§7 МАТЕМАТИЧЕСКОЕ И ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ НОВЫХ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

Бахрушин В.Е.

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ МЕТОДОВ АНАЛИЗА НЕЛИНЕЙНЫХ СТАТИСТИЧЕСКИХ СВЯЗЕЙ В СИСТЕМЕ R

Аннотация: *Существующие программные средства статистического анализа данных (SPSS, Statistica и др.) обычно предлагают для поиска корреляции лишь методы, пригодные для выявления линейной связи между числовыми данными, а также некоторые показатели связи для ранговых, качественных и смешанных данных. Однако реальная связь между количественными данными часто бывает нелинейной. Это приводит к тому, что имеющиеся средства не позволяют выявлять такие связи и могут приводить к ошибочным выводам об отсутствии корреляции. Универсальным показателем наличия статистической связи между двумя рядами числовых данных является выборочный коэффициент детерминации. Для его определения используют два подхода, один из которых базируется на аппроксимации неизвестной функции связи кусочно-постоянной функцией, а второй - на сглаживании имеющихся данных. В работе предложена программная реализация обоих методов средствами системы R. Достоинством этой системы является возможность использования большого числа специализированных библиотечных функций, предназначенных для статистического анализа, а также написания авторских программ для решения нестандартных задач. Тестирование разработанных приложений на модельных примерах показало их корректную работу и возможность использования для решения прикладных задач нелинейного корреляционного анализа.*

Ключевые слова: *Нелинейная связь, Коэффициент детерминации, Программное обеспечение, Язык R, Сглаживание данных, Корреляционное отношение, Коэффициент корреляции Пирсона, Тестирование, Группирование данынх, Кусочно-постоянная функция*

Введение

Проверка гипотез о существовании статистической связи между рядами данных от-

DOI: 10.7256/2305-6061.2014.2.11477

носится к числу основных задач статистического анализа. Для ее решения используют большое число различных методов, которые учитывают специфику конкретных задач, типов, объемов и структур данных и т.д.[1,2]. Для числовых данных в качестве показателя силы статистической связи наиболее часто используют коэффициент парной корреляции Пирсона. Однако он пригоден лишь для выявления линейных связей. В случае нелинейных связей этот показатель дает заниженные оценки силы связи между данными, либо вообще показывает ее отсутствие. Альтернативами коэффициенту корреляции Пирсона являются выборочный коэффициент детерминации, корреляционное отношение и индекс корреляции. Однако процедуры их вычисления отсутствуют в стандартных программных пакетах статистического анализа данных, таких как SPSS, Statistica и др.[3,4]. Поэтому на практике для решения подобных задач чаще используют показатели связи порядковых признаков – коэффициенты ранговой корреляции Спирмена и Кендалла [1,5,6]. Их существенным недостатком является то, что они могут дать корректную оценку силы связи лишь для монотонных зависимостей. В связи с этим актуальной является проблема разработки программных средств, базирующихся на использовании более адекватных показателей силы нелинейных статистических связей. Некоторые результаты в этом направлении были представлены в работе [7].

Целью данной работы являлась разработка программных модулей для оценивания выборочного коэффициента детерминации средствами языка программирования R.

Коэффициент детерминации и его использование при оценивании нелинейных статистических связей

Основными показателями, которые используют при оценивании нелинейных статистических связей, являются выборочный коэффициент детерминации, а также корреляционное отношение, или индекс корреляции (квадратный корень из коэффициента детерминации) [1,2]. Все они тесно связаны между собой, поэтому в дальнейшем мы будем рассматривать лишь выборочный коэффициент детерминации, являющийся наиболее универсальным из этих показателей. Он отражает долю вариации зависимой переменной, которая объясняется рассматриваемой моделью связи. В отличие от коэффициента корреляции Пирсона и коэффициентов ранговой корреляции, он не может принимать отрицательных значений и изменяется в пределах от 0 до 1. Это является следствием того, что нелинейные связи могут быть немонотонными. И в этом случае теряет смысл оценивание направления этих связей, отображаемое знаком показателя. Близость коэффициента детерминации к единице свидетельствует о наличии сильной, близкой к строго функциональной, связи между изучаемыми признаками. В этом случае используемая модель объясняет практически 100% вариации зависимой переменной, и лишь небольшая часть этой вариации приходится на неучтенные в модели или случайные факторы. Близость коэффициента детерминации к нулю, говорит о том, что связь между анализируемыми данными практически отсутствует.

В общем случае идея вычисления коэффициента детерминации базируется на гипо-

DOI: 10.7256/2305-6061.2014.2.11477

тезе, что мы имеем некоторую известную или предполагаемую модель связи. Тогда его значение можно рассчитать по формуле:

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

где s_{err}^2 - оценка дисперсии остатков этой модели, а s_{tot}^2 - общая дисперсия зависимой переменной y. Далее возможны два варианта. Если модель связи задана явно в виде функции $y = f(X)$, где X – вектор значений независимых переменных, то дисперсию остатков можно рассчитать по формуле:

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

где n – количество элементов в рассматриваемых выборках, y_i – значение зависимой переменной в i-ой точке. Формулу (2) используют при оценивании коэффициентов детерминации моделей регрессии, а также при разработке методов вычисления коэффициента детерминации для разных типов моделей связи.

Если же функция связи неизвестна, то приходится использовать ее различные аппроксимации. В случае одной независимой переменной для этого наиболее часто используют ее представление в виде кусочно-постоянной функции. С этой целью значения независимой переменной упорядочивают по возрастанию и группируют их по интервалам. Далее используют для вычисления коэффициента детерминации оценку:

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

где m – число интервалов, ν_j – число точек в j-ом интервале, y_{ij} – значения зависимых переменных для точек, которые попали в j-й интервал, $\overline{y_j}$ – их средние арифметические. Формулу (3) можно применять и в случае нескольких независимых переменных, но формирование интервалов и интерпретация результатов при этом значительно усложняются. Очевидно, что выбор способа формирования интервалов и их количества может существенно влиять на результат вычисления коэффициента детерминации. Особенно значимым это влияние оказывается в случае сильно нелинейных или немонотонных связей.

Существует другой подход к вычислению выборочного коэффициента детерминации [8]. Значения неизвестной функции парной связи в рассматриваемых точках предлагается заменить оценками, получаемыми в результате сглаживания имеющихся эмпирических данных. При сглаживании методом простых скользящих средних [9], получаем формулу

DOI: 10.7256/2305-6061.2014.2.11477

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

где $d = 2p+1$ – длина интервала сглаживания. Как и число интервалов, она определяется субъективно. Однако получаемые оценки оказываются значительно более устойчивыми к выбору этого параметра. Дополнительным преимуществом такого подхода является то, что он не требует предварительного упорядочивания данных. Поэтому появляется возможность оценивания неоднозначных зависимостей, для которых значение зависимой переменной определяется не только значениями независимых переменных, но и ее собственными предыдущими значениями. Такие зависимости достаточно часто возникают в различных прикладных задачах. Их примерами являются кривые гистерезиса, *S*-образные вольт-амперные характеристики и др.

В то же время для ряда задач использование формулы (4) оказывается нецелесообразным. В частности, при оценивании нелинейных авто- и кросс-корреляций временных рядов [10,11] изменение лага не приводит к "исчезновению" или "появлению" связи. Если связь имеется, то при изменении лага будет изменяться лишь вид соответствующей функции, но не сила связи. Поэтому оценки коэффициента детерминации практически не будут зависеть от выбранного лага. По этой причине оказывается невозможным решение одной из главных задач такого анализа – выявление цикличности изучаемых процессов. Метод группировки, напротив, может успешно использоваться при решении таких задач.

Оценивание коэффициента детерминации методом сглаживания

Для программной реализации задачи оценивания выборочных коэффициентов детерминации был выбран язык *R* [12], который в последнее время стал одним из основных средств статистических вычислений. К его основным преимуществам относятся:

- отсутствие необходимости больших затрат на первичную установку и обновления программного обеспечения, поскольку *R* является свободно распространяемым программным продуктом с открытым кодом;
- возможность использования большого числа специализированных статистических библиотек, в которых реализованы основные методы анализа данных;
- возможность самостоятельного написания программных продуктов для реализации отсутствующих в библиотеках методов.

Ниже приведен фрагмент скрипта для вычисления коэффициента детерминации методом сглаживания по формулам (1, 2, 4). Исходные данные были сформированы в предположении, что реальная функция связи является затухающей косинусоидой, и к ней добавлена нормально распределенная аддитивная погрешность.

DOI: 10.7256/2305-6061.2014.2.11477

```
# Создание массива данных тестовой зависимости
x=seq(-10,10,len=500)
e=rnorm(500, mean = 0, sd = 0.2)
y=e+cos(x)*exp(-abs(x)/3)

p=2 # задание параметра p для формулы 4

# Формирование вспомогательных векторов
mz=vector('numeric',length(y)-2*p) # вектор оценок функции связи
dz=vector('numeric',length(y)-2*p) # вектор остатков модели

# Вычисление дисперсии остатков по формулам 2, 4
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]
}
se=sum(dz*dz)

# Вычисление общей дисперсии y
y1=y[p:(length(y)-p)]
sy=sum((y1-mean(y1))^2)

# Вычисление коэффициента детерминации
Kd=1-se/sy
Kd

# Дополнительные характеристики силы связи
Corrat = sqrt(Kd)
Corrat
cor(x,y)
plot(x,y)
e1=e[p:(length(e)-p)]
KDTrue = 1 - var(e1)*length(y1 - 1)/sy
KDTrue
```

Некоторые комментарии к приведенному скрипту.

1. При вычислении дисперсий опущено деление на объем выборки;
2. При сглаживании теряются p левых и p правых точек исходных данных. Поэтому

DOI: 10.7256/2305-6061.2014.2.11477

используется перенумерация элементов массивов и уменьшается объем выборки, используемой для вычисления общей дисперсии.

3. Для проверки результатов и сравнения вычисляются корреляционное отношение (Corrat), коэффициент корреляции Пирсона (cor(x,y)) и "истинный" коэффициент детерминации (KDTrue); последний определяется подстановкой в (2) использованной при генерации данных модели связи. При этом для рассматриваемого метода "истинный" коэффициент детерминации часто оказывается несколько меньше расчетного. Это обусловлено тем, что из-за наличия случайной погрешности подбираемая сглаживанием модель оказывается локально лучшей, чем заданная исходная модель связи.
4. Также строится корреляционное поле, позволяющее визуально оценить силу связи между изучаемыми переменными.

Для приведенного примера получены следующие результаты:

```
> Kd
[1] 0.7393334
> Corrat
[1] 0.859845
> cor(x,y)
[1] -0.01059034
> KDTrue
[1] 0.7042699
```

На рис. 1 показано корреляционное поле для использованных данных, подтверждающее наличие достаточно сильной связи.

Аналогичные результаты получены для неоднозначной функции связи. В этом случае в качестве модели была взята зависимость, обратная функции

$$x = 0,5y + \frac{100}{\exp\left((y+4)/2\right) + \exp\left((-y-4)/2\right)}.$$

К ней добавляли нормально распределенную аддитивную погрешность. При этом получены такие результаты:

```
> Kd
 [1] 0.9942788
> Corrat
[1] 0,997135
> cor(x,y)
[1] 0.7005568
> KDTrue
[1] 0.9931243
```


DOI: 10.7256/2305-6061.2014.2.11477

На рис. 2. показано корреляционное поле для использованных данных, подтверждающее наличие весьма сильной связи.

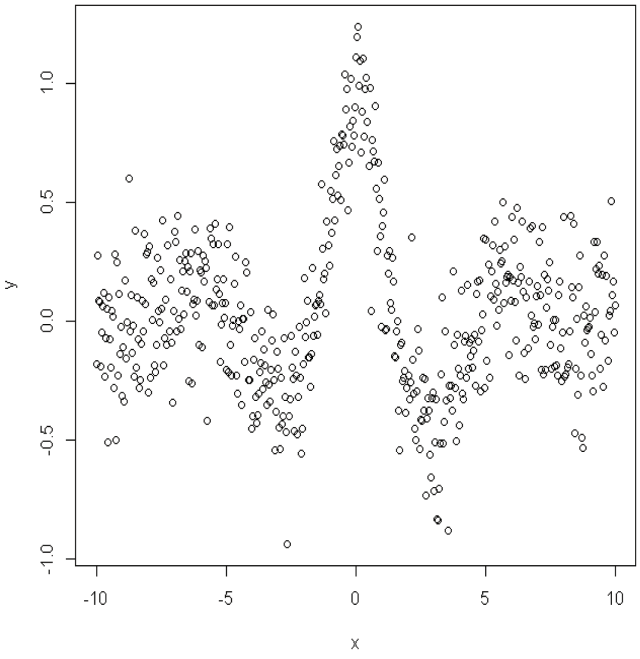

Рис. 1. Корреляционное поле для затухающей косинусоидальной зависимости

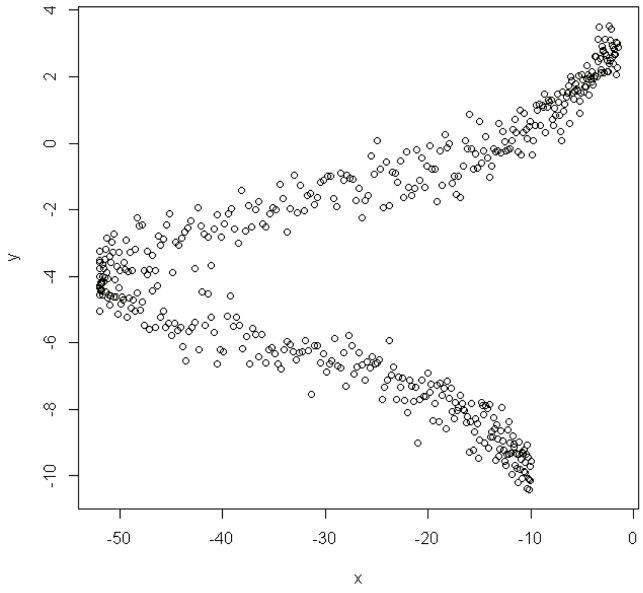

Рис. 2. Корреляционное поле для анализируемой неоднозначной зависимости

Были также рассмотрены примеры более простых (монотонных, квадратичных) типов связи. Во всех этих случаях имелось достаточно хорошее соответствие между рассчиты-

DOI: 10.7256/2305-6061.2014.2.11477

ваемыми показателями силы связи и визуально наблюдаемой степенью ее выраженности (последнюю регулировали, задавая различные параметры при генерации "погрешности" данных). В то же время при использовании в качестве значений зависимой переменной вектора случайных чисел и при моделировании связи с помощью хаотического (логистического) отображения, рассчитываемые значения коэффициента детерминации, как и следовало ожидать, были близки к нулю. Во всех случаях имелось хорошее совпадение получаемых оценок с результатами, которые рассчитывались "вручную" с помощью электронных таблиц MS Excel.

Оценивание коэффициента детерминации методом группировки данных

Ниже приведен фрагмент скрипта для вычисления коэффициента детерминации методом группировки данных по формулам (1 – 3) в случае рассмотренной в предыдущем разделе неоднозначной зависимости.

```
# Формирование массива исходных данных
nr = 25
nc = 20
nx = nr*nc
a=seq(-10,3,len=nx)
e=rnorm(nx, mean = 0, sd = 0.5)
x=-100/(exp((a+4)/2) + exp((-a-4)/2)) + 0.5*a
y = a + e

# Упорядочение массива исходных данных по возрастанию x
z = matrix(data = c(x, y), nrow = nx, ncol = 2)
h=z[order(z[,1]),]

# Вычисление дисперсии остатков
mat=matrix(data = h[,2], nrow = nr, ncol = nc) # группировка данных
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

# Вычисление общей дисперсии
sy=sum((y-mean(y))^2)/nx
```


DOI: 10.7256/2305-6061.2014.2.11477

```
# Вычисление коэффициента детерминации
Kd=1-se/sy
Kd

# Дополнительные характеристики силы связи
Corrat = sqrt(Kd)
Corrat
cor(x,y)
plot(x,y, type = “p”)
#KDTrue = 1 - var(e1)*length(y1 - 1)/sy
KDTrue = 1 - var(e)/sy
KDTrue
```

Для рассмотренного примера получены следующие результаты:

```
> Kd
[1] 0.4494663
> Corrat
[1] 0,670422
> cor(x,y)
[1] 0.2762725
> KDTrue
[1] 0.9824009
```

Видно, что значение коэффициента детерминации для анализируемой неоднозначной функции существенно оказалось меньшим, чем при оценивании методом сглаживания и хуже согласуется с “истинным” коэффициентом детерминации. Следует отметить, что в приведенном примере функция упорядочения массива фактически не работает, поскольку исходные данные уже упорядочены по x. Такие данные взяты для того, чтобы сопоставление результатов вычисления коэффициента детерминации разными методами выполнялось на одинаковых исходных данных. Однако в реальных ситуациях использованная опция может оказаться необходимой.

Аналогичные результаты получены для затухающей косинусоидальной зависимости:

```
> Kd
[1] 0.4064719
> Corrat
[1] 0.6375515
> cor(x,y)
[1] -0.01059034
> KDTrue
[1] 0.6734953
```


DOI: 10.7256/2305-6061.2014.2.11477

Здесь также значение коэффициента детерминации существенно ниже, чем при оценивании методом сглаживания, и хуже согласуется с "истинным" коэффициентом детерминации. Однако в этом случае наблюдается еще и большой разброс получаемых оценок коэффициента детерминации в зависимости от выбора числа интервалов группировки.

Для более простых моделей связи имелось достаточно хорошее соответствие между рассчитываемыми показателями силы связи и визуально наблюдаемой степенью ее выраженности. Также во всех случаях имелось хорошее совпадение получаемых оценок с результатами, которые рассчитывались "вручную" с помощью электронных таблиц MS Excel.

Выводы

На языке R разработаны программные скрипты для оценивания силы статистической связи между рядами данных на основе вычисления выборочных коэффициентов детерминации методами группировки и сглаживания данных. Тестирование показало, что разработанные программные модули корректно работают и могут быть использованы для решения задач, связанных с анализом нелинейных статистических связей.

Библиография :

1. Бахрушин В.Є. Методи аналізу даних. – Запоріжжя: КПУ, 2011. – 268 с.
2. Гайдышев И. Анализ и обработка данных. Специальный справочник – СПб.: Питер, 2001. – 752 с.
3. Бююль А., Цёфель П. SPSS: Искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. – СПб.: ДиаСофтЮП, 2005 – 608 с.
4. Халафян А.А. Statistica 6. Статистический анализ данных. – М.: ООО Бином-Пресс, 2008. – 512 с.
5. Кендэл М. Ранговые корреляции. – М.: Статистика, 1975. – 216 с.
6. Gauthier T. D. Detecting Trends Using Spearman's Rank Correlation Coefficient // Environmental Forensics. – 2001. No 2. – P. 359 – 362.
7. Бахрушин А.В., Бахрушин В.Е. Тестирование гипотез о нелинейных связях с использованием языка программирования R // Системные технологии: Регіональний міжвузівський збірник наукових праць. Дніпропетровськ, 2013. – № 3(86). – С. 168 – 172.
8. Бахрушин В.Е. Методы оценивания характеристик нелинейных статистических связей // Системні технології: Регіональний міжвузівський збірник наукових праць. Дніпропетровськ, 2011. – № 2(73). – С. 9 – 14.
9. Андерсон Т. Статистический анализ временных рядов. – М.: Мир, 1976. – 755 с.
10. Бахрушин В.Є., Павленко В.Є., Петрова С.В. Застосування показників нелінійної кореляції для побудови й аналізу крос-кореляційних функцій // Складні системи і процеси. – 2009, № 2. – С. 78 – 85.

DOI: 10.7256/2305-6061.2014.2.11477

11. Бахрушин В.Е., Павленко В.Е., Петрова С.В. Применение выборочного коэффициента детерминации для построения и анализа кросс-корреляционных функций // Фундаментальные физико-математические проблемы и моделирование технико-технологических систем / Под ред. Ю.М. Соломенцева, Б.Н. Четверушкина, А.В. Боголюбова и др. – М.: МГТУ "СТАНКИН", Янус-К, 2010. – Вып. 13. – С. 4
12. 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.

References:

1. Bakhrushin V.E. Metody analizu danykh. – Zaporizhzhya: KPU, 2011. – 268 s.
2. Gaidyshev I. Analiz i obrabotka dannykh. Spetsial'nyi spravochnik – SPb.: Piter, 2001. – 752 s.
3. Bühl A., Zöfel P.: Iskusstvo obrabotki informatsii. Analiz statisticheskikh dannykh i vosstanovlenie skrytykh zakonomernostei. – SPb.: DiaSoftYuP, 2005 – 608 s.
4. Khalafyan A.A. Statistica 6. Statisticheskii analiz dannykh. – M.: OOO Binom-Press, 2008. – 512 s.
5. Kendall M. Rangovye korrelyatsii. – M.: Statistika, 1975. – 216 s.
6. Gauthier T. D. Detecting Trends Using Spearman's Rank Correlation Coefficient // Environmental Forensics. – 2001. – No 2. – P. 359 – 362.
7. Bakhrushin A.V., Bakhrushin V.E. Testirovanie gipotez o nelineinykh svyazyakh s ispol'zovaniem yazyka programmirovaniya R // Sistemnye tekhnologii: Regional'nii mizhvuzivs'kii zbirnik naukovikh prats'. Dnipropetrovs'k, 2013. – № 3(86). – S. 168 – 172.
8. Bakhrushin V.E. Metody otsenivaniya kharakteristik nelineinykh statisticheskikh svyazei // Sistemni tekhnologii: Regional'nii mizhvuzivs'kii zbirnik naukovikh prats'. Dnipropetrovs'k, 2011. – № 2(73). – S. 9 – 14.
9. Anderson T. Statisticheskii analiz vremennykh ryadov. – M.: Mir, 1976. – 755 s.
10. Bakhrushin V.E., Pavlenko V.E., Petrova S.V. Zastosuvannya pokazntkiv neliniinoï korelyatsiï dlya pobudovy i analizu kros-korelyatsiinykh funktsii // Skladni sistemi i protsesy. – 2009, № 2. – S. 78 – 85.
11. Bakhrushin V.E., Pavlenko V.E., Petrova S.V. Primenenie vyborochnogo koeffitsienta determinatsii dlya postroeniya i analiza kross-korrelyatsionnykh funktsii // Fundamental'nye fiziko-matematicheskie problemy i modelirovanie tekhniko-tekhnologicheskikh sistem / Pod red. Yu.M. Solomentseva, B.N. Chetverushkina, A.V. Bogolyubova i dr. – M.: MGTU "STANKIN", Yanus-K, 2010. – Vyp. 13. – S. 4
12. Statisticheskii analiz dannykh v sisteme R / A.G. Bukhovets, P.V. Moskalev, V.P. Bogatova, T.Ya. Biryuchinskaya; Pod red. prof. Bukhovtsa A.G. – Voronezh: VGAU, 2010. – 124 s.

DOI: 10.7256/2305-6061.2014.2.11477