DOI 10.25699/SSSB.2020.29.56926

05.13.18

МЕТОД АНТИРОБАСТНОГО ОЦЕНИВАНИЯ ПАРАМЕТРОВ ЛИНЕЙНОЙ РЕГРЕССИИ: ЧИСЛО МАКСИМАЛЬНЫХ ПО МОДУЛЮ ОШИБОК АППРОКСИМАЦИИ

Носков С.И.

Иркутский государственный университет путей сообщения, г. Иркутск

В работе рассматривается метод антиробастного оценивания (МАО) параметров линейного регрессионного уравнения, основанный на минимизации расстояния Чебышева между расчетными и фактическими значениями зависимой переменной. В отличие от метода наименьших модулей, который, по существу, игнорирует выбросы в данных, МАО, напротив, к ним тяготеет. Подтверждено, что, в соответствии с экспериментальными результатами, число максимальных по модулю ошибок аппроксимации уравнения не меньше числа параметров плюс единица.

Ключевые слова: регрессионное уравнение, антиробастное оценивание параметров, ошибки аппроксимации.

ВВЕДЕНИЕ

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 – зависимая. x_i – i-ая независимая переменные; α_i – i-ый оцениваемый параметр; ε_k – k-ая ошибка аппроксимации. k – номер наблюдения. n – число наблюдений.

Представим уравнение (1) в векторной 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}.

В настоящее время разработан значительный арсенал методов оценивания параметров уравнения (1) (см., например. [1-6]).

Весьма интересная сфера приложения усилий ученых, занимающихся регрессионным анализом, связана с изучением ошибок аппроксимации ε в уравнении (2) (см., в частности, [7-11]). К ней относится и настоящая работа.

ОСНОВНАЯ ЧАСТЬ

Значительный класс методов оценивания параметров уравнения (2) связан с расчетом соответствующих нормам Гельдера так называемых L_ν – оценок. рассчитываемых посредством минимизации функций потерь вида[2]:

$$J_\nu(\alpha) = \sum_{k=1}^{n} |\varepsilon_k|^\nu. \qquad (3)$$

Каждая из этих оценок характеризуется различной реакцией на аномальные наблюдения. рассогласованные с выборкой в целом. При этом чем больше значение ν, тем сильнее L_ν – оценка реагирует на эти аномалии. В регрессионном анализе методы оценивания, слабо реагирующие на такие выбросы, или вообще их игнорирующие, называют робастными. Случаю ν=1 соответствует именно такой метод – метод наименьших модулей (МНМ). Более того, как показано в работе [12], МНМ обладает одним замечательным свойством – если матрица X не имеет особенностей, то число нулевых ошибок аппроксимации в (1) равно m. то есть совпадает с числом параметров уравнения. Тем самым оставшиеся n-m наблюдений выборки неявным образом полагаются выбросами.

Противоположным по смыслу по отношению к МНМ является соответствующий расстоянию Чебышева между расчетными и фактическими значениями зависимой переменной метод антиробастного оценивания (МАО), сильно реагирующий на выбросы, для которого $\nu \to \infty$.

Таким образом, МАО предполагает решение задачи:

$$J_\infty(\alpha) = \max_{k=1,n} |\varepsilon_k| \to \min, \qquad (4)$$

которая, в свою очередь, сводится к следующей задаче линейного программирования (ЛП) [2]:

$$X\alpha^1 - X\alpha^2 + u - \upsilon = y, \qquad (5)$$

$$u_k + \upsilon_k - r \le 0, \qquad k = \overline{1,n}, \tag{6}$$

$$\alpha^1 \ge 0, \quad \alpha^2 \ge 0, \quad u \ge 0, \quad \upsilon \ge 0, \quad r \ge 0, \tag{7}$$

$$r \to \min, \tag{8}$$

после решения которой вектор параметров уравнения (2) представим в виде:

$$\alpha^{MAO} = \alpha^1 - \alpha^2$$

Введем в рассмотрение число P максимальных по модулю ошибок аппроксимации при использовании α^{MAO}. Как показывает опыт практического применения МАО, он тоже, как и МНМ. обладает одним интересным свойством по отношению к ошибкам аппроксимации, а именно (см., в частности, [13]), справедливо неравенство:

$$P \ge m + 1, \tag{9}$$

где P – число максимальных по модулю ошибок аппроксимации уравнения (1).

Рассмотрим конкретный пример.

В работе [6] представлена статистическая информация результатов деятельности Красноярской железной дороги за 2000-2014 г.г. по показателям:

y - грузооборот;

x1 - прием груженых вагонов;

x2 – прием порожних вагонов;

x3 – динамическая нагрузка;

x4 – среднесуточный пробег локомотива;

x5 – эксплуатируемый парк локомотивов;

x6 – рабочий парк вагонов.

Для этих данных с помощью МАО было построено 67 линейных регрессий (1), объединенных в 7 групп:

1-ая – одно шестифакторное уравнение со свободным членом;

2-ая – шесть пятифакторных уравнений со свободным членом;

3-ья – пятнадцать четырехфакторных уравнений со свободным членом;

4-ая – восемнадцать трехфакторных уравнений со свободным членом;

5-ая – пятнадцать двухфакторных уравнений со свободным членом;

6-ая – шесть однофакторных уравнений со свободным членом;

7-ая – шесть однофакторных уравнений без свободного члена.

Приведем по одному представителю из каждой группы.

y=-27110.8+13.45x1+2.58x2+
368.21x3+9.12x4+28.6x5+0.06x6,

ε =(1224.08,408.12,-1224.08,-182.57,-1224.08,-502.50,1224.08,-1224.08, 1224.08, -1224.08,-911.62,-142.44, -942.69,-1224.08,-111.04), P=8,

y=-27091+13.6268x1+5.5543x2+385.34x3+ 8.66x5+ 27.99x6,

ε =(1224.54.424.41.-1224.54.-184.21.-1224.54. -516.64.1224.54. -1224.54.1224.54.-1224.54. 172.7.919.09.332.29.404.23.1130.59), P=7,

y= -26798+16.84x1+3.58x2+612.92x3+3.4x5,

ε =(1252.51,1113.91,-1252.51,-381.24,-1252.51,-1142.25, 1252.51, -861.75,1252.51, -102.62, 565.61,937.43,1252.51,412.1,1080.53), P=6,

y= -25765+15.69x1+5.84x2+527.85x3,

ε =(1298.03,728.18,-1298.03,-171.29,-1298.03,-917.85,1123.11,-847.78,1025.91,-1298.03,-151.24,1260.33,1298.03,405.14,890.26), P=5,

y=-20023.9+252.29x5-0.23x6,

ε =(-7123.55,-4634.89,-4687.49,-3026.61, 1409.72,4102.1,6447.98,517.3,7123.55, 3345.5,5297.46,-2489.36, -7123.55,704.34,7123.55), P=4,

y=49151+1.35x6,

ε =(-13532.61,-9046.27,-10587.8, -5130.92,-1565.09, -792.34, 3224.74, 7930.39, 13532.61, 8302.94, -10998.7,-5391.25,-6689.52,-13532.61, 1654.01), P=3,

y=2.93x6,

ε =(13484.79,17483.02,15521.3,20622.76, 23886.88,24451.31,28172.8,31726.89,36198.09,32003.08,-16823.63,-12126.05,-19054.16,-36198.09, -11786.37), P=2.

Анализ полученных уравнений показал, что в 66 случаях из 67 имеет место равенство P=m+1 и лишь для одного из них выполняется строгое неравенство (9):

y=92,37+15.12x1+8.92x2-11.3x5+0.09x6,

ε =(1432.1,252.32,-645.63,-97.91,-1432.1,-1432.1, 1432.1, -449.63,1353.95, -1432.1,-1432.1, 1432.1,1334.64,-276.9,238.93), P=7.

ЗАКЛЮЧЕНИЕ

В своих последующих работах автор намерен заняться дальнейшим изучением свойств оценок параметров линейного регрессионного уравнения, полученных по методу антиробастного оценивания.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Носков С.И.. Базилевский М.П. Построение регрессионных моделей с использованием аппарата линейно-булевого программирования. – Иркутск: ИрГУПС. 2018. – 176 с.

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

3. Базилевский М.П.. Носков С.И. Алгоритм построения линейно-мультипликативной регрессии//Современные технологии. Системный анализ. Моделирование. 2011. № 1 (29). С. 88-92.

4. Базилевский М.П.. Носков С.И. Программный комплекс построения линейной регрессионной модели с учётом критерия согласованности поведения фактической и расчётной траекторий изменения значений объясняемой переменной // Вестник ИрГТУ. – Иркутск. 2017. – Т.21 – №9. – С. 37-44.

5. Базилевский М.П.. Носков С.И. Формализация задачи построения линейно-мультипликативной регрессии в виде задачи частично-булевого линейного программирования // Современные технологии. Системный анализ. Моделирование. – Иркутск. - 2017. – №3. – С.101-105.

6. Носков С.И.. Врублевский И.П. Регрессионная модель динамики эксплуатационных показателей функционирования железнодорожного транспорта // Современные технологии. Системный анализ. Моделирование. - 2016. - №2. - С. 192-197.

7. Малова Н.Н. Об одном подходе к расчету средней ошибки аппроксимации регрессионных моделей // Международный технико-экономический журнал. 2017. № 5. С. 54-57..

8. Cornbleet PJ, Gochman N. Incorrect least-squares regression coefficients in method comparisons// Clin Chem 1979;25:432-438.

9. Cochran W. G. Errors of measurement in statistics // Technometrics. Vol. 10 (1968), pp. 637-666.

10. . Lipponen I, Kolehmainen V, Romakkaniemi S. Correction of approximation errors with Random Forests applied tomodelling of cloud droplet formation // Geosci. Model Dev., 6, 2087–2098, 2013.

11. Arridge S., Kaipio J., Kolehmainen,V., Schweiger M.,Somersalo, E., Tarvainen T., Vauhkonen M. Approximation errors and model reduction with an application inoptical diffusion tomography // , Inverse Probl., 2006, 22, 175–195.

12. 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.

13. Earle A. Minimum l_∞ Norm Solutions To Finite Dimensional Algebraic Underdetermined Linear Systems.-2014.-134 p..

Носков Сергей Иванович, доктор тех. наук, профессор, профессор кафедры «Информационные системы и защита информации» Иркутского государственного университета путей сообщения, 664074. Россия. г. Иркутск. ул. Чернышевского. д. 15. тел.: +79149022494, эл. почта: sergey.noskov.57@mail.ru

METHOD OF ANTIROBAST ESTIMATION OF LINEAR REGRESSION PARAMETERS: NUMBER OF MAXIMUM ON THE MODULE OF APPROXIMATION ERRORS

S.I. Noskov

Irkutsk State Transport University. Irkutsk

The paper considers the method of antirobast estimation (MAO) of parameters of linear regression equation, based on minimization of Chebyshev distance between calculated and actual values of dependent variable. Unlike the least modules method, which essentially ignores emissions in data, MAO, on the contrary, gravitates towards them. It is shown that, according to experimental results, the number of modulo-maximum approximation errors of the equation is not less than the number of parameters plus one.

Keywords: regression equation, antirobast parameter estimation, approximation errors.

REFERENCES

1. Noskov S.I. Basilevsky M.P. Construction of regression models using linear-Boolean programming apparatus. - Irkutsk: IrGUPS. 2018. - 176 p.

2. Noskov S.I. Technology of modeling objects with unstable functioning and uncertainty in data. Irkutsk: Oblique. 1996 .- 320 s.

3. Basilevsky M.P... Noskov S.I. Algorithm of linear-multiplicative regression construction//Modern technologies. System analysis. Modeling. 2011. № 1 (29). Page 88-92.

4. Basilevsky M.P... Noskov S.I. Software complex of construction of linear regression model taking into account the criterion of consistency of behavior of actual and calculated paths of change of values of the explained variable//Journal of IrGTU. - Irkutsk. 2017. – T.21 – NO. 9. – PAGE 37-44.

5. Basilevsky M.P... Noskov S.I. Formalization of the task of building linear-multiplicative regression in the form of the task of partial-boolean linear programming//Modern technologies. System analysis. Modeling. - Irkutsk. - 2017. - № 3. - C.101-105.

6. Noskov S.I. Vublevsky I.P. Regression Model of Dynamics of Ex-Plumation Indicators of Railway Transport Functioning//Modern Technolo

gies. System analysis. Modeling. - 2016. - № 2. - S. 192-197.

7. Malova N.N. About one approach to calculating the average approximation error of regression models // International technical and economic journal. 2017. No 5. S. 54-57.

8. Cornbleet PJ, Gochman N. Incorrect least-squares regression co-efficients in method comparisons// Clin Chem 1979;25:432-438.

9. Cochran W. G. Errors of measurement in statistics // Technomet-rics. Vol. 10 (1968), pp. 637-666.

10. . Lipponen I, Kolehmainen V, Romakkaniemi S. Correction of approximation errors with Random Forests applied tomodelling of cloud droplet formation // Geosci. Model Dev., 6, 2087–2098, 2013.

11. Arridge S., Kaipio J., Kolehmainen,V., Schweiger M.,Somersalo, E., Tarvainen T., Vauhkonen M. Approximation errors and model reduction with an application inoptical diffusion tomography // , Inverse Probl., 2006, 22, 175–195.

12. Lakeev A.V., Noskov S.I. Method of smallest modules for linear regression: number of zero approximation errors//Modern technologies. System analysis. Modeling. - 2012. - № 2 (34). - S. 48-50.

13. .Earle A. Minimum Norm Solutions To Finite Dimensional Algebraic Underdetermined Linear Systems.-2014.-134 p..

Sergey I. Noskov.- Doctor of Technical Science, Professor. the Subdepartment Information systems and information security. Irkutsk State Transport University.15. Chernyshevsky st.. Irkutsk. Russia. 664074.Phone:+79149022494.Email: sergey.noskov.57@mail.ru