УДК 519.23

А.В. Бахрушин, В.Е. Бахрушин

ТЕСТИРОВАНИЕ ГИПОТЕЗ О НЕЛИНЕЙНЫХ СВЯЗЯХ С ИСПОЛЬЗОВАНИЕМ ЯЗЫКА ПРОГРАММИРОВАНИЯ R

Аннотация. Средствами языка программирования R разработаны модули для вычисления выборочного коэффициента детерминации, являющегося универсальным показателем наличия статистической связи для числовых данных. Проведено тестирование программ, показавшее корректность вычисления коэффициента детерминации для разных типов связи.

Ключевые слова: коэффициент детерминации, статистическая связь, нелинейная связь, язык программирования R.

Введение

Одной из наиболее важных задач статистического анализа является проверка гипотез о существовании связей между исследуемыми выборками. На сегодня известно большое число методов, предназначенных для решения этой задачи при наличии данных различного типа, а также для проверки гипотез о множественной корреляции [1]. Однако далеко не все они реализованы в стандартном программном обеспечении. В частности, в специализированных статистических пакетах SPSS, Statistica и др. не реализованы методы проверки гипотез о наличии нелинейной статистической связи между выборками количественных данных. Язык программирования R в настоящее время стал неофициальным стандартом для статистических исследований [2]. Его преимуществами являются: наличие обширных библиотек, реализующих большое число классических и неклассических методов статистического анализа данных; возможность создания собственных программ, реализующих новые алгоритмы, а также открытая лицензия на использование среды разработки и имеющихся программ. В связи с этим целью данной работы являлась реализация методик проверки гипотезы о наличии нелинейной статистической связи средствами языка программирования R.

 ISSN 1562-9945

1. Нелинейные статистические связи в сложных системах и современные подходы к их оцениванию

Проверку гипотезы о наличии парной статистической связи целесообразно начинать с построения корреляционного поля или другого типа диаграмм, способных отображать проверяемую связь. Это дает возможность предварительно установить возможное наличие связи, а также ее тип. Грубой методической ошибкой является вывод о наличии или отсутствии статистической связи только на основании малых значений коэффициента парной корреляции Пирсона, который является мерой линейной связи и нечувствителен ко многим типам нелинейных связей.

Для проверки гипотезы о наличии нелинейной связи между количественными данными можно использовать выборочный коэффициент детерминации, корреляционное отношение и индекс корреляции [1]. Все эти показатели тесно связаны между собой. Поэтому в дальнейшем мы будем рассматривть только выборочный коэффициент детерминации.

В общем случае его величина определяется [3] как

$$K_d = 1 - \frac{s^e_{err}}{s^2_{tot}}, \qquad (1)$$

где s^2_{err} - оценка дисперсии остатков используемой модели связи, а s^2_{tot} - оценка полной дисперсии зависимой переменной. Эта величина показывает долю общей вариации зависимой переменной, которая может быть объяснена рассматриваемой моделью.

Для оценивания дисперсии остатков используют два подхода [1]. Если модель связи $f(X)$ задана в явном виде, то

$$s^2_{err} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f\left(X_i\right)\right), \qquad (2)$$

где n – объем выборки, y_i – значение зависимой переменной для i–ой точки, а $f(X_i)$ – его оценка, даваемая используемой моделью.

Если модель связи не задана, то данные предварительно распределяют по m интервалам. После этого, дисперсию остатков можно оценить по формуле:

$$s^2_{err} = \frac{1}{m}\sum_{j=1}^{m}\frac{1}{\nu_j}\sum_{i=1}^{\nu_j}\left(y_{ij} - \overline{y_j}\right)^2, \qquad (3)$$

где v_j - число точек в j-ом интервале, y_{ij} – значения зависимых переменных для точек, попавших в j-й интервал, а $\overline{y_j}$ - их среднее арифметическое. Для случая парной корреляции такая оценка соответствует приближению неизвестной модели связи кусочно-постоянной функцией. При этом выбор интервалов группирования может существенно влиять на результат.

В [4] было предложено использовать вместо (3) оценки неизвестной функции связи, получаемые сглаживанием эмпирической зависимости методом скользящего среднего. В этом случае в (2) используется величина:

$$f(X_i) = \frac{\sum_{j=i-p}^{i+p} y_j}{2p+1}, \qquad (4)$$

где $d = 2p + 1$ - длина интервала сглаживания, которая также является субъективно определяемым параметром. Однако, как показано в [4], в этом случае оценки в меньшей степени зависят от его выбора. Кроме того, отсутствие необходимости в предварительном упорядочивании данных дает возможность использовать получаемую величину для оценивания тесноты связи в случае неоднозначных функций.

2. Программная реализация оценивания коэффициента детерминации

На рис. 1 приведены фрагменты программ, которые реализуют две методики оценки коэффициента детерминации. В обоих случаях исходный массив данных формировали как синусоидальную зависимость, к которой добавлена нормально распределенная аддитивная погрешность с нулевым средним значением и стандартным отклонением 0,1 (рис. 2). Программа 1а оценивает коэффициент детерминации Kd по формулам 1, 2, 4, а программа 1б – по формулам 1, 3. Видно, что полученные значения сильно различаются. При этом во втором случае полученное значение гораздо лучше соответствует имеющейся связи.

 ISSN 1562-9945

```
R Console
> e=rnorm(nx, mean = 0, sd = 0.1)
> y=e+sin(2*3.1416*x/4)
>
> p=2
> mz=vector('numeric',length(y)-2*p)
> dz=vector('numeric',length(y)-2*p)
> d2z=vector('numeric',length(y)-2*p)
> for(i in (p+1):(length(y)-p))
+ {mz[i-p]=0
+ for (j in (i-p):(i+p))
+ {mz[i-p]=mz[i-p]+y[j]
+ }
+ mz[i-p]=mz[i-p]/(2*p+1)
+ dz[i-p]=y[i]-mz[i-p]
+ }
> d2z=dz*dz
> se=sum(d2z)
> y1=y[p:(length(y)-p)]
> sy=sum((y1-mean(y1))^2)
> Kd=1-se/sy
> Kd
[1] 0.9749843
> cor(x,y)
[1] 0.1537285
> plot(x,y)
> KDTrue = 1 - var(e)/sy
> KDTrue
[1] 0.9998161
```

а

```
R Console
> nx = nr*nc
> x = seq(xmin, xmax, len = nx)
>
> e=rnorm(nx, mean = 0, sd = 0.1)
> y=e+sin(2*3.1416*x/4)
> mat=matrix(data = y, nrow = nr, ncol = nc)
> su=vector('numeric', nc)
> se=0
> for(i in (1:nc))
+ {su[i]=sum((mat[,i]-mean(mat[,i]))^2)/nr
+ se=se+su[i]
+ }
> se=se/nc
> sy=sum((y-mean(y))^2)/nx
> Kd=1-se/sy
> Kd
[1] 0.5183555
> cor(x,y)
[1] 0.1664784
> plot(x,y, type = "o")
> KDTrue = 1 - var(e)/sy
> KDTrue
[1] 0.9836567
>
```

б

Рисунок 1 - Программные модули для расчета коэффициента детерминации по формулам: а – 1, 2, 4; б – по формулам 1, 3

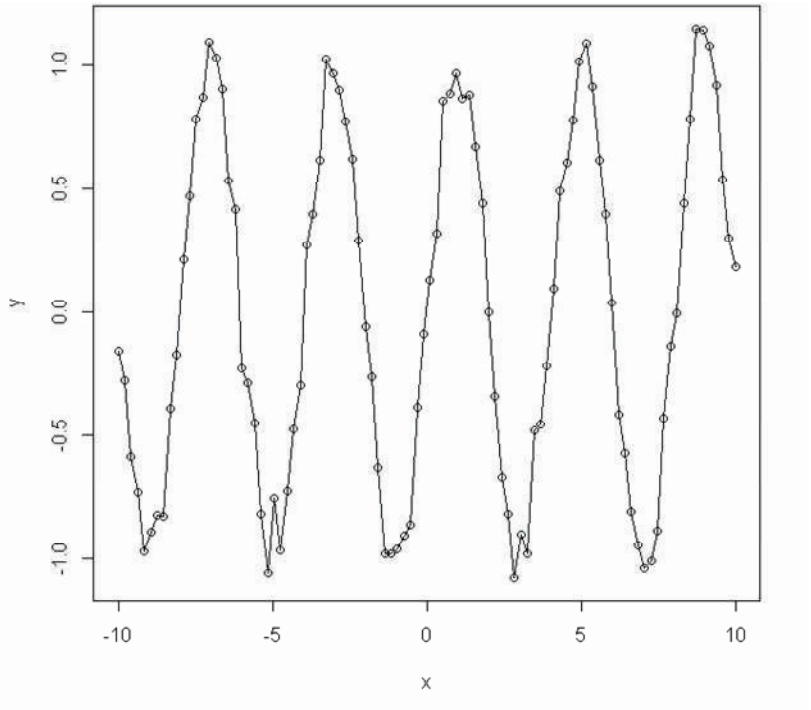

Рисунок 2 - График анализируемой зависимости

Приведенные на рис. 1 программные модули также выводят величину коэффициента корреляции Пирсона cor(x,y), которая показывает отсутствие значимой линейной связи, и величину KDTrue, характеризующую «истинный» коэффициент детерминации, рассчитываемый относительно использованной модели связи. Видно, что результаты расчета по формулам 1, 2, 4 значительно лучше согласуются с этой величиной, что подтверждает целесообразность ее использования. В то же время традиционный подход вычисления коэффициента детерминации более пригоден для решения задач нелинейного авто- и кросс-корреляционного анализа временных рядов [5].

Проведенное тестирование показало корректность работы написанных программных модулей и возможность их использования для получения оценок выборочного коэффициента детерминации.

Выводы

Предложены написанные на языке R программные модули, реализующие различные подходы к расчету выборочного коэффициента деитерминации. Они могут быть использованы для проверки гипотез о наличии нелинейных статистических связей в сложных системах.

ЛИТЕРАТУРА

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

2. Статистический анализ данных в системе R / А.Г. Буховец, П.В. Москалев, В.П. Богатова, Т.Я. Бирючинская; Под ред. проф. Буховца А.Г. – Воронеж: ВГАУ, 2010. – 124 с.

3. Everitt B.S. The Cambridge Dictionary of Statistics. – Cambridge University Press, 2002. – 420 p.

4. Бахрушин В.Е. Методы оценивания характеристик нелинейных статистических связей / В.Е. Бахрушин // Системні технології: Регіональний міжвузівський збірник наукових праць. Дніпропетровськ, 2011. - № 2(73). – С. 9 – 14.

5. Бахрушин В.Є. Застосування показників нелінійної кореляції для побудови й аналізу крос-кореляційних функцій / В.Є. Бахрушин, В.Є. Павленко, С.В. Петрова // Складні системи і процеси. – 2009, № 2. – С. 78 – 85.

 ISSN 1562-9945