Мальчик Ю.Н.

ОБ ОДНОМ СПОСОБЕ ОБРАБОТКИ СТАТИСТИЧЕСКИХ ДАННЫХ

Yu.N. Malchik

ONE METHOD PROCESSING OF STATISTICAL DATA

УДК: 519.2

Рассматривается задача аппроксимации статистических данных полиномами.

Коэффициенты этих полиномов предлагается искать с помощью MathCAD и MS Excel.

Solution of problem of approximation data of statistic of polynomials. This coefficients of polynomials search with the help of MathCAD and MS Excel

1. Введение

Пусть функция $f(t)$ задана табличным способом (см. табл 1)

Таблица 1

t	t_0	t_1	t_2	. . .	t_{n-1}	t_n
$f(t)$	y_0	y_1	y_2	. . .	y_{n-1}	y_n

Нужно выбрать критерий, согласно которому полином k – й степени

$$P(t) = C_0 + C_1 \cdot t + C_2 \cdot t^2 + \ldots + C_{k-1} \cdot t^{k-1} + C_k \cdot t^k ,$$

$$k < n, \quad t_0 \le t \le t_n , \qquad (1.1)$$

наилучшим образом аппроксимирует исходную кривую $f(t)$.

Парабола (1.1) не сможет пройти через $n + 1$ точку таблицы (1.1) и в действительности не пройдет ни через одну из них.

Хорошую аппроксимацию исходной функции $f(t)$ можно получить с помощью критерия наименьших квадратов, который широко применяется для обработки результатов эксперимента или обработки статистических данных.

В этом случае задача состоит в том, чтобы определить значения $C_0, C_1, C_2 \ldots C_k$ коэффициентов полинома (1.1) при которых сумма квадратов ошибок

$$S = \sum_{i=0}^{n} \left[y_i - \left(C_0 + C_1 \cdot t_i + C_1 \cdot t_i^{\,2} + \cdot \cdots + C_1 \cdot t_i^{\,k} \right) \right]^2 = \sum_{i=0}^{n} \left[y_i - P(t) \right]^2 \qquad (1.2)$$

минимальна.

Коэффициенты полинома $P(t)$ могут быть найдены из необходимых условий минимума S

$$\frac{\partial S}{\partial C_0} = 0, \quad \frac{\partial S}{\partial C_1} = 0, \ldots, \frac{\partial S}{\partial C_k} = 0 . \qquad (1.3)$$

Тогда средняя величина квадрата ошибок равна [1]

$$M = \frac{1}{n-1} \cdot \sum_{i=0}^{n} \left[y_i - \left(C_0 + C_1 \cdot t_i + C_1 \cdot t_i^{\,2} + \cdot \cdots + C_1 \cdot t_i^{\,k} \right) \right]^2$$

2. Способ определения коэффициентов, обеспечивающий минимум S

В литературе [1, с 701] приведена система линейных алгебраических уравнений для определения коэффициентов C_0 , C_1, C_2 . . . C_k , обеспечивающий минимум S. Однако эта система, на наш взгляд сложна.

Возможен другой способ для определения коэффициентов C_0 , C_1, C_2,. . ., C_k , обеспечивающий минимум суммы (1.2). Составим систему из $n + 1$ линейных алгебраических уравнений с $k + 1$ неизвестными

$$T \cdot C = y , \qquad (2.1)$$

где

$$T = \begin{pmatrix} 1 & t_0 & \cdots & t_0^{\,k} \\ 1 & t_1 & \cdots & t_1^{\,k} \\ \vdots & \vdots & \cdots & \vdots \\ 1 & t_n & \cdots & t_n^{\,k} \end{pmatrix}, \quad C = \begin{bmatrix} C_0 \\ C_1 \\ \vdots \\ C_k \end{bmatrix}, \quad y = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Умножим слева правую и левую части системы (2.1) на транспонированную матрицу T^T для матрицы T, получим

$$T^T T \cdot C = T^T y \qquad (2.2)$$

систему $k + 1$ линейных алгебраических уравнений для определения коэффициентов C_0 , C_1, C_2, . . . , C_k .

Если определитель матрицы $(T^T \cdot T)$ не равен нулю, то систему (2.2) можно решить матричным методом. Для матрицы $(T^T \cdot T)$ находим обратную матрицу $(T^T \cdot T)^{-1}$, и умножив слева систему (2.2) на матрицу $(\mathrm{T}^{\mathrm{T}} \cdot \mathrm{T})^{-1}$, получим

$$C = \left(T^T T \right)^{-1} \cdot \left(T^T y \right) \qquad (2.3)$$

решения C_0 , C_1, . . . , C_k .

Окончательный результат проще всего получить из (2.3) с помощью системы MathCAD.

2.1 Определение коэффициентов C_0 , C_1, . . . , C_k с помощью MathCAD

В [2] приведена таблица статистических данных за девять лет (см. табл.2).

Таблица 2

t (годы)	Курс (сом/$)	Внешторг (млн.$10)	Экспорт (млн.$10)	Импорт (млн.$10)
1999	45,429	105,35	45,38	59,97
2000	48,304	105,86	50,45	55,41
2001	47,719	94,33	47,61	46,72
2002	46,095	107,22	48,55	58,67
2003	44,190	129,87	58,17	71,70
2004	41,625	165,98	71,88	94,10
2005	35,499	177,33	67,20	110,13
2006	38,124	251,23	79,41	171,82
2007	35,499	355,19	113,42	241,70

Как и в [2] введем новую систему координат. За нуль примем 2000 год. Курс киргизской валюты сом по отношению к доллару обозначим y. Аппроксимируем y полиномом третьей степени

$$y1(t) = C_0 + C_1 \cdot t + C_2 \cdot t^2 + C_3 \cdot t^3 , \; -1 \le t \le 7 . \qquad (2.4)$$

Введем векторы t и y

$t := (-1 \;\; 0 \;\; 1 \;\; 2 \;\; 3 \;\; 4 \;\; 5 \;\; 6 \;\; 7) \quad t := t^T$

$y := (45.429 \;\; 48.304 \;\; 47.719 \;\; 46.095 \;\; 44.19 \;\; 41.625 \;\; 35.499 \;\; 38.124 \;\; 35.499) \quad y := y^T$

Векторы t и y потранспонировали поскольку Math CAD работает с векторами – столбцами. Введем матрицу T следующим образом

$$T := \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 1 & 0 & 1 & 2^2 & 3^2 & 4^2 & 5^2 & 6^2 & 7^2 \\ -1 & 0 & 1 & 2^3 & 3^3 & 4^3 & 5^3 & 6^3 & 7^3 \end{pmatrix} \quad T := T^T$$

Тогда по формуле (2.3) получим

$C^T = (47.9565 \;\; 1{,}1301 \;\; -1.1787 \;\; 0.1105)$.

Таким образом, полином (2.4) примет вид

$$y1(t) = 47.9565 + 1.13.1 \cdot t \; - 1.1787 \cdot t^2 + 0.1105 \cdot t^3 . \qquad (2.5)$$

Коэффициент корреляции двух векторов y и $y1(t)$ равен [3]

$corr(y,y1(t)) = 0.9713$. Следовательно, корреляция хорошая. Это очевидно из графиков (рис.1)

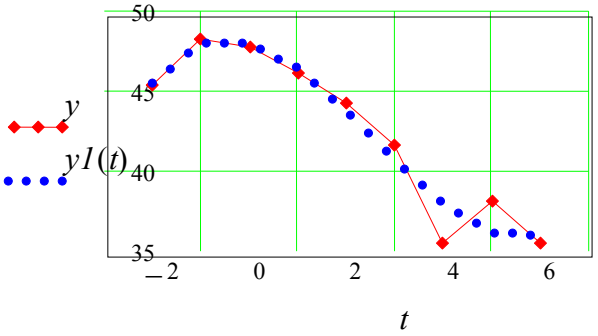

Рис 1. Графики Курс y и его аппроксимация $y1(t)$.

Аппроксимации статистических данных таб.2: Втешторг, Экспорт и Импорт могут быть найдены аналогично.

Аппроксимируем данные по экспорту Y полиномом

$$y(t) = C_0 + C_1 \cdot t + C_2 \cdot t^2 + C_3 \cdot t^3 , \; -1 \le t \le 7 . \qquad (2.6)$$

Введем вектор Y

$Y := (45.38 \;\; 50.45 \;\; 47.61 \;\; 48.55 \;\; 58.17 \;\; 71.88 \;\; 67.20 \;\; 79.41 \;\; 113.42) \quad Y := Y^T$.

Записываем формулу (2.3)

$$C = \left(T^T T\right)^{-1} \cdot \left(T^T Y\right)$$

Тогда получим

$C^T = (48.7101 \;\; 2.7123 \;\; 08335 \;\; 0.2401)$

Таким образом, полином (2.6) примет вид

$y(t) = 48.7101 + 2.7123 \cdot t - 0.8335 \cdot t^2 + 0.2401 \cdot t^3$

Коэффициент корреляции двух векторов y и $y1(t)$ равен [3]

$corr(y,y1(t)) = 0.969$. Следовательно, корреляция хорошая. Это очевидно из графиков (рис.2)

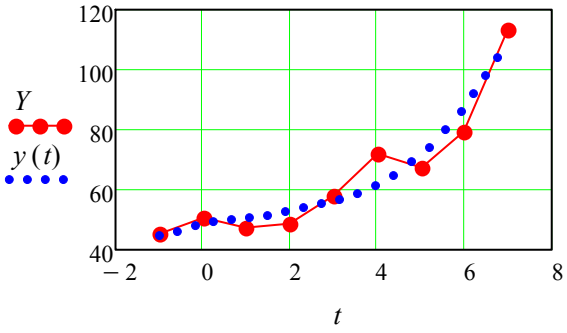

Рис 2. Графики Экспорт Y и его аппроксимация $y(t)$.

2.2 Определение коэффициентов C_0, C_1, . . . , C_k с помощью MS Excel

Внешнеторговый оборот обозначим V.

Аппроксимируем V полиномом третьей степени

$$v(t) = C_0 + C_1 \cdot t + C_2 \cdot t^2 + C_3 \cdot t^3 , \; -1 \le t \le 7 . \qquad (2.7)$$

Для отыскания коэффициентов C_i уравнения (2.7) используем стандартный набор функций «Анализ данных» табличного процессора MS Excel. Из этого набора функций нам потребуется функция «Регрессия». Далее нам потребуется Таблица 3 и функция «Регрессия».

Таблица 3

t	t^2	t^3	V	v(t)
-1	1	-1	105,35	103,2055
0	0	0	105,86	104,0688
1	1	1	94,33	104,6686
2	4	8	107,22	109,4057
3	9	27	129,87	122,6808
4	16	64	165,98	148,8946
5	25	126	177,33	192,4480
6	36	216	251,23	257,7417
7	49	343	355,12	349,1764

Для программы «Регрессия» входной параметр Y – столбец V таблицы 3, а входной параметр X – столбцы t, t^2 , t^3. Запустив программу «Регрессия» получим: искомые коэффициенты:

$C_0 = 104{,}0688$, $C_1 = -0{,}0019$, $C_2 = -0{,}13175$, $C_3 = -0{,}13175$,

функцию подбора $v(t)$ и коэффициент корреляции векторов V и $v(t)$ –

$R = 0{,}9936$. Пятый столбец таблицы 3 содержит значения функции подбора $v(t)$, подсчитанные по формуле (2.7).

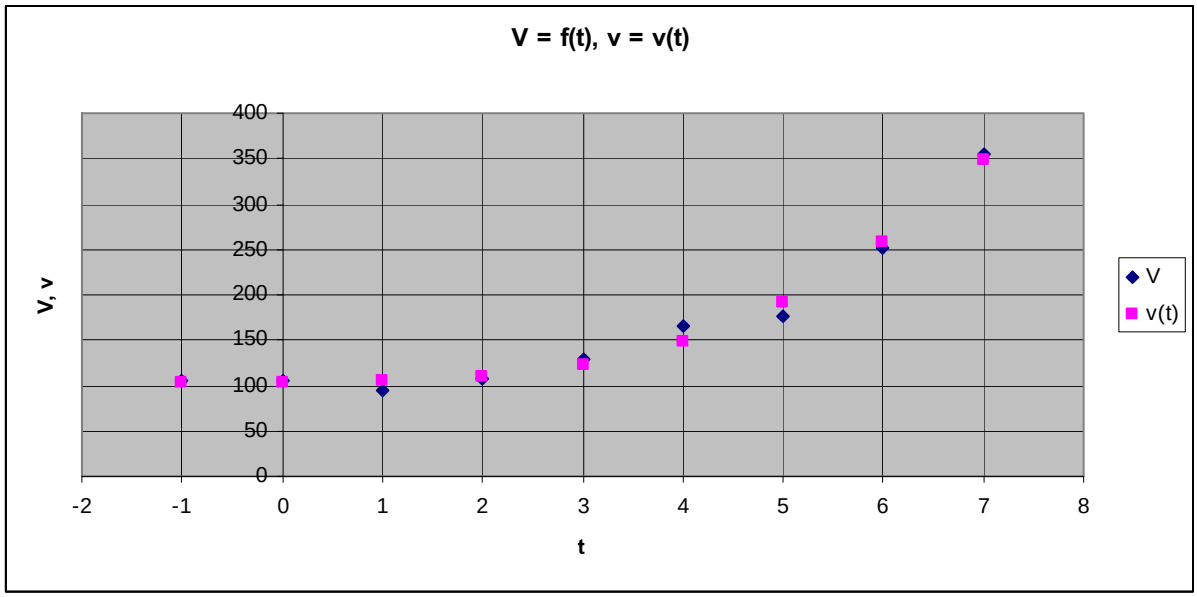

Рис. 3. Графики Внешторг – V и подбора $v(t)$.

Аналогично, используя таблицу 4 и программу «Регрессия»,

Таблица 4

t	t^2	t^3	I	i(t)
-1	1	-1	59,97	58,2813
0	0	0	55,41	55,3587
1	1	1	46,72	53,8397
2	4	8	58,67	56,6845
3	9	27	71,70	66,8536
4	16	64	94,10	87,3073
5	25	126	110,13	121,0059
6	36	216	171,82	170,9097
7	49	343	241,70	239,9793

находим коэффициенты

$C_0 = 55{,}3587$, $C_1 = -2{,}7142$, $C_2 = 0{,}7018$

$C_3 = 0{,}4934$, функции подбора $i(t)$

$$i(t) = C_0 + C_1 \cdot t + C_2 \cdot t^2 + C_3 \cdot t^3 \,, \; -1 \le t \le 7 \quad (2.8)$$

для показателя I – Импорт и коэффициент корреляции векторов I и $i(t)$ – $R = 0{,}9964$. Пятый столбец таблицы 4 содержит значения функции подбора $i(t)$, подсчитанные по формуле (2.8).

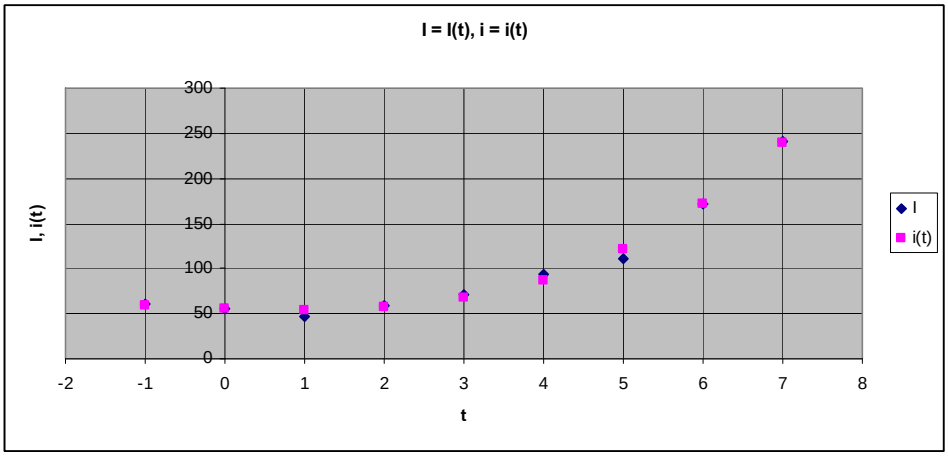

Рис. 4. Графики Импорт – I и подбора $i(t)$.

3. Заключение

Задача аппроксимации статистических данных полиномами k –й степени решена методом наименьших квадратов. Коэффициенты этих полиномов предлагается искать с помощью MathCAD и Excel.

Коэффициенты корреляции r_i между найденными полиномами и соответствующими статистическими характеристиками находятся в интервале $0{,}96 < r_i < 1$.

Литература

1. Анго А. Математика для электро- и радиоинженеров. М: Наука, 1967, 701 с.

2. Аттокурова Н.С. Проблемы развития внешнеполитических связей Кыргыской Республики в условиях международной интеграции: Диссертация на соискание ученой степени доктора экономических наук. Бишкек, 2009, 400 с.

3. Макаров Е.Г. Инженерные расчеты в MathCAD. Учебный курс. СПб: Питер, 2003, 448 с.

Рецензент: д.т.н., профессор Усманов СФ.