Программа построения нелинейной регрессии для математического моделирования распределения ДНК по экспериментальным данным

Галкина С.М.

математика, информатика, биология

студентка 2 курса МОМК № 1, Ногинский филиал

Научный руководитель: Туманов В.Е., кандидат химических наук, преподаватель физики и химии, МОМК № 1, Ногинский филиал

Цель настоящего исследования – разработать прототип программы построения парной нелинейной регрессии для анализа распределений ДНК, полученных из экспериментальных данных, с учетом возможности фиксировать значения некоторых параметров распределений при добавлении дополнительного биологического материала.

Разработан прототип такой программы на языке программирования Pascal ABC.

Предложено четыре вида функции для построения регрессионной модели.

Приведен результат тестовых испытаний программы, которые показал хорошее соответствие аппроксимирующей кривой с экспериментальными данными с коэффициентом детерминации, равным 0.9999.

Приведен результат работы программы на экспериментальных данных с использование двух видов функций. Показано, что работа программы дает удовлетворительные результаты.

Эта работа способствует развитию междисциплинарных связей предметов, изучаемых на 1-2 курсах медицинских колледжей.

Ключевые слова: анализ распределений ДНК, нелинейная парная регрессия, компьютерная программа, Pascal ABC

Введение

Линейные и нелинейные регрессии являются инструментом в проведении исследований в эконометрике, в естественно-научных и гуманитарных дисциплинах. Построению линейных и нелинейных регрессионных моделей

посвящено много монографий и учебных изданий, в частности монография [1]. Работа [2] была одной из первых работ, посвященных математическому анализу ДНК распределений, полученных в эксперименте.

Существует коммерческое программное обеспечение, которое реализует математическую обработку таких экспериментальных данных [3, 4]. Несмотря на то, что в таких программах предоставлен широкий выбор моделей, при внесении дополнительного биологического материала некоторые параметры распределений пересчитываются, что не совсем отвечает течению биохимических процессов. Решение этой проблемы является актуальной научно-практической задачей.

Цель настоящего исследования – разработать прототип программы построения парной нелинейной регрессии для анализа распределений ДНК, полученных из экспериментальных данных, с учетом возможности фиксировать значения некоторых параметров распределений при добавлении дополнительного биологического материала.

Задачами исследования являются:

- Изучить алгоритм Левенберга-Марквардта построения парной нелинейной регрессии;
- Разработать прототип программы построения парной нелинейной регрессии для набора функций, которые используются или могут быть использованы для анализа распределений ДНК.

Гипотеза исследования. Анализ учебной и научной литературы показывает, что поставленная цель может быть достигнута путем разработки программы и ее тестирования.

Объектом исследования является аппроксимация функции распределения ДНК методом построения парной регрессии.

Предметом исследования является реализация аппроксимации функции распределения ДНК путем построения нелинейной парной регрессии на языке программирования Pascal ABC.

Основная часть

Постановка задачи.

Математически задача построения парной нелинейной регрессии сводится к следующему [1]. Пусть задана табличная зависимость переменной y от переменной x, $\{y_i, x_i\}$, где $i = 1, 2, \ldots, M\}$. Необходимо аппроксимировать данную зависимость функцией $f(x, c_1, \ldots, c_n)$, вид которой известен, но ее параметры неизвестны:

$$y_i = f(x_i, c_1, \ldots c_n) + \varepsilon_i, \text{ где i} = 1, \ldots, M.$$

По методу наименьших квадратов минимизируется сумма остатков:

$$S = \sum_{1}^{M} \varepsilon_i^2$$

Для решения вычислительной задачи в этой работе использован алгоритм Левенберга-Марквардта [5], который сводится на каждом шаге итерации к решению системы уравнений

$$(J^T J + \mu^2 I)\Delta c = -J^T \varepsilon$$

решение которой определяет следующий шаг итерации

$$c^{(k+1)} = c^{(k)} - \left(J^T\left(c^{(k)}\right)J\left(c^{(k)}\right) + \mu^2 I\right)^{-1} J^T(c^{(k)})\varepsilon(c^{(k)})$$

J есть матрица Якоби (матрица первых частных производных функции S по параметрам c_i) [5], а J^T - транспонированная матрица Якоби, I – единичная матрица, $\mu^{(k)}$ - параметр регуляризации, который регулирует не только длину шага, но и направление поиска.

Условие завершение алгоритма

$$\left\|J^T(c^{(k)})\varepsilon(c^{(k)})\right\| < \delta$$

где ||*|| - евклидова норма.

Блок схема алгоритма, реализованный в программе, программы приведена на Рис. 1.

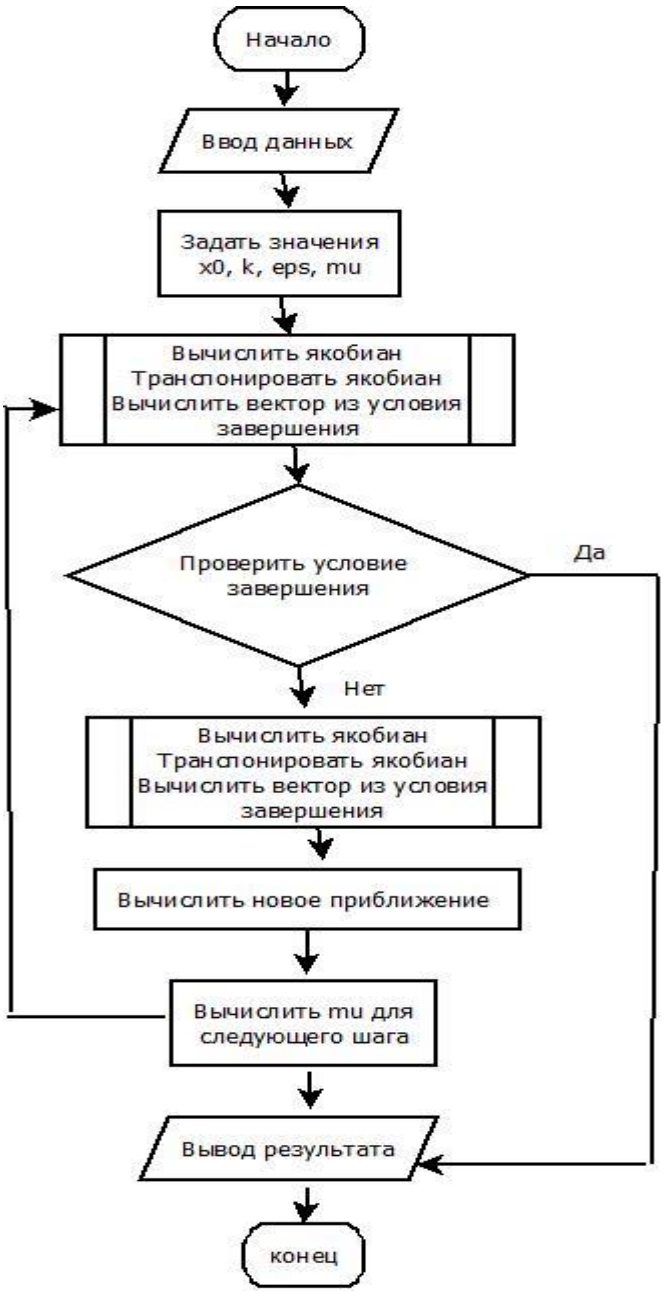

Рис. 1. Блок схема алгоритма Левенберга-Марквардта.

<u>Выбор функций аппроксимации.</u>

Вид функций аппроксимации обсуждался в работах [2, 6]. Нами были рассмотрены и другие функции. В программе могут быть использованы функции вида:

$$f(x) = \frac{c_1}{\sqrt{2\pi c_2}} e^{\frac{-(x-c_3)}{2c_2}} + \frac{c_4}{\sqrt{2\pi c_5}} e^{\frac{-(x-c_6)}{2c_5}} + c_7 + c_8 x + c_9 x^2 \quad (1)\ [2]$$

$$f(x) = \frac{c_1}{\sqrt{2\pi c_2}} e^{\frac{-(x-c_3)}{2c_2}} + \frac{c_4}{\sqrt{2\pi c_5}} e^{\frac{-(x-c_6)}{2c_5}} \quad (2)[6]$$

$$f(x) = \frac{c_1}{\sqrt{2\pi}c_2} e^{\frac{-(x-c_3)}{2c_2}} + \frac{c_4}{\sqrt{2\pi}c_5} e^{\frac{-(x-c_6)}{2c_5}} + \frac{c_{11}}{(1+e^{-c_7(x-c_9)})(1+e^{c_9(x-c_{10})})} \quad (3)$$

$$f(x) = \frac{c_1}{\sqrt{2\pi}c_2} e^{\frac{-(x-c_3)}{2c_2}} + \frac{c_4}{\sqrt{2\pi}c_5} e^{\frac{-(x-c_6)}{2c_5}} + \frac{c_7}{1+(\left|\frac{x-c_8}{c_9}\right|)^{2c_{10}}} \quad (4)$$

Результаты и обсуждение

Тестирование программы было выполнено по данным из работы [2] с функцией вида (2). Были получены следующие оценки коэффициентов: c_1 = 67401.88±1.08, c_2 = 1.41±2.57x10^{-05}, c_3 = 30.64±2.91x10^{-05}, c_4 = 9907.36±1.65, c_5 = 2.60±0.01, c_6 = 62.07±0.01 (c_1 = 67400.00±0.14, c_2 = 1.41±0.02, c_3 = 30.64±0.03, c_4 = 9910.00±1.50, c_5 = 2.60±0.26, c_6 = 62.07±0.27 по данным [2]) . На Рис. 2 приведен график полученной аппроксимации. Коэффициент детерминации [7], равный 0.9999, показывает хорошее соответствие аппроксимирующей кривой с экспериментальными данными.

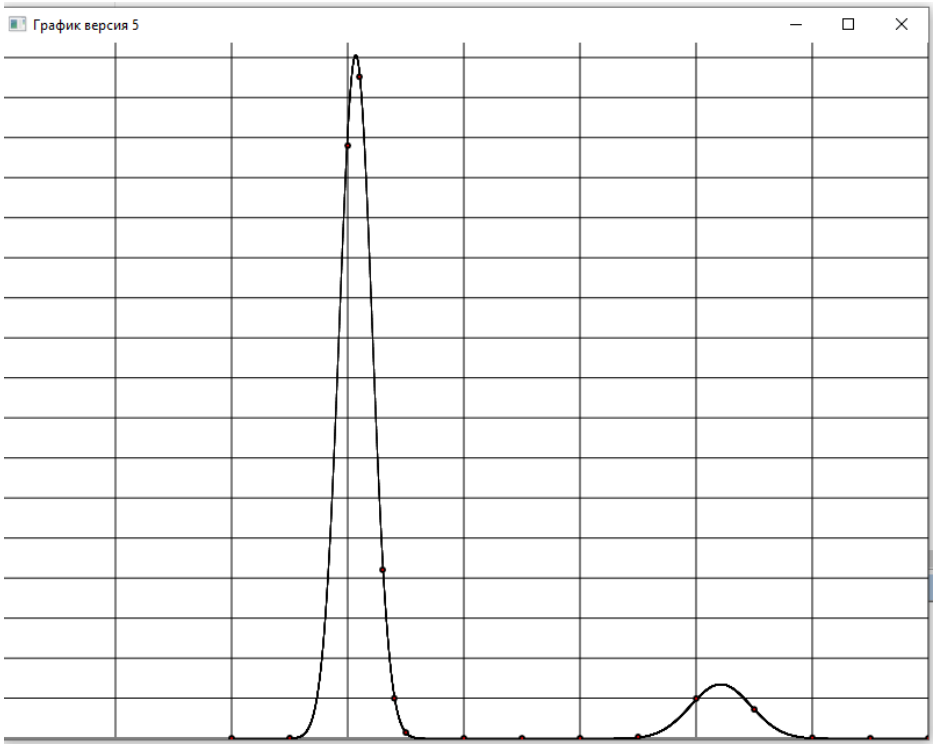

Рисунок 2. График полученной аппроксимации. Шаг сетки по горизонтальной оси 10, а по вертикальной – 5000.

Программа была опробована с функцией вида (2) на экспериментальном наборе из 1024 точек. Были получены следующие оценки коэффициентов: c_1 = 1600.36±16.15, c_2 = 17.07±0.29, c_3 = 285.01±0.23, c_4 = 1800.03±47.22, c_5 = 60.13±3.34, c_6 = 520.80±2.90. На Рис. 3 приведен график полученной аппроксимации. Коэффициент детерминации, равный 0.9376, показывает хорошее соответствие аппроксимирующей кривой с экспериментальными данными.

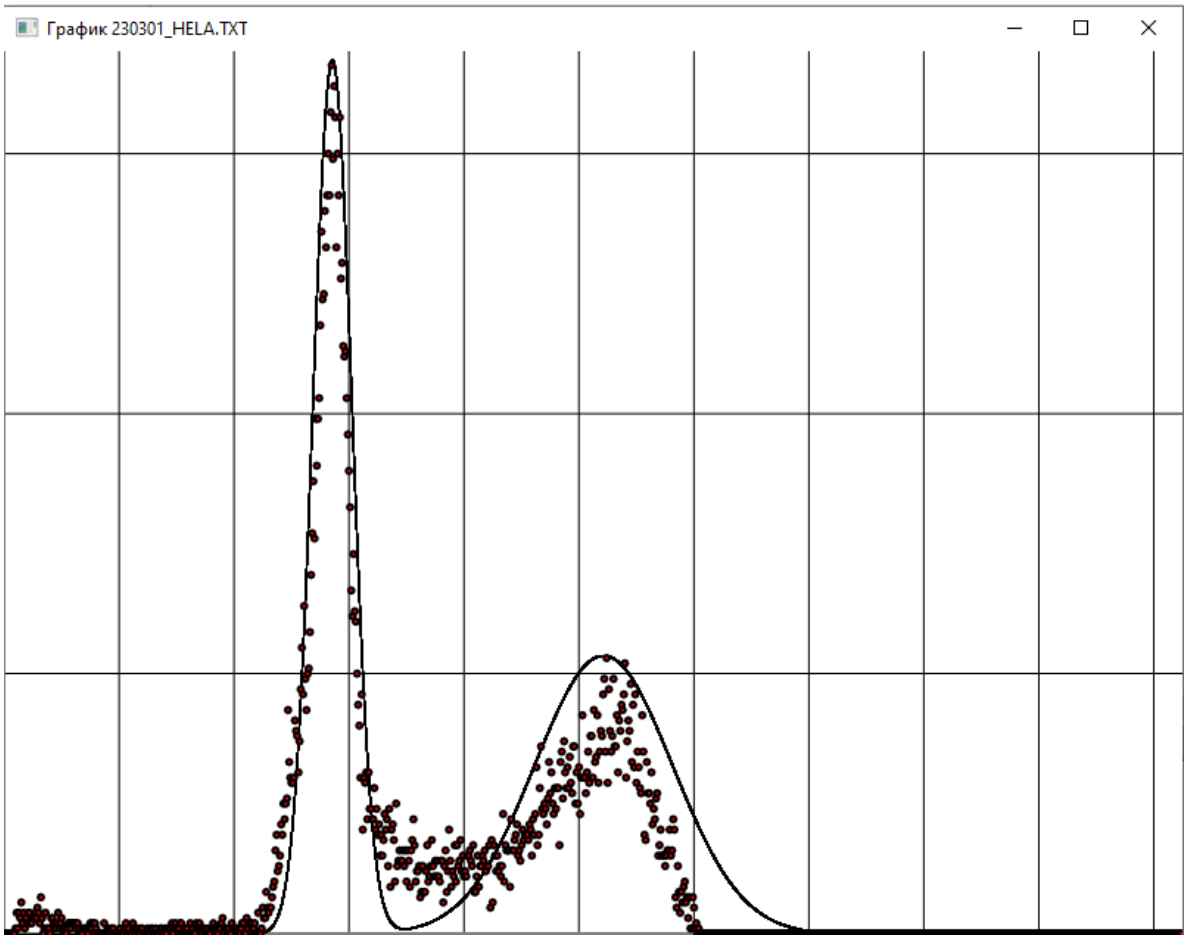

Рис. 3. График полученной аппроксимации. Шаг сетки по горизонтальной оси 100, а по вертикальной – 50.

Программа была выполнена с функцией вида (3) на том же экспериментальном наборе из 1024 точек. Параметры c_1-c_6 такие же, как и для функции вида (2) выше, c_7 = 1.01±0.03, c_8 = 300.0±8.21, c_9 = 2.03±0.13, c_{10} = 400.11±5.03, c_{11} = 10.0±1.63. На Рис. 4 приведен график полученной аппроксимации. Коэффициент детерминации, равный 0.9370, показывает хорошее соответствие аппроксимирующей кривой с экспериментальными данными.

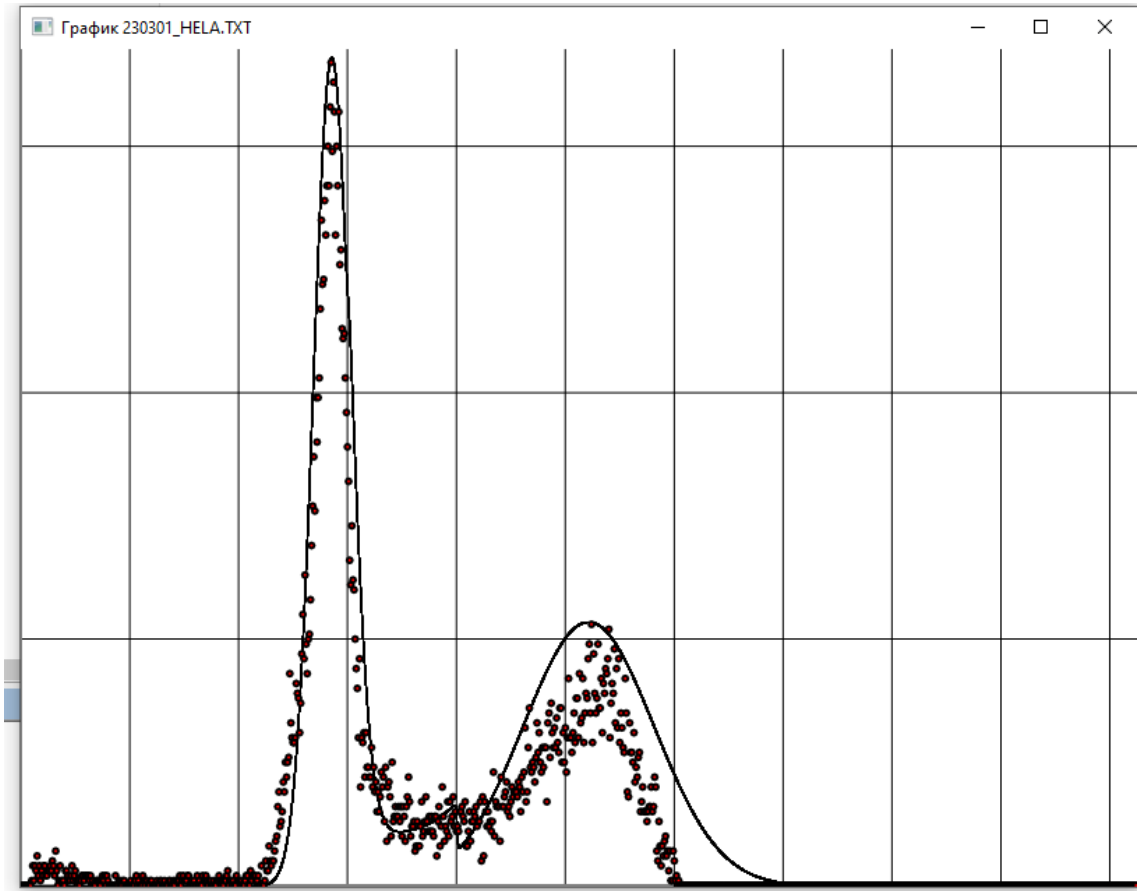

Рис. 4. График полученной аппроксимации. Шаг сетки по горизонтальной оси 100, а по вертикальной – 50.

Выводы

В статье представлен алгоритм прототипа программы построения парной регрессии. Программа предназначена для математического анализа распределений ДНК, полученных с помощью проточной микрофлуорометрии. Предложено четыре вида функции для построения регрессионной модели. Приведен результат тестовых испытаний программы, которые показал хорошее соответствие аппроксимирующей кривой с экспериментальными данными с коэффициентом детерминации, равным 0.9999.

Приведен результат работы программы на экспериментальных данных с использование двух видов функций. Показано, что работа программы дакт удовлетворительные результаты.

Эта работа способствует развитию междисциплинарных связем предметов, изучаемых на 1-2 курсах медицинских колледжей..

Литература

1. Демиденко Е. З. Линейная и нелинейная регрессии. / Е. З. Демиденко. — М.: «Финансы и Статистика». 1981. — 302 с.
2. Dean P.N., Jett J. H. Mathematical analysis of DNA distributions derived from flow microfluorometry // J. Cell Biol., 1974. V. 60. N. 2. P. 523-527. doi: 10.1083/jcb.60.2.523.
3. Watson J.V., Chambers S.H., Smith P.J. A pragmatic approach to the analysis of DNA histograms with a definable G1 peak // Cytometry, 1987. V. 8, N. 1. P. 1-8. doi: 10.1002/cyto.990080101.
4. FCS EXPRESS 7 // [Электронный ресурс]. URL: https:// https://denovosoftware.com/ (дата обращения: чч.мм.гггг).
5. Дэннис Дж., мл. Численные методы безусловной оптимизации и решения нелинейных уравнений ; пер. с англ. / Дэннис Дж., мл., Р. Шнабель. – М. : Мир, 1988. 440 с.
6. Fox M.H. A model for the computer analysis of synchronous DNA distributions obtained by flow cytometry // Cytometry, 1980. V. 1(1). P. 71-77. doi: 10.1002/cyto.990010114.
7. Гмурман В. Е. Теория вероятностей и математическая статистика : Учеб. пособие для вузов / В.Е. Гмурман. - 9-е изд., стер. - Москва : Высшая школа, 2003. - 478 с.