УДК 518.852+518.853.14

С.И. Носков[1]

[1] *Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация*

О МЕТОДЕ СМЕШАННОГО ОЦЕНИВАНИЯ ПАРАМЕТРОВ ЛИНЕЙНОЙ РЕГРЕССИИ

Аннотация. *В статье рассматривается метод смешанного оценивания параметров (МСО) линейного регрессионного уравнения, состоящий в использовании разных функций потерь на разных участках выборки. Подробно он описан для функций потерь, соответствующих методам наименьших модулей и антиробастного оценивания. Эти методы ведут себя по разному по отношению к выбросам – наблюдениям, не согласующимся с выборкой в целом. Первый выбросы игнорирует, второй к ним тяготеет. Рассмотрен численный пример.*

Ключевые слова. *Линейная регрессия, метод смешанного оценивания, метод наименьших модулей, антиробастное оценивание.*

S.I. Noskov[1]

[1] *Irkutsk state University of railway engineering, Russian Federation.*

ABOUT THE METHOD OF MIXED ESTIMATION OF PARAMETERS OF LINEAR REGRESSION

Annotation. *The article discusses a method mixed estimating of parameters (MME) of a linear regression equation, which consists in using different loss functions in different parts of the sample. It is described in detail for the loss functions corresponding to the methods of smallest modules and anti-robust estimation. These methods behave differently with respect to emissions - observations that are not consistent with the sample as a whole. The first emissions ignores, the second to them is lured. A numerical example is considered.*

Keywords. *Linear regression, the method of mixed estimation, the method of least modules, anti-robust evaluation.*

Рассмотрим линейное регрессионное уравнение

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

где y − эндогенная (объясняемая, зависимая), а x_i − i-ая экзогенная (объясняющая, независимая) переменные; α_i − i-ый подлежащий оцениванию параметр; ε − ошибки аппроксимации, k − номер наблюдения, n − их число.

Представим уравнение (1) в матричной форме:

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

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

Методам оценивания параметров уравнения (2) посвящена обширная литература (см., например, [1-10]).

Широкий класс методов оценивания параметров уравнения (2) связан с поиском так называемых L_ν −оценок посредством минимизации функций потерь вида[2, 7-9]:

$$J_\nu^P(\alpha) = \sum_{k \in P} |\varepsilon_k|^\nu, P = \{1, 2....n\}.$$

Каждая из этих оценок характеризуется реакцией на так называемые выбросы, то есть наблюдения, плохо согласующиеся (либо не согласующиеся вообще) со всей выборкой в целом. При этом чем больше значение ν, тем сильнее L_ν – оценка реагирует на выбросы. В регрессионном анализе методы оценивания, слабо реагирующие на выбросы, или вообще их игнорирующие, принято называть устойчивыми, или робастными.

Методом оценивания параметров уравнения (2), соответствующим $\nu = 2$, является всем хорошо известный и наиболее часто используемый в регрессионном анализе метод наименьших квадратов (МНК). При $\nu = 1$ − это метод наименьших модулей (МНМ), соответствующий городскому, или манхэттэнскому, расстоянию, при $\nu \to \infty$ − метод антиробастного оценивания (МАО), соответствующий расстоянию Чебышева. Отметим, что имеет место известное соотношение:

$$J_\infty = \max_{k \in P} |\varepsilon_k|.$$

В [2] впервые выдвинута идея поиска вектора параметров линейной регрессии (2) посредством минимизации суммы разных функций потерь на разных участках выборки. Назовем этот метод методом смешанного оценивания (МСО). Рассмотрим формальную постановку такой задачи.

Пусть исходная выборка с номерами наблюдений из множества P из соображений содержательного (или какого-либо другого) характера разбита на I подвыборок с номерами

из множеств N_i, i=1,2,…,I, для каждой из которой исследователь использует свою функцию потерь $J_{\nu(N_i)}^{N_i}(\alpha)$ с разными значениями $\nu(N_i)$. При этом должны выполняться естественные условия:

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

Тогда задача смешанного оценивания параметров линейной регрессии (2) имеет вид:

$$\min_{\alpha} \sum_{i=1}^{I} J_{\nu(N_i)}^{N_i}(\alpha). \qquad (3)$$

Разумеется, в общем случае задача (3) представляет собой весьма сложную задачу нелинейного программирования.

Рассмотрим существенно более простой случай, сводящийся к линейно-программной задаче.

Пусть исходная выборка с номерами из множества P разбита на две подвыборки с номерами из множеств N_1 и N_2 примерно равной длины (поскольку n может быть нечетным), то есть I=2, а в качестве функций потерь используются $J_{1(N_1)}^{N_1}(\alpha)$ и $J_{\infty(N_2)}^{N_2}(\alpha)$, соответствующие МНМ и МАО. Представим вектора α и ε в виде разностей их положительных a, u и отрицательных b, v частей соответственно:

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

Тогда, используя приемы, описанные, например, в [2, 3, 9], задача (3) может быть представлена в виде задачи линейного программирования:

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

$$u_k + \upsilon_k - r \leq 0, \qquad k \in N_2,$$

$$\sum_{k \in N_1} (u_k + \upsilon_k) / s + r \to \min.$$

Здесь s – мощность множества N_1.

Рассмотрим простой численный пример. Пусть дана выборка:

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

Построим для нее двухфакторную линейную регрессию

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon$$

четырьмя упомянутыми в работе методами.

При этом исходная выборка разобьется на две подвыборки с множествами номеров наблюдений N_1={1, 2, 3} и N_2={4,5,6}.

В результате получим следующие результаты.

1. α (МНК) = (0.501, 0.588), ε =(3.06, 2.14, -2.59, 0.23, -0.61, -2.21).
2. α (МНМ) = (0.566, 0.49), ε =(3.42, 1.94, -2.89, 0, 0, -1.75).
3. α (МАО) = (0.387, 0.714), ε =(2.65, 2.65, -2.04, 0.756, -1.39, -2.65).
4. α (МСО) = (0.511, 0.395), ε =(4, 2.813, -2.465, 0.72, -0.72, -0.72).

В своих последующих публикациях автор намерен заняться анализом свойств МСО.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. М.: Финансы и статистика. 1981.Т.1. 366 с., Т. 2. 351с.

2. Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. Иркутск: Облинформпечать.-1996. -320с.

3. Носков С.И., Баенхаева А.В. Множественное оценивание параметров линейного регрессионого уравнения // Современные технологии. Системный анализ. Моделирование. 2016. № 3 (51). -С. 133-138.

4. Носков С.И., Быкова О.В., Некипелова О.Е., Соколова Л.Е. Возможный способ поиска компромиссного решения в задаче линейного программирования с векторной целевой функцией // Фундаментальные исследования. 2014. № 6-3.- С. 502-505.

5. Носков С.И. Критерий «согласованность поведения» в регрессионном анализе //Современные технологии. Системный анализ. Моделирование. 2013. № 1 (37).- С. 107-110.

6. Лакеев А.В., Носков С.И. Метод наименьших модулей для линейной регрессии: число нулевых ошибок аппроксимации // Современные технологии. Системный анализ. Моделирование. 2012. № 2 (34). -С. 48-50.

7. Носков С.И. Проблема единственности Парето-оптимального решения в задаче линейного программирования с векторной целевой функцией // Современные технологии. Системный анализ. Моделирование. 2011. № S-4 (32). -С. 283-285.

8. Носков С.И. Точечная характеризация множества Парето в линейной многокритериальной задаче // Современные технологии. Системный анализ. Моделирование. 2008. № 1 (17).- С. 99-101.

9. Носков С.И. L-множество в многокритериальной задаче оценивания параметров регрессионных уравнений // Информационные технологии и проблемы математического моделирования сложных систем, 2004. № 1. -С. 64 - 69.

10. Носков С.И. Построение эконометрических зависимостей с учетом критерия «согласованность поведения» // Кибернетика и системный анализ, 1994. № 1. -С. 177 - 181.

REFERENCES

1. Dreyper N., Smith G. Applied regression analysis. M.: Finance and statistics. 1981. V.1. 366 P., V. 2. 351 P.

2. Noskov S.I. Technology of modeling of objects with unstable functioning and uncertainty in data. Irkutsk: Regional Information Printing.-1996.-320 P.

3. Noskov S.I., Bayenkhayeva A.V. Multiple estimation of parameters of the linear regression equation//Modern technologies. Systems analysis. Modeling. 2016. N. 3 (51). – P.P. 133-138.

4. Noskov S.I., Bykova O.V., Nekipelova O.E., Sokolova L.E. A possible way of search of a compromise solution in a problem of linear programming with vector target function//Basic researches. 2014. N. 6-3. – P.P. 502-505.

5. Noskov S.I. Criterion "coherence of behavior" in regression analysis//Modern technologies. Systems analysis. Modeling. 2013. N. 1 (37).-P.P. 107-110.

6. Lakeev A.V., S.I. Metod Socks of the smallest modules for a linear regression: number of zero errors of approximation//Modern technologies. Systems analysis. Modeling. 2012. N. 2 (34). – P.P. 48-50.

7. Noskov S.I. A problem of uniqueness of a pareto-optimal solution in a problem of linear programming with vector target function//Modern technologies. Systems analysis. Modeling. 2011. N. 2-4 (32). – P.P. 283-285.

8. Noskov S.I. Point characterization of a Pareto set in a linear multicriteria problem//Modern technologies. Systems analysis. Modeling. 2008. N. 1 (17).-P.P. 99-101.

9. Noskov S.I. A L-set in a multicriteria problem of estimation of parameters of the regression equations//Information technologies and problems of mathematical modeling of complex systems, 2004. N. 1. – P.P. 64 - 69.

10. Noskov S.I. Creation of econometric dependences taking into account criterion "coherence of behavior"//Cybernetics and systems analysis, 1994. -N. 1. – P.P. 177 - 181.

Информация об авторе

Сергей Иванович Носков – д.т.н., профессор, профессор кафедры «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: noskov_s@irgups.ru

Author

Sergey Ivanovich Noskov, Doctor of Technical Science, Professor, the Subdepartment Information systems and information security, Irkutsk State Transport University, Irkutsk, e-mail:noskov_s@irgups.ru

Для цитирования

Носков С.И. О методе смешанного оценивания параметров линейной регрессии// Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2019. – №1. – С. 14-20 – Режим доступа: http://ismm-irgups.ru/toma/11-2019, свободный. – Загл. с экрана. – Яз. рус., англ. (дата 20 обращения: 01.02.2019)

For citation

Noskov S.I. About the method of mixed estimation of parameters of linear regression// Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal], 2019. N. 1. P. 14-20. [Accessed 01/02/19]