УДК 518.852+518.853

С.И. Носков [1]

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

ОБОБЩЕННЫЙ КРИТЕРИЙ СОГЛАСОВАННОСТИ ПОВЕДЕНИЯ В РЕГРЕССИОННОМ АНАЛИЗЕ

Аннотация. *В статье приводится модификация введенного автором ранее нового критерия адекватности регрессионных уравнений – критерия «согласованность поведения», или так называемого СП-критерия. Он базируется, в отличие от традиционных в регрессионном анализе критериев адекватности, не на анализе ошибок аппроксимации уравнения, а на соотнесении знаков приростов соседних по номерам наблюдений фактических и расчетных значений зависимой переменной уравнения. Поэтому даже для уравнений с высокими значениями классических верификационных критериев СП-критерий может обладать низкой значимостью. В отличие от СП-критерия предлагаемый в настоящей статье обобщенный критерий согласованности поведения (ОСП-критерий) предполагает соотнесение знаков указанных приростов для пар наблюдений с произвольными номерами, что позволяет выявлять полную картину в соответствии поведения фактических и расчетных значений зависимой переменной уравнения на всей выборке с учетом всевозможных перекрестных связей. Кроме того, в работе предлагается алгоритм максимизации значения ОСП-критерия с фиксированным или несколько ухудшенным значением выбранной исследователем функции потерь в виде суммы абсолютных значений ошибок аппроксимации, соответствующей манхэттэнскому расстоянию, или методу наименьших модулей. Этот алгоритм позволяет свести данную задачу к задаче частично-булевого линейного программирования невысокой размерности. Он также предусматривает возможность комбинирования ОСП-критерия с функцией потерь посредством формирования их линейной свертки. При этом существует возможность придавать каждому из ее компонент различный вес в зависимости от того, какой критерий лицо, принимающее решение, считает более или менее важным. При программной реализации указанного алгоритма может быть эффективно использована размещенная в Интернете в свободном доступе программа LPsolve.*

Ключевые слова: *регрессионное уравнение, критерии адекватности, методы оценивания параметров, частично-булевое линейное программирование.*

S.I. Noskov[1]

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

GENERALIZED CRITERION OF COORDINATION OF BEHAVIOR INREGRESSION ANALYSIS

Abstract. *In the article, a modification of the new criterion of the adequacy of regression equations introduced by the author-the "consistency of behavior" criterion or the so-called CB-criterion is introduced. It is based, unlike the traditional adequacy criteria in the regression analysis, not on the analysis of the approximation errors of the equation, but on the correlation of the signs of the increments of the actual and calculated values of the dependent variable of the equation by the observation numbers. Therefore, even for equations with high values of classical verification criteria, the CB criterion may have low significance. In contrast to the SP criterion, the generalized criterion for consistency of behavior (GCB criterion) proposed in this article assumes the correlation of the indicated increments for pairs of observations with arbitrary numbers, which*

makes it possible to reveal the complete picture in accordance with the behavior of the actual and calculated values of the dependent variable of the equation throughout the sample, all possible cross-links. In addition, the paper proposes an algorithm for maximizing the value of an GCB test with a fixed or slightly degraded value chosen by the researcher for the loss function as a sum of the absolute values of the approximation errors corresponding to the Manhattan distance or the method of the smallest modules. This algorithm allows us to reduce this problem to the problem of partially-boolean linear programming of low dimensionality. It also provides for the possibility of combining the GCB criterion with the loss function by forming their linear convolution. In this case, it is possible to give each of its components a different weight, depending on which criterion the decision-maker considers more or less important. With the software implementation of this algorithm, the LPsolve program can be effectively used on the Internet.

Keywords: *regression equation, adequacy criteria, parameters estimation methods, partially-boolean linear programming.*

Регрессионный анализ является признанным инструментом построения качественных математических моделей сложных систем различного характера и масштаба, в частности, эконометрических (см., например, [1, 5-12]). В рамках этой научной дисциплины разработано значительное число критериев адекватности регрессионных моделей – множественной детерминации, Фишера, Стьюдента, Дарбина-Уотсона, средних относительных ошибок аппроксимации и прогноза, смещения и т.д (см., например, обзор в [2]). Каждый из них "отвечает" за ту или иную частную характеристику модельного описания исследуемого объекта или процесса и формально выражается, как правило, через рассчитанные по модели ошибки аппроксимации. Значимость этих критериев бесспорна в силу их глубокой теоретической обоснованности и повсеместного использования в различных комбинациях при построении практически всех известных статистических моделей. Не вызывает, однако, сомнений тезис о том, что адекватность модели – понятие многогранное, заключающее в себе множество самых различных частных характеристик, число которых в результате проводимых в этой области исследований постоянно увеличивается. Так, существует важный аспект в оценке качества статистических зависимостей, не связанный напрямую с точностью аппроксимации, а отражающий степень согласованности в характере изменения (поведении) расчетных и фактических значений зависимой переменной на различных наблюдениях выборки. Ниже предлагаются некоторые способы формализации отражающего этот аспект критерия "согласованность поведения" и корректировки оценок параметров регрессий на его основе. При этом в данной работе не ставится цель изучения качественных свойств полученных оценок. Отметим, что впервые этот критерий введен и описан в работе [3].

Рассмотрим обязательный элемент любой эконометрической модели - регрессионное уравнение общего вида:

$$y_k = F(a; x_{k1}, x_{k2}, \ldots, x_{km}) + \varepsilon_k, \qquad k = \overline{1,n}, \tag{1}$$

где y - зависимая переменная, x_i, $i = \overline{1,m}$ - независимые переменные, a - вектор оцениваемых параметров, F - аппроксимирующая вещественная функция, ε_k - ошибки аппроксимации, n - длина выборки.

В регрессионном анализе оценка параметров a регрессии (1) определяется посредством минимизации выбранной функции потерь:

$$I(a) = \sum_{k=1}^{n} \varphi(\varepsilon_k).$$

Вещественная функция φ является монотонно неубывающей (как правило, выпуклой) и принимающей неотрицательные значения. К наиболее известным функциям потерь должны быть отнесены в частности, функции [2] Хубера, Андрюса, Мешалкина, а также функции вида:

$$I_v(a)=\sum_{k=1}^{n}|\varepsilon_k|^v,\ v\ge 1.$$

При этом одному из наиболее популярных в регрессионном анализе методу наименьших модулей (МНМ) соответствует значение $v=1$, а методу наименьших квадратов (МНК) - $v=2$.

Обозначим через $\hat{y}_k,\ k=\overline{1,n}$ расчетные значения зависимой переменной, вычисленные с помощью найденной на основе использования выбранной функции потерь $I(a)$оценки $\hat{a}$:

$$\hat{y}_k=F(\hat{a};x_{k1},x_{k2},\ldots,x_{km}),\ k=\overline{1,n}.$$

При построении статистических моделей могут возникать ситуации, когда даже для "почти функциональных" регрессий с малыми значениями функций потерь "поведение" расчетных и фактических траекторий, характеризующих изменение значений зависимых переменных, не согласовано. Это может быть выражено, в частности, в несовпадении для некоторых пар номеров наблюдений k и $k+1$ знаков приращений $y_{k+1}-y_k$ и $\hat{y}_{k+1}-\hat{y}_k$, что, безусловно, снижает качество такой модели, в частности, ее прогностические возможности, поскольку она в этом случае не в достаточной степени "объясняет" исследуемый процесс. Причиной низкой «согласованности поведения» является либо отсутствие в числе независимых переменных регрессии существенных (значимых) факторов, либо неверный выбор вида аппроксимирующей функции F или функции потерь $I(\alpha)$.

Критерий "согласованность поведения" (СП-критерий), позволяющий выявлять подобные ситуации, может быть представлен в виде принимающей целые значения и подлежащей максимизации функции Φ_1:

$$\Phi_1(\alpha)=\sum_{k=1}^{n-1}sign[(\hat{y}_{k+1}-\hat{y}_k)(y_{k+1}-y_k)]. \qquad (2)$$

Значение $\Phi_1(\alpha)=n-1$ указывает на полную "согласованность" векторов y и $\hat{y}$ в смысле критерия (2). Если среди компонент суммы (2) присутствуют только значения 0 и 1, эти вектора будем считать почти "согласованными".

Следует отметить, СП-критерий, некоторые его модификации и способы оперирования ими описаны в работах [2-4].

Понятие согласованности поведения может быть расширено следующим образом. В идеале знаки приращений расчетных и фактических значений зависимой переменной должны совпадать не только для соседних наблюдений, но и для **всех** возможных их пар. Формализация такого обобщенного СП–критерия (будем называть его ОСП – критерием) может быть по аналогии с (2) произведена следующим образом:

$$\Phi_2(\alpha)=\sum_{k=1}^{n-1}\sum_{s=k+1}^{n}sign[(\hat{y}_k-\hat{y}_s)(y_k-y_s)]. \qquad (3)$$

ОСП–критерий (как, впрочем, и СП–критерий), безусловно, не может рассматриваться в качестве альтернативного по отношению к функции потерь и другим перечисленным выше критериям показателя качества регрессии, поскольку наиболее важной интегрирующей характеристикой адекватности модели исследуемому объекту или процессу является все-таки точность аппроксимации. Вместе с тем, имеет смысл использовать ОСП - критерий в качестве вспомогательного для корректировки уже найденной посредством минимизации выбранной функции потерь $I(\alpha)$ оценки α. Такая корректировка может быть произведена следующим образом.

Пусть I^* - найденное минимальное значение функции потерь для регрессии (1), а α^* - соответствующая ему оценка параметров. Предположим, что исследователь (разработчик модели) может назначить некоторую величину ΔI^*, на которую допустимо увеличение значения I^* без существенного ухудшения качества аппроксимации. Тогда задача повышения согласованности поведения представима в форме:

$$\Phi_2(\alpha) \to \max_{\alpha \in A}, \tag{4}$$
$$A = \{\alpha \mid I(\alpha) \le I^* + \Delta I^*\}.$$

Пусть a^{**} - решение задачи (4). Для того, чтобы несколько "подтянуть" эту оценку к *a**, не уменьшая значение функционала в (4), необходимо решить задачу:

$$I(\alpha) \to \min_{\alpha \in B}, \tag{5}$$
$$B = \{\alpha \mid \Phi_2(\alpha) = \Phi_2(\alpha^{**})\}.$$

В случае, когда регрессия (1) линейна, а функция потерь имеет вид $I_1(\alpha)$, то есть соответствует МНМ, задачи (4) и (5) могут быть сведены к одной задаче частично-целочисленного линейного программирования (ЧЦЛП). Воспользуемся для этого приемом, описанным, например, в [2], который позволяет свести задачу с альтернативными условиями к задаче математического программирования с частью булевых переменных. Применим также способ сведения задачи определения оценок параметров линейной регрессии с помощью МНМ к задаче линейного программирования (ЛП), впервые описанный, по-видимому, в [1].

В связи с наличием в (3) операции sign введем в рассмотрение булевы переменные σ_{ks} следующим образом:

$$\sigma_{ks} = \begin{cases} 1, \ sign[(y_k - y_s)(\hat{y}_k - \hat{y}_s)] = 1 \\ 0, \ \textit{в противном случае.} \end{cases}$$

Введем также неотрицательные вещественные переменные u_k, υ_k, характеризующие соответственно положительные и отрицательные значения ошибок аппроксимации ε_k в (1) в случае линейности функции F:

$$u_k = \begin{cases} y_k - \sum_{i=1}^{m} \alpha_i x_{ki}, \quad y_k > \sum_{i=1}^{m} \alpha_i x_{ki} \\ 0, \ \textit{в противном случае.} \end{cases}$$

$$\upsilon_k = \begin{cases} -y_k + \sum_{i=1}^{m} \alpha_i x_{ki}, \quad y_k < \sum_{i=1}^{m} \alpha_i x_{ki} \\ 0, \ \textit{в противном случае.} \end{cases}$$

Тогда заменяющая (4) и (5) задача ЧЦЛП примет вид:

$$D_1(\alpha) = -\sum_{k=1}^{n-1} \sum_{s=k+1}^{n} \sigma_{ks} + r \sum_{k=1}^{n} (u_k + v_k) \to \min, \tag{6}$$

$$\sum_{i=1}^{m} \alpha_i x_{ki} + u_k - \upsilon_k = y_k, \quad k = \overline{1,n}, \tag{7}$$

$$(y_k - y_s) \sum_{i=1}^{m} \alpha_i (x_{ki} - x_{si}) + M\sigma_{ks} \ge M, \tag{8}$$

$$\sigma_{ks} = 0,1, \ \ k = \overline{1,n-1}, \ s = \overline{k+1,n}, \tag{9}$$

$$\sum_{k=1}^{n} (u_k + \upsilon_k) \le I_1^* + \Delta I_1^*, \tag{10}$$

$$u_k \ge 0, \quad \upsilon_k \ge 0, \quad k = \overline{1,n}. \tag{11}$$

Здесь M – заранее выбранное большое отрицательное число.

В качестве константы r в функционале (6) может быть выбрано любое положительное число. Наличие второго слагаемого в (6) позволяет достичь совместного решения задач (4) и

(5) и, кроме того, обеспечивает выполнение условия $u_k \upsilon_k = 0$ для всех $k = \overline{1,n}$, необходимость реализации которого вытекает из определения переменных u_k и υ_k. Ограничение (10) из задачи можно и исключить, решая в таком случае сразу задачу одновременной оптимизации функции потерь и ОСП-критерия. При этом чем меньше значение r, тем большую значимость приобретает ОСП-критерий.

Неизвестными в задаче ЧЦЛП (6) - (11) являются вектора a, u, υ, σ с общей размерностью $m + 2n + n(n-1)/2$.

Для того, чтобы иметь возможность сравнения по критерию (3) регрессий, построенных на различных выборках, ему необходимо придать относительный характер посредством введения величины $\tilde{\Phi}_2(a) = 2\Phi_2(a) \cdot 100\% / (n(n-1))$.

Результаты проведенного сравнительного анализа эффективности использования МНК и МНМ, а также предложенного выше способа корректировки параметров на основе ОСП-критерия для ряда конкретных моделей показали следующее. Вычисленные в результате решения задачи (6) - (11) оценки параметров при возможном некотором ухудшении аппроксимационных характеристик, как правило, приводят к повышению точности прогноза зависимой переменной, значительно повышая при этом согласованность поведения фактической и расчетной траекторий изменения значений y как на обучающей, так и на экзаменующей выборках, что позволяет лучше «объяснять» исследуемый объект или процесс средствами моделирования.

Необходимость решения при этом задачи ЧЦЛП не является существенным препятствием вследствие наличия для современных ЭВМ большого количества соответствующих эффективных программных средств. Можно, в частности, указать на размещенную в Интернете в свободном доступе программу LPsolve.

Отметим, что ОСП - критерий позволяет помимо статистической учитывать также и экспертную информацию, выражающуюся, например, в назначении наблюдений выборки, установление согласованности для которых особенно необходимо. В этом случае переменные σ_{ks} в (6) следует умножить на соответствующие весовые множители.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Мудров В.И., Кушко В.А. Методы обработки измерений. Квазиправдоподобные оценки.- М.: Радио и связь, 1983. -304с.

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

3. Носков С.И. Построение эконометрических зависимостей с учетом критерия «согласованность поведения» // Кибернетика и системный анализ.-1994.-№1.-С.177-180.

4. Носков С.И. Критерий «согласованность поведения» в регрессионном анализе//Современные технологии. Системный анализ. Моделирование.-2013.-№1.-С.107-111.

5. Jingfei Yang M. Sc. Power System Short-term Load Forecasting: Thesis for Ph.d degree. Germany, Darmstadt, Elektrotechnik und Informationstechnik der Technischen Universitat, 2006. 139 p.

6. Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA Models / A.J. Conejo [at al.] // IEEE transaction on power systems. 2005, Vol. 20, No. 2. P. 1035 – 1042.

7. Armstrong J.S. Forecasting for Marketing // Quantitative Methods in Marketing. London: International Thompson Business Press, 1999. P. 92 – 119.

8. Draper N., Smith H. Applied regression analysis. New York: Wiley, In press, 1981. 693 p.

9. Pradhan R.P., Kumar R. Forecasting Exchange Rate in India: An Application of Artificial Neural Network Model // Journal of Mathematics Research. 2010, Vol. 2, No. 4. P. 111 – 117.

10. Yildiz B., Yalama A.,Coskun M. Forecasting the Istanbul Stock Exchange National 100 Index Using an Artificial Neural Network // An International Journal of Science, Engineering and Technology. 2008, Vol. 46. P.36 – 39.

11. Zhu J., Hong J., Hughes J.G. Using Markov Chains for Link Prediction in Adaptive Web Sites // 1st International Conference on Computing in an Imperfect World, UK, London, 2002. P. 60 – 73.

12. Singh S. Pattern Modelling in Time-Series Forecasting // Cybernetics and Systems-AnInternational Journal. 2000, Vol. 31, No. 1. P. 49 – 65.

REFERENCES

1. Mudrov VI, Kushko V.A. Methods for processing measurements. Quasi-like estimates .- M .: Radio and Communication, 1983.-304p.

2. Noskov S.I. The technology of modeling objects with unstable functioning and uncertainty in the data. Irkutsk: Oblinformpechat, 1996.-320 p.

3. Noskov S.I. Construction of econometric dependencies taking into account the criterion of "consistency of behavior" / / Cybernetics and system analysis.-1994.-No.1-P.177-180.

4. Noskov S.I. The criterion of "consistency of behavior" in regression analysis // Modern technologies. System analysis. Modeling.-2013.-No.1-P.107-111.

5. Jingfei Yang M. Sc. Power System Short-term Load Forecasting: Thesis for Ph.d degree. Germany, Darmstadt, Elektrotechnik und Informationstechnik der Technischen Universitat, 2006. 139 p.

6. Day-Ahead Electricity Price Forecasting Using the Wavelet Transform and ARIMA Models / A.J. Conejo [at al.] // IEEE transaction on power systems. 2005, Vol. 20, No. 2. P.P. 1035 – 1042.

7. Armstrong J.S. Forecasting for Marketing // Quantitative Methods in Marketing. London: International Thompson Business Press, 1999. P.P. 92 – 119.

8. Draper N., Smith H. Applied regression analysis. New York: Wiley, In press, 1981. 693 p.

9. Pradhan R.P., Kumar R. Forecasting Exchange Rate in India: An Application of Artificial Neural Network Model // Journal of Mathematics Research. 2010, Vol. 2, No. 4. P.P. 111 – 117.

10. Yildiz B., Yalama A.,Coskun M. Forecasting the Istanbul Stock Exchange National 100 Index Using an Artificial Neural Network // An International Journal of Science, Engineering and Technology. 2008, Vol. 46. P.P.36 – 39.

11. Zhu J., Hong J., Hughes J.G. Using Markov Chains for Link Prediction in Adaptive Web Sites // 1st International Conference on Computing in an Imperfect World, UK, London, 2002. P.P. 60 – 73.

12. Singh S. Pattern Modelling in Time-Series Forecasting // Cybernetics and Systems-AnInternational Journal. 2000, Vol. 31, No. 1. P.P. 49 – 65.

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. Обобщенный критерий согласованности поведения в регрессионном анализе // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2018. – №1. – С. 14-20 – Режим доступа:

http://ismm-irgups.ru/toma/11-2018, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 01.10.2018)

For citation

Noskov S.I. Obobshchennyj kriterij soglasovannosti povedeniya v regressionnom analize [Generalized criterion of coordination of behavior in regression analysis] // 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], 2018. No. 1. P. 14-20. [Accessed 01/10/18]