УДК 519.862.6

С.И. Носков, А.С. Вергасов

Прогнозирование по регрессионной модели с применением элементов теории сходства

Рассматривается одно из основных направлений практического применения математических моделей регрессионного типа, связанное с прогнозированием будущих значений зависимых переменных. Для повышения точности этих значений предлагается при оценивании неизвестных параметров регрессионной модели вместо обычного метода наименьших квадратов (МНК) использовать взвешенный метод наименьших квадратов (ВМНК). При этом при расчете весов наблюдений периода основания прогноза привлекается разработанная профессором Ю.А. Ворониным теория сходства, в соответствии с которой, чем более сходен вектор значений независимых переменных наблюдения периода упреждения с соответствующим вектором наблюдения периода основания, тем большим весом последнее должно обладать. Это соображение положено в основу предлагаемого в работе алгоритма вычисления весов. Подробно рассмотрен численный пример.

Ключевые слова: регрессионная модель, метод наименьших квадратов, взвешенный метод наименьших квадратов, теория сходства, прогнозирование.

doi: 10.21293/1818-0442-2019-22-3-67-70

Одним из основных направлений практического использования регрессионной модели является прогнозирование будущего развития исследуемого объекта.

Рассмотрим составной элемент такой модели – линейное регрессионное уравнение

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

где y_k и x_{ki} – k-е значения соответственно зависимой и i-й независимой переменных; $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_m)^T$ – вектор подлежащих оцениванию параметров; ε_k – ошибки аппроксимации; n – количество наблюдений (длина выборки).

Представим уравнение (1) в векторной форме:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}, \qquad (2)$$

где $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots \varepsilon_n)^T$, $\mathbf{X} = \|x_{ki}\|, k = \overline{1,n}, i = \overline{1,m}$.

Методам оценки параметров уравнения (1) посвящена весьма обширная литература (см., в частности, [1–9]). В рамках регрессионного анализа глубоко проработаны вопросы оценки адекватности моделей (см., например, [10–12]).

При известных оценках параметров уравнения (1) процесс прогнозирования будущих значений зависимой переменной состоит в подстановке в уравнение значений независимых переменных и последующем простом расчете $\mathbf{y}$. Формально эта процедура выглядит следующим образом.

Пусть известны значения независимых переменных на периоде упреждения прогноза $\tilde{x}_{ki}, k = \overline{n+1, n+\tau}, i = \overline{1,m}$. Тогда прогнозные значения зависимой переменной рассчитываются по формуле

$$\tilde{y}_k = \sum_{i=1}^{m} \alpha_i \tilde{x}_{ki}, k = \overline{n+1, n+\tau}, i = \overline{1,m}, \qquad (3)$$

где τ – длина периода упреждения.

Такая простая расчетная схема прогнозирования применяется для большинства известных регрессионных моделей.

Вместе с тем? эта схема не учитывает одно крайне важное обстоятельство, а именно, в какой мере векторы $\tilde{\mathbf{x}}_k = (\tilde{x}_{k1}, \ldots, \tilde{x}_{km}),\ k = \overline{n+1, n+\tau}$ соответствуют (подобны, сходны) каждому из векторов предыстории процесса $\mathbf{x}_k = (x_{ki}, \ldots, x_{kn}),\ k = \overline{1,n}$, то есть каждой строке матрицы $\mathbf{X}$, поскольку вполне естественно полагать, что чем выше такое соответствие, тем в большей степени тенденции, характерные для данного наблюдения исходной выборки, проявятся и на периоде упреждения прогноза.

Формализовать такой учет соответствия тенденций на периодах основания $\{1, 2, \ldots, n\}$ и упреждения $\{n+1, n+2, \ldots, n+\tau\}$ можно, используя взвешенные методы оценивания параметров уравнения (1), например, взвешенный метод наименьших квадратов (ВМНК) и теорию сходства, разработанную Ю.А. Ворониным [13].

Вектор параметров $\boldsymbol{\alpha}$ в уравнении (2) при использовании ВМНК рассчитывается по формуле

$$\tilde{\boldsymbol{\alpha}} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}, \qquad (4)$$

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

Выявление упомянутого сходства можно осуществлять следующим образом. Заметим, что впервые идея такого подхода предложена в [14].

Введем в рассмотрение матрицу $\mathbf{Z}_{ki}, k = \overline{1, n+\tau}$, $i = \overline{1,m}$ элементы которой имеют вид

$$z_{ki} = x_{ki}, k = \overline{1,n}, i = \overline{1,m},$$

$$z_{ki} = \tilde{x}_{ki}, k = \overline{n+1, n+\tau}, i = \overline{1,m}.$$

Введем также в рассмотрение переменные $z_i^{\min}$ и $z_i^{\max}$ по правилу

$$z_i^{\min} = \min_{k=\overline{1,n+\tau}} z_{ki}, \quad z_i^{\max} = \max_{k=\overline{1,n+\tau}} z_{ki}.$$

Преобразуем матрицу **Z** в $\tilde{\mathbf{Z}}$ следующим образом:

$$\tilde{z}_{ki} = \frac{z_{ki} - z_i^{\min}}{z_i^{\max} - z_i^{\min}}, k = \overline{1,n+\tau}, i = \overline{1,m}.$$

Очевидно, что $\tilde{z}_{ki} \in [0,1]$ для всех k и i.

Для того чтобы оценить, в какой степени s-е наблюдение периода упреждения прогноза сходно (подобно) с k-м наблюдением периода основания, воспользуемся одной из описанных в [13] мер сходства. Например, такой:

$$\omega(s,k) = 1 - \frac{1}{m}\sum_{i=1}^{m} |\tilde{z}_{si} - \tilde{z}_{ki}|,$$

$$s \in \{n+1,...,n+\tau\}, k \in \{1,...,n\}.$$

Отметим, что в [13] приведено еще девять возможных мер сходства.

После этого для расчета прогнозного значения зависимой переменной y_s необходимо вначале рассчитать оценку $\tilde{\boldsymbol{\alpha}}$ по формуле (4) с матрицей весов **W**, где

$$w_k = \omega(s,k), \tag{5}$$

с фиксированной s, а затем вычислить $\tilde{y}_s$ по формуле (3) с этой оценкой, т.е. при $\boldsymbol{\alpha} = \tilde{\boldsymbol{\alpha}}$.

Эту операцию необходимо проделать τ раз для всего периода упреждения прогноза.

Рассмотрим пример применения предложенного способа прогнозирования с использованием опубликованных в работе [15] известных данных, которые описывают работу установки для окисления аммиака в азотную кислоту и состоят из 21 наблюдения. Переменными модели при этом являются так называемый Stackloss (**y**), который зависит от скорости работы установки x_1, температуры охлаждающей воды на входе x_2 и концентрации кислоты x_3.

Статистическая информация представлена в табл. 1.

Прежде всего разобьем эту выборку на периоды основания и упреждения прогноза с номерами 1–13 и 14–21 соответственно. То есть положим n = 13, τ =8. На первой из них построим трехфакторную регрессию без свободного члена с помощью МНК:

$$y = 0{,}83x_1 + 1{,}17x_2 - 0{,}65x_3, \tag{6}$$

R = 0,98, F = 180, **T** = (4,1; 2,3; –6,4),

где R – критерий множественной детерминации, F – критерий Фишера, **T** – вектор значений критерия Стьюдента для каждого из параметров. Анализ значений этих критериев указывает на высокую адекватность уравнения (6).

Применим далее для прогнозирования значений зависимой переменной на периоде упреждения с использованием как обычного МНК, так и ВМНК на основе вычисления весов наблюдений по формуле (5). Весовые коэффициенты для каждого из 13 наблюдений периода основания и каждого наблюдения периода упреждения приведены в табл. 2.

Таблица 1

Статистическая информация

№	y	x_1	x_2	x_3
1	42	80	27	89
2	37	80	27	88
3	37	75	25	90
4	28	62	24	87
5	18	62	22	87
6	18	62	23	87
7	19	62	24	93
8	20	62	24	93
9	15	58	23	87
10	14	58	18	80
11	14	58	18	89
12	13	58	17	88
13	11	58	18	82
14	12	58	19	93
15	8	50	18	89
16	7	50	18	86
17	8	50	19	72
18	8	50	19	79
19	9	50	20	80
20	15	56	20	82
21	15	70	20	91

Таблица 2

Весовые коэффициенты наблюдений

ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	ω_{11}	ω_{12}	ω_{13}
0,29	0,27	0,46	0,61	0,68	0,65	0,77	0,77	0,71	0,63	0,86	0,80	0,68
0,36	0,34	0,46	0,61	0,68	0,64	0,56	0,56	0,69	0,68	0,91	0,85	0,73
0,28	0,31	0,38	0,64	0,70	0,67	0,48	0,48	0,71	0,75	0,83	0,82	0,80
0,13	0,14	0,23	0,46	0,52	0,49	0,36	0,36	0,53	0,75	0,60	0,59	0,71
0,16	0,18	0,26	0,50	0,57	0,54	0,36	0,36	0,58	0,85	0,63	0,63	0,80
0,20	0,22	0,29	0,55	0,62	0,58	0,40	0,40	0,63	0,84	0,61	0,60	0,79
0,25	0,27	0,36	0,65	0,72	0,68	0,50	0,50	0,74	0,85	0,72	0,71	0,90
0,56	0,53	0,73	0,64	0,70	0,67	0,69	0,69	0,61	0,46	0,70	0,64	0,52

В табл. 3 и 4 приведены нижние и верхние значения весовых коэффициентов для наблюдений периодов основания и упреждения.

В табл. 5 приведены восемь значений вектора параметров уравнения (6), оцененных по ВМНК.

Таблица 3

Нижние и верхние значения весовых коэффициентов наблюдений периода основания

Вес	ω_1	ω_2	ω_3	ω_4	ω_5	ω_6	ω_7	ω_8	ω_9	ω_{10}	ω_{11}	ω_{12}	ω_{13}
Нижняя граница	0,13	0,14	0,23	0,46	0,52	0,49	0,36	0,36	0,53	0,46	0,6	0,59	0,68
№ наблюдения	4	4	4	4	4	4	4,5	4,5	4	8	4	4	1
Верхняя граница	0,56	0,53	0,73	0,65	0,72	0,68	0,77	0,77	0,74	0,85	0,91	0,85	0,9
№ наблюдения	8	8	8	7	7	7	1	1	7	5,7	2	2	7

Таким образом, значения параметров при использовании ВМНК лежат в интервалах:

$$\alpha_1 \in [0{,}71; 0{,}88],$$

$$\alpha_2 \in [1{,}07; 1{,}23],$$

$$\alpha_3 \in [-0{,}65; -0{,}62].$$

В табл. 6 приведены фактические и расчетные прогнозные $\tilde{y}_k$ (МНК), $\tilde{y}_k$ (ВМНК), $k=\overline{14,21}$ значения зависимой переменной, вычисленные с использованием обычного МНК и ВМНК.

Таблица 4

Нижние и верхние значения весовых коэффициентов наблюдений периода упреждения

№ наблюдения	1	2	3	4	5	6	7	8
Нижняя граница	0,27	0,34	0,28	0,13	0,16	0,20	0,25	0,53
Верхняя граница	0,86	0,91	0,83	0,75	0,85	0,84	0,90	0,73

Таблица 5

Значения параметров уравнения

α_1	α_2	α_3
0,84	1,09	–0,64
0,84	1,08	–0,64
0,81	1,12	–0,63
0,77	1,20	–0,62
0,77	1,22	–0,63
0,78	1,23	–0,64
0,79	1,21	–0,64
0,86	1,07	–0,65

Таблица 6

Фактические и расчетные значения *y*

№	y_k (факт.)	$\tilde{y}_k$ (МНК), (расч.)	$\tilde{y}_k$ (ВМНК), (расч.)
14	12	9,535	9,809
15	8	4,331	4,545
16	7	6,297	6,425
17	8	16,640	16,405
18	8	12,053	11,980
19	9	12,568	12,494
20	15	16,249	15,976
21	15	21,999	22,617

Из последней таблицы следует, что для семи наблюдений периода упреждения из восьми ВМНК дает более точный прогноз, чем МНК. И лишь для последнего, 21-го, наблюдения – это не так.

Анализ результатов расчетов указывает на то, что предложенный в работе способ взвешивания наблюдений выборки, основанный на теории сходства, позволяет получать более точные прогнозы по сравнению с обычным МНК. Вместе с тем авторы отдают себе отчет в том, что на других данных такое «подавляющее преимущество» ВМНК над МНК может и не иметь места. В любом случае чем более широким арсеналом методов моделирования исследователь располагает, тем более точную модель анализируемого объекта он построит.

Литература

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

2. Golovchenko V.B. Estimation of an econometric model using statistical data and expert information / V.B. Golovchenko, S.I. Noskov // Automation and Remote Control. – 1991. – No. 4. – P. 123–132.

3. Базилевский М.П. Идентификация неизвестных параметров линейно-мультипликативной регрессии / М.П. Базилевский, С.И. Носков // Современные наукоемкие технологии. – 2012. – № 3. – С. 14.

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

5. Носков С.И. Регрессионная модель динамики эксплуатационных показателей функционирования железнодорожного транспорта / С.И. Носков, И.П. Врублевский // Современные технологии. Системный анализ. Моделирование. – 2016. – № 2(50). – С. 192–197.

6. Носков С.И. Оценивание параметров аппроксимирующей функции с постоянными пропорциями // Современные технологии. Системный анализ. Моделирование. – 2013. – № 2 (38). – С. 135–136.

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

8. Головченко В.Б. Прогнозирование на основе дискретной динамической модели с использованием дискретной информации / В.Б. Головченко, С.И. Носков // Автоматика и телемеханика. – 1991. – № 4. – С. 140.

9. Ильина Н.К. Идентификация параметров некоторых негладких регрессий / Н.К. Ильина, С.А. Лебедева, С.И. Носков // Информационные технологии и проблемы математического моделирования сложных систем. – 2016. – № 17. – С. 111.

10. Носков С.И. Обобщенный критерий согласованности поведения в регрессионном анализе // Информационные технологии и математическое моделирование в управлении сложными системами. – 2018. – № 1 (1). – С. 14–20.

11. Баенхаева А.В. Выбор структурной спецификации регрессионной модели валового регионального продукта Иркутской области / А.В. Баенхаева, М.П. Базилевский, С.И. Носков // Информационные технологии и проблемы математического моделирования сложных систем. – 2016. – № 16. – С. 31–38.

12. Носков С.И. Построение регрессионных моделей с использованием аппарата линейно-булевого программирования / С.И. Носков, М.П. Базилевский. – Иркутск, 2018. – 176 с.

13. Воронин Ю.А. Начало теории сходства. – Новосибирск: ВЦ СО АН СССР, 1989. – 224 с.

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

15. Dodge Y. The guinea pig of multiple regression. In Robust Statistics, Data Analysis, and Computer Intensive Methods // Springer Lecture Notes in Statistics (New York, Springer-Verlag). – 1999. – Vol. 109. – P. 91–117.

Носков Сергей Иванович
Д-р тех. наук, профессор каф.
информационных систем и защиты информации
Иркутского государственного ун-та путей сообщения
Чернышевского ул., д. 15, г. Иркутск, Россия, 664074
Тел.: +7-914-902-24-94
Эл. почта: noskov_s@irgups.ru

Вергасов Александр Сергеевич
Ассистент каф. информационных систем
и защиты информации
Иркутского государственного ун-та путей сообщения
Чернышевского ул., д. 15, г. Иркутск, Россия, 664074
Тел.: +7-999-641-93-46
Эл. почта: tluck@inbox.ru

Noskov S.I., Vergasov A.S.

Predicting a regression model using elements of the theory of similarity

The article discusses one of the main directions of practical application of mathematical models of regression type, associated with the prediction of future values of dependent variables. To improve the accuracy of these values, it is proposed, when estimating unknown parameters of the regression model, instead of the usual least squares method (LSM), to use the weighted least squares method (HMSC). At the same time, when calculating the weights of observations of the base period of the forecast, the theory of similarity developed by Professor Yu.A. Voronin is used, according to which the more similar the vector of values of independent variables of the lead-time observation is with the corresponding vector of observation of the base period, the greater the weight the latter must possess. This consideration is the basis of the weighting algorithm proposed in the paper. Considered in detail a numerical example.

Keywords: regression model, least squares method, weighted least squares method, theory of similarity, forecasting.
doi: 10.21293/1818-0442-2019-22-3-67-70

References

1. Noskov S.I. Tochechnaya harakterizaciya mnozhestva Pareto v linejnoj mnogokriterial'noj zadache. [*Sovremennye tekhnologii. Sistemnyj analiz. Modelirovanie*], 2008, no. 1(17), pp. 99–101 (in Russ.).

2. Golovchenko V.B., Noskov S.I. Estimation of an econometric model using statistical data and expert information. *Automation and remote control*, 1991, no. 4, pp. 123–132 (in Russ.).

3. Bazilevsky M.P., Noskov S.I. Identification of unknown parameters of linear-multiplicative regression. Modern technologies. *System analysis. Modeling*, 2012, no. 3, 14 p. (in Russ.).

4. Lakeev A.V., Noskov S.I. The method of least modules for linear regression: the number of zero approximation errors. *Technologies. System analysis. Modeling*., 2012, no. 2(34), pp. 48–50 (in Russ.).

5. Noskov S.I., Vrublevsky I.P. Regression model of the dynamics of operational indicators of railway transport functioning. *Modern technologies. System analysis. Modeling*., 2016, no. 2 (50), pp. 192–197 (in Russ.).

6. Noskov S.I. Estimation of parameters of an approximating function with constant proportions. *Modern technologies. System analysis. Modeling*, 2013, no. 2(38), pp. 135–136 (in Russ.).

7. Noskov S.I., Baenkhaeva A.V. Mnozhestvennoe ocenivanie parametrov linejnogo uravneniya. *Modern technologies. System analysis. Modeling*., 2016, no. 3(51), pp. 133–138 (in Russ.).

8. Golovchenko V.B., Noskov S.I. Prognozirovanie na osnove diskretnoj dinamicheskoj mo-deli s ispol'zovaniem diskretnoj informacii *Automation and Remote Control*, 1991, no. 4, p. 140 (in Russ.).

9. Ilina N.K., Lebedeva S.A., Noskov S.I. Identification of parameters of some non-smooth regressions *[Informacionnye tekhnologii i matematicheskoe modeliro-vanie v upravlenii slozhnymi sistemami]*, 2016, no. 17, p. 111 (in Russ.).

10. Noskov S.I. The generalized criterion for the consistency of behavior in regression analysis [*Informacionnye tekhnologii i matematicheskoe modeliro-vanie v upravlenii slozhnymi sistemami*], 2018, no. 1 (1), pp. 14–20 (in Russ.).

11. Baenkhaeva A.V., Bazilevsky M.P., Noskov S.I. The choice of the structural specification of the regression model of the gross regional product of the Irkutsk region [I*nformacionnye tekhnologii i matematicheskoe modeliro-vanie v upravlenii slozhnymi sistemami*], 2016, no. 16, pp. 31–38 (in Russ.).

12. Noskov S.I. Bazilevsky M.P. *Building regression models using linear-boolean programming*. Monograph, Irkutsk, 2018, 176 p. (in Russ.).

13. Voronin Yu.A. [*Nachalo teorii skhodstva*], Novosibirsk: EC of the USSR Academy of Sciences Academy of Sciences, 1989. 224 p. (in Russ.).

14. Noskov SI, Vergasov AS Implementation of the weighted least squares method using measures of similarity. *Bulletin of science and education*, 2018, no. 18–1(54), pp. 29–32 (in Russ.).

15. Dodge Y. The guinea pig of multiple regression. In Robust Statistics, Data Analysis, and Computer Intensive Methods. Springer Lecture Notes in Statistics (New York, Springer-Verlag), 1996, vol. 109, pp. 91–117.

Sergey I. Noskov
Doctor of Engineering Science, Professor, Sub-department
Information Systems and Information Security,
Irkutsk State Transport University
15, Chernyshevsky st., Irkutsk, Russia, 664074
Phone: +7-914-902-24-94
Email: noskov_s@irgups.ru

Aleksandr S. Vergasov
Applicant, Department of Information Systems and
Information Protection, Irkutsk State Transport University
15, Chernyshevsky st., Irkutsk, Russia, 664074
Phone: +7-999-641-93-46
Email: tluck@inbox.ru