УДК 004.42:303.724.32
DOI: 10.25140/2411-5363-2021-2(24)-141-150

Сергій Точилін, Вадим Рибін

КРОСПЛАТФОРМНА КОМП'ЮТЕРНА ПРОГРАМА ДЛЯ ПРОСТОГО РЕГРЕСІЙНОГО АНАЛІЗУ ДАНИХ

За допомогою мови програмування Java розроблена кросплатформна комп'ютерна програма для простого регресійного аналізу даних, яка при функціонуванні використовує різні моделі регресії. Вона має графічний інтерфейс користувача і застосовує для аналізу метод найменших квадратів. При цьому для визначення параметрів регресійної моделі із системи лінійних рівнянь, які формуються при обробці статистичних даних, використовується метод Гаусса. Розроблений додаток для оцінки якості моделі розраховує середню помилку апроксимації та коефіцієнт детермінації або індекс детермінації, а для оцінки її значущості обчислює фактичне і критичне значення F-критерію Фішера. При розрахунку критичного значення F-критерію Фішера програма використовує функцію бета-розподілу.

Ключові слова: *регресія; аналіз даних; метод найменших квадратів.*
Рис.: 8. Бібл.: 10.

Актуальність тими дослідження. У наш час спеціалізовані комп'ютерні програми широко застосовуються для аналізу статистичних даних.

Регресійний аналіз є одним із найбільш поширених статистичних методів. Регресійний аналіз підрозділяється на однофакторний (простий) та багатофакторний. Особливості проведення простого регресійного аналізу статистичних даних наведені в роботах [1-4].

При цьому розробка комп'ютерних програм для простого регресійного аналізу даних, які під час аналізу використовують різні моделі регресії, є актуальним завадням.

Постановка проблеми. Регресійний аналіз здебільшого виконується 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.

Водночас при проведенні регресійного аналізу для вибору оптимального рівняння регресії визначають його якість та значущість [3; 4].

При цьому розробка кросплатформних комп'ютерних програм із GUI, які при простому регресійному аналізі даних використовують різні моделі регресії, а також оцінюють їхню якість та значущість, є актуальною проблемою.

Аналіз останніх досліджень і публікацій. Відповідно до [3], проста регресія являє собою модель, де середнє значення Y_r залежної (що пояснюється) змінної Y розглядається як функція однієї незалежної (пояснюючої) змінної X, тобто це модель (рівняння), що має вигляд:

$$Y_r = f(X). \tag{1}$$

При простому регресійному аналізі переважно використовують такі моделі регресії:

$$Y_r = C_0 + C_1 \cdot X\,, \tag{2}$$

$$Y_r = C_0 + C_1 \cdot X + C_2 \cdot X^2 + \ldots + C_n \cdot X_n^n\,, \tag{3}$$

$$Y_r = C_0 + C_1 \cdot \ln X\,, \tag{4}$$

$$Y_r = C_0 \cdot X^{C_1}\,, \tag{5}$$

$$Y_r = C_0 \cdot e^{C_1 \cdot X}\,, \tag{6}$$

де C_0, C_1, C_2, …, C_n – постійні коефіцієнти, n – ступінь полінома.

Вирази (2)-(6) описують лінійну, поліноміальну, логарифмічну, статечну та експонентну регресію відповідно.

Для визначення C_0, C_1, C_2, …, C_n можна використовувати метод найменших квадратів (МНК) [3; 4]. Опис МНК із прикладами комп'ютерних програм, які його реалізують, наведено в [5-7].

Оцінку якості рівняння регресії здійснюють за допомогою коефіцієнта детермінації (індексу детермінації при $n > 1$) R^2 та середньої помилки апроксимації ME [3; 4]. Відповідно до [3; 4] для простої регресії їх визначають за допомогою виразів:

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(Y_i - Y_{ri})^2}{\sum_{i=1}^{m}(Y_i - \overline{Y})^2}, \tag{7}$$

$$ME = \frac{1}{m}\sum_{i=1}^{m}\left|\frac{(Y_i - Y_{ri})}{Y_i}\right| \cdot 100\ (\%), \tag{8}$$

де Y_i – експериментальні значення Y при X_i, Y_{ri} – значення функції (1) при X_i, $\overline{Y}$ – середнє для Y_1, Y_2, …, Y_m.

Якість моделі регресії зростає, якщо значення R^2 прагне до 1. У той час значення $ME \le 7$ % також свідчить про задовільний вибір моделі до вихідних даних [3].

Тим часом, для оцінки значущості моделі регресії для даних з m пар чисел, зіставляють фактичне значення F-критерію Фішера F і критичне F_α для заданого рівня значущості α. Якщо F більше F_α, модель регресії визнається значущою для цього рівня значимості, інакше – не значущою. Фактичний F-критерій визначають у такий спосіб [3]:

$$F = \frac{R^2}{(1 - R^2)} \cdot \frac{m - k - 1}{k}. \tag{9}$$

Величина k в (9) характеризує число ступенів свободи f_1 для факторної суми квадратів, а $(m - k - 1)$- число ступенів свободи f_2 для залишкової суми квадратів [3].

Критичні значення F-критерію можна розрахувати за допомогою неповної функції бета-розподілу, що має вигляд [8; 9]:

$$I_x(a,b) = \frac{1}{B(a,b)}\int_0^x t^{a-1}(1-t)^{b-1}dt, \tag{10}$$

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

$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt. \tag{11}$$

При цьому критичне значення F_α, для рівня значимості α, при f_1 та f_2, можна знайти за допомогою виразу [9]:

$$I_x\left(\frac{f_1}{2}, \frac{f_2}{2}\right) = 1 - \alpha, \tag{12}$$

де x має значення:

$$x = \frac{f_1 F_\alpha}{f_2 + f_1 F_\alpha}. \tag{13}$$

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

Останнім часом на мові програмування Java розроблена програма ST_Regression [10] для регресійного аналізу з GUI. Цей додаток для оцінки якості моделі розраховує середню помилку апроксимації та коефіцієнт детермінації або індекс детермінації, а для оцінки її значущості обчислює фактичне і критичне значення F-критерію Фішера.

Виділення недосліджених частин загальної проблеми. Програма ST_Regression має суттєве обмеження. Вона може виконувати лише поліноміальний аналіз даних.

Постановка завдання. У цій роботі була поставлена задача розробки за допомогою мови програмування Java кросплатформного додатка для простого регресійного аналізу даних з GUI, який при функціонуванні використовує різні моделі регресії. Крім того, представляє результати роботи в графічному та табличному вигляді, а також оцінює якість і значимість рівняння регресії.

Виклад основного матеріалу. Для рішення поставленої задачі був створений Java-додаток DS_Regression із графічним інтерфейсом користувача.

На початку роботи із програмою вибиралася необхідна модель регресії для аналізу даних. Вибір здійснювався при включенні відповідного прапорця GUI, назва якого збігалася з необхідним рівнянням регресії. Потім за допомогою меню «File» вихідні дані для аналізу завантажувалися з *.csv файлів і заповнювали таблицю «Data».

За необхідності при натисканні на кнопку «Input» рядки таблиці «Data» могли заповнюватися або доповнюватися даними по X та Y, які попередньо розміщалися в полях вводу «X» і «Y or PV» (поля вводу «X» та «Y or PV» використовувалися і при прогнозуванні значень Y_r для відповідних аргументів X, які вводилися користувачем).

Рядок, що виділявся, можна було видалити за допомогою кнопки «Cut Row». Кнопка «Clear» застосовувалася для видалення всіх рядків у таблицях додатка. У той час список, що розкривається, дозволяв установити ступінь n для поліноміальної моделі регресії (3) у межах від 1 до 6.

Аналіз даних програмою DS_Regression здійснювався за допомогою МНК. При розрахунках рівняння (3)-(6) попередньо приводилися до лінійного вигляду. Значення коефіцієнтів C_0, C_1, а також при необхідності і C_2, …, C_n, знаходилися із системи лінійних рівнянь, які формувалися при обробці даних, за допомогою методу Гаусса.

Для оцінки якості моделі регресії додаток DS_Regression визначав середню помилку апроксимації та коефіцієнт детермінації (індекс детермінації для рівняння (3) при $n > 1$), а для оцінки значущості моделі використовував критичне значення F-критерію Фішера при $\alpha = 0{,}05$.

Запуск обробки даних здійснювався при натисканні на кнопку «Calculate». При цьому програма для обраного рівняння регресії розраховувала коефіцієнти f_1, f_2, C_0, C_1, при необхідності C_2, …, C_n, середню помилку апроксимації, а також індекс детермінації або коефіцієнт детермінації, їх значення з'являлися в таблицях «Values f_m», «Coefficients C_n», а також полях вводу «ME, %» і «R^2», відповідно.

Також програма для аргументу X, що уводив користувач, визначала прогнозовану величину PV моделі, яка цікавила користувача, і поміщала її в поле вводу «Y or PV».

Крім того, для моделі регресії, що використовувалася при обробці даних, при $\alpha = 0{,}05$ розраховувалися та зіставлялися фактичне і критичне значення F-критерію Фішера. Результат цього зіставлення з'являвся в полі вводу «F ? $F_{0.05}$». У той час значення Y_r та залишків регресії, які визначалися програмою для всіх X з вихідного набору даних, розміщалися в таблиці «Data», а вихідні дані по Y і розрахована по них залежність $Y_r = f(X)$ відображалися в графічному вигляді на панелі додатка.

Користувач програми за допомогою меню «File» мав можливість зберегти значення параметрів регресійної моделі у файлах формату *.csv. Водночас за допомогою меню «Tuning» при необхідності настроювалися область відображення графіка, а при використанні меню «Help» була доступна інформація про створений додаток та особливостях його роботи.

Нами було проведене тестування додатка DS_Regression і порівняння результатів його роботи з результатами побудови регресії табличним процесором LibreOffice Calc при обробці тих самих наборів даних.

Для тестування додатка DS_Regression була додатково розроблена Java-програма, що генерувала випадкові значення залежної змінної Y для дискретних значень незалежної змінної X, які змінювалися в інтервалі від 0,1 до 1,1 із кроком 0,1, і зберігала їх файлах формату *.csv.

При цьому дані по Y формувалися в циклі за допомогою виразів:

$$Y_i = c_0 + c_1 \cdot X_i + 0.5 \cdot Math.random(), \quad (14)$$

$$Y_i = (c_0 + c_1 \cdot \ln X_i) \cdot (0.95 + 0.1 \cdot Math.random()), \quad (15)$$

$$Y_i = c_0 \cdot X_i^{c_1} \cdot (0.95 + 0.1 \cdot Math.random()), \quad (16)$$

$$Y_i = c_0 \cdot e^{c_1 \cdot X_i} \cdot (0.95 + 0.1 \cdot Math.random()), \quad (17)$$

де $Math.random()$ – реалізація генератора випадкових чисел на мові програмування Java, c_0, c_1 – постійні коефіцієнти, які були рівні 1 і 2, відповідно.

Формула (14) використовувалася при генерації значень Y із метою тестування програми при аналізі даних на основі лінійної та поліноміальної моделі регресії, формули (15)-(17) – логарифмічної, статечної і експонентної, відповідно.

На рис. 1, 3, 5, 7 зображені вікна розробленого додатка при роботі в тестовому режимі при аналізі даних на основі статечної, поліноміальної, експонентної та логарифмічної моделі регресії, відповідно.

На рис. 2, 4, 6, 8 зображена побудова за допомогою LibreOffice Calc статечної, поліноміальної, експонентної та логарифмічної регресії відповідно.

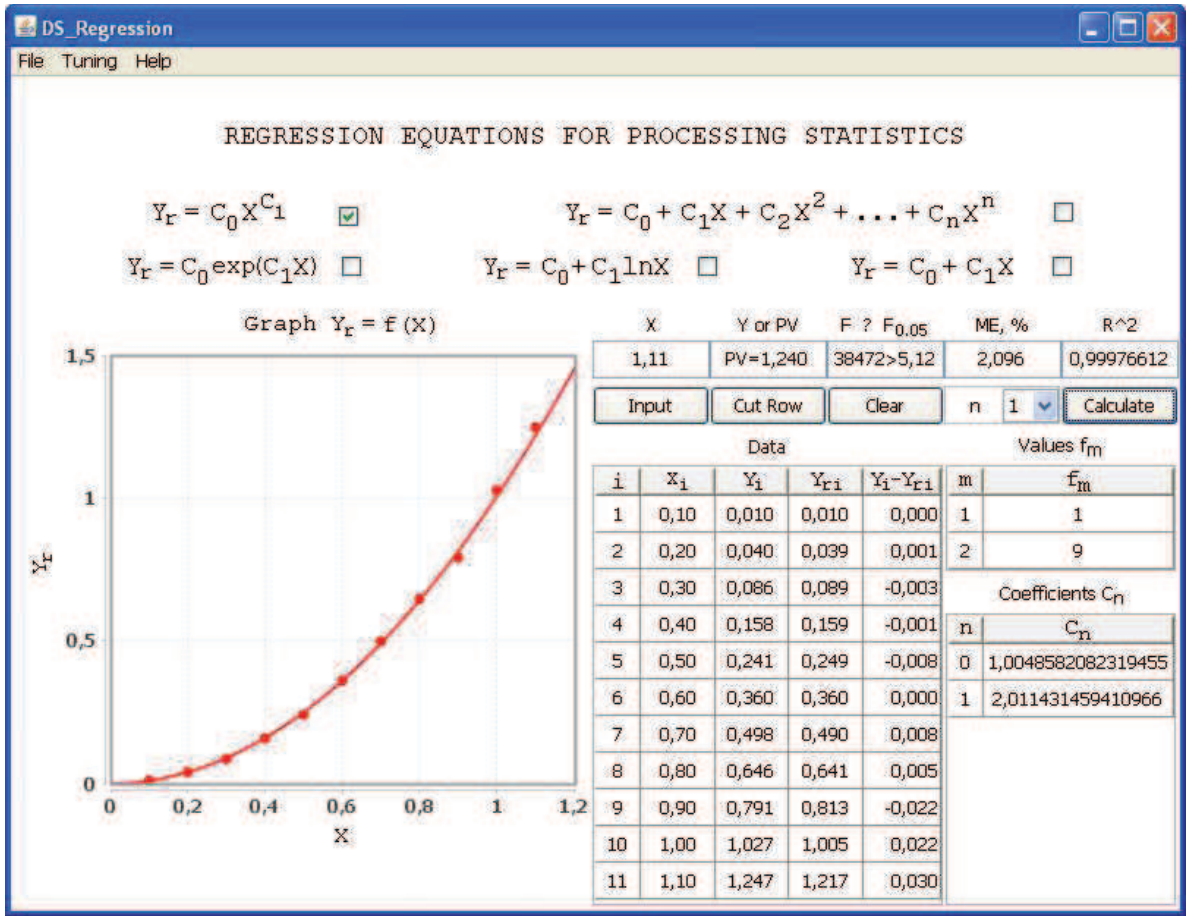

Рис. 1. Вікно програми DS_Regression під час аналізу даних на основі статечної моделі регресії

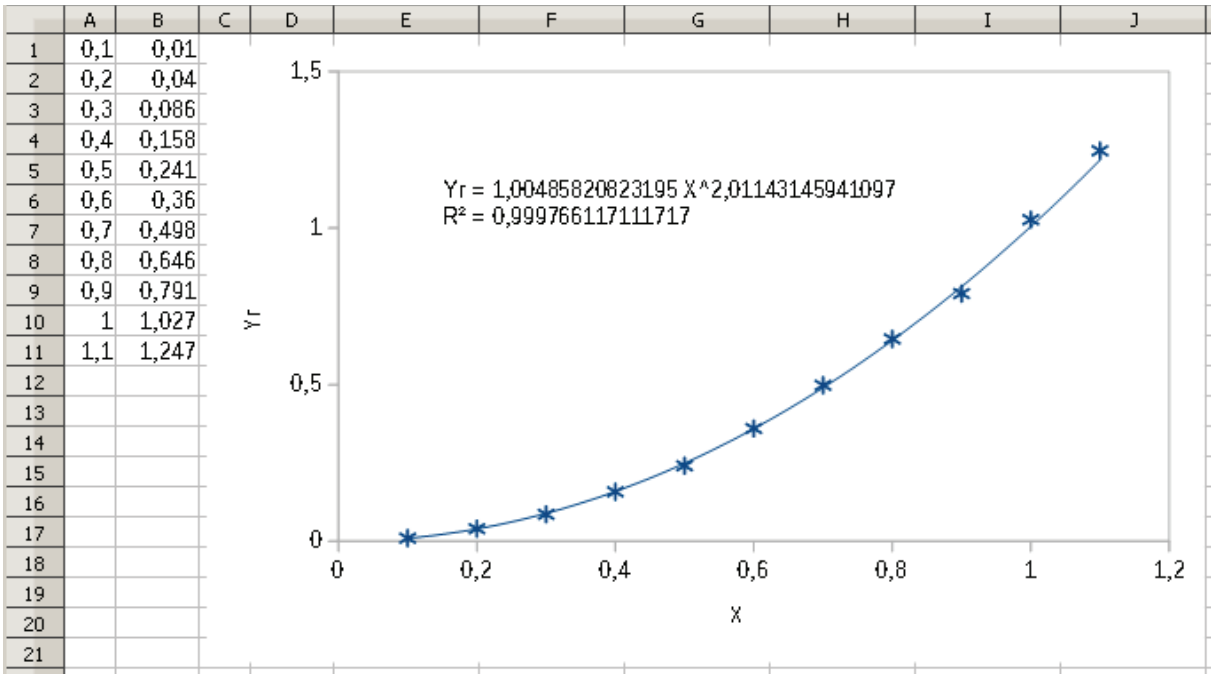

Рис. 2. Побудова статечної регресії за допомогою LibreOffice Calc

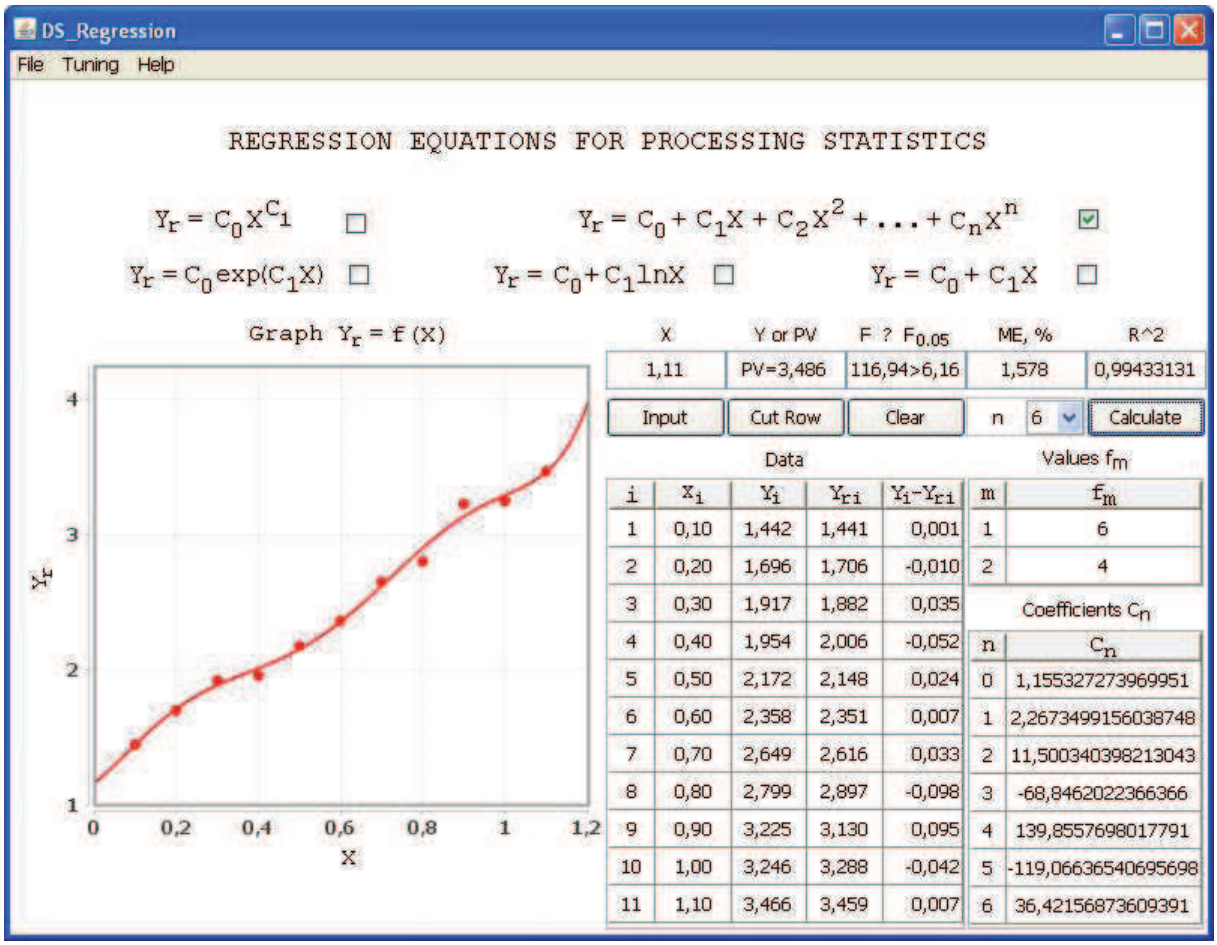

Рис. 3. Вікно програми DS_Regression під час аналізу даних на основі поліноміальної моделі регресії

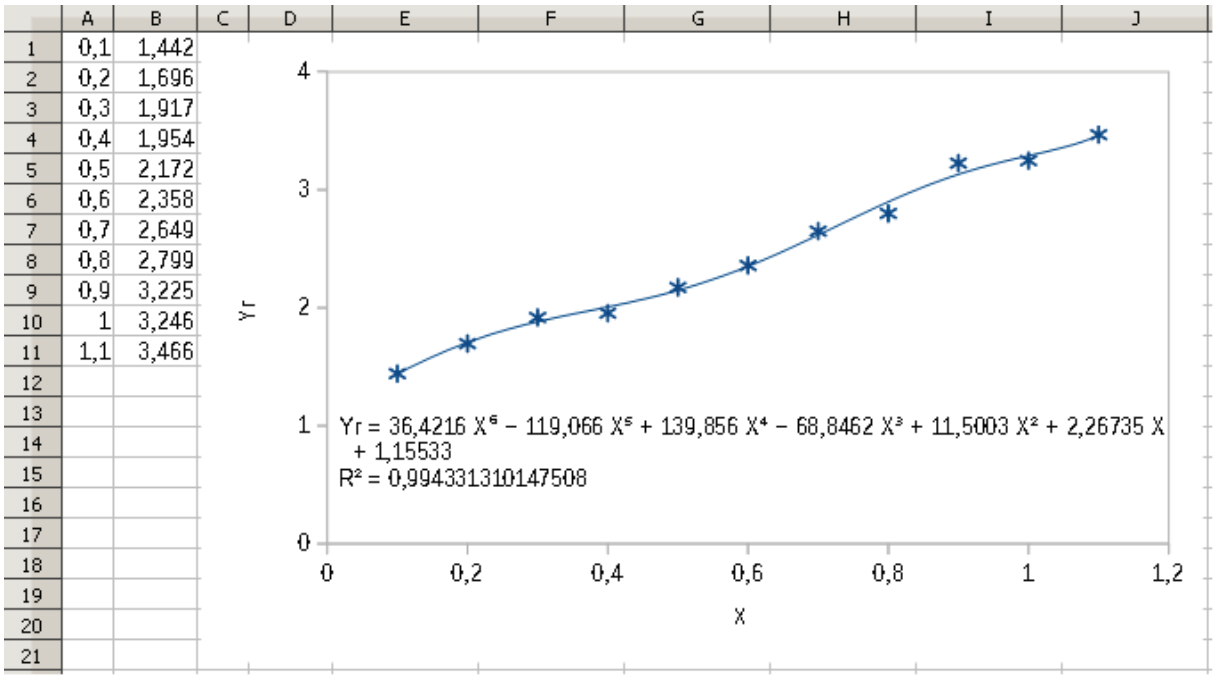

Рис. 4. Побудова поліноміальної регресії за допомогою LibreOffice Calc

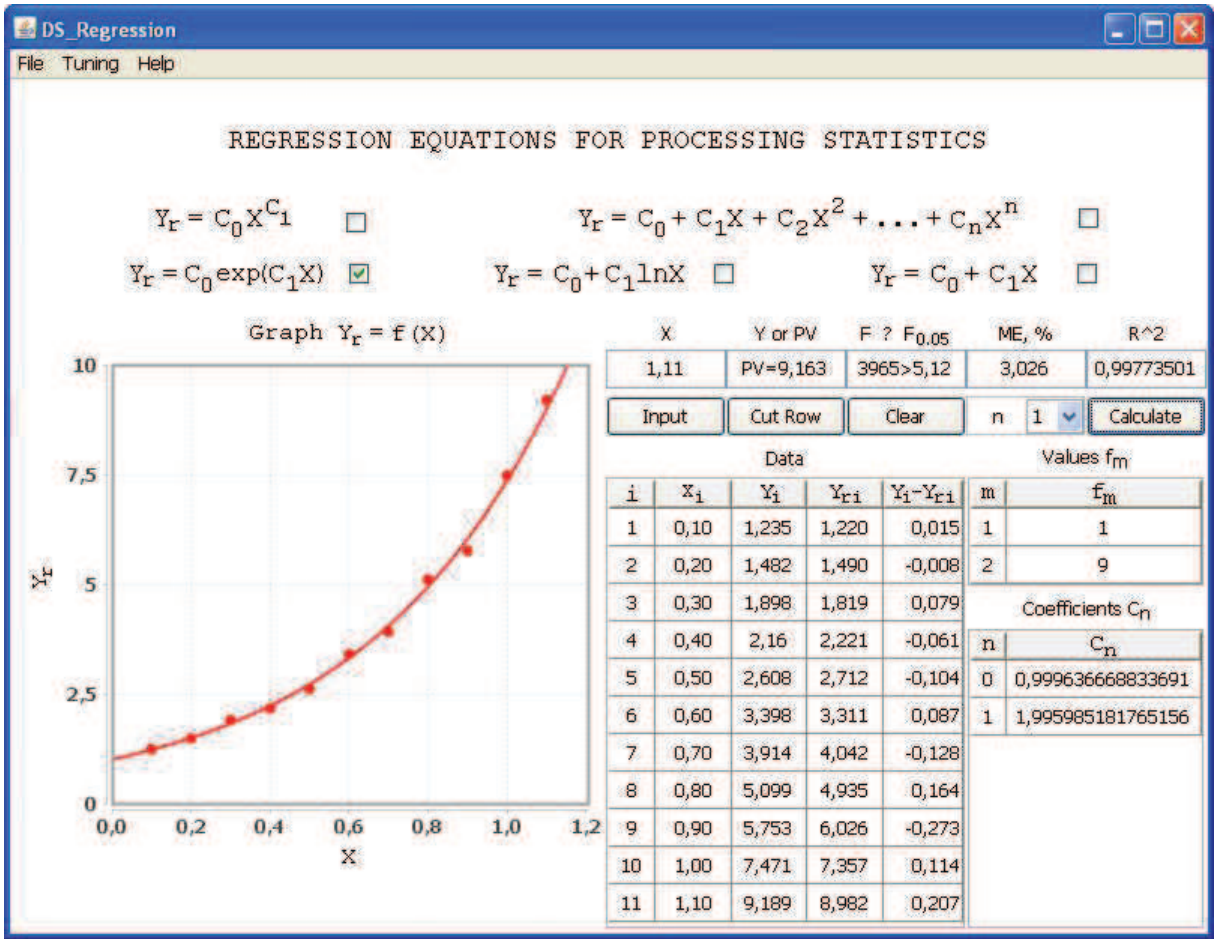

Рис. 5. Вікно програми DS_Regression під час аналізу даних на основі експонентної моделі регресії

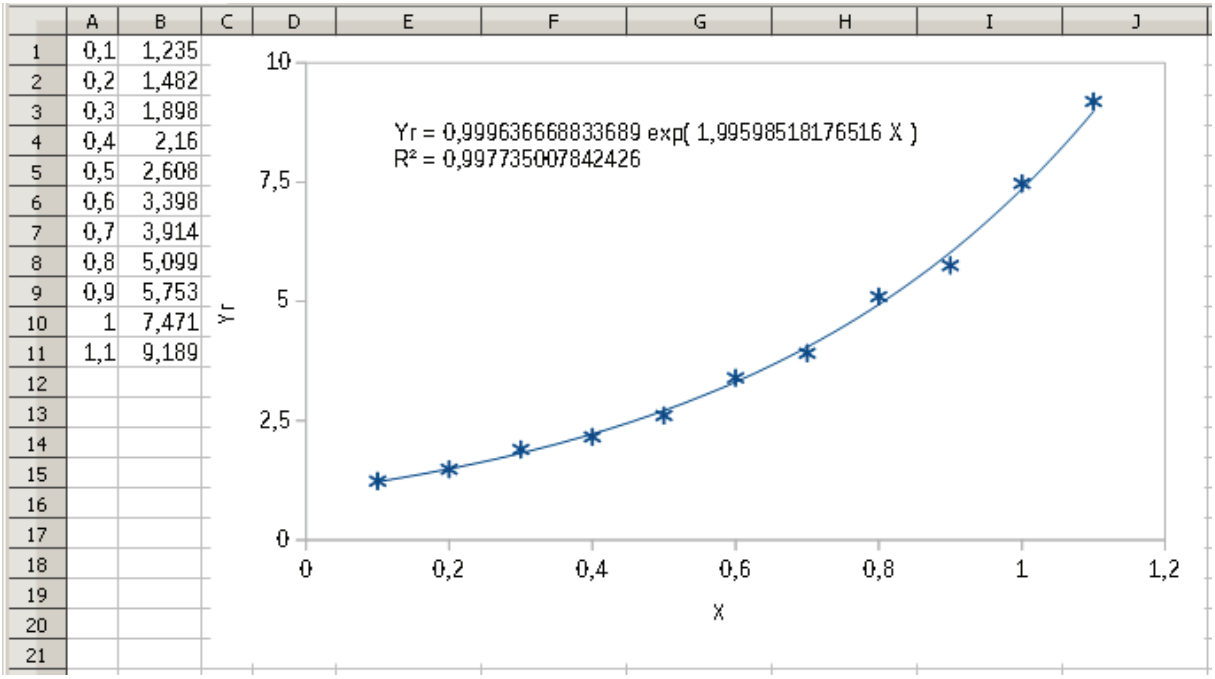

Рис. 6. Побудова експонентної регресії за допомогою LibreOffice Calc

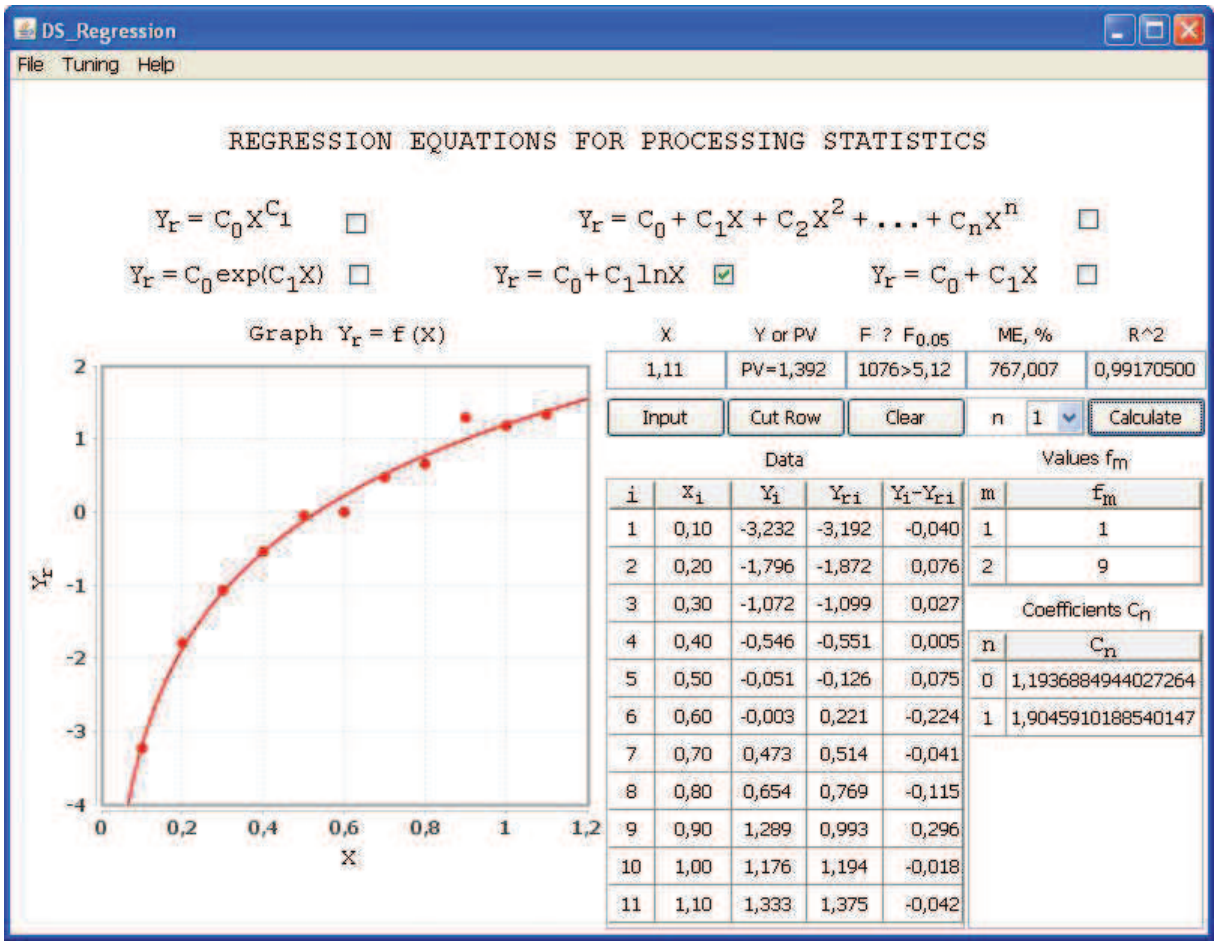

Рис. 7. Вікно програми DS_Regression під час аналізу даних на основі логарифмічної моделі регресії

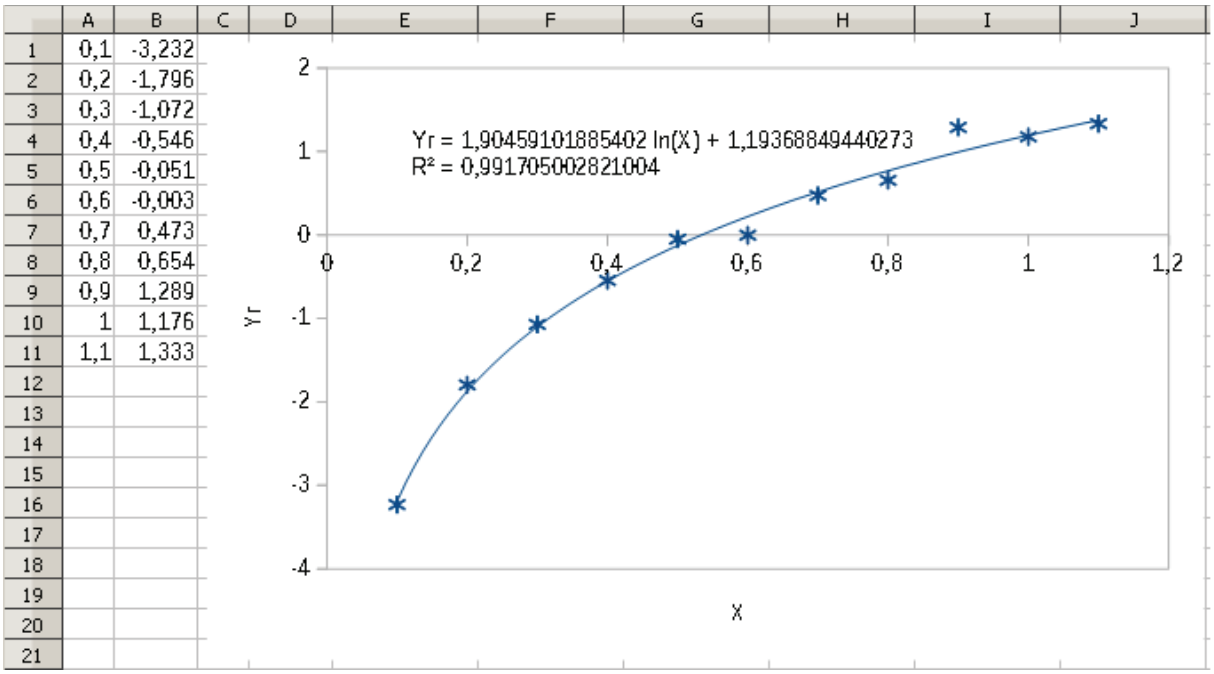

Рис. 8. Побудова логарифмічної регресії за допомогою LibreOffice Calc

Як з'ясувалося в результаті тестування програми, має місце задовільна згода результатів аналізу тих самих наборів даних додатком DS_Regression і табличним процесором LibreOffice Calc при використанні однакових моделей регресії.

Висновки. Таким чином, у цій роботі був розроблений кросплатформний Java-додаток DS_Regression для простого регресійного аналізу даних, який при функціонуванні використовує різні моделі регресії. Комп'ютерна програма має графічний інтерфейс користувача.

У процесі аналізу даних вона використовує метод найменших квадратів. Крім того, представляє результати аналізу в графічному та табличному вигляді, а також визначає параметри, які необхідні для оцінки якості та значущості рівняння регресії: коефіцієнт детермінації або індекс детермінації, середню помилку апроксимації, фактичне і критичне значення F-критерію Фішера.

Надалі передбачається модернізувати додаток DS_Regression, зокрема, при проведенні аналізу даних забезпечити можливість порівняння фактичного значення F-критерію і критичного при різних рівнях значущості.

Список використаних джерел

1. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Москва: Издательский дом «Вильямс», 2007. 912 с.
2. Бараз В. Р., Пегашкин В. Ф. Использование MS Excel для анализа статистических данных : учеб. пособие. Нижний Тагил : НТИ (филиал) УрФУ, 2014. 181 с.
3. Эконометрика : учебник / И. И. Елисеева и др. Москва : Финансы и статистика, 2007. 576 с.
4. Сажин Ю. В., Иванова И. А. Эконометрика : учебник. Саранск : Мордов. гос. ун-т. 2014. 316 с.
5. Мудров А. Е. Численные методы для ПЭВМ на языках Бейсик, Фортран и Паскаль. Томск : МП «РАСКО», 1991. 272 с.
6. Гайдышев И. Анализ и обработка данных: специальный справочник. Санкт-Петербург : Питер, 2001. 752 с.
7. Аппроксимация функций полиномом методом наименьших квадратов. URL: http://www.alexeypetrov.narod.ru/C/sqr_less_about.html.
8. Большов Л. Н., Смирнов Н. В. Таблицы математической статистики. Москва : Наука, 1983. 416 с.

9. Walck C. Hand-book on STATISTICAL DISTRIBUTIONS for experimentalists. Stockholm: University of Stockholm, 2000. 204 p.

10. Точилін С. Д., Рибін О. В. Кросплатформна комп'ютерна програма для поліноміального регресійного аналізу даних. *Вісник Херсонського національного технічного університету*. 2019. № 2. С. 154-158.

References

1. Dreyper, N., Smit, G. (2007). *Prikladnoy regressionnyy analiz [Applied regression analysis]*. ID «Vilyams».

2. Baraz, V.R., Pegashkin, V.F. (2014). *Ispolzovanie MS Excel dlia analiza statisticheskikh dannykh [Using MS Excel for statistical analysis].* STI (branch) UrFU.

3. Eliseeva, I.I., Kurysheva, S.V., Kosteeva, T.V. et al. (2007). *Ekonometrika [Econometrica]*. Finansy i statistika.

4. Sazhin, Yu. V., Ivanova, I. A. (2014). *Ekonometrika [Econometrica]*. Mordov. gos. un-t.

5. Mudrov, A.E. (1991). *Chislennye metody dlia PEVM na iazykah Beisik, Fortran i Paskal [Numerical methods for PC in Basic, Fortran and Pascal].* PASCO Publ.

6. Gaydyshev, I. (2001). *Analiz i obrabotka dannykh: Spetsialnyi spravochnik [Data analysis and processing: Special reference book]*. Piter

7. *Approksimatsiya funktsiy polinomom metodom naimen'shikh kvadratov [Approximation of functions by a polynomial by the method of least squares].* http://www.alexeypetrov.narod.ru/C/sqr_less_about.html.

8. Bolshev, L.N., Smirnov, N.V. (1983). *Tablitsy matematicheskoi statistiki [Mathematical statistics tables]* M.: Nauka [In Russian].

9. Walck, C. (2000). *Hand-book on STATISTICAL DISTRIBUTIONS for experimentalists.* Stockholm: University of Stockholm.

10. Tochilin, S.D., Rybin, O.V. (2019). Krosplatformna kompiuterna prohrama dlia polinomialnoho rehresiynoho analizu danykh [Cross-platform computer program for polynomial regression data analysis]. *Visnyk of Kherson National Technical University – Bulletin of Kherson National Technical University,* (2), 154-158.

UDC 004.42: 303.724.32

Sergei Tochilin, Vadim Rybin

CROSS-PLATFORM COMPUTER SOFTWARE FOR SIMPLE REGRESSION DATA ANALYSIS

Currently, specialized computer programs are widely used to analyze statistical data. At the same time, development of computer programs for regression data analysis is an urgent task.

Regression analysis is usually performed using special commercial computer programs that have a graphical user interface (GUI) and in many cases are rigidly tied to a specific platform.

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

Recently, a cross-platform program ST_Regression for regression analysis with GUI has been developed in the Java programming language.

The ST_Regression program has a significant limitation. It can only perform polynomial data analysis.

The task is to develop a Java GUI application for simple regression data analysis that uses different regression models to function. In addition, it presents the results of work in graphical and tabular form, and also assesses the quality and significance of the regression equation.

The features of functioning and the graphical user interface of a Java application that solves the problem are described. Examples of its use are given.

Using the Java programming language, a cross-platform computer program for simple regression analysis of data, which uses various regression models in operation has been developed. It has a graphical user interface and uses the least squares method for analysis. At the same time, the Gaussian method is used to determine the parameters of the regression model from a system of linear equations that are formed during processing of statistical data. The developed application for assessing the quality of the model calculates the average approximation error and the coefficient of determination or the index of determination, and to assess its significance calculates the actual and critical values of the Fischer F-criterion. When calculating the critical value of Fischer's F-criterion, the program uses the beta distribution function.

Keywords: *regression; data analysis; least squares method.*

Fig.: 8. References: 10.

Точилін Сергій Дмитрович – кандидат фізико-математичних наук, доцент, доцент кафедри комп'ютерних систем та мереж, Національний університет «Запорізька політехніка» (вул. Жуковського, 64, м. Запоріжжя, 69093, Україна).
Tochilin Sergei – PhD in Physico-Mathematical Sciences, Associate Professor, Associate Professor of Department of Computer Systems and Networks, Zaporizhzhia Polytechnic National University (64 Zhukovsky Str., 69063 Zaporizhzhya, Ukraine).
E-mail: tochnozp@gmail.com
ORCID: http//orcid.org/0000-0003-2010-6358
Scopus Author ID: 6602607112

Рибін Вадим Олегович – старший викладач кафедри комп'ютерних систем та мереж, Національний університет «Запорізька політехніка» (вул. Жуковського, 64, м. Запоріжжя, 69093, Україна).
Rybin Oleg – Senior Lecturer of Department of Computer Systems and Networks, Zaporizhzhia Polytechnic National University (64 Zhukovsky Str., 69063 Zaporizhzhya, Ukraine).
E-mail: V_rybin@i.ua
ORCID: http//orcid.org/0000-0001-5856-8844

Точилін С., Рибін В. Кросплатформна комп’ютерна програма для простого регресійного аналізу даних. *Технічні науки та технології.* 2021. № 2(24). С. 141-150.