Министерство образования и науки Российской федерации
Федеральное агентство по образованию

Государственное образовательное учреждение
высшего профессионального образования
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ имени академика С. П. КОРОЛЁВА»

ИСПОЛЬЗОВАНИЕ РЕГРЕССИОННОГО АНАЛИЗА ДЛЯ ИССЛЕДОВАНИЯ ПРОЦЕССОВ УПРАВЛЕНИЯ СЛОЖНЫМИ СИСТЕМАМИ

САМАРА 2005

Министерство образования и науки российской федерации
Федеральное агентство по образованию

Государственное образовательное учреждение
высшего профессионального образования
«САМАРСКИЙ ГОСУДАРСТВЕННЫЙ АЭРОКОСМИЧЕСКИЙ
УНИВЕРСИТЕТ имени академика С. П. КОРОЛЁВА»

ИСПОЛЬЗОВАНИЕ РЕГРЕССИОННОГО АНАЛИЗА ДЛЯ ИССЛЕДОВАНИЯ ПРОЦЕССОВ УПРАВЛЕНИЯ СЛОЖНЫМИ СИСТЕМАМИ

Методические указания

САМАРА 2005

Составители: А. Б. Прокофьев, Е. В. Шахматов

УДК 681.51

Использование регрессионного анализа для исследования процессов управления сложными системами: Метод. указания / СГАУ. Сост. А. Б. Прокофьев, Е. В. Шахматов. – Самара, СГАУ, 2005. – 23 с.

Изложены задачи регрессионного анализа и его приложение к исследованиям процессов при управлении сложными системами. Кратко представлены математические основы регрессионного анализа. Рассмотрены вопросы использования системы компьютерной математики MathCAD для решения задач регрессионного анализа.

Предназначены для студентов специальности 06.08.00 и специализации 13.02.09, изучающих курс «Теория управления сложными системами».

Печатаются по решению редакционно-издательского совета Самарского государственного аэрокосмического университета имени академика С. П. Королёва.

Рецензент: д. т. н., проф. Загузов И. С.

СОДЕРЖАНИЕ

1. ЗАДАЧИ И ПРИЛОЖЕНИЯ РЕГРЕССИОННОГО АНАЛИЗА

Важное значение для описания процессов управления играет математический анализ данных, полученных в процессе функционирования сложной системы. При этом на поведение системы большое влияние оказывают многочисленные внутренние и внешние параметры, которые могут быть, в том числе, и случайного характера. Поэтому точное математическое описание процессов управления в большинстве случаев невозможно. В связи с этим возникает задача представления сложных процессов управления набором простых функций, т.е. задача аппроксимации. На практике данные, полученные в результате регистрации рассматриваемых процессов, могут иметь погрешности, порою весьма значительные. В связи с этим возникает целый спектр задач обработки таких данных, например, сглаживание данных, очистка их от шума. Проведение аппроксимации с одновременной статистической обработкой данных обеспечивается методами регрессионного анализа.

Рассмотрим ряд приложений регрессионного анализа к процессам в сложных социально-экономических системах.

Прежде всего отметим, что регрессионная зависимость представляет собой аппроксимирующую функцию, график которой проходит в «облаке» экспериментальных (зарегистрированных) точек таким образом, что суммарная среднеквадратичная погрешность для всех точек является минимальной.

В работе [1] регрессионный анализ применяется для определения зависимости изменения цены от изменения технико-экономических параметров продукции. Регрессионная зависимость определяется по формуле:

$$\boldsymbol{P} = \boldsymbol{f}\left(\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}\right),$$

где $\boldsymbol{P}$ – цена продукции,

$\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ - параметры изделия.

Метод регрессионного анализа в этом случае позволяет моделировать изменение цен в зависимости от параметров изделий, определять аналитическую форму связи стоимости и характеристик продукции и использовать полученные уравнения регрессии для определения цен изделий, входящих в параметрический ряд.

Построение регрессионной зависимости изменения цены от технических параметров включает следующие этапы:

– отбор параметров, в наибольшей степени влияющих на цены изделий параметрического ряда;

– выбор формы изменения цен в зависимости от параметров;

– построение системы уравнений в соответствии с принятыми базовыми функциями;

– расчет коэффициентов регрессионной зависимости цен от параметров для параметрического ряда.

При этом могут быть использованы различные уравнения регрессии [2, 3]:

– линейное

$$y = a_0 + \sum_{i=1}^{n} a_i x_i \, ,$$

– степенное

$$y = a_0 \prod_{i=1}^{n} x_i^{a_i} \, ,$$

– параболическое

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} b_i x_i^2 \, .$$

Так, для некоторого семейства центробежных насосов регрессионная зависимость цены изделия от обеспечиваемой им подачи воды имеет вид:

$$P = 390{,}7 + 204{,}7\, x_1 \, ,$$

где x_1 - подача воды центробежным насосом, м^3/ч.

Исходя из полученной регрессионной зависимости, можно определить цену центробежного насоса, обеспечивающего подачу воды $x_1 = 580$ *м^3/ч:*

$$P = 390{,}7 + 204{,}7 \cdot 580 = 119100 \text{ руб.}$$

Метод регрессионного анализа успешно применяется в рыночной экономике. Предположим, фирма, производящая автомобили, разработала новую модель легкового автомобиля. Перед тем как запустить эту модель в производство, фирма желает определить будущую прибыль. Для этого она должна определить будущую цену своего автомобиля, которую «примет» рынок. Допустим, в данный момент на рынке реализуется 30 моделей автомобилей. Используя данные по этим моделям, можно построить уравнение регрессии, характеризующее зависимость цены от основных потребительских параметров. Полученную регрессию фирма может использовать для прогноза цены на свою модель и для определения прибыльности ее производства. Более того, она может использовать это уравнение регрессии при установлении первоначальной (пробной) цены на свою новую модель. Возможно, эта цена окажется завышенной, а объем продаж – ниже планируемого фирмой. В этом случае фирма может несколько понизить цену, либо улучшить модель при неизменной цене, либо увеличить расходы на рекламу, либо снять модель с производства. Первоначальная цена может оказаться заниженной, и возникает дефицит автомобилей новой модели. В этом случае фирма может повысить цену.

Другим примером использования регрессионного анализа является определение связи итогов голосования за власть с показателями, учитывающими уровень экономического развития региона [4]. В результате анализа была получена регрессионная зависимость:

$$Y = -29{,}68 + 5{,}77N + 229500D^3 - 4 \cdot 10^{-7}Z^3 + 1{,}33\log_2 S + 0{,}77Nt + 9{,}37St,$$

где Y – результаты голосования;

N – средние налоговые поступления в бюджет за 11 месяцев до выборов;

D – дебиторская задолженность за 2 месяца до выборов;

Z – среднее значение за 10 месяцев до выборов задолженности по заработной плате (в процентах) к фонду заработной платы;

S – число действующих предприятий с иностранным участием;

Nt – доля населения в трудоспособном возрасте;

St – численность студентов государственных высших учебных заведений на начало учебного года.

Построенная по рассматриваемым переменным математическая модель показывает, что на результаты выборов существенное влияние оказывает уровень развития региона. Чем лучше ситуация в регионе, чем больше в нем развита экономика, тем больше население поддерживает существующую власть. Таким образом, проведенный регрессионный анализ показал, что результаты голосования на выборах определяются экономической ситуацией в регионе. Чем лучше по сравнению с другими субъектами Российской Федерации развит регион, тем больше население поддерживает существующую власть.

Приведенные примеры позволяют оценить значимость регрессионного анализа при построении математических моделей процессов управления в сложных социально-экономических системах. Однако, для эффективного использования методов регрессионного анализа необходимо знание математических положений, на которых он построен.

2. МАТЕМАТИЧЕСКИЕ ОСНОВЫ РЕГРЕССИОННОГО АНАЛИЗА

Допустим, проведено n экспериментов, в каждом из которых учитывается m взаимно независимых входных переменных. Через x_{ij} обозначим точно определенное значение i-ой входной переменной, зарегистрированное в j-ом эксперименте. Предположим, что существует некоторая зависимость $f(x_i, a_0,$

a_i, $i = 1…m$), которая может рассматриваться как приближение к неизвестной, но объективно существующей зависимости $y(x_i$, $i = 1…m)$, точки которой представлены как $y_j(x_{ij}$, $i= 1…m$, $j = 1…n)$. Тогда можно записать:

$$y_j = f(x_{ij}, a_0, a_i, i = 1…m, j = 1…n) + \xi_j (j = 1…n).$$

В последнем выражении ξ_j – случайная величина, которая считается ошибкой эксперимента.

Задача регрессии заключается в том, чтобы найти параметры $a_i(i = 0…m)$ такими, при которых представление $y(x_i$, $i = 1…m)$ функцией $f(x_i$, a_0, a_i, $i = 1…m)$ имело бы наименьшую среднеквадратичную погрешность. Для этого в соответствии с методом наименьших квадратов необходимо минимизировать функцию

$$\Phi\left(a_i, i = 0...m\right) = \sum_{j=1}^{n}\left[f\left(x_{ij}, a_0, a_i, i = 1...m\right) - y_j\right]^2. \quad (1)$$

Рассмотрим случай линейной модели

$$\hat{y} = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m. \quad (2)$$

В последнем выражении через $\hat{y}$ обозначено модельное значение выходной величины.

Если в уравнение (2) вместо переменных x_i подставить их наблюдаемые значения, то получим систему уравнений:

$$\begin{cases} \hat{y}_1 = a_0 + a_1 x_{11} + a_2 x_{21} + ... + a_m x_{m1}, \\ \hat{y}_2 = a_0 + a_1 x_{12} + a_2 x_{22} + ... + a_m x_{m2}, \\ . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \\ \hat{y}_n = a_0 + a_1 x_{1n} + a_2 x_{2n} + ... + a_m x_{mn}. \end{cases} \quad (3)$$

Значения $\hat{y}_j$, полученные вычислением из (3), будут отличаться от измеренных значений y_j на величину случайной ошибки ξ_j.

Для случая линейной модели выражение (1) можно переписать в виде:

$$\Phi = \sum_{j=1}^{n} \left[\left(a_0 + a_1 x_{1j} + a_2 x_{2j} + \ldots + a_m x_{mj} \right) - y_j \right]^2 .$$

Таким образом, $\Phi = \Phi\left(a_i, i = 0 \ldots m\right)$. Минимум этой функции можно найти, используя необходимые условия экстремума функции нескольких переменных:

$$\frac{\partial \Phi}{\partial a_0} = 0, \qquad \frac{\partial \Phi_1}{\partial a_1} = 0, \qquad \ldots, \quad \frac{\partial \Phi_m}{\partial a_m} = 0. \qquad (4)$$

Если система (4) имеет единственное решение, то оно будет искомым.

Проведем соответствующие вычисления для случая, когда

$$\hat{y} = a_0 + a_1 x ,$$

тогда:

$$\Phi = \sum_{j=1}^{n} \left[\left(a_0 + a_1 x_j \right) - y_j \right]^2 = \left(a_0 + a_1 x_1 - y_1 \right)^2 + \ldots + \left(a_0 + a_1 x_n - y_n \right)^2 ,$$

откуда

$$\frac{\partial \Phi}{\partial a_0} = 2\left(a_0 + a_1 x_1 - y_1 \right) + \ldots + 2\left(a_0 + a_1 x_n - y_n \right),$$

$$\frac{\partial \Phi}{\partial a_1} = 2 x_1 \left(a_0 + a_1 x_1 - y_1 \right) + \ldots + 2 x_n \left(a_0 + a_1 x_n - y_n \right).$$

Приравняв частные производные к нулю и разделив на 2, получим систему двух линейных уравнений с двумя неизвестными:

$$\begin{cases} a_0 n + a_1 \sum_{j=1}^{n} x_j = \sum_{j=1}^{n} y_j, \\ a_0 \sum_{j=1}^{n} x_j + a_1 \sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} x_j y_j. \end{cases} \tag{5}$$

Решая эту систему, найдем a_0 и a_1. Из первого уравнения системы (5) можно записать:

$$a_0 = \frac{1}{n}\left(\sum_{j=1}^{n} y_j - a_1 \sum_{j=1}^{n} x_j \right). \tag{6}$$

Подставим это соотношение во второе уравнение системы (5):

$$\frac{1}{n}\sum_{j=1}^{n} y_j \sum_{j=1}^{n} x_j - \frac{a_1}{n}\left(\sum_{j=1}^{n} x_j \right)^2 + a_1 \sum_{j=1}^{n} x_j^2 = \sum_{j=1}^{n} x_j y_j ,$$

откуда

$$a_1 = \frac{\sum_{j=1}^{n} x_j y_j - \frac{1}{n}\sum_{j=1}^{n} x_j \sum_{j=1}^{n} y_j}{\sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j \right)^2}. \tag{7}$$

По величине a_1, определенной из (7), можно рассчитать по выражению (6) значение a_0.

Аналогичным образом можно получить выражения и для других видов регрессии: полиноминальной, экспоненциальной, логарифмической и т. д. Ввиду сложности соответствующих выражений они здесь не приводятся.

Мерой оптимальности подбора функции регрессии часто служит коэффициент взаимного соответствия данных, называемый коэффициентом корреляции. Близость его модуля к 1 означает, что данные хорошо коррелированы.

Для случая линейного уравнения регрессии

$$\hat{y} = a_0 + a_1 x$$

коэффициент корреляции определяется по соотношению:

$$r_{x,y} = \frac{\sum_{j=1}^{n}\left(x_i - x_{cp}\right)\left(y_i - y_{cp}\right)}{\sqrt{\sum_{j=1}^{n}\left(x_i - x_{cp}\right)^2 \sum_{j=1}^{n}\left(y_i - y_{cp}\right)^2}}, \qquad (8)$$

где $x_{cp} = \frac{1}{n}\sum_{j=1}^{n} x_j$,

$$y_{cp} = \frac{1}{n}\sum_{j=1}^{n} y_j .$$

Величина коэффициента корреляции изменяется в интервале от -1 до +1.

Если $r_{x,y} = 0$, то корреляционная связь между x и y отсутствует, т. е. x и y не связаны друг с другом (см. рис. 1,а).

Если $r_{x,y} > 0$, то корреляционная связь между x и y есть. При этом с увеличением x значения y увеличиваются (см. рис. 1,б).

Если $r_{x,y} < 0$, то корреляционная связь между x и y есть, но с увеличением величины x значения y уменьшаются (см. рис. 1,в).

Если $r_{x,y} = \pm 1$, то связь между x и y становится функциональной (см. рис. 1,г и 1,д).

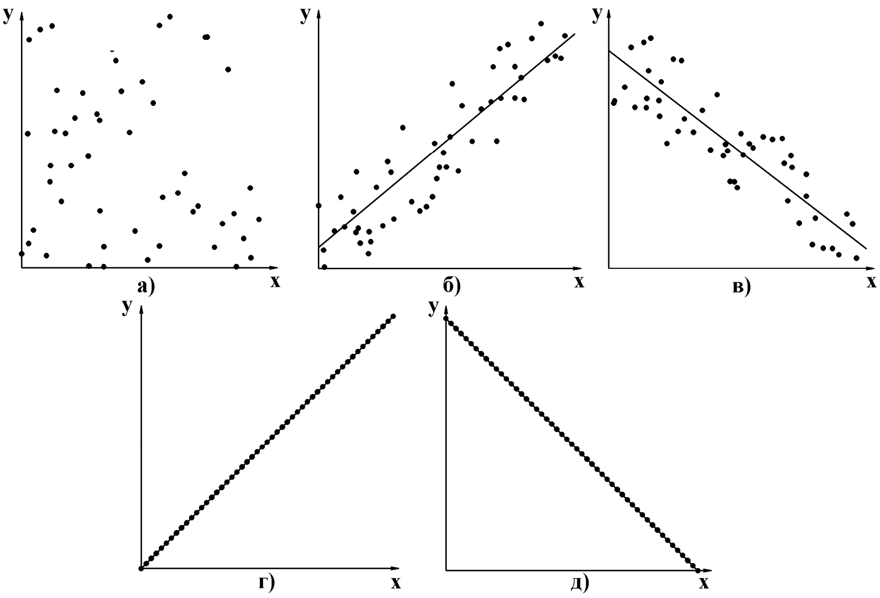

Рисунок 1. Аппроксимация экспериментальных точек линейной регрессионной зависимостью и коэффициент корреляции:
а) $r_{x,y} = 0$; б) $r_{x,y} \in (0;1)$; в) $r_{x,y} \in (-1;0)$; г) $r_{x,y} = 1$; д) $r_{x,y} = -1$.

3. ПРИМЕР ПОСТРОЕНИЯ ЛИНЕЙНОЙ РЕГРЕССИОННОЙ ЗАВИСИМОСТИ

Построим линейную регрессионную зависимость по экспериментальным данным, представленным в таблице 1.

Таблица 1. Исходные данные примера построения линейной регрессионной зависимости.

№ опроса	1	2	3	4	5	6	7	8	9	10	11	12
Значение независимой переменной x_i	3,7	9,7	10,3	11,3	12	13	22	28,8	38,2	40,1	42,4	43,3
Значение зависимой переменной y_i	16,27	17,18	15,45	17,18	15,94	15,2	18,06	18,91	19,46	20,21	20,44	18,21

Для решения этой задачи необходимо воспользоваться соотношениями (6) и (7). При этом промежуточные результаты вычислений целесообразно заносить в таблицу (см. таблицу 2), а сами расчеты проводить с использованием программы Excel.

$$a_1 = \frac{\sum_{j=1}^{n} x_j y_j - \frac{1}{n}\sum_{j=1}^{n} x_j \sum_{j=1}^{n} y_j}{\sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2} = \frac{5119{,}86 - \frac{1}{12}\cdot 274{,}8 \cdot 212{,}51}{8707{,}90 - \frac{1}{12}\cdot 274{,}8^2} = 0{,}105,$$

$$a_0 = \frac{1}{n}\left(\sum_{j=1}^{n} y_j - a_1 \sum_{j=1}^{n} x_j\right) = \frac{1}{12}\left(212{,}51 - 0{,}105 \cdot 274{,}8\right) = 15{,}306 .$$

Таким образом, искомая регрессионная зависимость имеет вид:

$$\hat{y} = 15{,}306 + 0{,}105x .$$

По соотношению (8) несложно определить коэффициент корреляции:

$$r_{x,y} = 0{,}862 .$$

Таблица 2. Промежуточные результаты вычислений линейной регрессионной зависимости.

№ опыта	x_i	y_i	$x_i y_i$	x_i^2
1	3,7	16,27	60,20	13,69
2	9,7	17,18	166,65	94,09
3	10,3	15,45	159,14	106,09
4	11,3	17,18	194,13	127,69
5	12,0	15,94	191,28	144,00
6	13,0	15,20	197,60	169,00
7	22,0	18,06	397,32	484,00
8	28,8	18,91	544,61	829,44
9	38,2	19,46	743,37	1459,24
10	40,1	20,21	810,42	1608,01
11	42,4	20,44	866,66	1797,76
12	43,3	18,21	788,49	1874,89
$\sum$	274,8	212,51	5119,86	8707,90
Сред. знач.	22,9	17,71		

4. ИСПОЛЬЗОВАНИЕ СИСТЕМЫ КОМПЬЮТЕРНОЙ МАТЕМАТИКИ MATHCAD ДЛЯ РЕШЕНИЯ ЗАДАЧ РЕГРЕССИОННОГО АНАЛИЗА

Предыдущий пример показал, что даже для получения простейшей линейной регрессионной зависимости приходится проводить хоть и несложные, но достаточно трудоемкие вычисления. Поэтому при решении задач регрессионного анализа целесообразно использование специализированных функций широко распространенных систем компьютерной математики, таких как, например, MathCAD.

4.1. Построение линейной регрессии

Для проведения линейной регрессии в систему MathCAD встроен ряд приведенных ниже специализированных функций.

● ***intercrpt (x, y)*** – возвращает значение $\boldsymbol{a_0}$ искомой зависимости $\boldsymbol{y = a_0 + a_1 x}$, характеризующее смещение линии регрессии по вертикали;

● ***slope (x, y)*** – возвращает значение $\boldsymbol{a_1}$ искомой зависимости $\boldsymbol{y = a_0 + a_1 x}$, характеризующее угол наклона регрессионной зависимости;

● ***corr (x, y)*** – возвращает значение коэффициента корреляции $\boldsymbol{r_{x,y}}$.

На рис. 2 показан пример проведения линейной регрессии для данных, представленных значениями элементов в векторах ***X*** и ***У***. При этом точки в векторе X могут отстоять друг от друга на различных расстояниях. Линия регрессии проходит в «облаке» исходных точек с максимальным среднеквадратичным приближением к ним. Рассчитанный с помощью системыMath CAD коэффициент корреляции $\boldsymbol{r_{x,y} = 0,862}$ совпадает с результатами вычислений по соотношению (8) в приведенном выше примере.

Вектор измеренных независимых переменных

Вектор измеренных зависимых переменных

$$X := \begin{pmatrix} 3.7 \\ 9.7 \\ 10.3 \\ 11.3 \\ 12 \\ 13 \\ 22 \\ 28.8 \\ 38.2 \\ 40.1 \\ 42.4 \\ 43.3 \end{pmatrix} \qquad Y := \begin{pmatrix} 16.27 \\ 17.18 \\ 15.45 \\ 17.18 \\ 15.94 \\ 15.2 \\ 18.06 \\ 18.91 \\ 19.46 \\ 20.21 \\ 20.44 \\ 18.21 \end{pmatrix}$$

Задание числа опытов или измерений

$N := 12$

$i := 0..N-1$

Вычисление коэффициентов регрессионной зависимости

$a := \mathrm{intercept}(X, Y)$ $\qquad b := \mathrm{slope}(X, Y)$

$a = 15.306$ $\qquad b = 0.105$

$f(x) := a + b \cdot x$

Вычисление коэффициента корреляции

$\mathrm{corr}(X, Y) = 0.862$

График регрессионной зависимости и исходных точек

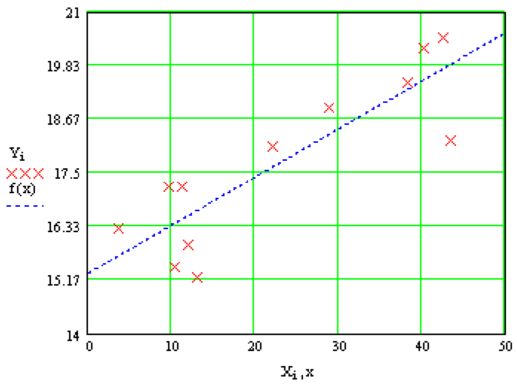

Рисунок 2. Вычисление линейной регрессионной зависимости вида $y = a + bx$ и коэффициента корреляции в системе компьютерной математики MathCAD.

4.2. Построение линейной регрессии общего вида

При построении линейной регрессии общего вида заданная совокупность экспериментальных точек аппроксимируется функцией

$$f(x, k_1, k_2 \ldots k_n) = k_1 \cdot f_1(x) + k_2 \cdot f_2(x) + \ldots + k_n f_n(x).$$

Таким образом, регрессионная зависимость является линейной комбинацией функций $f_1(x)$, $f_2(x)$,..., $f_n(x)$, причем сами эти функции могут быть нелинейными, что резко расширяет возможности такой аппроксимации и распространяет ее на многие нелинейные функции.

Для реализации линейной регрессии общего вида используется специальная функция $linfit(X, Y, F)$. Она возвращает вектор коэффициентов $k_1, k_2, \ldots, k_n$ линейной регрессии общего вида, при котором среднеквадратичная погрешность «облака» исходных точек будет минимальной.

Векторы X и Y, являющиеся входными переменными функции $l\inf it(X, Y, F)$, содержат соответствующие координаты исходных точек. Вектор F должен содержать базовые аппроксимирующие функции $f_1(x)$, $f_2(x)$,..., $f_n(x)$, записанные в символьном виде.

На рисунке 3 представлено проведение линейной регрессии общего вида с применением функции $l\inf it$. Координаты точек исходного вектора X могут быть расположены на произвольном расстоянии друг от друга, но в порядке возрастания. Вектор Y должен содержать координаты, соответствующие абсциссам в векторе X.

Регрессионная зависимость для аппроксимации исходных точек (см. рис. 3) ищется в виде

$$f(x) = k_1 \cdot \frac{1}{1+x} + k_2 x + k_3.$$

Проведение процедуры, представленной на рис. 3, позволяет получить регрессионную зависимость:

$$f(x) = \frac{95{,}08}{1+x} + 18{,}07x - 46{,}45.$$

$$X := \begin{pmatrix} -0.387 \\ -0.147 \\ -0.08 \\ 0.1 \\ 0.123 \\ 0.436 \\ 0.644 \\ 0.995 \\ 1.377 \\ 1.593 \\ 1.778 \\ 2.122 \\ 2.434 \end{pmatrix} \qquad Y := \begin{pmatrix} 112.1 \\ 54.1 \\ 43.2 \\ 44.1 \\ 35.9 \\ 34.4 \\ 21.1 \\ 22.3 \\ 24.5 \\ 27 \\ 20.1 \\ 22.1 \\ 15.5 \end{pmatrix}$$

Вектор базовых аппроксимирующих функций

$$F(x) := \begin{pmatrix} \dfrac{1}{1+x} \\ x \\ 1 \end{pmatrix}$$

$i := 0..\ 12$

$K := \mathrm{linfit}(X, Y, F)$

$g(x) := F(x) \cdot K$

Коэффициенты регрессионной зависимости

$$K = \begin{pmatrix} 95.083 \\ 18.066 \\ -46.454 \end{pmatrix}$$

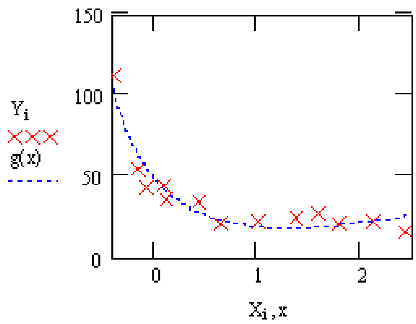

Рисунок 3. Вычисление линейной регрессионной зависимости общего вида в системе компьютерной математики MathCAD.

4.3. Построение нелинейной регрессии общего вида

Под нелинейной регрессией общего вида подразумевается нахождение n параметров $(k_1, k_2, ..., k_n)$ произвольной функции $f(x, k_1, k_2, ..., k_n)$, обеспечивающих минимальную среднеквадратичную погрешность приближения «облака» исходных точек.

Для проведения нелинейной регрессии общего вида используется функция $genfit(X, Y, s, F)$. Она возвращает вектор n параметров $(k_1, k_2, ..., k_n)$ функции $f(x, k_1, k_2, ...k_n)$, делающий минимальную среднеквадратичную погрешность приближения исходных данных.

Вектор F в функции $genfit(X, Y, s, F)$ должен быть вектором с символьными элементами, причем эти элементы должны представлять собой аналитические выражения для исходной функции и ее производных по всем параметрам k_i.

Вектор s в функции $genfit(X, Y, s, F)$ должен содержать начальные значения элементов вектора $k_1, k_2, ..., k_n$, необходимые для решения формируемой программой системы нелинейных уравнений регрессии итерационным методом.

На рис. 4 показан пример выполнения нелинейной регрессии общего вида с помощью нелинейной функции

$$f(x, a, b) = a\,sin(bx) + abe^{-\frac{b}{a}x}.$$

При решении этой задачи возникают 3 проблемы. Во-первых, надо вычислить значения производных по переменным a и b. В файле на рис. 4 это сделано средствами символьных операций.

$$X := \begin{pmatrix} 0 \\ 0.118 \\ 0.621 \\ 0.922 \\ 0.938 \\ 1.282 \\ 1.601 \\ 2.106 \\ 2.501 \\ 3.006 \\ 3.072 \\ 3.261 \end{pmatrix} \qquad Y := \begin{pmatrix} 64.63 \\ 62.51 \\ 15.31 \\ 23.95 \\ 28.31 \\ 6.81 \\ 12.6 \\ 5.77 \\ 12.12 \\ -3.06 \\ -8.33 \\ 8.16 \end{pmatrix}$$

$$i := 0..11$$

$$s := \begin{pmatrix} 2 \\ 5 \end{pmatrix}$$

$$f(x,a,b) := a \cdot \sin(b \cdot x) + a \cdot b \cdot e^{\frac{-b}{a} \cdot x}$$

Вычисление производных по переменным а и b

$$\frac{d}{da} f(x,a,b) \rightarrow \sin(b \cdot x) + b \cdot \exp\left(\frac{-b}{a} \cdot x\right) + \frac{1}{a} \cdot b^2 \cdot x \cdot \exp\left(\frac{-b}{a} \cdot x\right)$$

$$\frac{d}{db} f(x,a,b) \rightarrow a \cdot \cos(b \cdot x) \cdot x + a \cdot \exp\left(\frac{-b}{a} \cdot x\right) - b \cdot x \cdot \exp\left(\frac{-b}{a} \cdot x\right)$$

$$F(x,k) := \begin{bmatrix} k_0 \cdot \sin(k_1 \cdot x) + k_0 \cdot k_1 \cdot e^{\frac{-k_1}{k_0} \cdot x} \\ \sin(k_1 \cdot x) + k_1 \cdot \exp\left(\frac{-k_1}{k_0} \cdot x\right) + \frac{1}{k_0} \cdot (k_1)^2 \cdot x \cdot \exp\left(\frac{-k_1}{k_0} \cdot x\right) \\ k_0 \cdot \cos(k_1 \cdot x) \cdot x + k_0 \cdot \exp\left(\frac{-k_1}{k_0} \cdot x\right) - k_1 \cdot x \cdot \exp\left(\frac{-k_1}{k_0} \cdot x\right) \end{bmatrix}$$

$$P := \mathrm{genfit}(X, Y, s, F)$$

$$g(x) := f(x, P_0, P_1)$$

Вектор Р возвращает значения a=k и b=k для наилучшего среднеквадратичного приближения

$$P = \begin{pmatrix} 8.277 \\ 5.961 \end{pmatrix}$$

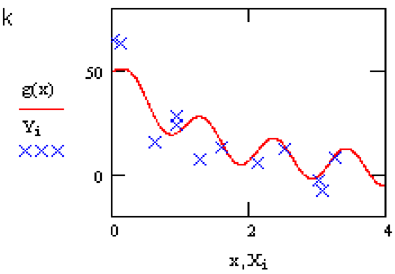

Рисунок 4. Вычисление нелинейной регрессионной зависимости в системе компьютерной математики MathCAD.

Вторая проблема связана с необходимостью применения функции ***genfit*** в ее стандартном виде. Поэтому при задании вектора ***F*** произведена замена параметра ***a*** на $\boldsymbol{k_0}$, а параметра ***b*** на $\boldsymbol{k_1}$.

Третья проблема определяется тем, что различные значения вектора ***s***, содержащие начальные условия для поиска итерационным методом величин $\boldsymbol{k_0}$ и $\boldsymbol{k_1}$, в принципе, могут привести к различным решениям.

Это объясняется полимодальным характером поверхности среднеквадратичных погрешностей. Сказанное иллюстрируется на рис. 5.

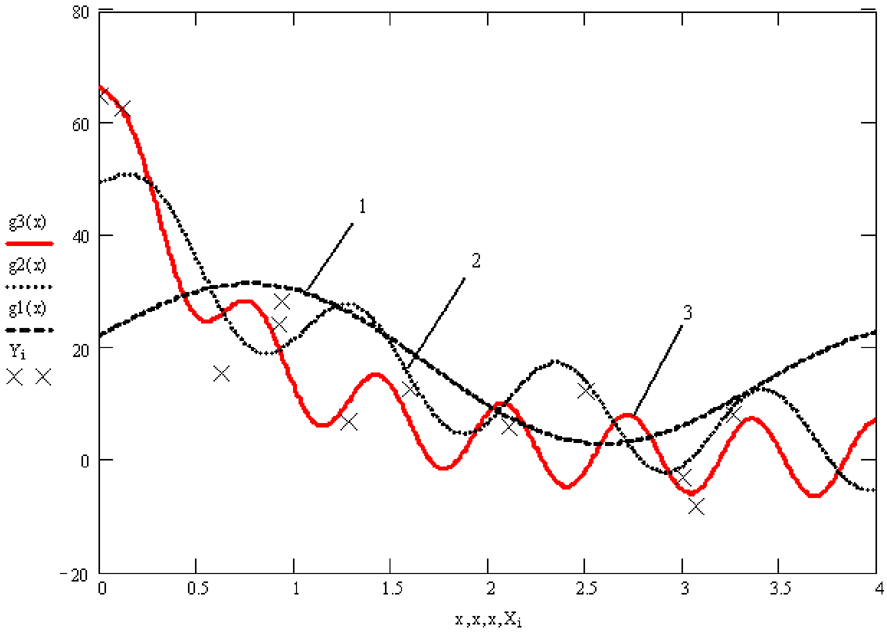

Рисунок 5. Влияние начальных условий итерационного поиска на вид нелинейной регрессионной зависимости.

Все 3 кривые получены при различных значениях вектора s :

1) кривая 1: $s = \begin{pmatrix} 6 \\ 2 \end{pmatrix}$

$$f(x) = 11{,}90\, sin(1{,}86x) + 22{,}13e^{-0{,}1563x}$$

2) кривая 2: $s = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$

$$f(x) = 8{,}277\, sin(5{,}961x) + 49{,}34e^{-0{,}7202x}$$

3) кривая 3: $s = \begin{pmatrix} 9 \\ 11 \end{pmatrix}$

$$f(x) = 6{,}77\, sin(9{,}814x) + 66{,}44e^{-1{,}450x}.$$

Для определения оптимальности той или иной полученной модели необходимо провести дополнительное исследование.

ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ

1. Изучить материалы, представленные в методических указаниях.
2. С использованием программы Excel и соотношений (6)-(8) для заданного варианта (см. приложение 1) построить линейную регрессионную зависимость и определить коэффициент корреляции.
3. Для заданного варианта (см. приложение 1) построить линейную регрессионную зависимость вида $\boldsymbol{y = ax + b}$ и определить коэффициент корреляции с использованием системы компьютерной математики MathCAD.
4. С использованием системы MathCAD для заданного варианта (см. приложение 2) построить линейную регрессионную зависимость общего вида.
5. С использованием MathCAD для заданного варианта (см. приложение 3) построить нелинейную регрессионную зависимость.
6. Оформить отчет.

Приложение 1.

Варианты заданий к линейному регрессионному анализу. Построение регрессионной зависимости вида $y=ax+b$.

№ вар.	№ опроса	1	2	3	4	5	6	7	8	9	10	11	12	13
1	Знач. незав. перем.	0.0	5.0	8.0	9.4	13.6	15.5	18.8	18.9	20.3	23.2	27.4	29.8	33.5
	Знач. завис. перем.	10	6	-1	1	-13	-10	-21	-16	-24	-24	-35	-36	-42
2	Знач. незав. перем.	0.0	1.4	5.5	7.9	13.7	14.9	19.9	22.0	22.6	23.7	30.6	31.4	
	Знач. завис. перем.	11.4	18.2	22.8	26.8	40.7	39.9	53.6	55.7	52.6	59.7	70.1	67.7	
3	Знач. незав. перем.	3.1	3.5	7.1	10.6	13.8	14.3	14.6	16.9	20.5	23.9	25.8		
	Знач. завис. перем.	-4	-18	-53	-72	-107	-94	-105	-119	-143	-176	-187		
4	Знач. незав. перем.	0.12	0.38	0.39	1.18	1.39	1.95	2.06	2.81	3.36	3.79	4.49	4.93	5.50
	Знач. завис. перем.	7	6	9	1	-1	-9	-6	-14	-14	-17	-23	-23	-28
5	Знач. незав. перем.	8.0	10.3	19.5	26.1	31.1	36.0	41.1	48.0	54.1	54.2	55.1	63.8	71.2
	Знач. завис. перем.	40	48	89	115	147	158	186	194	225	218	220	268	296
6	Знач. незав. перем.	2.2	5.2	5.3	9.5	13.5	15.1	15.2	15.7	21.7	27.4	32.2	35.9	36.9
	Знач. завис. перем.	-3	-19	-9	-44	-71	-78	-75	-83	-116	-150	-166	-200	-198

Приложение 2.

Варианты заданий к линейному регрессионному анализу общего вида.

№ вар.	№ опроса	1	2	3	4	5	6	7	8	9	10	11	12	13
Искомая регрессионная зависимость - $y(x)=k_1\sqrt{x}+k_2\frac{1}{\sqrt{x}}+k_3x^2$														
1	Знач. незав. перем.	0.14	0.15	0.26	0.37	0.53	0.67	0.83	0.86	0.92	0.98	1.00	1.17	1.29
	Знач. завис. перем.	16.5	20.2	7.2	6.3	12.3	6.1	8.1	25.4	25.7	13.1	17.4	25.2	24.4
Искомая регрессионная зависимость - $y(x)=k_1x+k_2+\frac{k_3}{x}$														
2	Знач. незав. перем.	0.17	0.45	0.65	0.80	0.94	1.10	1.30	1.49	1.50	1.52	1.78	2.00	2.12
	Знач. завис. перем.	43.3	23.7	20.5	25.0	20.8	24.1	15.5	19.1	15.6	14.7	21.9	22.2	19.4
Искомая регрессионная зависимость - $y(x)=k_1\ln x+k_2x^{0.25}+k_3x$														
3	Знач. незав. перем.	0.00	0.16	0.20	0.21	0.22	0.29	0.30	0.33	0.36	0.54	0.58	0.71	0.89
	Знач. завис. перем.	-42	-10	-7	-7	-6	-5	-5	-3	-4	0	0	3	5

Приложение 3.

Варианты заданий к нелинейному регрессионному анализу.

№ вар.	№ опроса	1	2	3	4	5	6	7	8	9	10	11	12	13
Искомая регрессионная зависимость - $y(x) = ab + ax^b$														
1	Знач. незав. перем.	0.00	0.04	0.07	0.25	0.35	0.59	0.63	0.84	0.89	1.09	1.13	1.31	1.38
	Знач. завис. перем.	9.0	15.4	12.0	7.3	12.7	12.4	15.6	14.3	16.4	11.4	14.3	20.0	22.3
Искомая регрессионная зависимость - $y(x) = ab + x^{\frac{a}{b}}$														
2	Знач. незав. перем.	0.01	0.17	0.40	0.56	0.82	1.02	1.21	1.29	1.38	1.41	1.65	1.69	1.71
	Знач. завис. перем.	11.1	0.1	2.5	16.4	11.5	9.0	12.4	16.1	24.9	25.8	52.7	67.2	89.9
Искомая регрессионная зависимость - $y(x) = ax + x^b$														
3	Знач. незав. перем.	0.00	0.17	0.25	0.26	0.48	0.54	0.70	0.74	0.95	1.11	1.24	1.44	1.57
	Знач. завис. перем.	1.6	2.7	0.3	4.1	-5.8	-5.9	-0.1	-7.2	-6.5	11.5	12.5	4.5	13.8

СПИСОК ЛИТЕРАТУРЫ

1. Крючкова О.Н., Попов Е.В. Классификация методов ценообразования // Маркетинг в России и за рубежом. – М.: Финпресс, 2002, №4.
2. Цены и ценообразование / Под ред. В.Е. Есипова. – СПб.: Питер, 2000.
3. Построение математических моделей систем с использованием регрессионного анализа: Метод. указания. Сост. Крючков А.Н., Перлов Ю.К., Шахматов Е.В. – Самара, 1994.
4. Мау В., Кочеткова О., Жаворонков С. Экономические факторы электорального поведения (опыт России 1995-1996 годов) – М.: Институт экономики переходного периода, 1998.

Учебное издание

ИСПОЛЬЗОВАНИЕ РЕГРЕССИОННОГО АНАЛИЗА ДЛЯ ИССЛЕДОВАНИЯ ПРОЦЕССОВ УПРАВЛЕНИЯ СЛОЖНЫМИ СИСТЕМАМИ

Составители: *Прокофьев Андрей Брониславович*
Шахматов Евгений Владимирович

Государственное образовательное учреждение высшего профессионального образования «Самарский государственный аэрокосмический университет имени академика С. П. Королёва»
443086, Самара, Московское шоссе, 34