

С.А. Миловидов¹, М.А. Нифедов¹, Г.Д. Гефан¹

¹ *Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация*

РАЗРАБОТКА ПРИЛОЖЕНИЯ «СТАТИСТИЧЕСКИЙ КАЛЬКУЛЯТОР ДЛЯ КОРРЕЛЯЦИОННО-РЕГРЕССИОННОГО АНАЛИЗА ДАННЫХ»

Аннотация. В работе описаны инструменты языка программирования Python, использованные для разработки приложения. Приведена инструкция использования "Статистического корреляционно-регрессионного калькулятора", разработанного для автоматизации и упрощения процесса корреляционно-регрессионного анализа данных. В данной статье были рассмотрены методы анализа линейной парной регрессии и линейной множественной регрессии. Представлены формулы для нахождения коэффициентов уравнений линии и плоскости регрессии на основе использования метода наименьших квадратов.

Ключевые слова: корреляция, регрессия, корреляционно-регрессионный анализ, метод наименьших квадратов, статистическое приложение, Python.

S.A. Milovidov¹, M.A. Nifedov¹, G.D. Gefan¹

¹ *Irkutsk State Transport University, Irkutsk, Russia*

DEVELOPMENT OF THE APPLICATION «STATISTICAL CALCULATOR FOR CORRELATION AND REGRESSION ANALYSIS OF DATA»

Abstract. This paper describes what tools of the Python programming language were used to develop the application. The instructions for using the "Statistical Correlation and Regression Calculator", designed to automate and simplify the process of correlation and regression analysis of data, are given. In this article, the methods of analysis of linear pair regression and linear multiple regression were considered. Formulas for finding the coefficients of the regression line and plane equations using the least squares method are presented.

Keywords: correlation, regression, correlation-regression analysis, least squares method, statistical application, Python.

Введение

В современном мире человечество оперирует огромным количеством данных. На первый взгляд может показаться, что связи между этими данными нет. Но это не совсем так: чаще всего эта связь присутствует и, более того, тесноту и силу этой связи можно измерить количественно. Такая связь (не функциональная, а вероятностная или статистическая) называется корреляцией. Сопоставив и проанализировав зависимость одних данных от других, мы получаем возможность сделать прогноз для вновь поступающих данных. Именно такая возможность весьма значима во многих важных сферах жизни: экономика, физика, IT-технологии и т.д.

К примеру, для решения отдельных задач селекции животных существенное значение имеет определение генетических корреляций [1], которые отражают процессы, происходящие на уровне генома. Также для оценки финансового состояния предприятий используются коэффициенты ликвидности, структуры капитала, оборачиваемости и рентабельности. На основе корреляционного анализа данных можно получить сравнительный рейтинг финансового состояния промышленных предприятий с учётом базисного стандартизированного показателя, как показано в статье [2]. Состояние кризиса российской экономики является на сегодняшний день определяющим фоном большинства теоретических разработок в области макроэкономического анализа, важным показателем

такого анализа является ВВП (внутренний национальный продукт), для измерения которого применяются методы регрессионного анализа, описанные в работе [3], на основе существующих статистических данных.

Стоит отметить, что проведение корреляционно-регрессионного анализа больших данных осуществляется благодаря объёмным математическим вычислениям, что может привести к определённым трудностям. Для упрощения этих вычислений и более наглядного представления информации нами было решено разработать приложение с помощью высокоуровневого языка программирования Python с использованием дополнительных графических библиотек языка.

Цель данной работы – разработка понятного и удобного инструмента для реализации корреляционно-регрессионного анализа данных.

1. Корреляционно-регрессионный анализ данных и постановка задачи

Ниже представлен список задач, которые были выполнены для получения конечного результата:

1. Изучить методы корреляционно-регрессионного анализа данных:
 - парной корреляции и регрессии;
 - множественной регрессии.
2. Реализовать методы программно.
3. Разработать пользовательский графический интерфейс для удобной работы.

Начнём с определения *функциональной* зависимости. Если величины x и y связаны так, что каждому значению x соответствует одно определённое значение y , то говорят, что имеет место функциональная зависимость $y = f(x)$. Напротив, корреляционная связь – это такая вероятностная или статистическая зависимость, которая не имеет строгого функционального характера [4]. Применительно к данной работе рассматриваются статистические зависимости.

Приведём пример корреляционной зависимости данных: стоимость квартиры зависит от её площади. Благодаря анализу данных, мы можем предположить, что с увеличением площади квартиры увеличивается её цена. Обратим внимание на график, построенный с помощью разработанного приложения (рис.1).

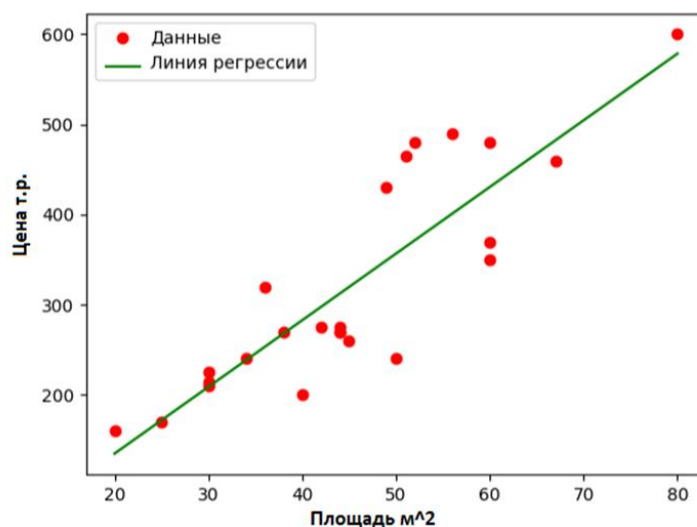


Рис.1. График линии регрессии и введенных данных

По оси абсцисс – площадь квартиры, по оси ординат – её цена. Точка на графике – квартира соответствующей площади и цены. Благодаря такому графику мы можем наглядно

37 проследить имеющуюся связь и сделать прогноз для новых данных, используя уравнение регрессии. Данное уравнение представлено ниже:

$$\bar{y}_x = ax + b . \quad (1)$$

На графике это уравнение имеет вид прямой, что соответствует случаю парной линейной регрессии. Чтобы найти коэффициенты уравнения прямой, используем метод наименьших квадратов. Нужно подобрать такие коэффициенты, при которых сумма квадратов отклонений всех точек от прямой будет минимальной, т.е. должно выполняться следующее условие:

$$S(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2 \rightarrow \min . \quad (2)$$

Воспользуемся необходимыми условиями экстремума функции $S(a, b)$ и получим следующую систему:

$$\begin{cases} \frac{dS}{da} = 2 \sum_{i=1}^n (ax_i + b - y_i)x_i = 0 \\ \frac{dS}{db} = 2 \sum_{i=1}^n (ax_i + b - y_i) = 0 \end{cases} . \quad (3)$$

Для простоты опускаем индексацию, получим систему уравнений:

$$\begin{cases} (\sum x^2)a + (\sum x)b = \sum xy \\ (\sum x)a + nb = \sum y \end{cases} . \quad (4)$$

После деления на n система принимает вид:

$$\begin{cases} \bar{x}^2 a + \bar{x}b = \bar{xy} \\ \bar{x}a + b = \bar{y} \end{cases} . \quad (5)$$

Решение системы может быть легко найдено с помощью метода Крамера:

$$a^* = \frac{\bar{xy} - \bar{x}\bar{y}}{\bar{x}^2 - (\bar{x})^2} , \quad (6)$$

$$b^* = \bar{y} - a^* \bar{x} . \quad (7)$$

Чтобы определить, насколько тесна связь между величинами, используют коэффициент корреляции, который рассчитывается по представленным ниже формулам:

$$r_{xy} = \frac{\bar{xy} - \bar{x}\bar{y}}{\sigma_x \sigma_y} , \quad (8)$$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i , \quad (9)$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i , \quad (10)$$

$$\overline{xy} = \frac{1}{n} \sum_{i=1}^n x_i y_i, \quad (11)$$

$$\sigma_x^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2, \quad (12)$$

$$\sigma_y^2 = \frac{1}{n} \sum_{i=1}^n y_i^2 - (\bar{y})^2. \quad (13)$$

Занимаясь корреляционно-регрессионным анализом, нужно учитывать, что некоторая величина может зависеть не от одной, но от двух и более величин. Выше мы рассмотрели пример зависимости цены квартиры от её площади. Но без внимания остались другие обстоятельства: удалённость от центра города, этаж, количество комнат и т.д. Между тем, без всяких вычислений ясно, что роль этих факторов весьма существенна.

Чтобы учесть все необходимые данные, понадобится уравнение множественной регрессии:

$$\overline{y_{x_1, x_2, \dots, x_k}} = a_1^* x_1 + a_2^* x_2 + \dots + a_k^* x_k + b^* \quad (14)$$

Для нахождения оценки параметров этого уравнения используют (как и для парной регрессии) метод наименьших квадратов:

$$S(a_1, a_2, \dots, a_k, b) = \sum_{i=1}^n (y_i - a_1 x_{i1} - a_2 x_{i2} - \dots - a_k x_{ik} - b)^2 \rightarrow \min. \quad (15)$$

Запишем необходимые условия экстремума функции S (индексацию по номерам наблюдения i опускаем):

$$\left\{ \begin{array}{l} \frac{dS}{da_1} = -2 \sum (y - a_1 x_1 - \dots - a_k x_k - b) x_1 = 0 \\ \frac{dS}{da_2} = -2 \sum (y - a_1 x_1 - \dots - a_k x_k - b) x_2 = 0 \\ \dots \\ \frac{dS}{da_k} = -2 \sum (y - a_1 x_1 - \dots - a_k x_k - b) x_k = 0 \\ \frac{dS}{db} = -2 \sum (y - a_1 x_1 - \dots - a_k x_k - b) = 0 \end{array} \right. \quad (16)$$

Или получим:

$$\left\{ \begin{array}{l} (\sum x_1^2) a_1 + (\sum x_2 x_1) a_2 + \dots + (\sum x_k x_1) a_k + b \sum x_1 = \sum x_1 y \\ (\sum x_1 x_2) a_1 + (\sum x_2^2) a_2 + \dots + (\sum x_k x_2) a_k + b \sum x_2 = \sum x_2 y \\ \dots \\ (\sum x_1 x_k) a_1 + (\sum x_2 x_k) a_2 + \dots + (\sum x_k^2) a_k + b \sum x_k = \sum x_k y \\ (\sum x_1) a_1 + (\sum x_2) a_2 + \dots + (\sum x_k) a_k + b n = \sum y \end{array} \right. \quad (17)$$

После деления на n получим:

$$\begin{cases} \overline{x_1^2} a_1 + \overline{x_2 x_1} a_2 + \dots + \overline{x_k x_1} a_k + \overline{x_1} b = \overline{x_1 y} \\ \overline{x_1 x_2} a_1 + \overline{x_2^2} a_2 + \dots + \overline{x_k x_2} a_k + \overline{x_2} b = \overline{x_2 y} \\ \dots \\ \overline{x_1 x_k} a_1 + \overline{x_2 x_k} a_2 + \dots + \overline{x_k^2} a_k + \overline{x_k} b = \overline{x_k y} \\ \overline{x_1} a_1 + \overline{x_2} a_2 + \dots + \overline{x_k} a_k + b = \overline{y} \end{cases} \quad (18)$$

Получаем систему $k+1$ линейных алгебраических уравнений с $k+1$ неизвестными. При большом числе неизвестных обычно используют методы линейной алгебры, записывая систему в матричном виде.

Для приложения решено было разработать режим с двумя объясняющими переменными. Тогда уравнение плоскости регрессии принимает следующий вид.

$$y_i = a_1 x_{1i} + a_2 x_{2i} + b + \varepsilon_i \quad (19)$$

Система уравнений на параметры функции регрессии имеет вид:

$$\begin{cases} \overline{x_1^2} a_1 + \overline{x_2 x_1} a_2 + \overline{x_1} b = \overline{x_1 y} \\ \overline{x_1 x_2} a_1 + \overline{x_2^2} a_2 + \overline{x_2} b = \overline{x_2 y} \\ \overline{x_1} a_1 + \overline{x_2} a_2 + b = \overline{y} \end{cases} \quad (20)$$

Выразим из системы неизвестную b :

$$b = \overline{y} - \overline{x_1} a_1 - \overline{x_2} a_2 \quad (21)$$

Подставим вместо b формулу (21) в систему (20):

$$\begin{cases} [\overline{x_1^2} - (\overline{x_1})^2] a_1 + [\overline{x_1 x_2} - \overline{x_1} \overline{x_2}] a_2 = \overline{x_1 y} - \overline{x_1} \overline{y} \\ [\overline{x_1 x_2} - \overline{x_1} \overline{x_2}] a_1 + [\overline{x_2^2} - (\overline{x_2})^2] a_2 = \overline{x_2 y} - \overline{x_2} \overline{y} \end{cases} \quad (22)$$

Коэффициентами системы (22) являются выборочные дисперсии и ковариации. Перепишем систему в следующем виде:

$$\begin{cases} D_{x_1} a_1 + \mu_{x_1 x_2} a_2 = \mu_{x_1 y} \\ \mu_{x_1 x_2} a_1 + D_{x_2} a_2 = \mu_{x_2 y} \end{cases}, \quad (23)$$

где символ D обозначает дисперсии, а μ - ковариации, т.е.:

$$D_{x_1} = \overline{x_1^2} - (\overline{x_1})^2 \quad (24)$$

$$D_{x_2} = \overline{x_2^2} - (\overline{x_2})^2 \quad (25)$$

$$\mu_{x_1 x_2} = \overline{x_1 x_2} - \overline{x_1} \overline{x_2} \quad (26)$$

$$\mu_{x_1 y} = \overline{x_1 y} - \overline{x_1} \overline{y} \quad (27)$$

$$\mu_{x_2 y} = \overline{x_2 y} - \overline{x_2} \overline{y} \quad (28)$$

Если $D_{x_1} D_{x_2} \neq \mu_{x_1 x_2}^2$, то система (23) имеет единственное решение:

$$a_1^* = \frac{\mu_{x_1 y} D_{x_2} - \mu_{x_2 y} \mu_{x_1 x_2}}{D_{x_1} D_{x_2} - \mu_{x_1 x_2}^2} \quad (29)$$

$$a_2^* = \frac{\mu_{x_2 y} D_{x_1} - \mu_{x_1 y} \mu_{x_1 x_2}}{D_{x_1} D_{x_2} - \mu_{x_1 x_2}^2} \quad (30)$$

$$b^* = \overline{y} - a_1^* \overline{x_1} - a_2^* \overline{x_2} \quad (31)$$

2. О разработке и использовании приложения

Был выбран высокоуровневый язык программирования Python версии 3.9.5, так как он поддерживает многофункциональные инструменты для работы с большими объёмами данных.

Были использованы следующие дополнительные библиотеки языка:

- Math [5] – содержит функции для произведения математических операций;
- KIVY [6] – библиотека для реализации пользовательского графического интерфейса;
- Matplotlib [7] – используется для визуализации данных (построение графиков).

Размер памяти, занимаемый приложением, составляет 82.2 Мб (рис.2), что для современных компьютеров является достаточно небольшим, например, по сравнению с MS Excel.

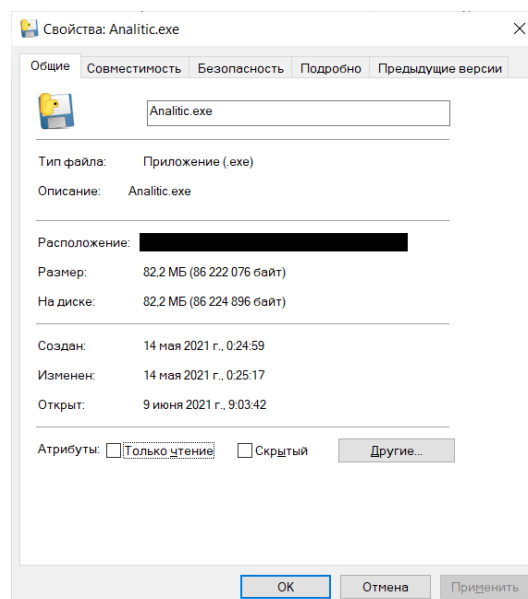


Рис.2. Информация о размере приложения

Для запуска приложение нужно 2 раза кликнуть левой кнопкой мыши по иконке, показанной на рисунке 3.

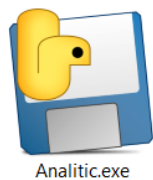


Рис.3. Иконка приложения

После нажатия приложение загрузиться, и откроется стартовое окно разработанного продукта (рис.4).

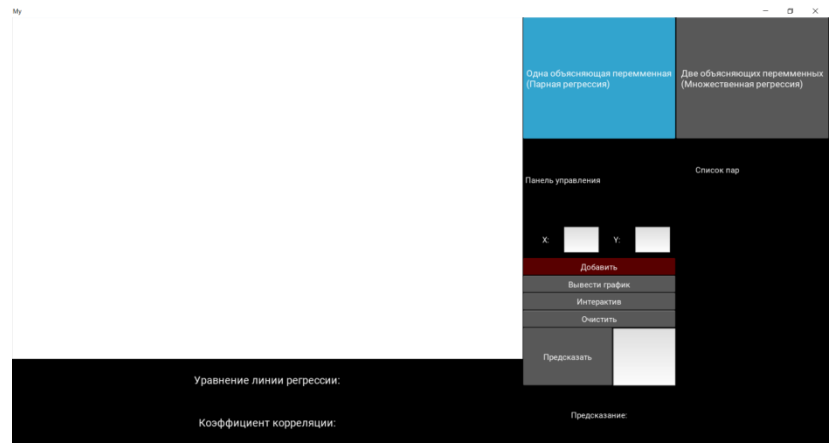


Рис.4. Стартовое окно приложения.

В верхнем правом углу находятся 2 кнопки с подписями «Одна объясняющая переменная (Парная регрессия)» и «Две объясняющих переменных (Множественная регрессия)». Эти кнопки нужны для переключения режимов работы приложения - парная и множественная регрессия соответственно.

Ниже этих кнопок располагается панель управления. В поля «X» и «Y» вводятся данные, после чего необходимо нажать кнопку «Добавить» и пара значений отобразится в правой части приложения (с соответствующим порядковым номером). В качестве примера возьмём набор из 9 пар – объем оперативной памяти в ПК и его стоимость, интерфейс приложения примет вид, показанный на рисунке 5.

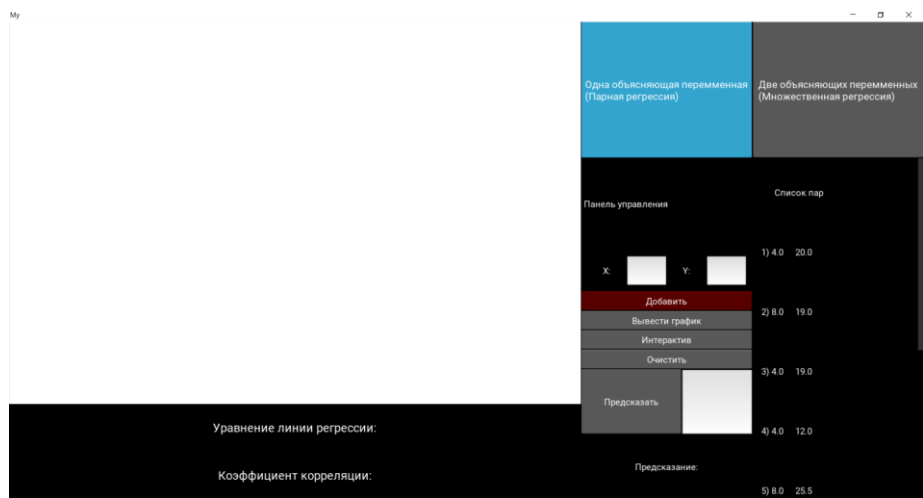


Рис.5. Приложение после введения данных

Когда все данные будут введены, необходимо нажать кнопку «Вывести график». В левой части приложения появится сам график и вычисленные уравнение регрессии и коэффициент корреляции, как показано на рисунке 6.

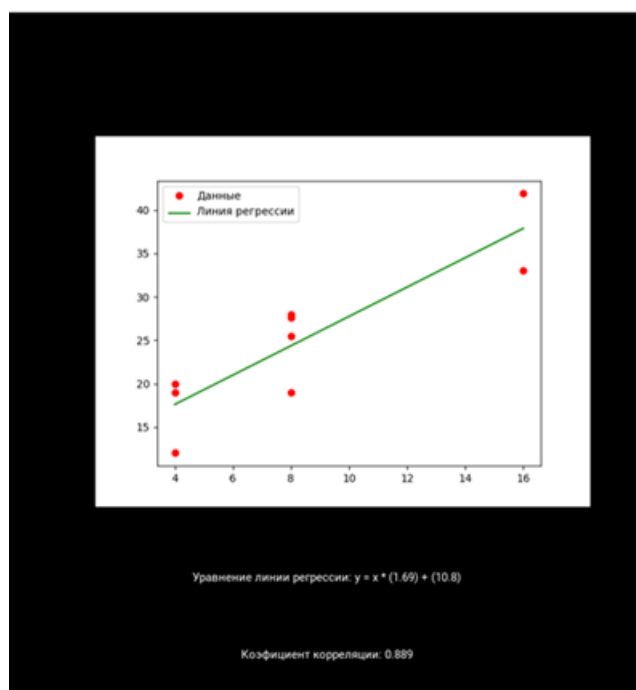


Рис.6. Полученный график, уравнение линии регрессии, коэффициент корреляции

Зелёная линия – линия регрессии, красные точки – введённые данные.

Из полученного результата работы программы видно, что связь между этими данными есть – чем больше объём оперативной памяти, тем выше цена ПК.

Кнопка «Интерактив» позволяет открыть график в интерактивном режиме (рис.7), где есть дополнительные функции для работы с ним:

- сохранить изображение графика;
- увеличить/уменьшить график;
- получить координаты курсора по осям;
- перемещение на графике.

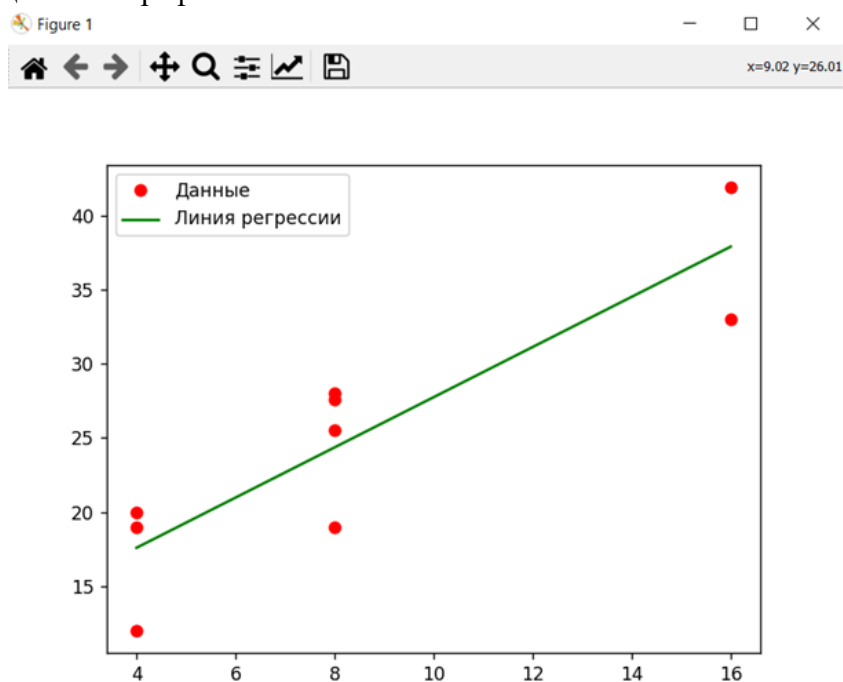


Рис.7. Интерактивный режим графика

Когда график, уравнение и коэффициент получены, появляется возможность воспользоваться функцией предсказания. Предсказание означает, что для конкретного X вычисляется соответствующий Y . Для этого в правой части экрана необходимо ввести значение, для которого необходимо получить предсказание, и нажать кнопку «Предсказать» (рисунок 8).

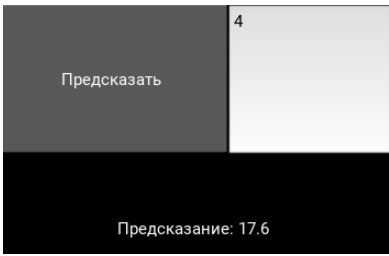


Рис.8. Предсказание y по x

Если возникнет необходимость дополнить данные новыми значениями, их можно добавить по такому же принципу, как и предыдущие, после чего программа вычислит новые значения.

Нажав на кнопку «Очистить», все данные стираются, после чего можно начать работу с новыми данными.

Перейдём в режим множественной регрессии, нажав на кнопку «Две объясняющих переменных (Множественная регрессия)».

Интерфейс для этого режима остаётся практически неизменным, кроме полей для ввода новых данных и полей для предсказания (рис.9).

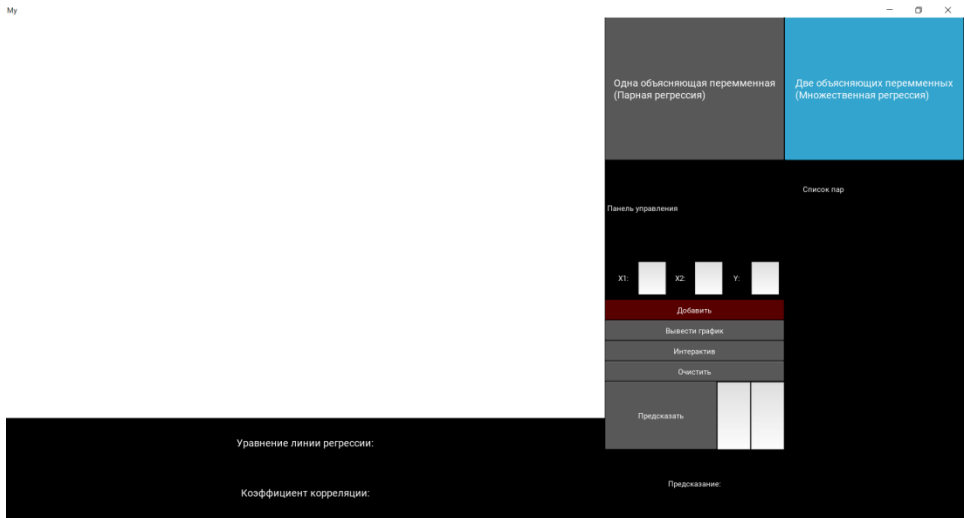


Рис.9. Приложение в режиме «множественная регрессия»

Для примера будем использовать тот же набор значений, дополненный объёмом жёсткого диска ПК, после чего выведем график (рис.10).

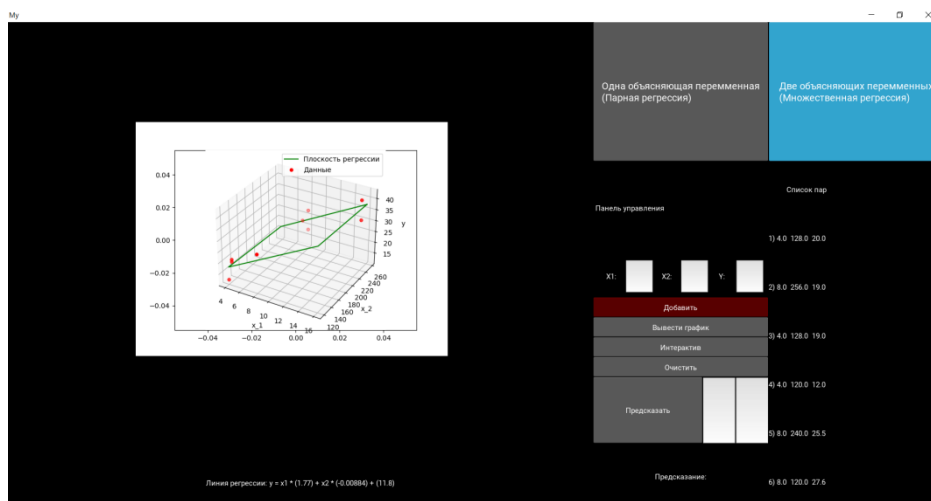


Рис.10. Построенный график по введённым данным

Зелёный много угольник – плоскость регрессии, красные точки – введённые данные. Попробуем сделать предсказание на основе полученных результатов (рис.11).

Предсказать	4	128
	Предсказание: 17.8	

Рис.11. Предсказание для двух объясняющих переменных

Для работы с плоскостью регрессии наиболее удобным представляется интерактивный режим, так как в нём имеется возможность рассмотреть полученный график под разными углами (рис.12).

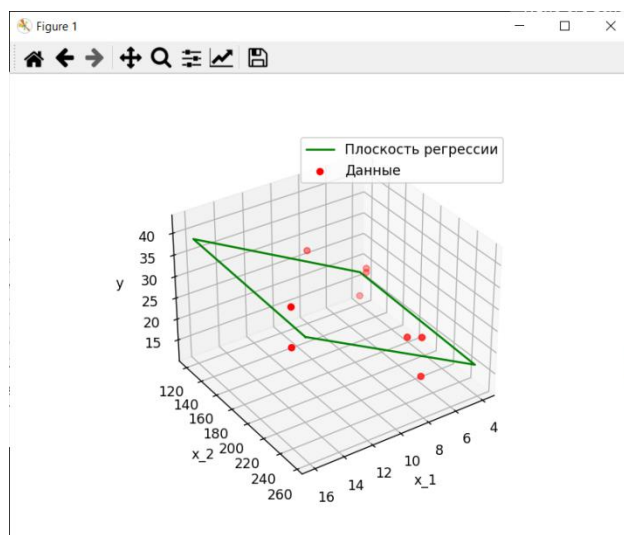


Рис.12. Интерактивный режим для графика

Если точка имеет яркий красный цвет, то она расположена ближе к камере, если же точка имеет бледный цвет – она находится дальше от камеры.

Заключение. Разработано компьютерное приложение для корреляционно-регрессионного анализа данных с двумя режимами:

- с одной объясняющей переменной;
- с двумя объясняющими переменными.

Приложение имеет интерактивный режим для линии и плоскости регрессии. Представлена инструкция по использованию приложения. В дальнейшем планируется разработать мобильную версию данного приложения, например, для использования его в учебных целях.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Горлов, А. И. Определение генетических корреляций селекционных признаков через частные корреляции / А. И. Горлов, Е. А. Ивина, И. А. Мокеев // Сборник научных трудов Ставропольского научно-исследовательского института животноводства и кормопроизводства. – 2009. – Т. 2. – № 2-2. – С. 25-29.
2. Яшина, Н. И. Инструментарий прогнозирования финансового состояния организаций на основе теории регрессионного анализа, методов Парето и ранговой корреляции / Н. И. Яшина, С. Н. Яшин // Финансы и кредит. – 2004. – № 5(143).
3. Дергунов, В. В. Анализ динамики ВВП методом линейной регрессии / В. В. Дергунов // Вестник Финансовой академии. – 1999. – № 4(12). – С. 98-108.
4. Гефан Г.Д. Статистический метод и основы его применения: Учебное пособие. – Иркутск: ИрГУПС, 2003. - 208 с
5. Math — Mathematical functions [Электронный ресурс]. – Режим доступа: <https://docs.python.org/3/library/math.html> (дата обращения 07.06.2021)
6. KIVY documentation [Электронный ресурс].– Режим доступа: <https://kivy.org/doc/stable/> (дата обращения 07.06.2021)
7. Matplotlib: Visualization with Python [Электронный ресурс].– Режим доступа: <https://matplotlib.org/> (дата обращения 07.06.2021)

REFERENCES

1. Gorlov, A. I. Opredelenie geneticheskikh korrelyacij selekcionnykh priznakov cherez chastnye korrelyacii [Determination of genetic correlations of breeding traits through partial correlations]. Sbornik nauchnykh trudov Stavropol'skogo nauchno-issledovatel'skogo instituta zhivotnovodstva i kormoproizvodstva [Collection of scientific papers of the Stavropol Research Institute of Animal Husbandry and Feed Production.]. 2009, T.2, no. 2-2, pp. 25-29.
2. Yashina, N. I. *Instrumentarij prognozirovaniya finansovogo sostoyaniya organizacij na osnove teorii regressionnogo analiza, metodov Pareto i rangovoj korrelyacii* [Tools for forecasting the financial condition of organizations based on the theory of regression analysis, Pareto methods and rank correlation]. *Finansy i kredit* [Finance and Credit]. 2004. no. 5, vol. 143.
3. Dergunov, V. V. *Analiz dinamiki VNP metodom linejnoy regressii* [Analysis of DNP dynamics by linear regression method]. *Vestnik Finansovoj akademii* [Bulletin of the Financial Academy]. 1999, no. 4, vol. 12, pp. 98-108.
4. Gefan G.D. *Statisticheskij metod i osnovy ego primeneniya* [Statistical method and the basics of its application]. *Uchebnoe posobie* [Student textbook]. Irkutsk: IrGUPS, 2003, p. 208.
5. Math — Mathematical functions [Electronic resource]. – Mode of access: <https://docs.python.org/3/library/math.html> (accessed 07.06.2021)
6. KIVY documentation [Electronic resource]. – Mode of access: <https://kivy.org/doc/stable/> (accessed 07.06.2021)
7. Matplotlib: Visualization with Python [Electronic resource]. – Mode of access: <https://matplotlib.org/> (accessed 07.06.2021)

Информация об авторах

Миловидов Святослав Александрович - студент гр. ПИ.1-19-2, Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: ponik_mira2@mail.ru

Нифедов Максим Александрович - студент гр. ПИ.1-19-2, Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: ziggy.zen@yandex.ru

Гефан Григорий Давыдович - к.ф.-м.н., доцент кафедры «Математика», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: grigef@rambler.ru

Authors

Svyatoslav Aleksandrovich Milovidov – student, Irkutsk State Transport University, Irkutsk, e-mail: ponik_mira2@mail.ru

Maksim Aleksandrovich Nifedov – student, Irkutsk State Transport University, Irkutsk, e-mail: ziggy.zen@yandex.ru

Grigory Davydovich Gefan – candidate of physical and mathematical sciences, associate professor of department of mathematics, Irkutsk State Transport University, Irkutsk, e-mail: grigef@rambler.ru

Для цитирования

Миловидов С.А., Нифедов М.А., Гефан Г.Д. Разработка приложения «Статистический калькулятор для корреляционно-регрессионного анализа данных» // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – . – 2021. – №3(11). – С. 35-46 – DOI: 10.26731/2658-3704.2021.3(11).35-46 – Режим доступа: <http://ismm-irgups.ru/toma/311-2021>, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 01.09.2021)

For citation

Milovidov S. A., Nifedov M. A., Gefan G. D. Development of the application «Statistical calculator for correlation and regression analysis of data»// *Informacionnye tekhnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: elektronnyj nauchnyj zhurnal* [Information technologies and mathematical modeling in the management of complex systems: electronic scientific journal], 2021. No. 3(11). P. 35-46. DOI: 10.26731/2658-3704.2021.3(11).35-46. [Accessed 01/09/21]