. P. 103614. DOI:10.1016/j.resourpol.2023.103614.
5. *Haug C.J., Drazen J.M.* Artificial intelligence and machine learning in clinical medicine // New England Journal of Medicine. 2023. Vol. 388. No. 13. P. 1201–1208. DOI:10.1056/NEJMra2302038.
6. *Kumar S., Gopi T., Harikeerthana N., Gupta M.K., Gaur V., Krolczyk G.M., Wu C.* Machine learning techniques in additive manufacturing: a state of the art review on design, processes and production control // Journal of Intelligent Manufacturing. 2023. Vol. 34. No. 1. P. 21–55. DOI: 10.1007/s10845-022-02029-5.
7. *Molnar C.* Interpretable machine learning. Lulu. com, 2020.
8. *Nie B., Du Y., Du J., Rao Y., Zhang Y., Zheng X., Ye N., Jin H.* A novel regression method: Partial least distance square regression methodology // Chemometrics and Intelligent Laboratory Systems. 2023. Vol. 237. P. 104827. DOI:10.1016/j.chemolab.2023.104827.
9. *Журавлев Ю.И., Сенько О.В., Докукин А.А., Киселева Н.Н., Саенко И.А.* Двухуровневый метод регрессионного анализа, использующий ансамбли деревьев с оптимальной дивергенцией // Доклады Российской академии наук. Математика, информатика, процессы управления. 2021. Т. 499. С. 63–66. DOI:10.31857/S2686954321040172.

10. *Базилевский М.П.* Двухкритериальное оценивание линейных регрессионных моделей методами наименьших модулей и квадратов // International Journal of Open Information Technologies. 2024. Т. 12. № 6. С. 76–81.
11. *Базилевский М.П.* Отбор информативных операций при построении линейно-неэлементарных регрессионных моделей // International Journal of Open Information Technologies. 2021. Т. 9. № 5. С. 30–35.
12. *Носков С.И.* Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. Иркутск: РИЦ ГП Облинформпечать, 1996. 321 с.
13. *Miller A.* Subset selection in regression. Chapman and hall/CRC, 2002.
14. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. М.: ЮНИТИ, 1998. 1005 с.
15. *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.
16. *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.
17. *Базилевский М.П.* Сравнительный анализ эффективности методов построения вполне интерпретируемых линейных регрессионных моделей // Моделирование и анализ данных. 2023. Т. 13. № 4. С. 59–83.
18. *Шунина Ю.С.* Влияние способа формирования обучающей и тестовой выборок на качество классификации // Вестник Ульяновского государственного технического университета. 2015. № 2 (70). С. 43–46.
19. *Мун Д.Е., Савченко Д.Ю.* Проблемы подготовки обучающих выборок для построения системы скоринга персонала // Современные проблемы экономического развития предприятий, отраслей, комплексов, территорий. 2020. С. 390–394.
20. *Парасич В.А., Парасич И.В., Волович Г.И., Некрасов С.Г., Парасич А.В.* Переобучение в машинном обучении: проблемы и решения // Вестник Южно-Уральского государственного университета. Серия: Компьютерные технологии, управление, радиоэлектроника. 2024. Т. 24. № 2. С. 18–27. DOI:10.14529/ctcr240202.
21. *Фёрстер Э., Рёнц Б.* Методы корреляционного и регрессионного анализа. М.: Финансы и статистика, 1983. 303 с.
22. *Charnes A., Cooper W.W., Ferguson R.O.* Optimal estimation of executive compensation by linear programming // Management science. 1955. Vol. 1. No. 2. P. 138–151. DOI: 10.1287/mnsc.1.2.138.

Optimization Problem of Constructing Linear Regressions with a Minimum Value of the Mean Absolute Error on Test Sets

Mikhail P. Bazilevskiy*
Irkutsk State Transport University (ISTU), Irkutsk, Russia,
ORCID: https://orcid.org/0000-0002-3253-5697
e-mail: mik2178@yandex.ru

This article is devoted to the problem of selecting a given number of the most informative regressors in linear regressions. When using the ordinary least squares method, the exact solution to this problem by the criterion of maximizing the coefficient of determination when using the entire data set can be obtained as a result of solving a specially formulated mixed 0–1 integer linear programming problem. However, in machine learning, an important stage in creating a reliable and efficient model is its construction based on the training set and checking the accuracy of its prediction based on the test set. Therefore, in this article formulates an optimization problem for subset selection in linear regressions based on the criterion of minimizing the mean absolute error on the test set. The formulation is based on a well-known technique, according to which absolute errors should be presented as the difference between two non-negative variables. Computational experiments were carried out using the statistical data on athletes' salaries stored into the Gretl package and the LPSolve optimization problem solver. For this purpose, the training set was formed from 70%, 75%, and 80% of observations. In all these cases, the average decrease in the value of the coefficient of determination of the models was 24.76%, 18.4%, and 12.22%, but the mean absolute error decreased by 24.8%, 26.3%, and 21.05%, respectively. Experiments showed that the average time to solve problems when minimizing the mean absolute error on test sets was 2.33–2.85 times higher than the time to solve problems when maximizing the coefficient of determination on training sets.

Keywords: machine learning, regression analysis, ordinary least squares method, subset selection in regression, coefficient of determination, mean absolute error, training set, test set, mixed 0–1 integer linear programming problem.

For citation:
Bazilevskiy M.P. Optimization Problem of Constructing Linear Regressions with a Minimum Value of the Mean Absolute Error on Test Sets. *Modelirovanie i analiz dannykh = Modelling and Data Analysis,* 2024. Vol. 14, no. 4, pp. 91–103. DOI: https://doi.org/10.17759/mda.2024140406 (In Russ., abstr. in Engl.).

Mikhail P. Bazilevskiy, PhD (Engineering), Associate Professor, Department of Mathematics, Irkutsk State Transport University (ISTU), Irkutsk, Russia, ORCID: https://orcid.org/0000-0002-3253-5697, e-mail: mik2178@yandex.ru

References

1. Rashka S. *Python i mashinnoe obuchenie* [Python and Machine Learning]. Moscow, DMK Press, 2017. 418 p.
2. Janiesch C., Zschech P., Heinrich K. Machine learning and deep learning, *Electronic Markets*, 2021, vol. 31, no. 3, pp. 685–695. DOI:10.1007/s12525-021-00475-2.
3. Mhlanga D. Artificial intelligence and machine learning for energy consumption and production in emerging markets: a review, *Energies*, 2023, vol. 16, no. 2, pp. 745. DOI:10.3390/en16020745.
4. Xu Z., Mohsin M., Ullah K., Ma X. Using econometric and machine learning models to forecast crude oil prices: Insights from economic history, *Resources Policy*, 2023, vol. 83, pp. 103614. DOI:10.1016/j.resourpol.2023.103614.
5. Haug C.J., Drazen J.M. Artificial intelligence and machine learning in clinical medicine, *New England Journal of Medicine*, 2023, vol. 388, no. 13, pp. 1201–1208. DOI:10.1056/NEJMra2302038.
6. Kumar S., Gopi T., Harikeerthana N., Gupta M.K., Gaur V., Krolczyk G.M., Wu C. Machine learning techniques in additive manufacturing: a state of the art review on design, processes and production control, *Journal of Intelligent Manufacturing*, 2023, vol. 34, no. 1, pp. 21–55. DOI: 10.1007/s10845-022-02029-5.
7. Molnar C. *Interpretable machine learning*. Lulu. com, 2020.
8. Nie B., Du Y., Du J., Rao Y., Zhang Y., Zheng X., Ye N., Jin H. A novel regression method: Partial least distance square regression methodology, *Chemometrics and Intelligent Laboratory Systems*, 2023, vol. 237, pp. 104827. DOI:10.1016/j.chemolab.2023.104827.
9. Zhuravlev Yu.I., Sen'ko O.V., Dokukin A.A., Kiseleva N.N., Saenko I.A. Dvukhurovnevyy metod regressionnogo analiza, ispol'zuyushchiy ansambli derev'ev s optimal'noy divergentsiey [Two-level regression method using ensembles of trees with optimal divergence], *Doklady Mathematics*, 2021, vol. 499, pp. 63–66. DOI:10.31857/S2686954321040172.
10. Bazilevskiy M.P. Dvukhkriterial'noe otsenivanie lineynykh regressionnykh modeley metodami naimen'shikh moduley i kvadratov [Two-criteria estimation of linear regression models using least absolute deviations and squares], *International Journal of Open Information Technologies*, 2024, vol. 12, no. 6, pp. 76–81.
11. Bazilevskiy M.P. Otbor informativnykh operatsiy pri postroenii lineyno-neelementarnykh regressionnykh modeley [Selection of informative operations in the construction of linear non-elementary regression models], *International Journal of Open Information Technologies*, 2021, vol. 9, no. 5, pp. 30–35.
12. Noskov S.I. *Tekhnologiya modelirovaniya ob"ektov s nestabil'nym funktsionirovaniem i neopredelennost'yu v dannykh* [Technology for modeling objects with unstable operation and uncertainty in data]. Irkutsk, RITs GP «Oblinformpechat'», 1996. 320 p.
13. Miller A. *Subset selection in regression*. Chapman and hall/CRC, 2002.
14. Ayvazyan S.A., Mkhitaryan V.S. *Prikladnaya statistika i osnovy ekonometriki* [Applied Statistics and Basics of Econometrics]. Moscow, YuNITI, 1998. 1005 p.
15. Bazilevskiy M.P. Svedenie zadachi otbora informativnykh regressorov pri otsenivanii lineynoy regressionnoy modeli po metodu naimen'shikh kvadratov k zadache chastichno-bulevogo lineynogo programmirovaniya [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, vol. 6, no. 1 (20), pp. 108–117.
16. Bazilevskiy M.P. Otbor optimal'nogo chisla informativnykh regressorov po skorrektirovannomu koeffitsientu determinatsii v regressionnykh modelyakh kak zadacha chastichno tselochislennogo

lineynogo programmirovaniya [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, no. 2, pp. 41–54.

17. Bazilevskiy M.P. Sravnitel'nyy analiz effektivnosti metodov postroeniya vpolne interpretiruemykh lineynykh regressionnykh modeley [Comparative analysis of the effectiveness of methods for constructing quite interpretable linear regression models], *Modelling and Data Analysis*, 2023, vol. 13, no. 4, pp. 59–83.
18. Shunina Yu.S. Vliyanie sposoba formirovaniya obuchayushchey i testovoy vyborok na kachestvo klassifikatsii, *Bulletin of the Ulyanovsk State Technical University*, 2015, no. 2 (70), pp. 43–46.
19. Mun D.E., Savchenko D.Yu. Problemy podgotovki obuchayushchikh vyborok dlya postroeniya sistemy skoringa personala [Problems of preparation of training samples for building a personnel scoring system], *Modern Problems of Economic Development of Enterprises, Industries, Complexes, Territories*, 2020, pp. 390–394.
20. Parasich V.A., Parasich I.V., Volovich G.I., Nekrasov S.G., Parasich A.V. Pereobuchenie v mashinnom obuchenii: problemy i resheniya [Overfitting in machine learning: problems and solutions], *Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics*, 2024, vol. 24, no. 2, pp. 18–27. DOI:10.14529/ctcr240202.
21. Ferster E., Rents B. *Metody korrelyatsionnogo i regressionnogo analiza* [Methods of correlation and regression analysis]. Moscow, Finance and Statistics, 1983. 303 p.
22. Charnes A., Cooper W.W., Ferguson R.O. Optimal estimation of executive compensation by linear programming, *Management science*, 1955, vol. 1, no. 2, pp. 138–151. DOI: 10.1287/mnsc.1.2.138.

Получена 23.09.2024 Received 23.09.2024
Принята в печать 15.11.2024 Accepted 15.11.2024