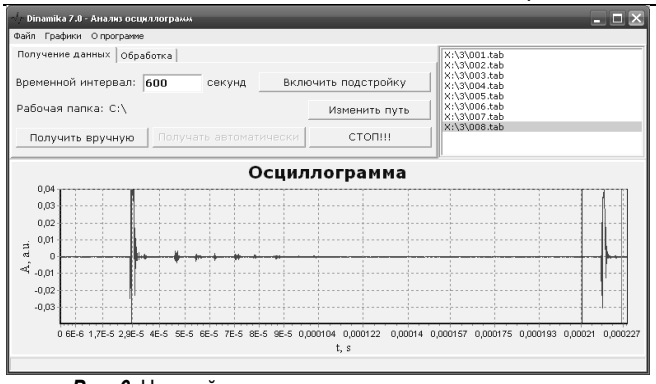

Рис. 6. Настройка автоматического извлечения данных

На экране, показанном на рис. 6, также предусмотрена возможность автоматической подстройки интервала между измерениями. Если данный режим включен, добавление осциллограммы в последовательность сопровождается уменьшением интервала, а отказ от нее в связи со слабым изменением сигнала увеличивает интервал. При этом численно заданное значение является предельным верхним значением длительности интервала. Такой подход в ряде случаев позволяет более эффективно исследовать процессы с неравномерной скоростью протекания [7].

СПИСОК ЦИТИРОВАННЫХ ИСТОЧНИКОВ

1. Костюк, Д.А. Виртуальная лаборатория диагностики диссипативных сред // Современные информационные компьютерные технологии: сб. науч. ст. / Д.А. Костюк, Л.Н. Николаюк. – Гродно: ГрГУ, 2006. – С. 58–63.
2. Костюк, Д.А. Аномальное отражение продольного ультразвука от сильно диссипативной среды // Инженерно-физический журнал / Д.А. Костюк, Ю.А. Кузавко. – 2004. – Т. 77. – № 5. – С. 161–169.
3. Козак, А.Ф. Программно-аппаратный комплекс акустического спектрального анализа диссипативно-дисперсионных сред // Проблемы проектирования и производства радиоэлектронных средств: сб-к материалов V Междунар. науч.-тех. конф. / А.Ф. Козак, Д.А. Костюк, Л.Н. Николаюк. – Новополоцк, 29–30 мая 2008: в 3 т. / Полоц. гос. ун-т. – Новополоцк, 2008. – Т. III. Информатика. – С. 220–223.
4. Костюк, Д.А. Компьютеризированная установка акустического спектрального анализа диссипативных сред // Современные информационные компьютерные технологи: сб. науч. ст. / Д.А. Костюк, Л.Н. Николаюк. – Гродно: ГрГУ, 2006. – С. 58–63.
5. Козак, А.Ф. Приборное решение акустического спектрального анализа для диагностики вязких сред // Современные методы и приборы контроля качества и диагностики состояния объектов. Материалы 2-й международной научно-технической конференции / А.Ф. Козак, Д.А. Костюк, Ю.А. Кузавко. – Могилев, 2006. – С. 54–56.
6. Kozak A., Kostiuk D., Kuzavko Y., Nikolayuk L., Tomassi P. The acoustic spectral analysis of metal corroding surfaces // Proc. of the Internat. Conf. CORROSION 2005 „Science & Economy". Poland, Warsaw, 8–10 June 2005, Inżynieria Powierzchni, 2005, 2A, 63–70.
7. Данилевский, В.П. Акустические спектроскопические методы и средства диагностики материалов и веществ // Материалы, технологии, инструменты / В.П. Данилевский, Д.А. Костюк, Н.В. Кудинов, Ю.А. Кузавко. – № 3. – Т. 8, 2003. – С. 104–112.

Материал поступил в редакцию 11.11.09

KOSTIUK D.A., GRISEVICH L.N. Software System For Automated Experimental Researches Of Applied Acoustical Spectroscopy

A software system for carrying out automated experiments is presented, designed as a part of software-hardware complex of the dissipative-dispersion medium acoustic spectral analysis. The software structure is considered as well as its practical implementation, tasks and specific features of the operation.

УДК 519.23/.25

Дереченник С.С., Дмитриева А.В., Дереченник С.С.- мл.

ИНТЕГРАЛЬНАЯ ОЦЕНКА КАЧЕСТВА РЕГРЕССИОННЫХ МОДЕЛЕЙ

Введение. Типичной задачей обработки данных является установление функциональной зависимости некоторой величины (отклика) от одной или нескольких переменных (факторов). В теории вероятностей функция, приближенно представляющая статистическую зависимость случайных величин, определяется как регрессия, в частности – средняя квадратическая регрессия. Для вычисления коэффициентов регрессии (в простейшем случае – линейной) обычно применяется метод наименьших квадратов (МНК), созданный в 1806 году Гауссом и Лежандром. Благодаря ряду преимуществ МНК (простота и удобство применения, эффективность получаемых оценок и др.), этот универсальный математический инструмент, помимо статистики и теории ошибок, используется также при аппроксимации функций, численном решении уравнений, нахождении псевдообратных матриц и т.д.

При построении моделей регрессии с нелинейной связью факторов и отклика могут возникнуть определенные трудности, т.к. применяемые обычно приемы линеаризации не всегда корректны [1]. Нелинейное преобразование шкалы отклика зачастую приводит к тому, что случайные аддитивные остатки модели становятся мультипликативными, а это нарушает условие их гомоскедастичности. Аналогичные действия над шкалой фактора не влекут подобных проблем, однако корректность математических операций над данными ограничена типом шкалы измерений, в которой они получены. Применительно к шкалам интервалов, периодической и отношений, допустимыми, в этом смысле, являются полиномиальные модели регрессии. Нелинейные (логарифмические, показательные, экспоненциальные и т.п.) преобразования фактора вполне корректны лишь при условии его представления в абсолютной (безразмерной) шкале измерений.

Отсутствие физической размерности у фактора регрессионной модели характерно для задач аппроксимации вероятностных функций распределения некоторых случайных, например, метеорологических величин. Так, при долгосрочном прогнозировании экстремальных значений снеговой нагрузки на поверхности земли, эмпирическое распределение в области больших значений отклика ("хвостовой" части распределения) проверяется на асимптотическую принадлежность к одному из типов – Гумбеля, Фреше или Вейбулла [2]. Значение фактора регрессионной модели в данном случае вычисляется путем логарифмирования эмпирической оценки вероятности, поэтому его отсчеты не являются равноотстоящими – с увеличением значения расстояние между соседними отсчетами фактора прогрессивно возрастает. Таким образом, большинство отсчетов

***Дмитриева Анна Владимировна**, магистр технических наук, аспирант кафедры ЭВМиС Брестского государственного технического университета.*

***Дереченник–мл. Станислав Станиславович,** студент 3 курса факультета электронно-информационных систем Брестского государственного технического университета.*

Беларусь, БрГТУ, 224017, г. Брест, ул. Московская, 267.

группируется в области малой и средней вероятности, в области же высокой (близкой к единице) вероятностной обеспеченности, наиболее важной, с точки зрения прогнозирования экстремальных значений отклика, оказывается весьма небольшая часть эмпирических точек. Однако классические инструменты МНК предусматривают вычисление сумм квадратов отклонений от регрессионной зависимости по всем эмпирическим точкам, без учета регулярности их расположения. Мы полагаем, что указанное обстоятельство может влиять, по крайней мере, на значение коэффициента детерминации получаемых регрессионных моделей.

В данной работе предложен и реализован подход, основанный на интегральной оценке отклонений эмпирических данных от регрессионной зависимости на всем интервале изменения фактора.

Построение коэффициента интегральной детерминации регрессионной модели. Рассмотрим однофакторную регрессию, в которой отклик $y = f(x)$ представлен в числовой шкале интервалов или отношений, а фактор x – в абсолютной безразмерной шкале. Пусть имеется выборка $\left\{(X_i, Y_i),\ X_i < X_{i+1},\ i = \overline{1,n}\right\}$ эмпирических наблюдений, относительно которых полагаем, что $Y_i = f(X_i) + \varepsilon_i$, где случайные остатки ε_i независимы и одинаково распределены, т.е. обладают гомоскедастичностью. Ограничив возможные виды функции регрессии типовыми двухпараметрическими зависимостями: линейной $y = ax + b$, логарифмической $y = a\ln x + b$ и экспоненциальной $y = b\exp(ax)$, доопределим исходную функцию, заданную на дискретном множестве эмпирических точек, одним из возможных кусочно-гладких приближений того же вида:

$$f^*(x) = \sum_{i=1}^{n-1} f_i^*(x),$$

$$f_i^*(x) = \begin{cases} \left\langle \begin{array}{l} A_i x + B_i \\ A_i \ln x + B_i \\ B_i \exp(A_i x) \end{array} \right\rangle, & X_i \le x < X_{i+1} \\ 0, & x \notin \left[X_i, X_{i+1}\right). \end{cases} \quad (1)$$

На интервале $\left[X_1, X_n\right)$ найдем интегральную квадратичную ошибку регрессии (остаточную дисперсию):

$$J_E = \int_{X_1}^{X_n} \left[f(x) - f^*(x)\right]^2 dx = \sum_{i=1}^{n-1} \int_{X_i}^{X_{i+1}} \left[f(x) - f_i^*(x)\right]^2 dx \quad (2)$$

и полный квадрат отклонения кусочно-гладкого приближения эмпирических данных:

$$J = \int_{X_1}^{X_n} \left[f^*(x) - \overline{f^*}\right]^2 dx = \sum_{i=1}^{n-1} \int_{X_i}^{X_{i+1}} \left[f_i^*(x) - \overline{f^*}\right]^2 dx, \quad (3)$$

где интегральное среднее отклика регрессии определяется как

$$\overline{f^*} = \int_{X_1}^{X_n} f^*(x)dx \Big/ \int_{X_1}^{X_n} dx = \frac{1}{X_n - X_1} \sum_{i=1}^{n-1} \int_{X_i}^{X_{i+1}} f_i^*(x)dx. \quad (4)$$

Подобно известному коэффициенту детерминации (качества) регрессии R-квадрат, определенному на дискретном множестве отсчетов, введем более точный, определенный на интервале, *коэффициент интегральной детерминации* (definite determinative). Он вычисляется как доля полного квадрата отклонения кусочно-гладкого приближения эмпирических данных, которая объяснена регрессионной моделью:

$$R_{DD}^2 = \frac{J - J_E}{J} = 1 - \frac{J_E}{J}. \quad (5)$$

Возможный диапазон значений нового коэффициента аналогичен диапазону значений традиционного коэффициента R-квадрат – от нуля (статистическая зависимость отклика и фактора отсутствует, при этом параметры регрессии принимают значения $a = 0$, $b = \overline{f^*}$) до единицы (точная регрессионная зависимость, $\varepsilon_i \equiv 0$ для всех отсчетов выборки). Коэффициент R_{DD} -квадрат пригоден, таким образом, для измерения качества регрессионного приближения, в том числе нелинейного, на всем интервале изменения фактора, вне зависимости от расположения отдельных его отсчетов. Заметим также, что данный коэффициент практически отождествляется с коэффициентом R-квадрат в случае простой линейной регрессии с равноотстоящими отсчетами фактора.

В ходе исследований нами установлено, что в ряде случаев возможна аналитическая минимизация интегральной квадратичной ошибки (2), позволяющая находить оценки параметров регрессии при произвольном расположении отсчетов фактора.

Минимизация интегральной квадратичной ошибки регрессии. Общая схема нахождения параметров регрессии и оценки ее качества соответствует классической схеме МНК. На примере *линейной регрессии* $y = ax + b$ покажем возможность аналитической минимизации квадратичной ошибки на интервале, а также порядок вычисления коэффициента интегральной детерминации.

Коэффициенты A_i, B_i уравнения (1) определяются для отсчетов исходной выборки соотношениями: $A_i = (Y_{i+1} - Y_i)/(X_{i+1} - X_i)$, $B_i = Y_i - A_i X_i$. Интегральная квадратичная ошибка линейной регрессии:

$$J_E = \sum_{i=1}^{n-1} \int_{X_i}^{X_{i+1}} \left[ax + b - (A_i x + B_i)\right]^2 dx =$$

$$= \sum_{i=1}^{n-1} \frac{\left[(a - A_i)X_{i+1} + b - B_i\right]^3 - \left[(a - A_i)X_i + b - B_i\right]^3}{3(a - A_i)},$$

для ее минимизации найдем две частных производных:

$$\frac{\partial J_E}{\partial a} = \frac{2}{3}a\left(X_n^3 - X_1^3\right) + b\left(X_n^2 - X_1^2\right) -$$

$$-\frac{2}{3}\sum_{i=1}^{n-1} A_i\left(X_{i+1}^3 - X_i^3\right) - \sum_{i=1}^{n-1} B_i\left(X_{i+1}^2 - X_i^2\right),$$

$$\frac{\partial J_E}{\partial b} = a\left(X_n^2 - X_1^2\right) + 2b(X_n - X_1) -$$

$$-\sum_{i=1}^{n-1} A_i\left(X_{i+1}^2 - X_i^2\right) - 2\sum_{i=1}^{n-1} B_i(X_{i+1} - X_i).$$

Введем ряд обозначений:

$$\Delta_1 = X_n - X_1;\ \Delta_2 = X_n^2 - X_1^2;\ \Delta_3 = X_n^3 - X_1^3;$$

$$c = \frac{2}{3}\sum_{i=1}^{n-1} A_i\left(X_{i+1}^3 - X_i^3\right) + \sum_{i=}^{n-1} B_i\left(X_{i+1}^2 - X_i^2\right);$$

$$d = \sum_{i=1}^{n-1} A_i\left(X_{i+1}^2 - X_i^2\right) + 2\sum_{i=1}^{n-1} B_i(X_{i+1} - X_i),$$

необходимых для удобства записи искомого решения:

$$\begin{cases} \frac{\partial J_E}{\partial a} = 0, \\ \frac{\partial J_E}{\partial b} = 0 \end{cases} \Rightarrow \begin{cases} \frac{2a}{3}\Delta_3 + b\Delta_2 = c, \\ a\Delta_2 + 2b\Delta_1 = d \end{cases} \Rightarrow \begin{cases} a = \frac{d - 2b\Delta_1}{\Delta_2}, \\ b = \frac{c\Delta_2 - 2d\Delta_3/3}{\Delta_2^2 - 4\Delta_1\Delta_3/3}. \end{cases}$$

Таблица 1. Статистические характеристики параметров линейных регрессионных моделей

Параметр и его характеристика	Модель с равноотстоящими отсчетами фактора		Модель с нерегулярными отсчетами фактора	
	МНК	Минимизация интегральной квадратичной ошибки	МНК	Минимизация интегральной квадратичной ошибки
Коэффициент a				
интервальная оценка	3,988±0,034	3,982±0,035	4,000±0,042	4,033±0,053
размах	3,54…4,42	3,46…4,43	3,48…4,46	3,25…4,57
Коэффициент b				
интервальная оценка	5,42±0,52	5,48±0,53	5,26±0,33	4,64±0,57
размах	– 0,2…11,4	– 0,3…11,8	1,4…8,6	– 1,9…12,7
Коэффициент детерминации	R-квадрат	R_{DD} -квадрат	R-квадрат	R_{DD} -квадрат
среднее значение	0,9623	0,9735	0,9332	0,9826
размах	0,947…0,979	0,960…0,989	0,908…0,966	0,962…0,995

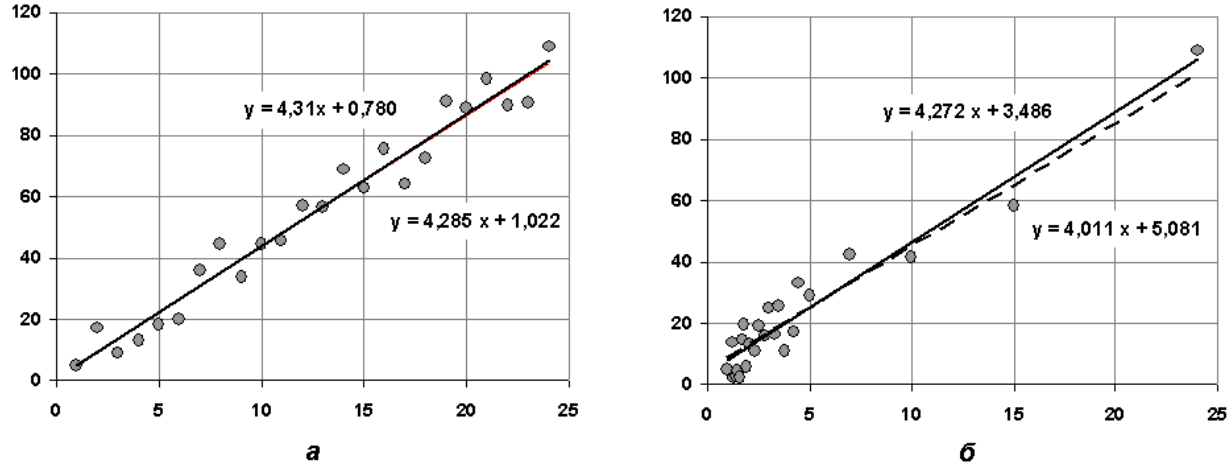

Рис. 1. Модели линейной регрессии для равноотстоящих (а) и нерегулярных (б) отсчетов фактора. Сплошная линия – регрессия, найденная МНК (уравнение сверху); пунктирная линия – регрессия с минимизацией интегральной квадратичной ошибки (уравнение снизу)

Значение интегрального среднего отклика регрессии:

$$\overline{f^*} = \sum_{i=1}^{n-1} \int_{X_i}^{X_{i+1}} \left(A_i x + B_i\right) dx \Big/ \int_{X_1}^{X_n} dx =$$

$$-\frac{1}{X_n - X_1} \sum_{i=1}^{n-1} \left[\frac{A_i}{2} \left(X_{i+1}^2 - X_{i+1}^2 \right) + B_i \left(X_{i+1} - X_i \right) \right],$$

Полный квадрат отклонения кусочно-линейного приближения эмпирических данных:

$$J = \sum_{i=1}^{n-1} \int_{X_i}^{X_{i+1}} \left[\left(A_i x + B_i\right) - \overline{f^*} \right]^2 dx ,$$

где отдельные слагаемые вычисляются следующим образом:

$$\int_{X_i}^{X_{i+1}} \left[\left(A_i x + B_i\right) - \overline{f^*} \right]^2 dx =$$

$$= \begin{cases} \dfrac{\left(A_i X_{i+1} + B_i - \overline{f^*}\right)^3 - \left(A_i X_i + B_i - \overline{f^*}\right)^3}{3A_i}, & A_i \neq 0 \\ \dfrac{A_i}{2} \left(X_{i+1}^2 - X_{i+1}^2 \right) + B_i \left(X_{i+1} - X_i \right), & A_i = 0. \end{cases}$$

С учетом найденных параметров a и b, рассчитываем величину интегральной квадратичной ошибки линейной регрессии J_E (см. начало параграфа) и, окончательно, значение коэффициента интегральной детерминации $R_{DD}^2 = 1 - J_E / J$.

Все необходимые решения в замкнутой аналитической форме существуют также для *логарифмической регрессии* $y = a \ln x + b$.

В случае *экспоненциальной регрессии* $y = b \exp(ax)$ аналитически определяются коэффициенты A_i, B_i, интегральное среднее и полный квадрат отклонения кусочно-гладкого приближения, а также интегральная квадратичная ошибка и обе ее производные. Параметр b, аналитически выраженный через параметр a из соотношения $\partial J_E / \partial b = 0$, подставляется в соотношение $\partial J_E / \partial a = 0$. Полученное трансцендентное уравнение с одним неизвестным решается численно, при этом хорошим начальным приближением может служить подходящим образом найденное среднее коэффициентов A_i кусочно-гладкого приближения.

Сравнение моделей регрессии, построенных различными методами. Рассмотрим тестовый пример линейной регрессии. Сгенерируем две выборки исходных точек $\left\{ (X_i, Y_i),\ X_i < X_{i+1},\ i = \overline{1,n} \right\}$ одинакового объема $n = 24$ так, что первую выборку составляют регулярные (равноотстоящие) отсчеты X_i , вторую – нерегулярные отсчеты (величина интервала $X_{i+1} - X_i$ возрастает с увеличением номера *i*). Значения отклика соответствуют исходному уравнению $Y_i = 4X_i + 5 + \varepsilon_i$, где последовательные значения (наборы) равномерно распределенной случайной ошибки $\varepsilon_i \in [-10;10]$ для обеих выборок идентичны. Статистически обработанные данные о параметрах регрессионных моделей по серии из 100 экспериментов сведены в таблицу, а на

рис. 1 приведены результаты построения моделей для одного из наборов случайных величин $\{\varepsilon_i\}$.

Линии регрессии для отдельных случайных наборов исходных данных, найденные классическим и предлагаемым способами, в случае с равноотстоящими точками фактора весьма близки (в представленном варианте – практически совпадают). Это свидетельствует о корректности способа минимизации интегральной квадратичной ошибки. Регрессии отличаются более заметно, если интервалы между отсчетами фактора неодинаковы. В среднем же различие оценок коэффициентов уравнения для сравниваемых способов менее существенно. В каждом эксперименте значение коэффициента интегральной детерминации R_{DD} -квадрат превышало значение классического коэффициента детерминации. При переходе к нерегулярной модели коэффициент R-квадрат существенно снижается, а коэффициент интегральной детерминации, напротив, возрастает.

Ранее различные авторы уже отмечали недостаточную адекватность классической меры оценки качества регрессии, по крайней мере, применительно к моделям с нелинейными зависимостями. Например, в разделе "О типах линий тренда" официального сайта Microsoft (в отношении табличного процессора MS Excel) без каких-либо дополнительных пояснений указано, что "…отображаемое вместе с линией тренда значение величины R-квадрат не является корректным" [3]. Мы полагаем, что подобная некорректность проявляется также и в случае нерегулярного расположения эмпирических точек.

Преложенный подход был апробирован на практической задаче прогнозирования экстремальных значений временных рядов снеговой нагрузки [2]. Одной из особенностей задачи является выраженная нерегулярность расположения отсчетов фактора, представленного в нелинейной (дважды логарифмической) шкале. На рис. 2 приведен пример аппроксимации эмпирического вероятностного распределения годовых максимумов нагрузки на метеостанции Гродно. Исследовалась хвостовая часть распределения, включающая 12 наибольших отсчетов вариационного ряда наблюдений, с целью установления ее принадлежности к одному из двух типов – Гумбеля либо Вейбулла (соответствует, в представленных координатах, линейной либо логарифмической аппроксимации).

Традиционный анализ МНК приводит к выводу о наличии линейной регрессии, так как коэффициент детерминации $R^2 = 0{,}9745$ наилучшей логарифмической модели $s = 0{,}85 \ln x + 0{,}235$ (на рис. 2 не показана) заметно ниже. Если же применить способ минимизации интегральной квадратичной ошибки, то предпочтительна, напротив, логарифмическая регрессия, поскольку для линейной модели $s = 0{,}343x + 0{,}122$, найденной этим же способом, $R^2_{DD} = 0{,}9887$. Отметим также близкое сходство линейных моделей, полученных различными способами.

Заключение. Регрессии, найденные классическим методом наименьших квадратов, равно как и соответствующий им коэффициент детерминации R-квадрат, оказываются недостаточно корректными в случаях нелинейной зависимости отклик-фактор и/или нерегулярного расположения отсчетов на шкале фактора.

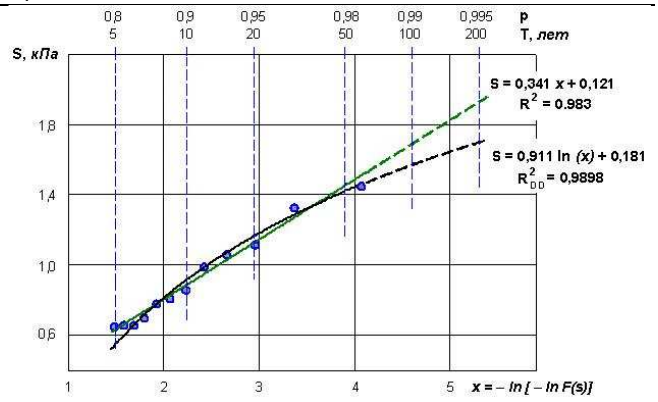

Рис. 2. Наилучшие аппроксимации хвостовой части вероятностного распределения годовых максимумов снеговой нагрузки на поверхности земли (метеостанция Гродно). Линейная регрессия получена МНК, логарифмическая регрессия – минимизацией интегральной квадратичной ошибки

Прогнозируемые значения снеговой нагрузки определяются с помощью рис. 2 по точкам пересечения экстраполированной регрессии с вертикальными пунктирными линиями, соответствующими заданному периоду повторяемости (сроку прогноза). Выбор конкретной модели регрессии при этом чрезвычайно важен, поскольку расхождение результатов прогнозирования для различных типов аппроксимации быстро нарастает с увеличением срока прогноза. На основании представленных выше результатов, учитывая также иные физические аспекты рассмотренной практической задачи, мы считаем более адекватными модели, получаемые минимизацией интегральной квадратичной ошибки.

Для оценки качества регрессии предлагается более точный коэффициент интегральной детерминации R_{DD} -квадрат (definite determinative), равный доле полного квадрата отклонения кусочно-гладкого приближения эмпирических данных, которая объяснена регрессионной моделью. Вычисления при этом выполняются на интервале изменения фактора.

Для некоторых типовых зависимостей фактор-отклик, например линейной, логарифмической и, частично, экспоненциальной, возможна аналитическая минимизация интегральной квадратичной ошибки, что позволяет находить таким способом адекватную регрессию при произвольном расположении отсчетов на шкале фактора.

Корректность предлагаемого способа вычисления коэффициентов регрессии и меры оценки ее качества подтверждены тестированием на модельном примере. Показана применимость нового подхода к решению практической задачи долговременного прогнозирования экстремальных значений временного ряда.

СПИСОК ЦИТИРОВАННЫХ ИСТОЧНИКОВ

1. Шитиков, В.К. Количественная гидроэкология: методы системной идентификации / В.К. Шитиков, Г.С. Розенберг, Т.Д. Зинченко – Тольятти: ИЭВБ РАН, 2003. – 463 с.
2. Тур, В.В. Нормирование снеговых нагрузок для территории Республики Беларусь / В.В. Тур, В.В. Валуев, С.С. Дереченник [и др.] // Строительная наука и техника. – 2008, № 2 (17). – С. 27–45.
3. О типах линий тренда. Значение R-квадрат // Microsoft Office online: Microsoft Office Excel [Электронный ресурс]. – 2009. – Режим доступа.
4. http://office.microsoft.com/ru-ru/excel/HP052000681049.aspx . – Дата доступа: 11.11.2009.

Материал поступил в редакцию 17.11.09

DERECHENNIK S.S., DMITRIEVA A.V., DERECHENNIK-jr. S.S. Integral quality assessment of regression models

The integral definite determinative – square R_{dd} – is proposed to measure quality of a regression. It equals to regression-explained fraction of the total square deviation of the sectionally smooth approximated empirical data. Calculations are carried out in a factor change interval. Despite traditional correlation coefficient – square R – the integral assessment correctness does not depend on both regression model linearity and factor references regularity in the interval of its change. The possibility of analytical minimization of integral square error is determined for some standard dependences of a factor-response type, particularly for linear, logarithmic and (partially) exponential ones. The new approach is shown to be applicable at tasks of predicting time series extreme values.