

Совместное применение методов максимальной согласованности и антиробастного оценивания параметров при построении регрессионных моделей

С.И. Носков, Ю.А. Бычков

Иркутский государственный университет путей сообщения,
664074, г. Иркутск, ул. Чернышевского, 15, Россия

Резюме. Цель. Целью исследования является решение задачи вычисления параметров линейной регрессионной модели на основе совместного применения методов антиробастного оценивания и максимальной согласованности между реальными и вычисленными значениями зависимой переменной в непрерывной форме. **Метод.** Расчет неизвестных параметров модели производится с помощью сведения исходной задачи к задаче линейного программирования. Ее решение не должно вызывать вычислительных трудностей вследствие значительного количества разработанных эффективных программных средств. **Результат.** Сформированная задача линейного программирования обладает допустимой для реальных ситуаций размерностью. **Вывод.** Результаты решения численного примера указывают на эффективность предложенного в работе способа вычисления параметров линейной регрессионной модели на основе совместного применения методов антиробастного оценивания и максимальной согласованности. Окончательный выбор значений параметров остается за разработчиком модели.

Ключевые слова: линейная регрессия, идентификация параметров, метод наименьших модулей, метод антиробастного оценивания, метод максимальной согласованности, задача линейного программирования, альтернативность

Для цитирования: С.И. Носков, Ю.А. Бычков. Совместное применение методов максимальной согласованности и антиробастного оценивания параметров при построении регрессионных моделей. Вестник Дагестанского государственного технического университета. Технические науки. 2024;51(3):117-122. DOI:10.21822/2073-6185-2024-51-3-117-122

Combined application of methods of maximum consistency and anti-robust parameter estimation in the construction of regression models

S.I. Noskov, Yu.A. Bychkov

Irkutsk State Transport University,
15 Chernyshevskogo Str., Irkutsk 664074, Russia

Abstract. Objective. The purpose of the study is to solve the problem of calculating the parameters of a linear regression model based on the joint application of anti-robust estimation methods and maximum consistency between the real and calculated values of the dependent variable in continuous form. **Method.** The unknown parameters of the model are calculated by reducing the original problem to a linear programming problem. Its solution should not cause computational difficulties due to the significant number of developed effective software tools. **Result.** The generated linear programming problem has a dimension acceptable for real situations. **Conclusion.** The results of solving a numerical example indicate the effectiveness of the method proposed in the work for calculating the parameters of a linear regression model based on the joint application of anti-robust estimation methods and maximum consistency. The final choice of parameter values remains with the model developer.

Keywords: linear regression, parameter identification, least modulus methods, antirobust estimation, maximum consistency, linear programming problem, alternativeness

For citation: S.I. Noskov, Yu.A. Bychkov. Combined application of methods of maximum consistency and anti-robust parameter estimation in the construction of regression models.

Herald of Daghestan State Technical University. Technical Sciences. 2024;51(3):117-122.
DOI:10.21822/2073-6185-2024-51-3-117-122

Введение. При построении математических моделей регрессионного типа для сложных систем любой природы исследователи часто совмещают различные методы. Так, в работе [1] констатируется, что проблема больших размерностей является одной из важнейших при разработке модели линейной регрессии.

В этой связи, например, при построении модели уровня загрязнения реки Тигр одновременно применялись методы плавного интегрирования подсчета абсолютного отклонения, а также метод FGLasso. В [2] при разработке модели рыночной стоимости акций крупной компании Baghdad Soft Drinks Company был использован метод штрафных наименьших квадратов для оценки параметров модели путем задействования различных штрафных функций, включая штрафы LASSO, SCAD, минимаксный вогнутый штраф в дополнение к штрафной функции Elastic-net. Статья [3] посвящена, в частности, построению модели параметрической регрессии для объема, диаметра и высоты деревьев, основанной на двухпараметрическом расширении экспоненциального гамма-распределения. При этом кратко описываются различные процедуры оценки, а именно метод максимального правдоподобия, оценка момента, оценка максимального произведения расстояний, обычная и взвешенная оценка наименьших квадратов и оценка Крамера-фон-Мизеса, а также проводится исследование результатов моделирования для сравнения производительности методов.

В [4] исследуются модели частичной регрессии, построенные с применением методов робастной оценки с вейвлет-порогом, Спикмана и методом сглаживания Надарьи-Уотсона с использованием прогнозных экспериментов с различными размерами выборки и коэффициентами засорения данных. В исследовании [5] отмечается, что модель обратной гауссовой регрессии применяется, когда переменная отклика является непрерывной, положительно асимметричной и хорошо соответствует обратному распределению Гаусса. В условиях мультиколлинеарности оценка максимального правдоподобия не является правильным выбором. Поэтому предлагается новая гребневая оценка главного компонента для такой регрессии, которая объединяет в себе методы главных компонент и обычную гребневую оценку. В работе [6] изучается географически взвешенная регрессия — это развитие теории наименьших квадратов по отношению к модели взвешенной регрессии, которая учитывает пространственные эффекты, в результате чего получается оценка параметров модели, которую можно использовать для прогнозирования местоположения объектов. В статье [7] предлагается построенная с помощью различных методов модель множественной регрессии для одновременного определения показателей преломления и температуры при использовании интерферометрического оптического датчика.

В [8] демонстрируется совместное применение методов наименьших модулей (МНМ) и нелинейных наименьших квадратов для построения нелинейных регрессий. Работа [9] посвящена прогнозированию индекса качества воздуха в городах района Цзинь-Цзинь-Цзи в Китае на основе линейной регрессионной модели панельных данных. При этом используются метод наименьших квадратов и его двухэтапный вариант, оценки квазикасимального правдоподобия и обобщенного момента, робастная оценка с фиксированным эффектом. В [10] представлена новая структура М-квантильной регрессии, в которой все коэффициенты описываются гибкими параметрическими функциями. На основе модели изучаются условия жизни в странах Европейского Союза и в трех регионах Италии.

Постановка задачи. Рассмотрим линейное регрессионное уравнение (модель) вида:

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

где: y — зависимая переменная,

а x_i — i -ая независимая переменные,

α_i — i -ый определяемый параметр,

ε_k – ошибки аппроксимации,
 k – номер наблюдения,
 n – число наблюдений (длина выборки).

Заметим, что настоящая работы выполнена в рамках логико-алгебраического (аппроксимационного) подхода к анализу данных (см., например, [11]), поэтому все переменные в (1) будем считать детерминированными величинами.

Линейную модель (1) можно представить в виде:

$$y_k = \hat{y}_k + \varepsilon_k, \quad k = \overline{1, n},$$

где y_k и $\hat{y}_k$ – соответственно фактические (наблюдаемые) и расчетные (вычисленные по модели) значения зависимой переменной y .

Наиболее распространенными методами оценивания параметров модели (1) являются методы наименьших квадратов и МНМ (см., например, [12]). Существенно менее часто применяются методы антиробастного оценивания (МАО) [12] и максимальной согласованности (ММС) [13] в поведении фактических и расчетных значений зависимой переменной на предыстории исследуемого процесса.

В настоящей работе рассмотрен алгоритмический способ совместного применения МАО и ММС при оценивании линейной регрессии (1).

Методы исследования. Отличительной особенностью МАО является его выраженный антиробастный характер, что проявляется в сильной реакции на выбросы – наблюдения, плохо согласующиеся с выборкой в целом. Реализация МАО приводит к необходимости решения следующей задачи линейного программирования (ЛП):

$$\sum_{i=1}^m \alpha_i x_{ki} + u_k - v_k = y_k, \quad k = \overline{1, n}, \quad (2)$$

$$u_k + v_k - r \leq 0, \quad k = \overline{1, n}, \quad (3)$$

$$u_k \geq 0, v_k \geq 0, \quad k = \overline{1, n}, \quad (4)$$

$$r + \rho \sum_{k=1}^n (u_k + v_k) \rightarrow \min, \quad (5)$$

где ρ – наперед заданное малое положительное число.

ММС направлен на максимизацию согласованности между заданными в выборке данных $y_k, k = \overline{1, n}$ и вычисленными по модели $\hat{y}_k, k = \overline{1, n}$ значениями зависимой переменной. Формально непрерывная форма ММС состоит в решении задачи оптимизации:

$$L = \sum_{k=1}^{n-1} \sum_{s=k+1}^n l_{ks} \rightarrow \min, \quad (6)$$

где,

$$l_{ks} = \begin{cases} |\hat{y}_k - \hat{y}_s|, & (y_k - y_s)(\hat{y}_k - \hat{y}_s) < 0 \\ 0, & \text{в противном случае.} \end{cases}$$

Рассчитаем числа $\omega_{ks}, k = \overline{1, n-1}, s = \overline{k+1, n}$ следующим образом:

$$\omega_{ks} = \begin{cases} 1, & y_k - y_s > 0 \\ -1, & y_k - y_s < 0 \\ 0, & y_k - y_s = 0. \end{cases}$$

Тогда, как показано в [14], задача (6) сводится к задаче ЛП с ограничениями (2), (4),

$$\omega_{ks} \sum_{i=1}^m \alpha_i (x_{ki} - x_{si}) + l_{ks} \geq 0, \quad k = \overline{1, n-1}, \quad s = \overline{k+1, n}, \quad (7)$$

$$l_{ks} \geq 0, \quad k = \overline{1, n-1}, \quad s = \overline{k+1, n} \quad (8)$$

и целевой функцией

$$d \sum_{k=1}^n (u_k + v_k) + (1-d) \sum_{k=1}^{n-1} \sum_{s=k+1}^n l_{ks} \rightarrow \min. \quad (9)$$

Здесь $d \in (0, 1]$ – заранее выбранное положительное число, задающее уровень компромисса в целевой функции (9) между функцией потерь $M = \sum_{k=1}^n |\varepsilon_k| = \sum_{k=1}^n (u_k + v_k)$, соответствующей МНМ и непрерывным критерием согласованности поведения (6).

Легко видеть, что при $d = 1$ задача ЛП (2), (4), (7) – (9) реализует МНМ, а при d , близком к 0, она позволяет определить оценки параметров модели (1), обеспечивающие

максимальную согласованность поведения расчетных и фактических значений выходной переменной, как это рекомендовано в работе [14].

Обсуждение результатов. Для совместного применения МСО и непрерывной формой ММС необходимо решить задачу ЛП с ограничениями (2) - (4), (7), (8) и целевой функцией

$$sr + \delta \sum_{k=1}^n (u_k + v_k) + (1 - s) \sum_{k=1}^{n-1} \sum_{s=k+1}^n l_{ks} \rightarrow \min, \quad (10)$$

где δ – наперед заданное положительное число, $s \in (0,1)$ – число, задающее уровень компромисса между МАО и ММС.

Оценивание параметров модели (1) путем решения задачи ЛП (2) - (4), (7), (8), (10) позволяет объединить достоинства МАО и непрерывной формы ММС.

Рассмотрим простой численный пример. Пусть задана выборка данных вида:

$$X = \begin{pmatrix} 4 & 6 \\ 8 & 4 \\ 3 & 7 \\ 9 & 2 \\ 7 & 8 \\ 5 & 2 \end{pmatrix}, \quad y = \begin{pmatrix} 3 \\ 6 \\ 9 \\ 8 \\ 1 \\ 7 \end{pmatrix}.$$

Построим на ее основе линейную регрессию

$$y_k = \alpha_1 x_{k1} + \alpha_2 x_{k2} + \varepsilon_k, \quad k = \overline{1,6}, \quad (11)$$

используя МНМ, МАО, ММС и предложенный выше способ, воспользовавшись специализированными программами [15, 16].

МНМ.

$$y_k = 0.913x_{k1} - 0.109x_{k2} + \varepsilon_k, \quad k = \overline{1,6},$$

$$M=15.07, r=7.02.$$

МАО.

$$y_k = 0.25x_{k1} + 0.5x_{k2} + \varepsilon_k, \quad k = \overline{1,6},$$

$$M=22, r=4.75.$$

Непрерывный ММС ($d=0.5$).

$$y_k = 0.469x_{k1} + 0.19x_{k2} + \varepsilon_k, \quad k = \overline{1,6},$$

$$M=19.25, r=6.28, L=12.375.$$

Совместное применение МАО и непрерывный ММС ($s=0.9, \delta=0.2$).

$$y_k = 0.29x_{k1} + 0.48x_{k2} + \varepsilon_k, \quad k = \overline{1,6},$$

$$M=21.52, r=4.8, L=15.33.$$

Анализ результатов моделирования указывает на следующее:

1. В трех из четырех случаях (кроме МНМ) коэффициенты при независимых переменных положительны.
2. Эти коэффициенты, как значения критериев адекватности, близки для МАО и рассмотренного выше способа, что можно объяснить близостью значения s к единице.
3. В любом случае, окончательный выбор значений параметров остается за разработчиком модели, который, руководствуясь своим опытом и знаниями, должен основывать свое решение на тех задачах, которые он намерен решить с ее помощью.

Вывод. В ходе проведенного исследования получены следующие результаты:

1. Поставленная задача сведена к задаче линейного программирования.
2. Сформированная задача линейного программирования допускает альтернативные решения путем задания разных значений константам, определяющим уровень значимости задействованных методов.
3. Представленный подход может быть успешно использован при комбинировании других методов моделирования при построении регрессионных конструкций.
4. Решен численный пример.

Библиографический список:

1. Mohammed Jasim Farhan, Ahmed Mahdi Salih. Using Feasible Graphical Lasso Regression Method to Estimate the Parameters of General Linear Regression Model Under High Dimensional Data with Application // *Al Kut Journal of Economics and Administrative Sciences* 2023, Volume 15, Issue 48, P. 321-331.
2. Zahraa Karim Fahd, Basem Shaliba Muslim. Estimating the parameters of the kink regression model in the presence of a high-dimensional problem with a practical application // *Al Kut Journal of Economics and Administrative Sciences* 2023, Volume 15, Issue 48, P. 533-560.
3. Kumar D., Sharma V.K. An Extension of Exponentiated Gamma Distribution: A New Regression Model with Application // *Lobachevskii Journal of Mathematics* 2022, Volume 43, Issue 9, P. 2525–2543,
4. Ekhllass Abdulameer Al-Azzawi, Lekaa Ali Al-Always. Robust Estimation OF The Partial Regression Model Using Wavelet Thresholding // *Journal of Economics and Administrative Sciences*, 2022 Vol. 28, No. 133, p. 97–113.
5. Akram M.N., Amin M., Lukman A.F., Afzal S. Principal component ridge type estimator for the inverse Gaussian regression model // *Journal of Statistical Computation and Simulation*, 2022, Vol. 92, Is. 10, P. 2060-2089.
6. Hasibuan D. O., Pau Teku H., Drostela Putri M. F., Setyawan Y. Dwi Bekt R. Application of Geographically Weighted Regression Method on the Human Development Index of Central Java Province // *Enthusiastic: International Journal of Applied Statistics and Data Science*, 2023, Volume 3, Issue 2, P. 189–201.
7. Guzman-Chavez A.D., Vargas-Rodrigue E. Application of a Multiple Regression Model for the Simultaneous Measurement of Refractive Index and Temperature Based on an Interferometric Optical System // *Journal of Sensors*, 2023, Volume 2023, Special Issue, Article ID 2820062, 14 pages.
8. Türkşen Ö. A novel perspective for parameter estimation of seemingly unrelated nonlinear regression // *Journal of Applied Statistics*, 2021, Volume 48, Issue 13-15: Special Issue: Recent Statistical Methods for Data Analysis, Applied Economics, Business & Finance, P. 2326-2347.
9. Ji A., Wei B., Xu L. Robust estimation of panel data regression models and applications // *Communications in Statistics - Theory and Methods*, 2023, Vol. 52, No. 21, P. 7647-7659.
10. Frumento P., Salvati N. Parametric Modelling of M-Quantile Regression Coefficient Functions with Application to Small Area Estimation. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 2020, Volume 183, Issue 1, Pages 229–250.
11. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Основы моделирования и первичная обработка данных. –М.: Финансы и статистика, 1983. –472 с.
12. Носков С.И., Попов Е.С., Середкин С.П., Тирских В.В., Торопов В.Д. Вариантное регрессионное моделирование производства электроэнергии в Российской Федерации // *Вестник Дагестанского государственного технического университета. Технические науки.* - 2023. - Т 50. - № 1. - С. 123-129.
13. Носков С.И. Метод максимальной согласованности в регрессионном анализе // *Известия Тульского государственного университета. Технические науки.* 2021. № 10. С. 380-385.
14. Носков С.И. Применение непрерывного критерия согласованности поведения при построении регрессионных моделей // *Известия ТулГУ. Технические науки.* 2021 № 6. С. 74-78.
15. Свидетельство о государственной регистрации программы для ЭВМ № 2023660087 Российская Федерация. Программа реализации ассоциирования непрерывной формы метода максимальной согласованности с антиробастным оцениванием параметров при построении линейной регрессионной модели: № 2023618290: заявл. 25.04.2023; опубл. 17.05.2023 / С. И. Носков, Ю. А. Бычков; заявитель Федеральное государственное бюджетное образовательное учреждение высшего образования «Иркутский государственный университет путей сообщения».
16. Свидетельство о государственной регистрации программы для ЭВМ № 2022681373 Российская Федерация. Программный комплекс регрессионного моделирования методом смешанного оценивания параметров с тремя альтернативными вариантами разбиения обрабатываемой выборки данных на подвыборки: № 2022680260: заявл. 28.10.2022; опубл. 14.11.2022 / С. И. Носков, К. С. Перфильева; заявитель Федеральное государственное бюджетное образовательное учреждение высшего образования «Иркутский государственный университет путей сообщения».

References:

1. Mohammed Jasim Farhan, Ahmed Mahdi Salih. Using Feasible Graphical Lasso Regression Method to Estimate the Parameters of General Linear Regression Model Under High Dimensional Data with Application. *Al Kut Journal of Economics and Administrative Sciences*. 2023;15(48):321-331.
2. Zahraa Karim Fahd, Basem Shaliba Muslim. Estimating the parameters of the kink regression model in the presence of a high-dimensional problem with a practical application. *Al Kut Journal of Economics and Administrative Sciences*. 2023;15(48):533-560.
3. Kumar D., Sharma V.K. An Extension of Exponentiated Gamma Distribution: A New Regression Model with Application. *Lobachevskii Journal of Mathematics*. 2022; 43(9):2525–2543.

4. Ekhlass Abdulameer Al-Azzawi, Lekaa Ali Al-Always. Robust Estimation OF The Partial Regression Model Using Wavelet Thresholding. *Journal of Economics and Administrative Sciences*, 2022;28(133): 97–113.
5. Akram M.N., Amin M., Lukman A.F., Afzal S. Principal component ridge type estimator for the inverse Gaussian regression mode. *Journal of Statistical Computation and Simulation* 2022; 92(10):2060-2089.
6. Hasibuan D. O., Pau Teku H., Drostela Putri M. F., Setyawan Y. Dwi Bekti R. Application of Geographically Weighted Regression Method on the Human Development Index of Central Java Province. *Enthusiastic: International Journal of Applied Statistics and Data Science*. 2023; 3(2):189–201.
7. Guzman-Chavez A.D., Vargas-Rodrigue E. Application of a Multiple Regression Model for the Simultaneous Measurement of Refractive Index and Temperature Based on an Interferometric Optical System. *Journal of Sensors*. 2023;2023, Special Issue: 14, Article ID 2820062.
8. Türksen Ö. A novel perspective for parameter estimation of seemingly unrelated nonlinear regression. *Journal of Applied Statistics*. 2021;48(13-15): 2326-2347. Special Issue: Recent Statistical Methods for Data Analysis, Applied Economics, Business & Finance.
9. Ji A., Wei B., Xu L. Robust estimation of panel data regression models and applications. *Communications in Statistics - Theory and Methods*. 2023; 52(21):7647-7659.
10. Frumento P., Salvati N. Parametric Modeling of M-Quantile Regression Coefficient Functions with Application to Small Area Estimation. *Journal of the Royal Statistical Society Series A: Statistics in Society*. 2020;183(1):229–250.
11. Ayvazyan S.A., Enyukov I.S., Meshalky L.D. Applied statistics. Basics of modeling and primary data processing. *M.: Finance and Statistics*. 1983;472. (In Russ)
12. Noskov S.I., Popov E.S., Seredkin S.P., Tirskikh V.V., Toropov V.D. Variant regression modeling of electricity production in the Russian Federation. *Herald of Daghestan State Technical University. Technical sciences*. 2023; 50(1):123-129. (In Russ)
13. Noskov S.I. Method of maximum consistency in regression analysis. *News of Tula State University. Technical science*. 2021. No. 10. P. 380-385. (In Russ)
14. Noskov S.I. Application of a continuous criterion for consistency of behavior in the construction of regression models. *Izvestia of Tula State University. Technical science*. 2021; 6:74-78. (In Russ)
15. Certificate of state registration of a computer program No. 2023660087 Russian Federation. Program for implementing the association of the continuous form of the maximum consistency method with antirobust parameter estimation when constructing a linear regression model: No. 2023618290: application. 04/25/2023: publ. 05/17/2023 / S. I. Noskov, Yu. A. Bychkov; applicant Federal State Budgetary Educational Institution of Higher Education "Irkutsk State Transport University".(In Russ)
16. Certificate of state registration of a computer program No. 2022681373 Russian Federation. Software package for regression modeling using the method of mixed parameter estimation with three alternative options for dividing the processed data sample into subsamples: No. 2022680260: application. 10/28/2022: publ. 11/14/2022 / S. I. Noskov, K. S. Perflyeva; applicant Federal State Budgetary Educational Institution of Higher Education "Irkutsk State Transport University".(In Russ)

Сведения об авторах:

Носков Сергей Иванович, доктор технических наук, профессор, профессор кафедры информационных технологий и защиты информации; sergey.noskov.57@mail.ru

Бычков Юрий Александрович, аспирант кафедры информационных технологий и защиты информации; nik24-11@mail.ru

Information about authors:

Sergey I. Noskov, Dr. Sci.(Eng.), Prof., Prof., Department of Information Technologies and Information Security; sergey.noskov.57@mail.ru

Yuri A. Bychkov, Graduate Student, Department of Information Technologies and Information Security; nik24-11@mail.ru

Конфликт интересов/Conflict of interest.

Авторы заявляют об отсутствии конфликта интересов/The authors declare no conflict of interest.

Поступила в редакцию/ Received 19.03.2024.

Одобрена после рецензирования/ Reviced 30.04.2024.

Принята в печать/ Accepted for publication 30.04.2024.