УДК 519.2:681.3

В.Е. Бахрушин

МЕТОДЫ ОЦЕНИВАНИЯ ХАРАКТЕРИСТИК НЕЛИНЕЙНЫХ СТАТИСТИЧЕСКИХ СВЯЗЕЙ

В работе предложен новый метод оценивания силы нелинейной статистической связи, основанный на оценивании значения коэффициента детерминации с использованием значений неизвестной функции регрессии, полученных методом скользящих средних. Такой подход позволяет повысить точность и устойчивость оценок и, таким образом, получить более надежную информацию о наличии и силе связи.

Ключевые слова: нелинейная связь, статистическая гипотеза, коэффициент детерминации, модель регрессии, скользящие средние.

Введение

В реальных динамических системах часто существуют разнообразные связи между различными параметрами. Их учет, помимо более корректного отображения свойств объекта моделью, позволяет существенно уменьшить проблемы, связанные с некорректностью моделей, плохо обусловленными системами уравнений, вычислительными трудностями и т.п. Все это обусловливает актуальность разработки методов выявления существующих связей. Для слабоформализованных систем, к которым относятся многие системы в металлургии и машиностроении, обычно применяются статистические методы проверки гипотезы о наличии или отсутствии связи. Однако для нелинейных связей используемые при этом показатели определяются неоднозначно, а получаемые оценки могут быть неустойчивыми [1].

Примерами задач, где возникает проблема поиска нелинейных связей, являются авто- и кросскорреляционный анализ временных рядов; исследование динамики параметров колебательных систем, в частности, систем с затуханием или наложением колебаний и т.п.

В связи с этим **целью** данной работы была разработка метода проверки гипотезы о наличии нелинейной статистической связи,

который позволял бы получать более надежные оценки характеристик такой связи.

Традиционные показатели нелинейной связи и методы их оценивания

Основными показателями, которые используют при оценивании нелинейных связей, являются коэффициент детерминации, индекс корреляции и корреляционное отношение [2 – 4].

Выборочный коэффициент детерминации некоторого параметра y по вектору независимых переменных **X** определяется [1, 2] по формуле:

$$K_d\left(y;\mathbf{X}\right)=1-\frac{s_\varepsilon^2}{s_y^2}, \tag{1}$$

где

$$s_y^2=\frac{1}{n}\sum_{k=1}^{n}\left(y_k-\overline{y}\right)^2, \tag{2}$$

n – количество наблюдений. Выборочную оценку дисперсии невязок вычисляют по одной из следующих формул:

$$s_\varepsilon^2=\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{f}\left(\mathbf{X}_i\right)\right)^2; \tag{3}$$

$$s_\varepsilon^2=\frac{1}{m}\sum_{j=1}^{m}\frac{1}{\nu_j}\sum_{i=1}^{\nu_j}\left(y_{ij}-\overline{y}_j\right)^2, \tag{4}$$

где $\hat{f}\left(\mathbf{X}_i\right)$ – оценка неизвестного значения функции регрессии в точке X_i; ν_j – количество точек, попавших в j-й интервал группирования; y_{ji} – значение i-го наблюдения в j-ом интервале; $\overline{y}_j=\sum_{i=1}^{\nu_j}y_{ji}\Big/\nu_j$ – среднее арифметическое наблюдений, попавших в j-й интервал, m – количество интервалов. Анализ выражения (4) показывает, что оно соответствует выражению (3) для случая аппроксимации неизвестной однофакторной модели регрессии кусочно – постоянной функцией, где интервалы постоянных значений y соответствуют выделенным интервалам группирования данных.

Индекс корреляции [3] связан с коэффициентом детерминации соотношением:

$$I\left(y;\mathbf{X}\right)=\sqrt{K_d\left(y;\mathbf{X}\right)} \tag{5}.$$

 ISSN 1562-9945

Корреляционное отношение рассчитывают [4] по формуле:

$$\rho_{yx}^2 = 1 - \frac{s_{\bar{y}(x)}^2}{s_y^2}, \qquad (6)$$

где $s_{\bar{y}(x)}^2 = \frac{1}{n}\sum_{j=1}^{m} \nu_j \left(\bar{y}_j - \bar{y}\right)^2$, $\bar{y} = \frac{1}{n}\sum_{j=1}^{m} \nu_j \bar{y}_j$.

Рассмотренные показатели имеют ряд недостатков, главным из которых является неоднозначность их определения, обусловленная возможностью выбора различных моделей регрессии при использовании формулы (3) и различных разбиений на интервалы в формулах (4, 6). Еще один существенный недостаток заключается в том, что для выборок ограниченного объема, которые наиболее часто встречаются на практике, существует проблема неустойчивости и недостаточной точности получаемых оценок. Кроме того, оценки, получаемые с использованием формул (4, 6), не являются симметричными. В частности $\rho_{yx}^2 \neq \rho_{xy}^2$ из-за того, что при группировании по разным переменным мы будем получать различные значения корреляционного отношения. Все это делает актуальной задачу разработки методики оценивания нелинейных связей, которая была бы в меньшей степени подвержена указанным недостаткам.

Применение методов сглаживания для оценивания коэффициента детерминации

В основу усовершенствованной методики оценивания коэффициента детерминации нами положены формулы (1 – 3), где оценки неизвестной функции регрессии $\hat{f}(\mathbf{X}_i)$ получаются методом скользящего среднего [5]:

$$\hat{f}(\mathbf{X}_i) = \frac{\sum_{j=i-p}^{i+p} y_j}{2p+1}, \qquad (1)$$

где $d = 2p+1$ – длина интервала сглаживания. Очевидно, что при использовании такого подхода должно существовать некоторое оптимальное с точки зрения точности получаемых оценок значение параметра d. При малых значениях значения оценок будут сильно зависеть от случайных ошибок значений y_j. При больших d

уменьшится эффективное число точек, используемых для вычисления коэффициента детерминации, которое равно n – 2p.

На рис. 1, 2 показаны результаты применения предлагаемого подхода для оценивания коэффициентов детерминации некоторых модельных зависимостей. В обоих случаях ε – это случайная величина, равномерно распределенная на отрезке [–0,2; 0,2]. Из приведенных графиков легко увидеть, что между переменными x и y имеется близкая к жестко детерминированной связь. Для зависимости, представленной на рис. 1, истинный коэффициент детерминации, рассчитанный по формулам (1 – 3), равен 0,980; коэффициент корреляции Пирсона R = – 0,77; коэффициент детерминации, вычисляемый по формулам (1, 2, 4), колеблется в интервале 0,94 – 0,97; коэффициент детерминации, определенный по предложенной методике, уменьшается от 0,985 при d = 3 до 0,980 при d = 9. Для зависимости, представленной на рис. 2, истинный коэффициент детерминации равен 0,969; коэффициент корреляции Пирсона R = – 0,17; коэффициент детерминации, вычисляемый по формулам (1, 2, 4), колеблется в интервале 0,25 – 0,77; коэффициент детерминации, определенный по предложенной методике, уменьшается от 0,981 при d = 3 до 0,874 при d = 9.

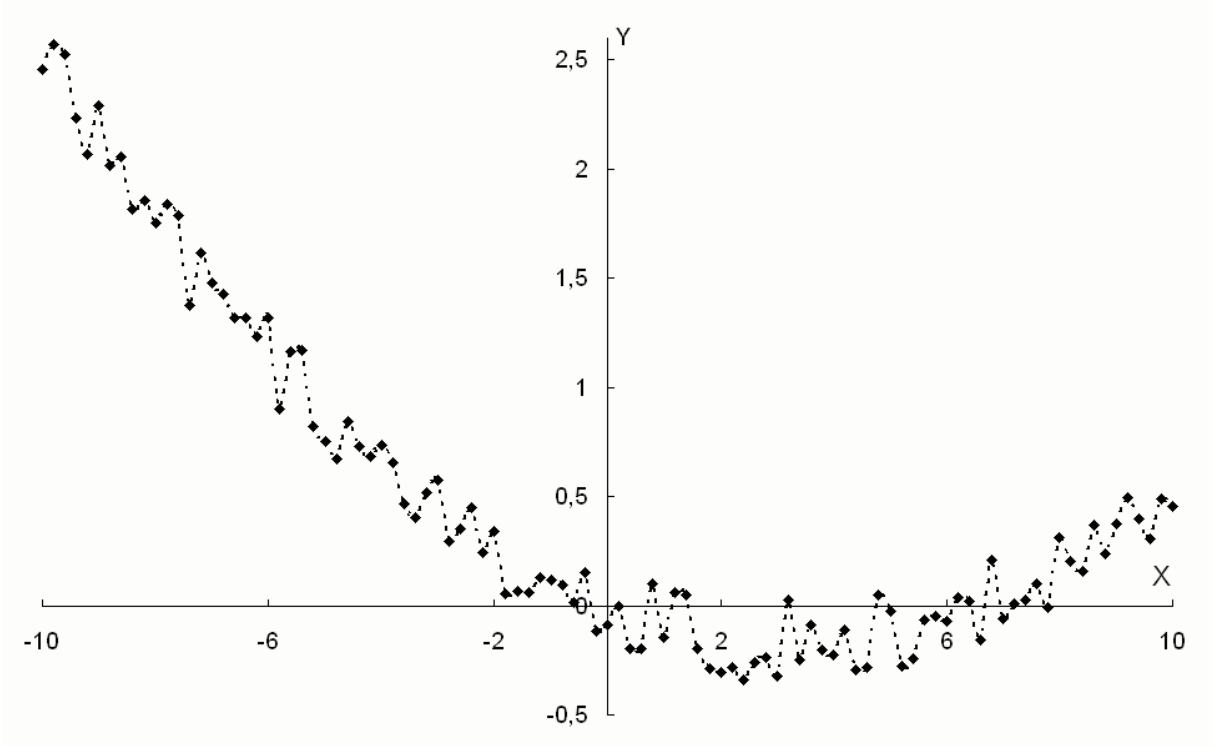

Рисунок 1 – График зависимости, соответствующей модели $y = 0,015x^2 - 0,1x + \varepsilon$

 ISSN 1562-9945

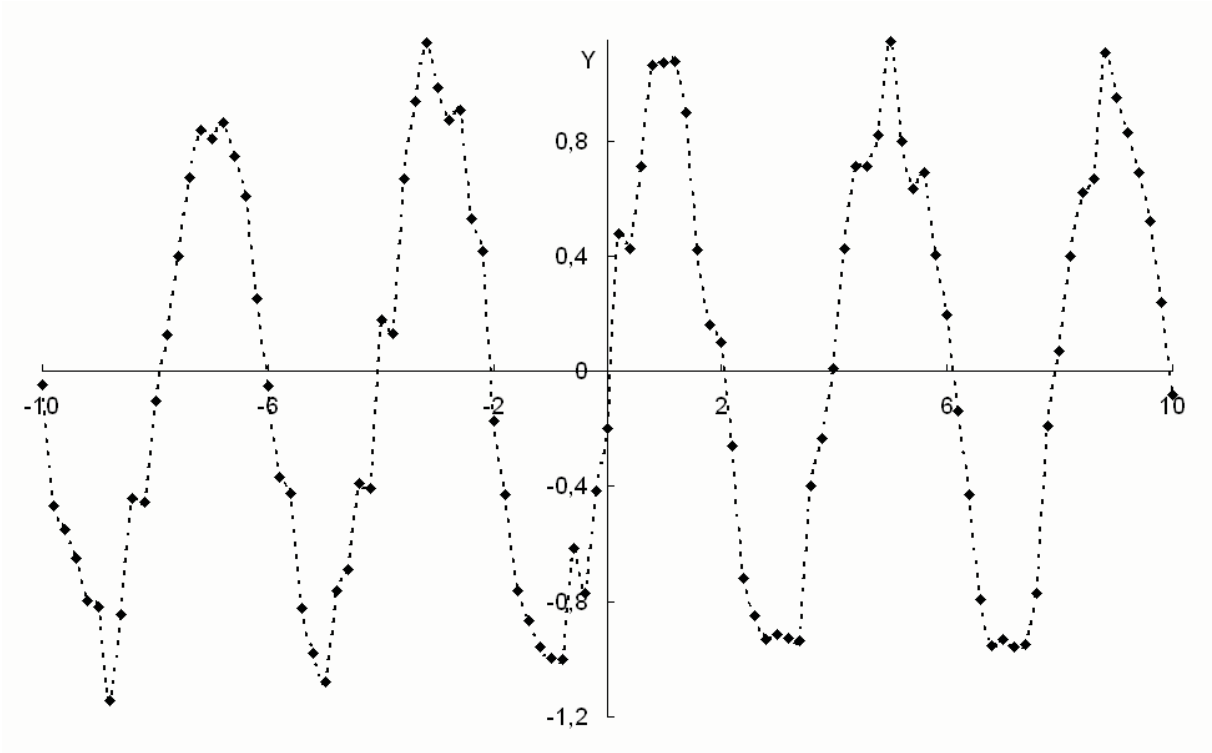

Рисунок 2 – График зависимости, соответствующей модели $y = \sin(\pi x / 2) + \varepsilon$

В таблице 1 представлены сводные результаты по результатам применения различных методов оценивания силы связи для различных модельных данных.

Таблица 1

Сравнительные характеристики показателей связи для различных модельных данных

Вид зависимости	ε	R	K_{d1}	K_{d2}	K_d
Случайная	0,1	0,13	0,10 – 0,16	0,05 – 0,12	0,02
	0,5	–0,17	0,09 – 0,17	0,00 – 0,06	0,00
	2	–0,13	0,14 – 0,17	0,12 – 0,13	0,03
Парабола	0,1	-0,78	0,96 – 0,98	0,994	0,994
	0,5	-0,73	0,82 – 0,84	0,83 – 0,84	0,85
	2	–0,27	0,17 – 0,21	0,08 – 0,09	0,09
Синусоида	0,1	0,16	0,28 – 0,78	0,89 – 0,98	0,99
	0,5	0,18	0,28 – 0,69	0,77 – 0,84	0,85
	2	0,11	0,15 – 0,30	0,25 – 0,28	0,29

В таблице использованы такие обозначения: ε – диапазон изменения равномерно распределенной на отрезке $[-\varepsilon;\varepsilon]$ случайной величины; K_{d1} – коэффициент детерминации, определенный по формулам (1, 2, 4); K_{d2} – коэффициент детерминации, вычисленный по предлагаемой методике, K_{d1} – коэффициент детерминации, рассчитанный по формулам (1 – 3).

Таким образом, видно, что предлагаемая методика позволяет существенно повысить надежность и устойчивость определения коэффициента детерминации. Наилучшие результаты получаются при выборе длины интервала сглаживания в интервале от 5 до 9.

Выводы

Предложена новая методика определения коэффициента детерминации, основанная на оценивании значений неизвестной модели регрессии методом скользящих средних. Она позволяет повысить точность и устойчивость оценок силы нелинейной связи. Наиболее сильно преимущества предложенной методики проявляются для сложных моделей связи, в частности для немонотонных и периодических зависимостей. Предлагаемый подход может быть использован для оценивания других показателей нелинейной связи в слабоформализованных системах.

ЛИТЕРАТУРА

1. Бахрушин В.Є. Методи аналізу даних. – Запоріжжя: КПУ, 2011. – 268 с.

2. Steel, R. G. D. and Torrie, J. H., Principles and Procedures of Statistics, New York: McGraw-Hill, 1960. – P. 187 – 287.

3. Венецкий И.Г., Венецкая В.И. Основные математико-статистические понятия и формулы в экономическом анализе. – М.: Статистика, 1974. – 277 с.

4. Рудакова Р.П., Букин Л.Л., Гаврилов В.И. "Статистика". – СПб.: Питер, 2008. – 288 с.

5. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.: ЮНИТИ, 1998. – 1000 с.

 ISSN 1562-9945