УДК 519.862.6

С.И. Носков[1], Н.И. Глухов [1], А.С. Вергасов[1]

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

АНАЛИЗ МЕР СХОДСТВА ПРИ ИСПОЛЬЗОВАНИИ ВЗВЕШЕННОГО МЕТОДА НАИМЕНЬШИХ КВАДРАТОВ

Аннотация. *В статье рассматривается задача оценивания параметров линейного регрессионного уравнения с использованием взвешенного метода наименьших квадратов. При этом в основу алгоритма назначения весов наблюдений предлагается положить разработанную Ю.А.Ворониным теорию сходства. В работе проводится анализ десяти таких мер применительно к статистической информации, касающейся двух реальных объектов.*

Ключевые слова. *Регрессионное уравнение, оценивание параметров, взвешенный метод наименьших квадратов, теория сходства.*

S.I. Noskov[1], N.I. Glukhov[1], A.S. Vergasov[1]

[1]Irkutsk State Transport University, Irkutsk, Russia

ANALYSIS OF SIMILARITY MEASURES WHEN USING THE WEIGHED METHOD OF THE LEAST SQUARES

Annotation. *The article deals with the problem of estimating the parameters of a linear regression equation using the weighted least squares method. In this case, the theory of assigning weights of observations is proposed to put the theory of similarity developed by Yu.A. Voronin. The paper analyzes ten such measures in relation to statistical information relating to two real objects.*

Keywords. *Regression equation, parameter estimation, weighted least squares method, theory of similarity.*

Рассмотрим элемент любой регрессионной модели - линейное уравнение

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

где y_k и x_{ki} – k-ые значения соответственно выходной и i-ой входной переменных, $\beta = (\beta_1, \dots, \beta_d)^T$ – вектор подлежащих оцениванию параметров, ε_k – ошибки аппроксимации, d – количество наблюдений выборки.

Представим уравнение (1) в векторной форме:

$$y = X\beta + \varepsilon, \quad (2)$$

где $y = (y_1, \dots, y_d)^T$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_d)$, $\mathrm{X} = \|x_{ki}\|, k = \overline{1,d}, i = \overline{1,n}$.

Методам оценивания неизвестных параметров уравнения (1) и критериям его верификации посвящена обширная литература (см., например, [1-14]).

Одно из основных направлений практического применения регрессионных моделей с составным элементом - уравнением (1), - является расчет прогнозных значений зависимых переменных при известных значениях независимых. При этом необходимо иметь в виду следующее важное обстоятельство. Часто функционирование исследуемого на модельном уровне объекта имеет динамический характер, что предполагает различие в информационном статусе наблюдений выборки. Для таких ситуаций вместо традиционно применяемых методов оценивания параметров – наименьших квадратов, модулей, антиробастного, - более эффективно использовать их «взвешенные» модификации, например, взвешенный метод наименьших квадратов (ВМНК), расчетная формула которого имеет вид:

$$\beta = (X^T W X)^{-1} X^T W y, \quad (3)$$

где $\mathrm{W} = \mathrm{diag}(\omega_k)$, $k = \overline{1,d}$, $\omega_k > 0$ – веса наблюдений выборки.

В работе [15] при расчете весов наблюдений авторами предлагается подход, основанный на постулате – чем ближе в некотором наперед заданном смысле вектор

значений независимых переменных прогнозного периода к соответствующему наблюдению периода основания прогноза, тем большим весом это наблюдение должно обладать, а, значит, тем больше должен быть его вес ω_k в (3).

Мера оценки такой близости может быть основана на разработанной профессором Ю.А. Ворониным теории сходства (см., например,[16]). В [16] рассмотрены десять возможных мер сходства. Приведем их формальные представления.

Пусть для некоторых s объектов задана матрица H характеризующих их поведение n признаков:

H= $\|h_{ki}\|, k = \overline{1,s}, i = \overline{1,n}.$

Введем обозначения:

$h_i^- = \min_k h_{ki}$, $h_i^+ = \max_k h_{ki}$.

Для каждого a-го объекта рассчитаем значения:

$f_i^a = (\mathrm{h}_{ai} - h_i^-) / (h_i^+ - h_i^-)$, $i = \overline{1,n}.$

Очевидно, что для всех a и i справедливы неравенства

$0 \le f_i^a \le 1, a = \overline{1,s}, i = \overline{1,n}.$

Тогда аналитические выражения для мер сходства между объектами k и l примут вид.

1) $1 - \sum_{i=1}^n \alpha_i \left| f_i^k - f_i^l \right|$.

2) $\frac{1 - \sqrt{\sum_{i=1}^n \alpha_i^2 (f_i^k - f_i^l)^2}}{\sqrt{\sum_{i=1}^n \alpha_i}}$.

3) $1 - \max_i \left| f_i^k - f_i^l \right|$.

4) $\sum_{i=1}^n \alpha_i \frac{\min(f_i^k, f_i^l)}{\max(f_i^k, f_i^l)}$.

5) $\frac{1}{1 + \sum_{i=1}^n |f_i^k - f_i^l|}$.

6) $1 - \frac{\sum_{i=1}^n |(f_i^k - f_i^l)| + |\sum_{i=1}^n (f_i^k - f_i^l)|}{2}$.

7) $\frac{\sum_{i=1}^n (f_i^k f_i^l)}{(\sum_{i=1}^n (f_i^k)^2)^{\frac{1}{2}} (\sum_{i=1}^n (f_i^l)^2)^{\frac{1}{2}}}$.

8) $1 - e^{-\left(\sum_{i=1}^n \left(f_i^k - f_i^l\right)^2\right)^{\frac{1}{2}}}$.

9) $\sum_{i=1}^n \alpha_i \left(1 - \left| f_i^k - f_i^l \right|\right) * \frac{\sum_{i=1}^n (f_i^k f_i^l)}{\left(\sum_{i=1}^n (f_i^k)^2\right)^{\frac{1}{2}} \left(\sum_{i=1}^n (f_i^l)^2\right)^{\frac{1}{2}}}$.

10) $\sum_{i=1}^{n_1} \alpha_i \left(1 - \left| f_i^k - f_i^l \right|\right) * \prod_{i=n_1}^{n} \left(1 - \left| f_i^k - f_i^l \right|\right)$,

$\alpha_i \ge 0, \sum_{i=1}^n \alpha_i = 1.$

Здесь α_i - весовые коэффициенты признаков, которые в простейшем случае могут быть приняты равными, например, 1/n.

Для того, чтобы определиться с тем, какую именно меру сходства из представленных десяти использовать при реализации ВМНК, проведем анализ их значений на конкретных числовых данных. Это значения пяти социально-экономических показателей Иркутской области за десять лет [17] и двадцати эксплуатационных показателей Красноярской железной дороги за пятнадцать лет [18].

В таблицах 1 и 2 приведены значения второй и пятой мер сходства на наблюдениях первого объекта.

Таблица 1. Мера сходства №2

1.000, 0.963, 0.959, 0.950, 0.926, 0.917, 0.901, 0.887, 0.870, 0.893, 0.880, 0.874, 0.860, 0.849, 0.849
1.000, 0.976, 0.968, 0.947, 0.937, 0.918, 0.904, 0.885, 0.910, 0.895, 0.891, 0.875, 0.863, 0.864
1.000, 0.965, 0.946, 0.936, 0.919, 0.906, 0.884, 0.917, 0.895, 0.890, 0.874, 0.864, 0.869
1.000, 0.975, 0.960, 0.938, 0.924, 0.901, 0.925, 0.901, 0.894, 0.879, 0.865, 0.870
1.000, 0.976, 0.953, 0.939, 0.914, 0.936, 0.906, 0.896, 0.883, 0.867, 0.876
1.000, 0.972, 0.957, 0.933, 0.944, 0.910, 0.898, 0.887, 0.868, 0.878
1.000, 0.980, 0.956, 0.954, 0.914, 0.901, 0.892, 0.871, 0.884
1.000, 0.970, 0.958, 0.917, 0.906, 0.899, 0.877, 0.893
1.000, 0.944, 0.914, 0.902, 0.900, 0.875, 0.890
1.000, 0.926, 0.914, 0.909, 0.892, 0.909
1.000, 0.965, 0.961, 0.946, 0.945
1.000, 0.976, 0.959, 0.952
1.000, 0.968, 0.957
1.000, 0.962
1.000

Таблица 2. Мера сходства №5

1.000, 0.244, 0.231, 0.180, 0.130, 0.111, 0.096, 0.083, 0.070, 0.085, 0.071, 0.067, 0.060, 0.056, 0.057
1.000, 0.317, 0.244, 0.164, 0.138, 0.115, 0.097, 0.079, 0.096, 0.080, 0.078, 0.069, 0.064, 0.064
1.000, 0.245, 0.163, 0.135, 0.113, 0.096, 0.081, 0.106, 0.081, 0.077, 0.068, 0.066, 0.066
1.000, 0.305, 0.209, 0.144, 0.114, 0.089, 0.116, 0.087, 0.080, 0.071, 0.066, 0.067
1.000, 0.321, 0.186, 0.138, 0.103, 0.143, 0.094, 0.084, 0.074, 0.067, 0.073
1.000, 0.281, 0.184, 0.127, 0.171, 0.100, 0.088, 0.079, 0.070, 0.076
1.000, 0.342, 0.189, 0.198, 0.106, 0.091, 0.084, 0.070, 0.077
1.000, 0.286, 0.214, 0.116, 0.097, 0.092, 0.074, 0.080
1.000, 0.179, 0.111, 0.095, 0.091, 0.073, 0.079
1.000, 0.140, 0.111, 0.104, 0.094, 0.104
1.000, 0.261, 0.227, 0.160, 0.150
1.000, 0.310, 0.202, 0.164
1.000, 0.258, 0.193
1.000, 0.227
1.000

Упорядочим по убыванию значения обеих мер для первого наблюдения. Оказывается, эти упорядочения совпадают: 2, 5, 3, 6, 4, 7, 10, 9, 8.

Проведем аналогичный полный анализ всех десяти мер для обоих объектов. По его результатам можно сделать следующие два основных вывода.

Шестая и десятая меры не могут быть использованы в качестве расчетных формул для вычисления весовых коэффициентов ВМНК, поскольку соответствующие таблицы содержат либо отрицательные элементы, либо диагональные компоненты, меньшие единицы, либо обладают значительной насыщенностью нулевыми элементами.

Остальные восемь мер дают сходные результаты, подобные приведенному выше. Поэтому любая из них в равной степени может быть использована при реализации взвешенного метода наименьших квадратов и давать при этом близкие результаты.

При анализе мер сходства может быть использован также математический аппарат, представленный в [19].

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. 320 с.

3. Носков С.И. Идентификация параметров кусочно-линейной функции риска. Транспортная инфраструктура Сибирского региона, 2017. Т. 1. С. 417-421.

4. Иванова Н.К., Носков С.И. Организация прогнозных расчетов по регрессионным моделям // Информационные технологии и проблемы математического моделирования сложных систем, 2017. № 18. С. 78-80.

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

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

7. 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.

8. Базилевский М.П., Носков С.И. Анализ систем программирования для решения вычислительной задачи проведения «конкурса» регрессионных моделей // Информационные технологии и проблемы математического моделирования сложных систем, 2011. № 9. С. 47-51.

9. Носков С.И., Зырянов С.И. Применение критерия смещения при построении регрессионных уравнений // Современные технологии. Системный анализ. Моделирование, 2004. № 2. С. 93.

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

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

12. Головченко В.Б., Носков С.И. Выбор класса линейной по параметрам регрессии на основе экспертных высказываний // Кибернетика и системный анализ, 1992. № 5. С. 109.

13. Носков С.И., Потороченко Н.А. Диалоговая система реализации «конкурса» регрессионных зависимостей // Управляющие системы и машины, 1992. № 2-4. С. 111.

14. Golovchenko V.B., Noskov S.I. Estimation of an econometric model using statistical data and expert information // Automation and Remote Control, 1991. V. 52. № 4. P.542-548.

15. Носков С.И., Вергасов А.С. Реализация взвешенного метода наименьших квадратов с использованием мер сходства. // Вестник науки и образования.-2018.-№18-1 (54). –С. 29-32.

16. Воронин Ю.А. Начала теории сходства. Новосибирск: ВЦ СО АН СССР, 1989. 120 с.

17. Баенхаева А.В. Моделирование валового регионального продукта Иркутской области на основе применения методики множественного оценивания / Базилевский М.П., Носков С.И. // Фундаментальные исследования.– 2016. –№10 (часть 1). – С. 9-14.

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

19. Носков С.И., Удилов В.П. Управление системой обеспечения пожарной безопасности на региональном уровне.-Иркутск: ВСИ МВД РФ, 2003.-151 с.

REFERENCES

1. Draper N., Smith G. Applied regression analysis. M .: Finance and Statistics, 1981.T. 1. 366 p., T. 2. 351s.

2. Noskov S.I. Object modeling technology with unstable operation and data uncertainty. Irkutsk: Oblinformphechat, 1996. 320 p.

3. Noskov S.I. Identification of parameters of a piecewise linear risk function. Transport infrastructure of the Siberian region, 2017. T. 1. S. p. 417-421.

4. Ivanova N.K., Noskov S.I. Organization of predictive calculations for regression models // Information technologies and problems of mathematical modeling of complex systems, 2017. No. 18. P. 78-80.

5. Noskov S.I., Baenkhaeva A.V. Multiple estimation of parameters of a linear regression equation // Modern technologies. System analysis. Modeling, 2016. № 3 (51). S. 133-138.

6. Noskov S.I. The criterion "consistency of behavior" in the regression analysis // Modern technologies. System analysis. Modeling, 2013. № 1 (37). Pp. 107-110.

7. Lakeev A.V., Noskov S.I. The method of least modules for linear regression: the number of zero approximation errors // Modern technologies. System analysis. Modeling, 2012. № 2 (34). Pp. 48-50.

8. Bazilevsky M.P., Noskov S.I. Analysis of programming systems for solving the computational problem of the “competition” of regression models // Information technologies and problems of mathematical modeling of complex systems, 2011. No. 9. P. 47-51.

9. Noskov S.I., Zyryanov S.I. Application of the displacement criterion in the construction of regression equations // Modern technologies. System analysis. Modeling, 2004. № 2. S. 93.

10. Noskov S.I. L-set in the multicriteria problem of estimating the parameters of regression equations // Information technologies and problems of mathematical modeling of complex systems, 2004. No. 1. P. 64.

11. Noskov S.I. Building econometric dependencies taking into account the criterion "consistency of behavior" // Cybernetics and Systems Analysis, 1994. No. 1. P. 177.

12. Golovchenko V.B., Noskov S.I. The choice of the class of linear regression on the basis of expert statements // Cybernetics and Systems Analysis, 1992. No. 5. P. 109.

13. S. Noskov, N.Potorochenko N.A. Dialogue system for the implementation of the “competition” of regression dependencies // Control systems and machines, 1992. № 2-4. P. 111.

14. Golovchenko V.B., Noskov S.I. Estimation of an econometric model using statistical data and expert information // Automation and Remote Control, 1991. V. 52. No. 4. P.542-548.

15. Noskov S.I., Vergasov A.S. Implement a weighted least squares method using similarity measures. // Bulletin of science and education.-2018.-№18-1 (54). -WITH. 29-32.

16. Voronin Yu.A. Beginning of the theory of similarity. Novosibirsk: Computing Center of Siberian Branch of the USSR Academy of Sciences, 1989. 120 p.

17. Baenkhaeva A.V. Simulation of the gross regional product of the Irkutsk region based on the application of the method of multiple estimation / Bazilevsky MP, Noskov SI // Basic research.– 2016. –№10 (part 1). - p. 9-14.

18. Wroblewski, I.P. Regression model of the dynamics of operational indicators of railway transport functioning / S.I. Noskov, I.P. Wroblewski // Modern technologies. System analysis. Modeling. - 2016.-№2.-C. 192-197.

19. Noskov SI, Udilov V.P. Management of the fire safety system at the regional level.- Irkutsk: All-Russian Central Executive Committee of the Ministry of Internal Affairs of the Russian Federation, 2003.-151 p.

Информация об авторах

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

Вергасов Александр Сергеевич – соискатель кафедры «Информационные системы и защиты информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: tluck@inbox.ru

Глухов Николай Иванович – к.э.н., доцент кафедры «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: gni1953@mail.ru

Authors Information

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

Vergasov Alexander Sergeevich – applicant, Department of Information Systems and Information Protection, Irkutsk State Transport University, Irkutsk, e-mail: tluck@inbox.ru

Glukhov Nikolay Ivanovich – Ph. D., associate Professor of the Department "Information systems and information protection", Irkutsk State Transport University, Irkutsk, e-mail: gni1953@mail.ru

Для цитирования

Носков С.И., Вергасов А.С., Глухов Н.И. Анализ мер сходства при использовании взвешенного метода наименьших квадратов // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2019. – №2. – С. 12-17 – Режим доступа: http://ismm-irgups.ru/toma/23-2019, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 19.06.2019)

For citations

Noskov S.I., Vergasov A.S., Glukhov N.I. *Analiz mer skhodstva pri ispol'zovanii vzveshennogo metoda naimen'shikh kvadratov* [Analysis of similarity measures when using the weighed method of the least squares] // *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. No. 2. P. 12-17 – Access mode: http://ismm-irgups.ru/toma/23-2019, free. – Title from the screen. – Language Russian, English. [Accessed 19/06/19]