УДК 519.65, 004.94

КОМПЬЮТЕРНАЯ РЕАЛИЗАЦИЯ НЕКОТОРЫХ АЛГОРИТМОВ РЕГРЕССИИ И СГЛАЖИВАНИЯ

Дашкина А.Р., Бобылёв Ю.Ю.

Научный руководитель: Сабитова Г.С.

Аннотация. На практике часто приходится сталкиваться с различными задачами, связанными с обработкой экспериментальных данных. Использование встроенных функций обработки данных программы MathCAD позволяет реализовать сложные и громоздкие методы. Целью данной статьи является решение с помощью MathCAD таких нетривиальных, в общем, задач, как интерполирование сплайнами или сглаживание.

Ключевые слова: аппроксимация, экспериментальные данные, полиномиальная регрессия, обобщенная регрессия, сглаживание, MathCAD.

COMPUTER IMPLEMENTATION OF SOME REGRESSION AND SMOOTHING ALGORITHMS

A.R. Dashkina, Y. Y. Bobylev

Scientific adviser: G. S. Sabitova

Annotation. In practice, one often has to deal with various tasks related to the processing of experimental data. Using the built-in data processing functions of the MathCAD program allows you to implement complex and cumbersome methods. The purpose of this article is to solve using MathCAD such non-trivial, in general, problems, such as interpolation with splines or smoothing.

Keywords: approximation, experimental data, polynomial regression, generalized regression, smoothing, MathCAD.

В практических задачах часто возникает ситуация, когда имеется таблично заданная функция, однако описать данную функцию аналитическим выражением не представляется возможным. Например, некоторая последовательность значений может быть получена в результате измерения какого-то сложного технического или физического процесса, формульного представления для которого не существует. Для анализа результата эксперимента, необходимо провести через имеющиеся точки максимально соответствующую им кривую. В итоге, станет возможным предсказание значения функции в промежуточных точках, в которых измерения не проводились. Также в таком случае можно будет дифференцировать и интегрировать экспериментальную зависимость. Таким образом, с эмпирической функцией можно будет выполнять те же операции, что и с заданной аналитически. Задача в том, как провести через точки «правильную» кривую? Этот вопрос может быть решен с помощью алгоритмов интерполяции. В компьютерной системе MathCAD есть некоторый набор функций интерполяции, позволяющих соединить точки экспериментальных данных кривыми различной степени гладкости. Кроме того, существует возможность интерполяции сплайнами [1].

Задача регрессионного анализа – установление параметров описывающего экспериментальную зависимость выражения с учетом того, что эмпирические точки получены с некоторой погрешностью. Как правило, это делается путем вычисления минимума тем или иным способом задаваемой функции ошибки. Практическая важность регрессии очень велика: так, например, любой корректно построенный график экспериментальной зависимости должен быть задан с ее помощью. Вычисление регрессии может быть очень трудоемким процессом в случае его проведения «вручную», ввиду большого объема вычислительной работы. В системе MathCAD, благодаря наличию многочисленных встроенных функций, задачи подобного рода можно решать достаточно быстро.

MathCAD включает ряд функций для вычисления регрессии. Функции отличаются прежде всего типом кривой, которую они используют, чтобы аппроксимировать данные. Классическим алгоритмом линейной регрессии является метод наименьших квадратов, идея которого сводится к поиску таких коэффициентов для уравнения прямой, чтобы сумма квадратов абсолютных ошибок $(b + a\,x_i - y_i)$ была минимальна [2].

Существуют и другие алгоритмы регрессии [3]. Для того, чтобы разобраться с применением регрессионных алгоритмов математической обработки экспериментальных данных, последовательно выполним ряд заданий в программе MathCAD версии 14.0.0.163.

Задание 1. Создайте таблицу экспериментальных данных:

$x_i = a + h \cdot i$, где $h = (b - a)/10$, $i = 0{,}1, \dots ,10$ на отрезке $[a, b]$.

Решение. Этапы выполнения задания в программе MathCAD приведены на рис. 1.

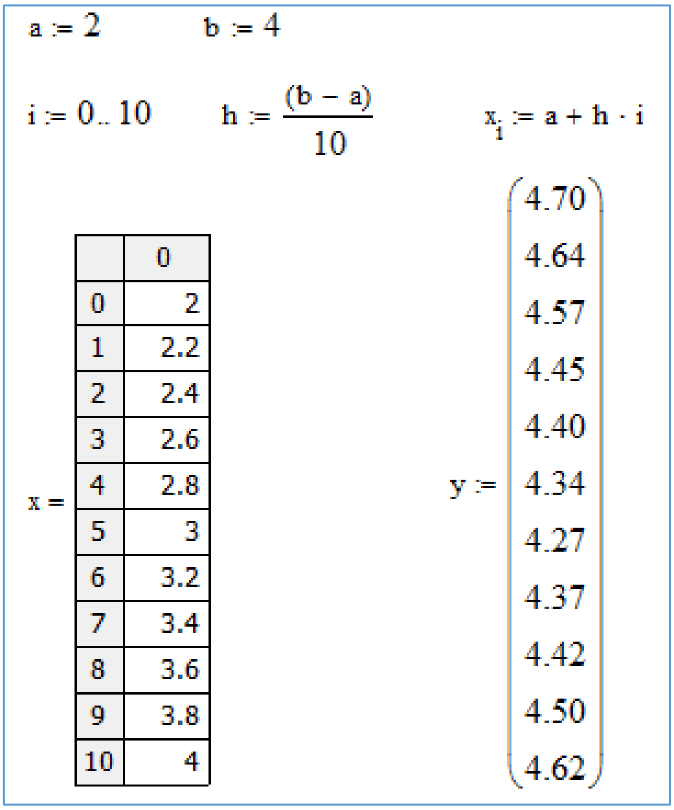

Рис. 1. Создание таблицы экспериментальных данных

Задание 2. Аппроксимировать данные из векторов x_i и y_i:

- полиномом 4-ой степени при помощи функций *regress* и *interp*;

• наборами полиномов второго порядка с помощью функций *loess* и *interp* (при *span* равном 0,5 и 2,5).

Отобразить графически результаты аппроксимации.

Решение. Полиномиальная регрессия в MathCAD реализована одним полиномом, отрезками нескольких полиномов, а также двумерной регрессией массива данных.

Полиномиальная регрессия означает приближение данных (x_i, y_i) полиномом m-й степени $\varphi(x) = a_m x^m + a_{m-1}x^{m-1} + \cdots + a_1 x + a_0$. При $m = 1$ полином является прямой линией, при $m = 2$ – параболой, при $m = 3$ – кубической параболой и т. д. На практике применяются $m < 5$.

В MathCAD полиномиальная регрессия выполняется комбинацией встроенной функции *regress* и полиномиальной интерполяции.

Следует использовать функцию *regress*, когда нужно получить единственный полином произвольной степени, чтобы приблизить все данные. Не рекомендуется делать степень аппроксимирующего полинома выше 4 – 6, поскольку погрешности реализации регрессии сильно возрастают.

Помимо приближения массива данных одним полиномом имеется возможность выполнить регрессию сшивкой фрагментов нескольких полиномов.

Встроенная функция *loess* облегчает эти проблемы, выполняя локальное приближение. Вместо одного полинома *loess* создает различные полиномы второго порядка в зависимости от расположения на кривой. Применение функции *loess* аналогично функции *regress*.

Параметр *span* определяет размер фрагментов полиномов. Параметр *span* задает степень сглаженности данных. При больших значениях *span* регрессия почти не отличается от регрессии одним полиномом.

Решение задачи аппроксимации полиномом и набором полиномов представлено на рис. 2-4

Способ 1. Провести аппроксимацию полиномом четвертой степени при помощи функций *regress* и *interp*.

$n := 4$ - степень полинома

$vs := regress(x, y, n)$

$re(t) := interp(vs, x, y, t)$ - результат полиномиальной регрессии

Отобразим графески результаты полиномиальной аппроксимации и совокупность точек векторов x_i и y_i.

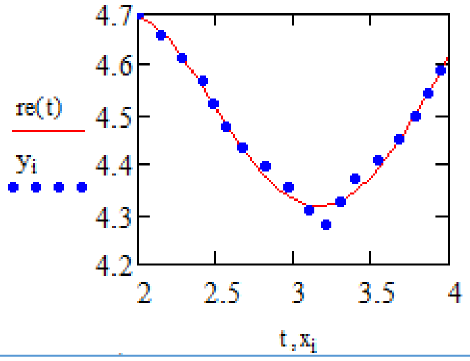

Рис. 2. Аппроксимация полиномом 4-ой степени и построение результирующего графика

Способ 2. Провести аппроксимацию наборами полиномов второго порядка с помощью функций *loess* и *interp* при *span* равном 2.5 и 0.5.

$span := 2.5$

$vs1 := loess(x, y, span)$

$lo(z) := interp(vs1, x, y, z)$ - результат полиномиальной регрессии

Отобразим графески результаты аппроксимации набором полиномов второго порядка при *span* = 2.5 и совокупность точек векторов x_i и y_i.

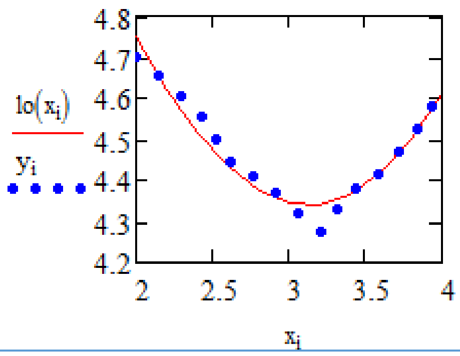

Рис. 3. Аппроксимация набором полиномов второго порядка при span = 2.5 и построение результирующего графика

$span := 0.5$

Отобразим графески результаты аппроксимации набором полиномов второго порядка при *span*=0.5 и совокупность точек векторов x_i и y_i.

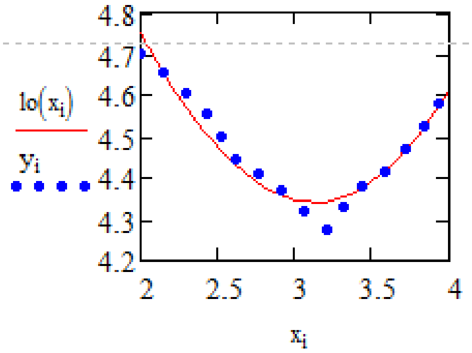

Рис. 4. Аппроксимация набором полиномов второго порядка при span = 0.5 и построение результирующего графика

Задание 3. Аппроксимировать экспериментальные данные из таблиц значений x_i и y_i линейной комбинацией функций:

$$f(x) = a_1 \cdot f_1(x) + a_2 \cdot f_2(x) + a_3 \cdot f_3(x).$$

Коэффициенты вектора a найти с помощью функции *linfit*. Отобразить графически совокупность точек векторов x_i и y_i и результаты проведенной *линейной регрессии общего вида.*

Решение. Аппроксимация методом линейной регрессии общего вида приведена на рис. 5-6.

$f1(x) := \dfrac{1}{1+x^2}$ $\quad f2(x) := \exp(\sin(x))$ $\quad f3(x) := x$ - заданные функции

$F(x) := \begin{pmatrix} f1(x) \\ f2(x) \\ f3(x) \end{pmatrix}$ - функция-вектор

$aa := \mathrm{linfit}(x, y, F)$ - вектор коэффициентов линейной регрессии общего вида

$z := 2, 2.01 .. 4$

$L(z) := F(z) \cdot aa$

Рис. 5. Аппроксимация обобщенной линейной регрессией

Отобразим графически совокупность точек векторов x_i и y_i и результаты проведенной *линейной регрессии общего вида*.

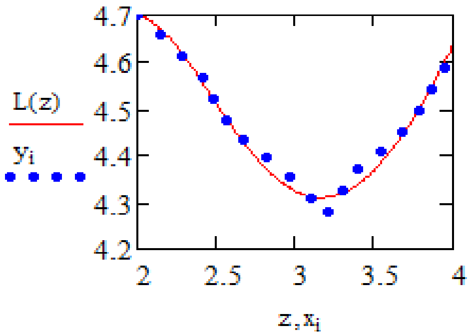

Рис. 6. Построение результирующего графика

Задание 4. Аппроксимировать экспериментальные данные из таблиц значений x_i и y_i функцией вида:

$$f(x) = e^{u_0 + u_1 \cdot x + u_2 \cdot x^2}$$

Решение. Если данные должны быть смоделированы в виде функции

$f(x) = f(x, u_0, u_1, \ldots, u_n)$,

то нужно использовать функцию *genfit*, чтобы найти неизвестные параметры u_i. Это *нелинейная регрессия общего вида*.

Решение задание нелинейной регрессией общего вида представлена на рис. 7-8.

$$D(x,u) := \begin{pmatrix} \exp\left(u_0 + u_1 \cdot x + u_2 \cdot x^2\right) \\ \exp\left(u_0 + u_1 \cdot x + u_2 \cdot x^2\right) \\ x \cdot \exp\left(u_0 + u_1 \cdot x + u_2 \cdot x^2\right) \\ x^2 \cdot \exp\left(u_0 + u_1 \cdot x + u_2 \cdot x^2\right) \end{pmatrix}$$

- первый элемент содержит приближающую функцию, последующие содержат ее частные производные по исходным параметрам

$$vg := \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix}$$

- вектор начальных значений для переметров

$P := \mathrm{genfit}(x, y, vg, D)$

$g(z) := D(z, P)_0$

Рис. 7. Проведение нелинейной регрессии общего вида

Отобразим графически совокупность точек векторов x_i и y_i и результаты проведенной *нелинейной регрессии общего вида.*

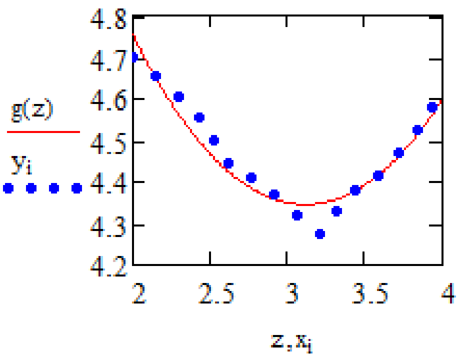

Рис. 8. Построение результирующего графика

Задание 5. Выполнить сглаживание экспериментальной функции, заданной таблицей значений x_i и y_i с помощью встроенных функций MathCAD: *medsmooth*, *ksmooth* и *supsmooth*. Результаты сглаживания отобразить графически.

Решение. Как правило, регрессия очень эффективна, когда заранее известен или угадывается закон распределения данных (x_i, y_i). Регрессия является частным случаем общей проблемы сглаживания данных. Помимо регрессии в MathCAD имеется несколько встроенных функций, реализующих различные алгоритмы сглаживания данных [4].

Сглаживание предполагает использование набора значений y (и, возможно, x) и возвращение нового набора значений y, который является более гладким, чем исходный набор. В отличие от регрессии и интерполяции, сглаживание приводит к новому набору значений y, а не к функции, которая может оценивать значения между заданными точками данных. Функция *medsmooth* выполняет сглаживание алгоритмом «бегущих медиан» (предполагает, что данные расположены равномерно), *ksmooth* на основе функции Гаусса, *supsmooth* – локальное сглаживание адаптивным алгоритмом, основанное на анализе ближайших соседей каждой пары данных.

Решение задач сглаживания представлено на рис. 9-10.

$t := 1$

$ksm := ksmooth(x, y, t)$ - n-мерный массив, созданный сглаживанием при помощи гауссовского ядра данных и n-мерного вектора y

$m := 5$

$meds := medsmooth(y, m)$ - n-мерный массив, созданный сглаживанием n-мерного вектора y с помощью скользящей медианы

$sups := supsmooth(x, y)$ - n-мерный массив, созданный локальным использованием симметричной линейной функции процедуры сглаживания МНК

Рис. 9. Создание массивов для решения задач сглаживания

Отобразим графически совокупность точек векторов x_i и y_i и результаты сглаживаний.

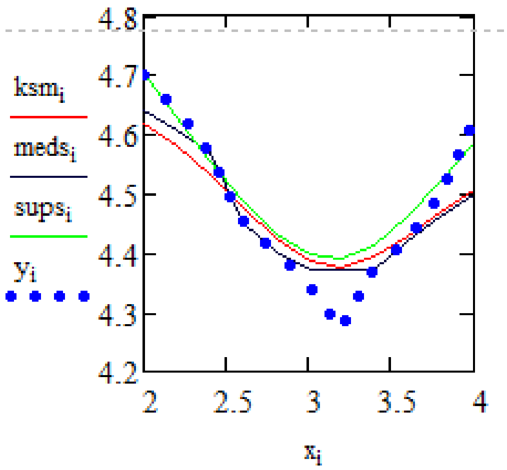

Рис. 10. Построение результирующего графика

Заключение. В статье рассмотрены задачи математической обработки экспериментальных данных, разрешаемые при помощи нескольких альтернативных методов: решение задач с помощью полиномиальной регрессии, обобщенной регрессии и сглаживания, что дает большую точность полученных результатов в области обработки эмпирических данных. Приведена реализация задач с использованием встроенных функций математического пакета MathCAD.

Список литературы

1. Дашкина, А. Р., Бобылев, Ю. Ю. Интерполяция и предсказание функций в MathCAD / А. Р. Дашкина, Ю. Ю. Бобылев // Kazakhstan Science Journal. –2019. – № 1 (14). – Т. 3. – С. 13-24.

2. Дашкина, А. Р., Бобылев, Ю. Ю. Решение задачи линейной регрессии в системе MathCAD / А. Р. Дашкина, Ю. Ю. Бобылев // Kazakhstan Science Journal. –2019. – № 4 (17). – Т. 3. – С. 17-26.

3. Сабитова, Г. С., Калиев, И. А. Практикум по информационным технологиям. Часть 1: учеб. пособие / Г. С. Сабитова, И. А. Калиев. – Стерлитамак: РИО Стерлитамакского филиала БашГУ. – 2015. – 127с.

4. Сабитова, Г. С. Численная обработка результатов экспериментальных данных / Г. С. Сабитова // Известия вузов Кыргызстана – Бишкек: ООО науч. журн. и дет. худ. лит., 2019. – № 8. – С. 17-23.

Сведения об авторах

Дашкина Алина Рустемовна – студент факультета математики и информационных технологий, Стерлитамакский филиал Башкирского государственного университета; Россия, Стерлитамак.

Бобылев Юрий Юрьевич – студент факультета математики и информационных технологий, Стерлитамакский филиал Башкирского государственного университета; Россия, Стерлитамак.

Сабитова Гульнара Сагындыковна – кандидат физико-математических наук, доцент кафедры прикладной информатики и программирования, Стерлитамакский филиал Башкирского государственного университета; Россия, г. Стерлитамак.

About authors

Dashkina Alina Rustemonva – Student of Faculty of Mathematics and Information Technology, Sterlitamak branch of Bashkir State University; Russia, Sterlitamak.

Bobylev Yuri Yurievich – Student of Faculty of Mathematics and Information Technology, Sterlitamak branch of Bashkir State University; Russia, Sterlitamak.

Sabitova Gulnara Sagyndykovna – Candidate of Physical and Mathematical Sciences, Associate Professor of the Department of Applied Informatics and Programming, Sterlitamak branch of the Bashkir State University; Russia, Sterlitamak.