УДК 519.852 ISSN 1995-5499
DOI: https://doi.org/10.17308/sait.2021.1/3377
Поступила в редакцию 08.01.2021
Подписана в печать 26.04.2021

МЕТОД СМЕШАННОГО ОЦЕНИВАНИЯ ПАРАМЕТРОВ ЛИНЕЙНОЙ РЕГРЕССИИ: ОСОБЕННОСТИ ПРИМЕНЕНИЯ

© 2021 С. И. Носков✉

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

Аннотация. Работа основана на предложенном ранее автором методе смешанного оценивания неизвестных параметров линейного регрессионного уравнения. Этот метод предполагает одновременную минимизацию разных функций потерь на разных участках обрабатываемой выборки данных. Основным достоинством такого подхода является совмещение привлекательных свойств каждого задействованного метода оценивания параметров при обработке одной выборки. В статье рассматриваются способы формирования подвыборок исходной выборки для функций потерь, соответствующих городскому расстоянию и расстоянию Чебышева. Эти функции по-разному реагируют на плохо согласующиеся с выборкой в целом наблюдения — первая их, по существу, игнорирует, вторая, наоборот, к ним крайне чувствительна. Показано, что реализация метода смешанного оценивания для такой комбинированной функции потерь сводится к задаче линейного программирования. При разбиении исходной выборки на подвыборки использованы следующие свойства методов оценивания параметров линейного регрессионного уравнения: метод наименьших модулей обеспечивает равенство числа нулевых ошибок аппроксимации числу параметров; при использовании метода антиробастного оценивания число максимальных по модулю ошибок аппроксимации не меньше числа параметров плюс единица. Рассмотрен численный пример с десятью наблюдениями и тремя независимыми переменными. Сравниваются оценки параметров и значения некоторых частных критериев адкватности при использовании методов наименьших квадратов, модулей, антиробастного и смешанного оценивания. Исходная выборка разбивается при этом на две подвыборки, на одной из которых метод смешанного оценивания тяготеет к игнорированию аномальных наблюдений, а на другой, напротив, неявным образом придает им больший вес, позволяя совместить тем самым преимущества методов наименьших модулей и антиробастного оценивания при использовании на одних данных, в целом способствуя повышению адекватности в их обработке.
Ключевые слова: линейное регрессионное уравнение, выборка данных, метод смешанного оценивания, городское расстояние, расстояние Чебышева, выбросы.

ВЕДЕНИЕ

Методы регрессионного анализа являются признанным инструментом научного анализа сложных, с множеством внутренних и внешних взаимосвязей объектов различной природы. Эти методы позволяют на модельном уровне формализовать закономерности, присущие этим объектам, посредством разработки их качественных абстрактных образов, что открывает широкие возможности в повышении эффективности вырабатываемых управляющих воздействий, поскольку при этом экспериментирование может проводиться не с «живой» системой, а с ее математической моделью.

Рассмотрим линейное регрессионное уравнение

$$y_k = \sum_{i=1}^{m} \alpha_i x_{ki} + \varepsilon_k, \ k \in P = \{1, 2, ..., n\}, \quad (1)$$

✉ Носков Сергей Иванович
e-mail: sergey.noskov.57@mail.ru

где y — объясняемая, а x_i — i-я объясняющая переменные, α_i — i-й подлежащий оцениванию параметр, ε_k — ошибки аппроксимации, k — номер наблюдения, n — их количество.

Представим уравнение (1) в векторной форме:

$$y = X\alpha + \varepsilon, \qquad (2)$$

где $y = (y_1,...,y_n)^T$, $\alpha = (\alpha_1,...,\alpha_m)^T$, $\varepsilon = (\varepsilon_1,...,\varepsilon_n)^T$, X — $(n \times m)$-матрица с компонентами x_{ki}.

Отметим, что данная работа выполнена в рамках логико-алгебраического подхода к анализу данных, не предполагающего вероятностной природы ошибок ε.

Ведем в рассмотрение вектор $\hat{y}$ расчетных значений объясняемой переменной:

$$\hat{y} = X\alpha = y - \varepsilon.$$

Методам оценивания вектора параметров α уравнения (2) посвящено весьма значительное количество работ. Некоторые из них рассмотрены в кратком обзоре, данном в статье [1] и основанном на давно уже ставших классическими монографиях [2–7].

Как правило, эти методы основаны на минимизации функций потерь

$$J(\alpha) = \rho(y, \hat{y}),$$

определяемых способом задания расстояния ρ между расчетными $\hat{y}_k$ и фактическими y_k значениями объясняемой переменной на выборке (X, y), $k \in P$. Причем, и это принципиально, минимизация $J(\alpha)$ производится на **всей** выборке.

Поскольку настоящая работа основана на разработанном ранее автором [8] методе смешанного оценивания (МСО) параметров α, рассмотрим его здесь более подробно.

МЕТОД СМЕШАННОГО ОЦЕНИВАНИЯ

Пусть исходная выборка с номерами наблюдений из множества P из соображений содержательного или какого-либо иного характера разбита (разделена, кластеризована) на I подвыборок с номерами из множеств N_i, $i = 1,2,\ldots,I$, для каждой из которой исследователь использует **свою** функцию потерь $J^i(\alpha)$, основанную на свойственном ей способе задания расстояния между векторами $\hat{y}$ и y. При этом должны выполняться естественные условия:

$$P = \bigcup_{i=1}^{I} N_i, \quad N_i \cap N_j = \varnothing, i \neq j.$$

Такое разбиение неявным образом предполагает, что каждая из подвыборок обладает какими-то своими уникальными свойствами.

Тогда задача смешанного оценивания параметров линейной регрессии (2) представима в виде:

$$\alpha = \arg\min_{\alpha \in R^m} \sum_{i=1}^{I} J^i(\alpha). \qquad (3)$$

В общем случае задача (3) представляет собой весьма сложную задачу нелинейного программирования.

В [8] рассмотрен также существенно более простой случай, сводящийся к линейно-программной задаче.

Пусть исходная выборка разбита на две подвыборки с номерами из множеств N_1 и N_2, то есть $I = 2$, а в качестве функций потерь используются следующие

$$J^1(\alpha) = \sum_{k \in N_1} |\varepsilon_k|, \qquad (4)$$

$$J^2(\alpha) = \max_{k \in N_2} |\varepsilon_k|. \qquad (5)$$

Функция $J^1(\alpha)$ соответствует городскому расстоянию между векторами y и $\hat{y}$, а $J^2(\alpha)$ — расстоянию Чебышева. Методами оценивания вектора параметров α, связанными с их минимизацией, являются соответственно методы наименьших модулей (МНМ) и антиробастного оценивания (МАО) [7].

Представим вектора α и ε в виде разностей их положительных a, u и отрицательных b, v частей соответственно:

$$\alpha = a - b, \ \varepsilon = u - v.$$

При этом справедливы соотношения

$$u_k v_k = 0, |\varepsilon_k| = u_k + v_k, k \in P,$$

$$a_i b_i = 0, |\alpha_i| = a_i + b_i, i = \overline{1,m}.$$

Тогда, используя приемы, описанные, например, в [9], задача (3) может быть представлена в виде задачи линейного программирования (ЛП):

$$\sum_{i=1}^{m} (a_i - b_i) x_{ki} + u_k - v_k = y_k, k \in P \qquad (6)$$

$$u_k + v_k - r \le 0,\ k \in N_2, \quad (7)$$

$$a_i \ge 0, b_i \ge 0, i = \overline{1,m},\ u_k \ge 0, v_k \ge 0, k \in P, \quad (8)$$

$$\sum_{k \in N_1} (u_k + v_k) / s + r +$$

$$+ h_1 \sum_{i=1}^{m} (a_i + b_i) + h_2 \sum_{k \in N_2} (u_k + v_k) \to \min \quad (9)$$

Здесь s — число элементов в множестве N_1, h_1 и h_2 — малые положительные константы. Присутствие в целевой функции (9) задачи ЛП (6)–(9) последних двух слагаемых обеспечивает равенство нулю произведений положительных и отрицательных частей введенных выше новых переменных, следующее из их определений.

Основная проблема, связанная с применением МСО при обработке каждой конкретной выборки, состоит в поиске ответа на вопрос: каким образом исходную выборку разбивать на подвыборки с номерами наблюдений из множеств N_1 и N_2?

СПОСОБЫ ФОРМИРОВАНИЯ МНОЖЕСТВ N_1 И N_2

Едва ли на этот вопрос существует единственный и формально строго обоснованный ответ.

Отметим, что проблеме выявления выбросов в данных посвящена значительная литература (см., в частности, [10–14]). Представляется, однако, что в основу формирования множеств N_1 и N_2 должны быть положены некоторые бесспорные эвристические соображения, базирующиеся на фундаментальных свойствах именно МНМ и МАО. Эти методы диаметрально противоположны по реакции на выбросы — аномальные наблюдения, не согласующиеся со всей выборкой в целом. МНМ к ним не чувствителен, МАО же, напротив, сильно на них реагирует.

а) Использование МНМ в качестве основы при разбиении исходной выборки на подвыборки.

В работе [15] доказано, что если выборка не имеет особенностей, то число нулевых ошибок аппроксимации равно m. По существу это означает, что МНМ из всей выборки длины n «выбирает» m наблюдений, точно через которые и проходит гиперплоскость регрессии. При этом остальные $n-m$ наблюдений попросту игнорируются, полагаются выбросами, хотя в строгом смысле таковыми могут и не являться.

Рассчитаем вектор ошибок аппроксимации ε после применения МНМ для оценивания вектора параметров α уравнения (2). Сформируем множество номеров наблюдений N_1, включив в него те номера $k \in P$, для которых $\varepsilon_k = 0$. Множество номеров наблюдений N_2 включит в свой состав номера оставшихся наблюдений:

$$N_2 = P \setminus N_1.$$

б) Использование МАО в качестве основы при разбиении исходной выборки на подвыборки.

Известно (см., например, [16]), что при использовании МАО число максимальных по модулю ошибок аппроксимации не меньше числа $m+1$. Это означает, что МАО неявным же образом «назначает» выбросами именно столько наблюдений выборки. Из их номеров и предлагается в этом случае формировать множество N_2, остальные номера составят множество N_1:

$$N_1 = P \setminus N_2.$$

в) Использование метода наименьших квадратов (МНК) в качестве основы при разбиении исходной выборки на подвыборки.

Если исследователь не располагает программными средствами построения регрессии (2), в которых реализованы МНМ и (или) МАО, при формировании множеств N_1 и N_2 можно воспользоваться и обычным МНК, который по реакции на выбросы занимает своего рода промежуточную позицию между МНМ и МАО. Сделать это можно, например, следующим образом.

Обозначим через E среднюю относительную ошибку аппроксимации линейной регрессии (1):

$$E = \sum_{k=1}^{n} (|\varepsilon_k| / y_k) n.$$

Пусть исследователь, исходя из своих знаний и опыта, может назначить величину $\delta > 0$, представляющую собой некий порог

приемлемости этой ошибки. Тогда при выполнении условия

$$|\varepsilon_k| / y_k > E + \delta$$

номер наблюдения k включается в множество N_2 в противном случае — в множество N_1.

Подчеркнем еще раз, что все перечисленные выше способы формирования множеств N_1 и N_2 имеют эвристический характер и не приводят к их единственно возможному составу для конкретной выборки. Эта проблема, по-видимому, требует дальнейшего исследования.

ПРИМЕР

Пусть дана выборка, представленная в табл. 1.

Построим на этих данных линейное регрессионное уравнение

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \varepsilon \qquad (10)$$

с помощью МНК, МНМ, МАО и МСО.

а) МНК.

$$\alpha = (-3.06, 1.35, 0.16, -0.11),$$
$$\varepsilon = (-0.46, 0.59, 0.05, 2.06, 1.51, -2.05, 0.67, 1.00, -4.63, 1.23),$$
$$E = 0.80.$$

Таблица 1. Исходные данные
[Table 1. Source data]

y	x_1	x_2	x_3
5	6	7	7
9	8	7	5
1	3	5	8
9	7	5	3
1	2	1	3
2	5	4	3
8	7	6	1
6	5	9	2
1	6	6	4
8	7	8	9

б) МНМ.

$$\alpha = (-3.29, 1.37, 0.30, -0.17),$$
$$\varepsilon = (-0.91, 0.75, 0.00, 0.00, 1.65, 1.75, -2.29, 0.00, 0.00, -5.11),$$
$$E = 0.84.$$

в) МАО.

$$\alpha = (-2.18, 1.73, -0.54, 0.09),$$
$$\varepsilon = (1.38, 3.12, 1.85, 2.35, 3.12, 3.12, -0.68, 1.77, 3.12, -3.12),$$
$$E = 1.09.$$

г) МСО.

За основу формирования множеств N_1 и N_2 примем МНМ:

$$N_1 = \{3, 4, 8, 9\}, \quad N_2 = \{1, 2, 5, 6, 7, 10\}.$$

Далее оценим вектор параметров уравнения (10) с помощью МСО:

$$\alpha = (-2.64, 1.06, 0.42, -0.20),$$
$$\varepsilon = (-0.21, 1.71, 1.26, 0.00, 2.75, 1.71, -1.71, 0.93, 0.00, -4.40),$$
$$E = 0.78.$$

Как следует из полученных результатов, средняя ошибка аппроксимации при использовании МСО несколько ниже, чем для МНК, МНМ и МАО, хотя для других данных этот вывод может быть и другим. Кроме того, на наблюдениях с номерами из множества $N_1 = \{3, 4, 8, 9\}$ МСО проявляет себя как МНМ, а на наблюдениях с номерами из множества $N_2 = \{1, 2, 5, 6, 7, 10\}$ — как МАО. При этом лишь на наблюдениях 3 и 9 ошибка аппроксимации равна нулю, а максимальная по модулю ошибка единственна — для наблюдения 10.

Безусловно, тот факт, что для данного примера МСО по критерию E оказался лучше, чем МНМ, МАО и МНК, отнюдь не является гарантией того, что для других данных это тоже будет именно так. Дело отнюдь не в этом. Основное достоинство МСО по отношению к данным из табл. 1 проявилось в том, что на подвыборке N_1 он тяготеет к игнорированию наблюдений, которые «полагает» выбросами, а на подвыборке N_2, наоборот, «стремится» неявным образом придать им больший вес, позволяя совместить тем самым преимущества МНМ и МАО при их одновременном использовании на одних данных, в целом способствуя повышению адекватности в их обработке.

ЗАКЛЮЧЕНИЕ

В работе предложены три возможных способа разбиения выборки данных на две подвыборки при применении метода смешанного оценивания параметров линейного регрессионного уравнения. Эти способы имеют эвристический характер и основаны на свойствах методов наименьших модулей, квадратов и антиробастного оценивания, касающихся их реакции на аномальные наблюдения.

КОНФЛИКТ ИНТЕРЕСОВ

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

СПИСОК ЛИТЕРАТУРЫ

1. *Носков С. И.* Компромиссные паретовские оценки параметров линейной регрессии // Математическое моделирование. – 2020. – Т. 32, № 11. – С.70–78.

2. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ. Множественная регрессия. – 3-е изд. – М. : Диалектика, 2007. – 912 с.

3. *Фёрстер Э., Рёнц Б.* Методы корреляционного и регрессионного анализа. – М. : Финансы и статистика, 1981. – 302 с.

4. *Айвазян С. А.* Прикладная статистика и основы эконометрики. – М. : Юнити, 2001. – 432 с.

5. *Винн Р., Холден И.* Введение в прикладной эконометрический анализ. – М. : Финансы и статистика, 1981. – 294 с.

6. *Демиденко Е. З.* Оптимизация и регрессия. – М. : Наука, 1989. – 296 с.

7. *Демиденко Е. З.* Линейная и нелинейная регрессии. – М. : Финансы и статистика, 1981. – 302 с.

8. *Носков С. И.* О методе смешанного оценивания параметров линейной регрессии // Информационные технологии и математическое моделирование в управлении сложными системами. – 2019. – № 1. – С. 41–45.

9. *Носков С. И.* Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. – Иркутск : Облинформпечать, 1996. – 320 с. https://www.researchgate.net/publication/340570185_Tehnologia_modelirovania_obektov_s_nestabilnym_funkcionirovaniem_i_neopredelennostu_v_dannyh

10. *Кузовлев В. И., Орлов А. О.* Выявление аномалий при прогнозном анализе данных // Вестник московского государственного технического университета им. Н. Э. Баумана. Серия приборостроение. – 2016. – № 5. – С. 75–85.

11. *Орлов А. О.* Проблема поиска расстояний между значениями категориальных атрибутов при обнаружении выбросов в данных // В мире научных открытий. – 2012. – № 8-1. – С. 142–155.

12. *Кузовлев В. И., Орлов А. О.* Методика выбора параметров и интерпретации результатов анализа выбросов в данных систем поддержки принятия решений // Инженерный журнал: наука и инновации. – 2013. – № 11. – С. 13.

13. *Шестерняк Л. В.* Методы обработки результатов вычислительного эксперимента // Устойчивое развитие науки и образования. – 2019. – № 11. – С. 282–285.

14. *Лютикова Л. А.* Построение логического алгоритма выявления выбросов в зашумленных данных // Моделирование, оптимизация и информационные технологии. – 2018. – Т. 6, № 4. – С. 132–142.

15. *Лакеев А. В., Носков С. И.* Метод наименьших модулей для линейной регрессии: число нулевых ошибок аппроксимации // Современные технологии. Системный анализ. Моделирование. – 2012. – № 2. – С. 48–50.

16 *Носков, С. И.* Метод антиробастного оценивания параметров линейной регрессии: число максимальных по модулю ошибок аппроксимации // Южно-Сибирский научный вестник. – 2020. – № 1 (29). – С. 51–54.

Носков Сергей Иванович — д-р техн. наук, проф., профессор кафедры «Информационные системы и защита информации» Иркутского государственного университета путей сообщения.
E-mail: sergey.noskov.57@mail.ru
ORCID iD: https://orcid.org/0000-0003-4097-2720

DOI: https://doi.org/10.17308/sait.2021.1/3377 ISSN 1995-5499
Received 08.01.2021
Accepted 26.04.2021

A METHOD FOR THE MIXED ESTIMATION OF LINEAR REGRESSION PARAMETERS: APPLICATION SPECIFICS

© 2021 S. I. Noskov✉

Irkutsk State University of Railways
15, Chernyshevsky Street, 664074 Irkutsk, Russian Federation

Annotation. The presented study is based on the method of the mixed estimation of unknown parameters of linear regression equations proposed earlier by the author. This method assumes the simultaneous minimisation of different loss functions in different parts of the processed data sample. The main advantage of this approach is that it combines the strengths of each parameter estimation method used when processing a single data sample. The article discusses the ways to form subsamples of the initial sample for the loss functions corresponding to the Manhattan and Chebyshev distances. These functions react differently to observations that are inconsistent with the sample – the former essentially ignores them, while the latter, on the contrary, is extremely sensitive to them. The article demonstrates that the implementation of the mixed estimation method for such a combined loss function is reduced to a linear programming problem. When dividing the initial sample into subsamples, we used the following advantages of the methods for estimating the parameters of linear regression equations: the least absolute deviation method ensures that the number of zero approximation errors equals the number of parameters; the anti-robust estimation method ensures that the number of maximum approximation errors in the module is no fewer than the number of parameters plus one. In the article, we consider a numerical example with ten observations and three independent variables. We compared the estimates of the parameters and values of certain adequacy criteria obtained when using the methods of least squares and modules, the anti-robust estimation method, and the mixed estimation method. In this case, the initial sample is divided into two subsamples. For one subsample, the method of mixed estimation tends to ignore outlying observations, and for the other, on the contrary, implicitly gives them more weight. It thus combines the advantages of the methods of least modules and anti-robust estimation when applied to the same data, generally enhancing the adequacy of the data processing.
Keywords: linear regression equation, data sampling, mixed estimation method, Manhattan distance, Chebyshev distance, outliers.

CONFLICT OF INTEREST

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

REFERENCES

1. *Noskov S. I.* (2020) Compromise Pareto's evaluation of parameters linear regression // Mathematical Models and Computer Simulations. No 11. P.70–78.

2. *Draper N. and Smith G.* (2007) Applied regression analysis. Multiple Regression. – 3rd ed. Moscow : Dialectics. 912 p.

3. *Förster E. and Rönz B.* (1981) Methods of correlation and regression analysis. Moscow : Finance and Statistics. 302 p.

4. *Ayvazyan S. A.* (2001) Applied statistics and foundations of econometrics. Moscow : Unity. 432p.

5. *Wynn R. and Holden I.* (1981) Introduction to Applied Econometric Analysis. Moscow : Finance and statistics. 294 p.

6. *Demidenko E. Z.* (1989) Optimization and regression. Moscow : Nauka. 296 p.

✉ Noskov Sergey I.
e-mail: sergey.noskov.57@mail.ru

7. *Demidenko E. Z.* (1981) Linear and nonlinear regression. Moscow : Finance and statistics. 302 p.

8. *Noskov S. I.* (2019) On the method of mixed estimation of linear regression parameters // Information technologies and mathematical modeling in the management of complex systems. No 1. P. 41–45.

9. *Noskov S. I.* (1996) A technology for modeling objects with unstable functioning and uncertainty in data. Irkutsk : Oblinformpechat. 320 p.

10. *Kuzovlev V. I. and Orlov A. O.* (2016) Anomaly detection in predictive data analysis // Bulletin of the Moscow State Technical University. N. E. Bauman. Instrumentation series. No 5. P. 75–85.

11. *Orlov A. O.* (2012) The problem of finding the distances between the values of categorical attributes when detecting outliers in the data. No 8-1. P. 142–155.

12. *Kuzovlev V. I. and Orlov A. O.* (2013) Methodology for the choice of parameters and interpretation of the results of the analysis of emissions in the data of decision support systems // Engineering journal: science and innovations. No 11. P. 13.

13. *Shesternyak L.V.* (2019) Methods for processing the results of a computational experiment // Sustainable Development of Science and Education. No 11. P. 282–285.

14. *Lyutikova L. A.* (2018) Construction of a logical algorithm for detecting outliers in noisy data // Modeling, optimization and information technologies. No 4. P. 132–142.

15. *Lakeev A. V. and Noskov S. I.* (2012) Least modulus method for linear regression: the number of zero approximation error // Modern technologies. System analysis. Modeling. No 2. P. 48–50.

16. *Noskov S. I.* (2020) Method of anti-robust estimation of linear regression parameters: the number of approximation errors maximal in modulus // Yuzhno-Siberian Scientific Bulletin. No 1. P. 51–54.

Noskov Sergey I. — DSc in Technical Sciences, Professor, Department of Information Systems and Information Security, Irkutsk State Transport University.
E-mail: sergey.noskov.57@mail.ru
ORCID iD: https://orcid.org/0000-0003-4097-2720