УДК 004.42

С.Д. ТОЧИЛІН, В.О. РИБІН
Запорізький національний технічний університет

КРОСПЛАТФОРМНА КОМП'ЮТЕРНА ПРОГРАМА ДЛЯ ПОЛІНОМІАЛЬНОГО РЕГРЕСІЙНОГО АНАЛІЗУ ДАНИХ

Поліноміальний регресійний аналіз широко використовується для побудови математичних залежностей по експериментальним статистичним даним. Поліноміальний регресійний аналіз даних виконується, як правило, за допомогою комерційних комп'ютерних програм, які мають графічний інтерфейс користувача. Однак ці програми в багатьох випадках не є кроссплатформними.

У наш час однією з найбільш популярних мов програмування є Java. Вона кросплатформна і має API, що вільно розповсюджується, який дозволяє створювати програми для статистичної обробки експериментальних даних із графічним інтерфейсом користувача.

У даній роботі за допомогою мови програмування Java була розроблена кроссплатформна комп'ютерна програма для поліноміального регресійного аналізу даних. Вона має графічний інтерфейс користувача і використовує при функціонуванні метод найменших квадратів. При цьому для визначення параметрів регресійної моделі із системи лінійних рівнянь, які формуються при обробці статистичних даних, застосовується метод Гаусса. Програма представляє результати аналізу даних у графічному та табличному вигляді. У графічному вигляді зображуються вихідні дані та відповідна їм регресійна залежність. У табличному вигляді відображаються параметри моделі регресії, а також дані, що аналізуються, і відповідні їм значення апроксимуючого полінома та залишків регресії.

Крім того, програма визначає параметри необхідні для оцінки якості та значимості поліноміальної моделі регресії, яка була отримана при обробці даних. Розроблений додаток для оцінки якості моделі розраховує середню помилку апроксимації та коефіцієнт детермінації або індекс детермінації, а для оцінки її значимості обчислює фактичне і теоретичне значення F-критерію Фішера. При розрахунку теоретичного значення F-критерію Фішера програма використовує функцію бета-розподілу.

Також розроблений додаток при аналізі даних дозволяє прогнозувати значення залежної змінної для значень регресора, які визначає користувач.

Ключові слова: регресія, аналіз даних, поліном, метод найменших квадратів.

С.Д. ТОЧИЛИН, В.О. РЫБИН
Запорожский национальный технический университет

КРОССПЛАТФОРМЕННАЯ КОМПЬЮТЕРНАЯ ПРОГРАММА ДЛЯ ПОЛИНОМИАЛЬНОГО РЕГРЕССИОННОГО АНАЛИЗА ДАННЫХ

Полиномиальный регрессионный анализ широко используется для построения математических зависимостей по экспериментальным статистическим данным. Полиномиальный регрессионный анализ данных выполняется, как правило, с помощью коммерческих компьютерных программ, которые имеют графический интерфейс пользователя. Однако эти программы во многих случаях не являются кроссплатформенными.

В настоящее время одним из наиболее популярных языков программирования является Java. Он кросплатформенный и имеет свободно распространяемый API, который позволяет создавать программы для статистической обработки экспериментальных данных с графическим интерфейсом пользователя.

В данной работе с помощью языка программирования Java была разработана кроссплатформенная компьютерная программа для полиномиального регрессионного анализа данных. Она имеет графический интерфейс пользователя и использует при функционировании метод наименьших квадратов. При этом для определения параметров регрессионной модели из системы линейных уравнений, которые формируются при обработке статистических данных, применяется метод Гаусса. Программа представляет результаты анализа данных в графическом и табличном виде. В графическом виде изображаются исходные данные и соответствующая им регрессионная зависимость. В табличном виде отображаются параметры модели регрессии, а также анализируемые данные и соответствующие им значения аппроксимирующего полинома и остатков регрессии.

Кроме того, программа определяет параметры необходимые для оценки качества и значимости полиномиальной модели регрессии, которая была получена при обработке данных. Разработанное приложение для оценки качества модели рассчитывает среднюю ошибку аппроксимации и коэффициент детерминации или индекс детерминации, а для оценки ее значимости

вычисляет фактическое и теоретическое значение F-критерия Фишера. При расчете теоретического значения F-критерия Фишера программа использует функцию бета-распределения.

Также разработанное приложение при анализе данных позволяет прогнозировать значения зависимой переменной для значений регрессора, которые определяет пользователь.

Ключевые слова: регрессия, анализ данных, полином, метод наименьших квадратов.

S.D. TOCHILIN, V.O. RYBIN
Zaporozhye National Technical University

CROSS-PLATFORM COMPUTER PROGRAM FOR POLYNOMIAL REGRESSION DATA ANALYSIS

Polynomial regression analysis is widely used to construct mathematical dependencies from experimental statistics. Polynomial regression data analysis is performed, as a rule, using commercial computer programs that have a graphical user interface. However, these programs in many cases are not cross-platform.

Currently, one of the most popular programming languages is Java. It is cross-platform and has a free API, which allows to create programs for statistical processing of experimental data with a graphical user interface.

In this paper, the cross-platform computer program for polynomial regression data analysis was developed using the Java programming language. It has a graphical user interface and uses the least squares method for its operation. In this case, the Gauss method is used to determine the parameters of the regression model from a system of linear equations that are formed during the processing of statistical data. The program presents the results of data analysis in graphical and tabular form. In a graphic form the initial data and the corresponding regression dependence are represented. The tabular form displays the parameters of the regression model, as well as the analyzed data and the corresponding values of the approximating polynomial and regression residuals.

In addition, the program determines the parameters necessary for assessing the quality and significance of the polynomial regression model that was obtained during data processing. The developed application for assessing the quality of the model calculates the average approximation error and the coefficient of determination or index of determination, and to estimate its significance it calculates the actual and theoretical value of the Fisher F-test. When calculating the theoretical value of the Fisher F-test, the program uses the beta distribution function.

Also, the developed application, when analyzing data, allows to predict the values of the dependent variable for the regressor values that the user determines.

Keywords: regression, data analysis, polynomial, least squares method.

Постановка проблеми

Поліноміальний регресійний аналіз є найпоширенішим статистичним методом, який використовується для побудови математичних залежностей по експериментальним даним [1-3].

Поліноміальний регресійний аналіз виконується, як правило, за допомогою спеціальних комерційних комп'ютерних програм, які мають графічний інтерфейс користувача (Graphical User Interface - GUI). Однак ці програми в багатьох випадках жорстко прив'язані до певної платформи.

У той час при проведенні регресійного аналізу для вибору оптимальної моделі регресії оцінюють її якість та значимість [4, 5].

При цьому розробка кросплатформних комп'ютерних програм із GUI для поліноміального регресійного аналізу даних, які визначають параметри моделі регресії, а також оцінюють її якість і значимість, є актуальною проблемою.

Аналіз останніх досліджень і публікацій

Якщо для величини Y, що випадковим чином залежить від величини X, отримані m експериментальних точок (m пар спостережень): (X_1,Y_1); (X_2,Y_2); … (X_m,Y_m), то для знаходження функціональної залежності Y від X при поліноміальному регресійному аналізі використовують рівняння багаточлена (поліному) Y_r, яке має вигляд:

$$Yr = f(X) = C_0 + C_1 \cdot X + C_2 \cdot X^2 + \ldots + C_n \cdot X_n^n, \quad (1)$$

де C_0, C_1, C_2, …, C_n – постійні коефіцієнти,

n – ступінь поліному.

Для визначення C_0, C_1, C_2, …, C_n в багатьох випадках застосовують метод найменших квадратів (МНК). Опис МНК з прикладами комп’ютерних програм, які його реалізують, наведено в [6-9].

Відзначимо також, що для оцінки якості моделі регресії використовують індекс детермінації (коефіцієнт детермінації при $n=1$) R^2 та середню помилку апроксимації ME [4, 5]. У відповідності з [4, 5] для поліноміальної регресії їх визначають за допомогою виразів:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(Y_i - Y_{ri})^2}{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}, \tag{2}$$

$$ME = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{(Y_i - Y_{ri})}{Y_i}\right| \cdot 100 \ (\%), \tag{3}$$

де Y_i - експериментальні значення Y при X_i,

Y_{ri} - значення поліному (1) при X_i,

$\bar{Y}$ - середнє для Y_1, Y_2, …, Y_m.

Чим ближче величина R^2 до одиниці тим більш якісніше модель регресії, у той час значення $ME \le 7$ % також свідчать про задовільний вибір моделі до вихідних даних [4].

Між тим, для оцінки значимості моделі регресії для даних із m пар чисел, зіставляють фактичне значення F-критерію Фішера F та теоретичне F_α при певному рівні значимості α. Якщо F більше F_α, модель регресії визнається значимою для цього рівня значимості, в протилежному випадку – не значимою. Фактичне F-співвідношення визначають у такий спосіб [4]:

$$F = \frac{R^2}{(1-R^2)} \cdot \frac{m-n-1}{n}. \tag{4}$$

Величина n в (4) характеризує число ступенів свободи ν_1 для факторної суми квадратів, а $(m-n-1)$ - число ступенів свободи ν_2 для залишкової суми квадратів [4].

Теоретичні значення F-критерію можна розраховувати за допомогою функції бета-розподілу [10-12]:

$$I_x(a,b) = \frac{1}{B(a,b)}\int_0^x t^{a-1}(1-t)^{b-1}dt, \tag{5}$$

де відрізок $0 \le x \le 1$ на якому визначається $I_x(a,b)$,

a, b - параметри,

$B(a,b)$ - бета-функція Ейлера:

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt. \tag{6}$$

Теоретичне значення F_α, для рівня значимості α, при ν_1 і ν_2, визначають при використанні функції бета-розподілу, яка у відповідності з [12] має вигляд:

$$I_x\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right) = 1 - \alpha, \tag{7}$$

де x має значення [12]:

$$x = \frac{\nu_1 F_\alpha}{\nu_2 + \nu_1 F_\alpha}. \tag{8}$$

Відзначимо також, що у наш час найбільш поширеною мовою програмування, у відповідності з індексом TIOBE [13], є Java™. Вона кросплатформна та має необхідні засоби для розробки прикладних програм із GUI. Також Java має розвинений API, що вільно розповсюджується, до якого входять як середовища розробки, так і різного роду пакети, які розширюють функціональність додатків.

Формулювання мети дослідження

У даній роботі була поставлена задача розробки за допомогою мови програмування Java кросплатформного додатка з GUI для поліноміального регресивного аналізу даних, який визначає параметри регресивної моделі, оцінює її якість та значимість, а також представляє результати роботи в графічному та табличному вигляді.

Викладення основного матеріалу дослідження

Для розв'язку поставленої задачі був створений Java-додаток ST_Regression із графічним інтерфейсом користувача.

Для знаходження параметрів моделі регресії програма використовувала МНК. При цьому для визначення C_0, C_1, C_2, …, C_n із системи лінійних рівнянь, що формувались при обробці даних, використовувався метод Гаусса, який реалізовувався на основі обчислювальної схеми єдиного поділу [14]. Для оцінки якості моделі регресії додаток ST_Regression розраховував середню помилку апроксимації та індекс детермінації (коефіцієнт детермінації при n=1), а для оцінки значимості моделі обчислював значення, як фактичного F-критерію Фішера F, так і теоретичного F_α при α = 0,05.

На початку роботи з програмою за допомогою меню «File» вихідні експериментальні дані для регресійного аналізу завантажувалися з *.csv або *.txt файлів і заповнювали таблицю «Data». При необхідності рядки таблиці «Data» могли заповнюватися або доповнюватися з використанням вихідних даних для X і Y, які попередньо розміщалися у полях вводу «X» та «Y or Yr», та кнопки «Input» (поля вводу «X» та «Y or Yr» використовуються і при прогнозуванні значень залежної змінної для значень регресора, які визначає користувач). Рядок, який виділявся, можна було видалити за допомогою кнопки «Cut Row». Кнопка «Clear» застосовувалася для видалення усіх рядків, як у таблиці «Data», так і у таблиці «Coefficients». Необхідний ступінь n полінома (1) для регресійного аналізу даних установлювався за допомогою списку, що розкривається, у межах від 1 до 6.

Запуск обробки даних здійснювався при натисканні на кнопку «Calculate». При цьому програма для поліноміальної моделі регресії розраховувала коефіцієнти C_0, C_1, C_2, …, C_n, середню помилку апроксимації та індекс детермінації (коефіцієнт детермінації при n=1), їх значення з'являлися в таблиці «Coefficients», а також в полях вводу «ME, %» і «R^2», відповідно. Крім того, для моделі регресії, що використовувалася при обробці відповідних даних, обчислювалися та зіставлялися значення фактичного і теоретичного F-критерію Фішера. Результат цього зіставлення з'являвся у полі вводу «F ? $F_{0.05}$». У той час значення полінома (1) та залишків регресії, які визначалися програмою для всіх X, із вихідного набору даних, розміщалися в таблиці «Data», а вихідні дані та розрахована для них поліноміальна залежність $Yr = f(X)$ відображалися в графічному вигляді на панелі додатка. Також, якщо в полі вводу «X» знаходилося значення аргументу X, яке попередньо ввів користувач, то програма визначала для нього значення функції $Yr = f(X)$ і поміщала його в поле вводу «Y or Yr».

Користувач програми за допомогою меню «File» мав можливість зберегти результати аналізу для параметрів (C_0, C_1, C_2, …, C_n) регресійної моделі у файлах формату *.csv або *.txt. Крім того, використовуючи меню «File», графік, який був отриманий при регресійному аналізі даних, можна було скопіювати в буфер обміну або відправити на друк. Тим часом за допомогою меню «Tuning» при необхідності настроювалися область відображення та шкали осей графіка $Yr = f(X)$, а також його орієнтація. Між тим інформація про створений Java-додаток та особливостях його роботи була доступна при використанні меню «Help».

Для тестування розробленої програми ми одержали експериментальні дані для величини струму колектора I_K (Y_i у мА) біполярного транзистора при різних значеннях струму бази $I_Б$ (X_i у мА) (транзистор працював у ключовому режимі в схемі з спільним емітером). Вікно програми ST_Regression при аналізі цих даних зображене на рис. 1.

Висновки

Таким чином, у даній роботі був розроблений кросплатформний Java-додаток ST_Regression для поліноміального регресійного аналізу даних.

Комп'ютерна програма має графічний інтерфейс користувача. При регресійному аналізі даних вона використовує метод найменших квадратів і дозволяє прогнозувати значення залежної змінної для значень регресора, які визначає користувач. Крім того, представляє результати аналізу у графічному та табличному вигляді, а також розраховує параметри необхідні для оцінки якості та значимості моделі регресії: коефіцієнт детермінації або індекс детермінації, середню помилку апроксимації, фактичне та теоретичне значення F-критерію Фішера.

Надалі передбачається модернізувати Java-додаток ST_Regression, зокрема, забезпечити можливість проведення регресійного аналізу даних, як при різних рівнях значимості, так і при використанні експонентної та логарифмічної моделі регресії.

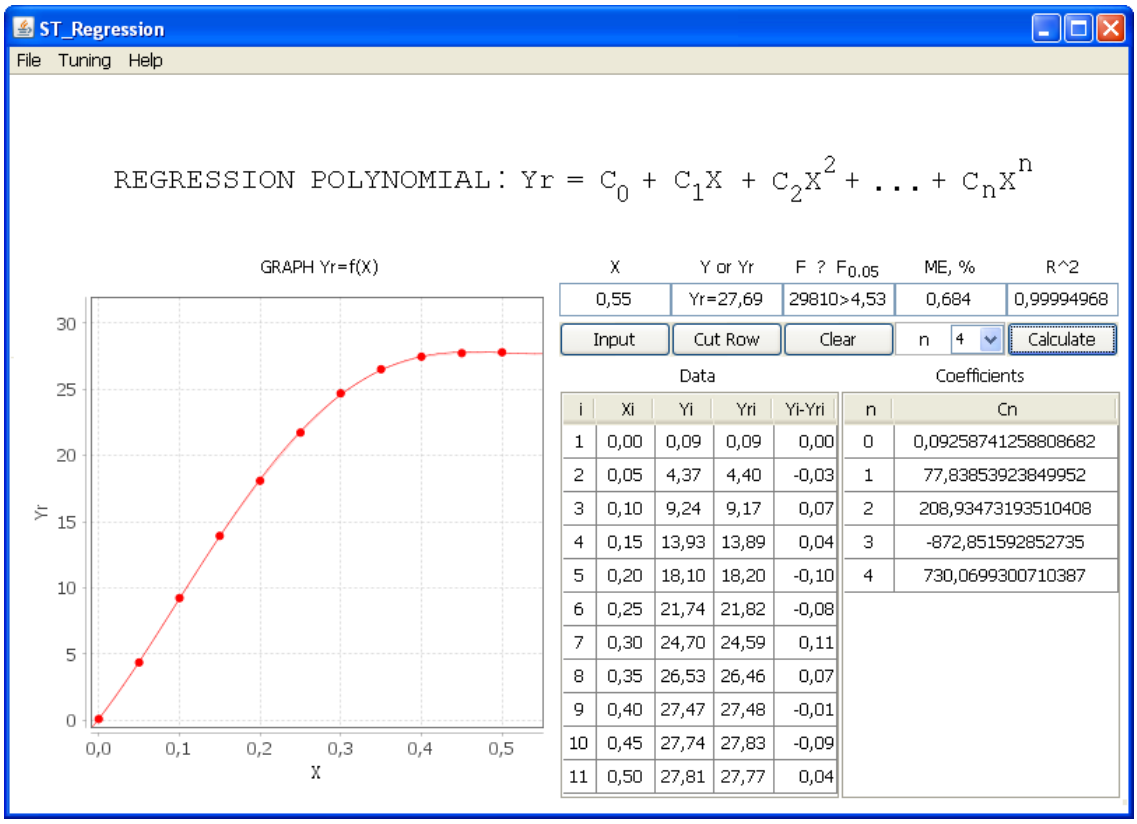

Рис. 1. Вікно програми ST_Regression при регресійному аналізі даних

Список використаної літератури

1. Омельченко А.В. Оценивание коэффициентов полиномиальной регрессии по совокупности реализаций / А.В. Омельченко, А.В. Федоров // Радиоэлектроника и информатика, 2009. – № 1 (44). – С.28-32.
2. Дрейпер Н. Прикладной регрессионный анализ / Н. Дрейпер, Г. Смит. – М.: Издательский дом «Вильямс», 2007. – 912 с.
3. Бараз В.Р. Использование MS Excel для анализа статистических данных : учеб. пособие / В. Р. Бараз, В. Ф. Пегашкин. − Нижний Тагил: НТИ (филиал) УрФУ, 2014. – 181 с.
4. Елисеева И.И. Эконометрика: учебник / И.И. Елисеева, С.В. Курышева, Т.В. Костеева и др. – М.: Финансы и статистика, 2007. – 576 с.
5. Сажин Ю.В., Иванова И.А. Эконометрика: учебник / Ю.В. Сажин, И.А. Иванова. − Саранск: Мордов. гос. ун-т. 2014. – 316 с.
6. Мудров А.Е. Численные методы для ПЭВМ на языках Бейсик, Фортран и Паскаль / А.Е. Мудров. – Томск: МП «РАСКО», 1991. – 272 с.
7. Гринчишин Я.Т. TURBO PASCAL: Чисельні методи в фізиці та математиці: Навчальний посібник / Я.Т. Гринчишин. – Тернопіль, 1994. – 121 с.
8. Гайдышев И. Анализ и обработка данных: специальный справочник / И. Гайдышев. – СПб: Питер, 2001. – 752 с.
9. Аппроксимация функций полиномом методом наименьших квадратов [Електронний ресурс]. – Режим доступу: http://www.alexeypetrov.narod.ru/C/sqr_less_about.html (дата звернення: 07.05.2019).
10. Большов Л.Н., Смирнов Н.В. Таблицы математической статистики / Л.Н. Большов, Н.В. Смирнов. – М.: Наука, 1983. – 416 с.
11. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников / А.И. Кобзарь. – М.: ФИЗМАТЛИТ, 2006. – 816 с.
12. Walck C. Hand-book on STATISTICAL DISTRIBUTIONS for experimentalists / C. Walck. Stockholm: University of Stockholm, 2000. – 204 p.
13. TIOBE Index for May 2019. WEB–сайт [Электронный ресурс] / Режим доступа: http://www.tiobe.com/tiobe-index/ (дата звернення: 07.05.2019).
14. Копченова Н.В. Вычислительная математика в примерах и задачах / Н.В. Копченова, И.А. Марон. – М.: Наука, 1972. – 368 с.