УДК 519.711

СПОСОБ УТОЧНЕНИЯ ИЕРАРХИЧЕСКОЙ ОДНОРОДНОЙ МАТЕМАТИЧЕСКОЙ МОДЕЛИ СТАТИСТИЧЕСКОГО ТИПА

[1]Носков С.И., [2]Торопов В.Д., [3]Носкова Н.С.

[1]ФГБОУ ВПО «Иркутский государственный университет путей сообщения», Иркутск, e-mail: sergey.noskov.57@mail.ru;

[2]ФГБОУ ВПО «Байкальский государственный университет экономики и права», Иркутск, e-mail: vdtor@yandex.ru;

[3]Тяньцзиньский университет науки и технологий, Вэй Цзинь Road, Нанкай района, Тяньцзинь, e-mail: nnadusha@mail.ru

Часто в практике моделирования исследователям приходится сталкиваться с ситуацией, когда объект анализа (система) естественным образом «распадается» на составляющие (подсистемы), поведение которых также представляет интерес. При этом, следуя канонам системного анализа, как правило, исходная система формализуется методами математического моделирования «лучше», адекватнее, чем её составные части. Более того, один из принципов системного анализа гласит: «Не элементы составляют систему, а система распадается на элементы». Это означает, что чем крупнее объект исследования, тем более точную модель можно построить, тем инерционнее будет этот объект, тем устойчивее будут закономерности его функционирования. Небольшой, несколько утрированный, но показательный пример: поведение водоема «смоделировать» гораздо проще, чем поведение отдельной молекулы.

Ключевые слова: моделирование, системный анализ, регрессионный анализ, критерий Фишера, критерий Дарбина – Уотсона, ошибка аппроксимации

WAY OF SPECIFICATION OF HIERARCHICAL UNIFORM MATHEMATICAL MODEL OF STATISTICAL TYPE

[1]Noskov S.I., [2]Toropov V.D., [3]Noskova N.S.

[1]FGBOU VPO «Irkutsk State University of Means of Communication», Irkutsk, e-mail: sergey.noskov.57@mail.ru;

[2]FGBOU VPO «Baikal state university of economy and right», Irkutsk, e-mail: vdtor@yandex.ru;

[3]Tianjin university of science and technology. PR China, Tianjin, Dagu Nan road, e-mail: nnadusha@mail.ru

Often in practice, simulation researchers have to deal with the situation when the object of analysis (system) in a natural way «breaks up» into components (subsystems), whose behavior is also of interest. However, following the canons of system analysis, as a rule, the original system is formalized methods of mathematical modeling «better», more appropriately than its constituent parts. Moreover, one of the principles of system analysis States: «the elements that constitute the system, and the system breaks down into elements». This means that the larger the object of study, the more accurate model can be built, inserzione will be this object, the stronger the laws governing its functioning. A small, somewhat exaggerated, but illustrative example: the behavior of the reservoir «model» is much simpler than the behavior of individual molecules.

Keywords: modeling, system analysis, regression analysis, Fischer's criterion, Darbin-Watson's criterion, approximation error

В практике моделирования исследователям часто приходится сталкиваться с ситуацией, когда объект анализа (система) естественным образом «распадается» на составляющие (подсистемы), поведение которых также представляет интерес.

При этом, следуя канонам системного анализа, как правило, исходная система формализуется методами математического моделирования «лучше», адекватнее, чем её составные части. Более того, один из принципов системного анализа гласит: «Не элементы составляют систему, а система распадается на элементы». Это означает, что чем крупнее объект исследования, тем более точную модель можно построить, чем инерционнее будет этот объект, тем устойчивее будут закономерности его функционирования.

Небольшой, несколько утрированный, но показательный пример: поведение водоема «смоделировать» гораздо проще, чем поведение отдельной молекулы.

Итак, пусть необходимо построить иерархическую (для определенности, двухуровневую) одномерную статистическую модель, описывающую переменную z регрессией

$$z_k = f\left(\alpha; x_k\right) + \varepsilon_k, \quad k = \overline{1,n}, \qquad (1)$$

где k – номер наблюдения обрабатываемой выборки длины n; f – вещественная аппроксимирующая функция; α – вектор оценива-

емых параметров; x_k – вектор экзогенных переменных модели; ε_k – ошибки аппроксимации. Пусть исследуемый объект (допустим, каскад электростанций; z – объем произведенной на нем электроэнергии) состоит из r составных частей (отдельных электростанций), для каждой из которых построена своя модель:

$$z_k^j = f^j(\alpha^j, x_k^j) + \varepsilon_k^j, \quad k = \overline{1,n}, \qquad j = \overline{1,r}. \quad (2)$$

Обозначения здесь очевидны.

При оценивании параметров регрессионного уравнения (2) и оценки его качества можно использовать, в частности, следующие работы [1–5, 7, 8, 10].

На предыстории процесса выполняется естественное равенство

$$z_k = \sum_{j=1}^{r} z_k^j, \quad k = \overline{1,n}. \quad (3)$$

Как правило, одно из основных направлений использования моделей (1), (2) – прогнозирование с их помощью будущего состояния объектов.

Обозначим через $\hat{z}_k, \hat{z}_k^j$, $j = \overline{1,r}$, $k > n$ прогнозные значения соответствующих переменных. При этом в общем случае равенство (3) может нарушаться, т.е.

$$\hat{z}_k \neq \sum_{j=1}^{r} \hat{z}_k^j, k > n.$$

Возникает вопрос: что делать с образующимся дисбалансом

$$\Delta z_k = \hat{z}_k - \sum_{j=1}^{r} \hat{z}_k^j ?$$

Для его решения воспользуемся идеей, высказанной в [5], а именно: дисбаланс Δz_k должен быть распределен между r объектами обратно пропорционально точности соответствующих моделей. То есть чем более адекватна j-я модель (2), тем меньшая часть Δz_k должна быть направлена на корректировку значения $\hat{z}_k^j$.

К настоящему времени в регрессионном анализе разработан широкий спектр критериев адекватности статистических моделей (достаточно представительный их перечень приведен в [8]). Это, в частности, коэффициент множественной детерминации, критерии Фишера, Стьюдента и Дарбина – Уотсона, средние относительные ошибки аппроксимации и прогноза, критерии смещения и согласованности поведения и другие. Приведем эти критерии более подробно, используя, в частности, работы [11–14]. При этом во избежание путаницы зависимую переменную будем обозначать через y.

а) критерий множественной детерминации R^2, выражающий степень согласованности вычисленных и фактических значений зависимой переменной, поскольку он представляет собой квадрат коэффициента корреляции между соответствующими векторами. Эквивалентная, по существу, трактовка R^2 такова: он показывает, какая доля дисперсии у объясняется регрессией (1).

Формула расчета R^2 имеет вид (в случае присутствия в (1) свободного члена):

$$R^2 = \frac{\sum_{k=1}^{n} (\hat{y}_k - \overline{y})^2}{\sum_{j=1}^{n} (y_j - \overline{y})^2};$$

или, что то же

$$R^2 = 1 - \frac{\sum_{k=1}^{n} \varepsilon_k^2}{\sum_{j=1}^{n} (y_j - \overline{y})^2},$$

где $\overline{y}_k$, $k = \overline{1,n}$ – вычисленные значения зависимой переменной, $\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ – среднее значение y. Легко видеть, что всегда $R^2 \in [0, 1]$.

Другая, формально более аргументированная интерпретация критерия множественной детерминации состоит в том [8], что он показывает, насколько регрессия (1) лучше модели среднего. Поэтому при описании динамических процессов с помощью регрессии, содержащей трендовую составляющую

$$y_k = a_0 + \sum_{i=1}^{m} a_i x_{ki} + a_{m+1} k + \varepsilon_k,$$

для того, чтобы сравнить, насколько такая регрессия лучше модели простого тренда $y_k = \beta_0 + \beta_1 k + \delta_k$, в [8] предлагается использовать в качестве аналога R^2 показатель R_T^2, рассчитываемый по формуле

$$R_T^2 = 1 - \frac{\sum_{k=1}^{n} \varepsilon_k^2}{\sum_{j=1}^{n} \delta_j^2}.$$

Существенным недостатком R^2 является то, что он не уменьшает своих значений при добавлении в (1) новых переменных. Поэтому всегда можно сделать R^2 как угодно близким к единице путем добавления в регрессию дополнительных независимых переменных. Для элиминирования этого недостатка часто вместо R^2 используют его

скорректированное на число степеней свободы значение $\overline{R}^2$:

$$\overline{R}^2 = 1 - \frac{n-1}{n-m}\left(1-R^2\right);$$

б) величина остаточной дисперсии s^2, определяющая меру вариации выходного показателя относительно регрессии:

$$s^2 = \frac{1}{n-m}\sum_{k=1}^{n}\varepsilon_k^2.$$

Для придания остаточной дисперсии относительного характера иногда пользуются следующей формулой:

$$s^2 = \frac{1}{n-m}\frac{\sum_{k=1}^{n}\varepsilon_k^2}{\overline{y}^2}\cdot 100\%;$$

в) F-критерий Фишера, показывающий отношение дисперсии фактических значений у к остаточной дисперсии. В зависимости от существующих вариантов интерпретации этого критерия он указывает на: отсутствие (или наличие) линейной связи зависимой переменной с одной из независимых; значимость критерия R^2; степень линейности уравнения. F-критерий рассчитывают по формуле

$$F = \frac{R^2\left(n-m\right)}{\left(1-R^2\right)\left(m-1\right)}.$$

F-критерий имеет статистический характер и требует использования соответствующих таблиц F-распределения. При превышении значения F над табличным первое считается удовлетворительным. В любом случае значение F-критерия тем лучше, чем оно выше;

г) критерий Дарбина –Уотсона d, указывающий на наличие или отсутствие корреляции (положительной или отрицательной) остатков ε:

$$d = \frac{\sum_{k=2}^{n}\left(\varepsilon_k - \varepsilon_{k-1}\right)^2}{\sum_{k=1}^{n}\varepsilon_k^2}.$$

Критерий d принимает значения на отрезке [0, 4] и также требует привлечения соответствующих статистических таблиц. Идеальное его значение, указывающее на отсутствие автокорреляции остатков, равно двум.

д) t-критерий, показывающий, во сколько раз оцененное значение каждого параметра регрессии (1) превышает его стандартную ошибку. Этот критерий служит мерой вариации каждого параметра и так же, как критерия f и d, требует привлечения соответствующих таблиц распределения, в данном случае Стьюдента. Формула для его расчета имеет вид

$$t_i = \frac{\hat{a}_i}{\sqrt{\left(X^TX\right)^{-1}_{ii}}\,s}, \quad i = \overline{1,m}$$

где $\hat{a}_i$ – оцененное значение i-го параметра регрессии (1), $\left(X^TX\right)^{-1}_{ii}$ – i-й диагональный элемент матрицы $\left(X^TX\right)^{-1}$.

Удовлетворительным считается значение t-критерия, превышающее единицу. В противном случае соответствующий регрессор считается незначимым;

е) средняя относительная ошибка аппроксимации E. Это обычно применяемый в инженерных расчетах показатель, вычисляемый по формуле

$$E = \frac{1}{n}\sum_{k=1}^{n}\left|\frac{y_k - \hat{y}_k}{y_k}\right|\cdot 100\ \%;$$

ж) средняя относительная ошибка прогноза $\overline{\varepsilon}$, рассчитанная по экзаменующей выборке следующим образом.

Вся выборка делится на две части – большую (обучающую) с номерами наблюдений 1, 2, …, τ и меньшую (экзаменующую) с номерами τ+1, τ+2,…,n. По наблюдениям обучающей выборки определяются параметры регрессии (1) $\hat{a}$, после чего значение $\overline{\varepsilon}$ вычисляется по формуле

$$\overline{\varepsilon} = \frac{1}{n-\tau}\sum_{k=\tau+1}^{n}\frac{\left|y_k - \sum_{i=1}^{m}\hat{a}_i x_{ki}\right|}{\left|y_k\right|}.$$

Данный критерий служит оценкой прогностических возможностей построенного уравнения. Естественно, после их проверки параметры регрессии пересчитываются по полной выборке;

з) критерий смещения $n_{см}$, представляющий собой меру «стабильности» оценок параметров относительно различных участков выборки.

Для расчета этого критерия вся выборка делится примерно пополам, и для каждой части определяются оценки параметров $\hat{a}^1$ и $\hat{a}^2$.

Тогда значение $n_{см}$ принимает вид

$$n_{см} = \frac{\sum_{k=1}^{n}\left(\sum_{i=1}^{m}\hat{a}_i^1 x_{ki} - \sum_{i=1}^{m}\hat{a}_i^2 x_{ki}\right)^2}{\sum_{k=1}^{n}y_k^2}.$$

Из этой формулы следует, что значение критерия $n_{см}$ чем меньше, тем лучше.

Необходимо отметить, что обычно в литературе использование приведенных

критериев носит традиционно пассивный характер, выражающийся в неформальном доказательстве удовлетворительности качества построенной модели в случае попадания значений критериев в требуемые интервалы.

Пусть в качестве основного нами выбран критерий «средняя относительная ошибка аппроксимации». Рассчитаем его значение для каждой из r моделей (2):

$$\mathrm{E}^j = \frac{1}{n}\sum_{k=1}^{n}\left|\frac{z_k^j - \hat{z}_k^j}{z_k^j}\right| \cdot 100\ \%.$$

Вычислим корректирующие элементы по очевидной формуле:

$$\omega^j = \frac{\mathrm{E}^j}{\sum_{i=1}^{r}\mathrm{E}^i}, \qquad j = \overline{1,r}.$$

После этого прогнозные значения $\hat{z}_k^j$ уточняются следующим образом:

$$\tilde{z}_k^j = \hat{z}_k^j + \omega^j \Delta z_k, \ \ j = \overline{1,r}, \ \ k > n.$$

В своих последующих работах авторы намерены продолжить развивать предлагаемый здесь подход.

Список литературы

1. Базилевский М.П. Носков С.И. Алгоритм формирования множества регрессионных моделей с помощью преобразования зависимой переменной // Международный журнал прикладных и фундаментальных исследований. – 2010. – № 3. – С. 159–160.

2. Базилевский М.П. Носков С.И. Алгоритм построения линейно-мультипликативной регрессии // Современные технологии. Системный анализ, Моделирование. – 2011. – № 1. – С. 88–92.

3. Базилевский М.П. Носков С.И. Идентификация неизвестных параметров линейно-мультипликативной регрессии // Современные наукоемкие технологии. – 2012. – № 3. – С. 14–18.

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

5. Матросов В.М., Головченко В.Б., Носков С.И. Моделирование и прогнозирование показателей социально-экономического развития области. – Новосибирск: Наука, 1991. – С. 144.

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

7. Носков С.И. Оценивание параметров аппроксимирующей функции с постоянными пропорциями // Современные технологии. Системный анализ, Моделирование. – 2013. – № 2. – С. 135–136.

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

9. Протопопов В.А. Носков С.И. Оценка уровня уязвимости объектов транспортной инфраструктуры: формализованный подход // Современные технологии. Системный анализ, Моделирование. – 2011. – № 4 (32). – С. 241–244.

10. A Description of the set of solutions of a linear equation with interval defined operator and right-hand side /A. V. Lakeev, S.I.Noskov // Doklad Mathematics. – 1993. – T.47, № 3.

11. Approximate linear algebra is intractable / V. Kreinovich, A.V. Lakeev, S.I. Noskov // Linear algebra and its Applications. – 1996. – T. 232, № 1–3. – P. 45–54.

12. Description of the solution set to linear equation with the intervally defined operator and right-hand side /A.V. Lakeev, S.I. Noskov // Doklad Mathematics. – 1993. – т. 330, № 4. – P. 430.

References

1. Bazilevskij M.P. Noskov S.I. Algoritm formirovanija mnozhestva regressionnyh modelej s pomoshhju preobrazovanija zavisimoj peremennoj. Mezhdunarodnyj zhurnal prikladnyh i fundamentalnyh issledovanij no. 3, 2010. pp. 159–160.

2. Bazilevskij M.P. Noskov S.I. Algoritm postroenija linejno-multiplikativnoj regressii Sovremennye tehnologii. Sistemnyj analiz, Modelirovanie no. 1, 2011. pp. 88–92.

3. Bazilevskij M.P. Noskov S.I. Identifikacija neizvestnyh parametrov linejno-multiplikativnoj regressii. Sovremennye naukoemkie tehnologii no. 3, 2012. pp. 14–18.

4. Lakeev A.V. Noskov S.I. Metod naimenshih modulej dlja linejnoj regressii: chislo nulevyh oshibok approksimacii. Sovremennye tehnologii. Sistemnyj analiz, Modelirovanie no. 2, 2012. pp. 48–50.

5. Matrosov V.M., Golovchenko V.B., Noskov S.I. Modelirovanie i prognozirovanie pokazatelej socialno-jekonomicheskogo razvitija oblasti. Novosibirsk: Nauka, 1991. pp. 144.

6. Noskov S.I. Kriterij «soglasovannost povedenija» v regressionnom analize. Sovremennye tehnologii. Sistemnyj analiz, Modelirovanie no. 1, 2013. pp. 107–111.

7. Noskov S.I. Ocenivanie parametrov approksimirujushhej funkcii s postojannymi proporcijami. Sovremennye tehnologii. Sistemnyj analiz, Modelirovanie no. 2, 2013. pp. 135–136.

8. Noskov S.I. Tehnologija modelirovanija obektov s nestabilnym funkcionirovaniem i neopredelennostju v dannyh. Irkutsk: Oblinformpechat, 1996. pp. 320.

9. Protopopov V.A. Noskov S.I. Ocenka urovnja ujazvimosti obektov transportnoj infrastruktury: formalizovannyj podhod. Sovremennye tehnologii. Sistemnyj analiz, Modelirovanie no. 4 (32), 2011. pp. 241–244.

10. A Description of the set of solutions of a linear equation with interval defined operator and right-hand side / A.V. Lakeev, S.I. Noskov / Doklad Mathematics, T.47, no. 3, 1993.

11. Approximate linear algebra is intractable / V. Kreinovich, A.V. Lakeev, S.I. Noskov / Linear algebra and its Applications, T.232, no. 1–3, 1996. pp. 45–54.

12. Description of the solution set to linear equation with the intervally defined operator and right-hand side /A.V. Lakeev, S.I. Noskov / Doklad Mathematics, t.330, no. 4, 1993. pp. 430.

Рецензенты:

Кузьмин О.В., д.ф.-м.н., профессор, заведующий кафедрой теории вероятностей и дискретной математики, Иркутский государственный университет, г. Иркутск;

Лакеев А.В., д.ф.-м.н., ведущий научный сотрудник, Институт динамики систем и теории управления СО РАН, г. Иркутск.

Работа поступила в редакцию 18.03.2015.