## Министерство образования Республики Беларусь

Учреждение образования

## БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Инженерно-экономический факультет

### Кафедра ЭИ

Отчет

по лабораторной работе №2

«Анализ зависимостей. Корреляционный и регрессионный анализ. Парная корреляция»

по курсу

«Машинное обучение»

Вариант 7

|  |  |
| --- | --- |
| Выполнил:  Ермаков К. Ю. | Проверил:  Милентьев В.А. |

Минск 2024

**Цель работы**: Изучить основные понятия регрессионных моделей, ознакомиться с методикой построения регрессионных моделей.

Задание 1:

1. Произвести корреляционный и регрессионный анализ данных с помощью Python.
2. Построить график зависимости и линию регрессии
3. Определить и записать уравнение регрессии
4. Исходя из величины коэффициента корреляции, сделать выводы о силе зависимости между переменными и ее направление
5. Сделать выводы о значимости корреляции



import numpy as np

import matplotlib.pyplot as plt

from scipy import stats

electric\_activity = np.array([0, 38.5, 59, 97.4, 119.2, 129.5, 198.7, 248.7, 318, 438.5])

vascular\_permeability = np.array([19.5, 15, 13.5, 23.3, 6.3, 2.5, 13, 1.8, 6.5, 1.8])

correlation\_coefficient, p\_value = stats.pearsonr(electric\_activity, vascular\_permeability)

print("Коэффициент корреляции:", correlation\_coefficient)

print("p-value:", p\_value)

slope, intercept, r\_value, p\_value, std\_err = stats.linregress(electric\_activity, vascular\_permeability)

print("Уравнение регрессии: y =", slope, "x +", intercept)

plt.scatter(electric\_activity, vascular\_permeability, label='Данные')

plt.plot(electric\_activity, slope \* electric\_activity + intercept, color='red', label='Линия регрессии')

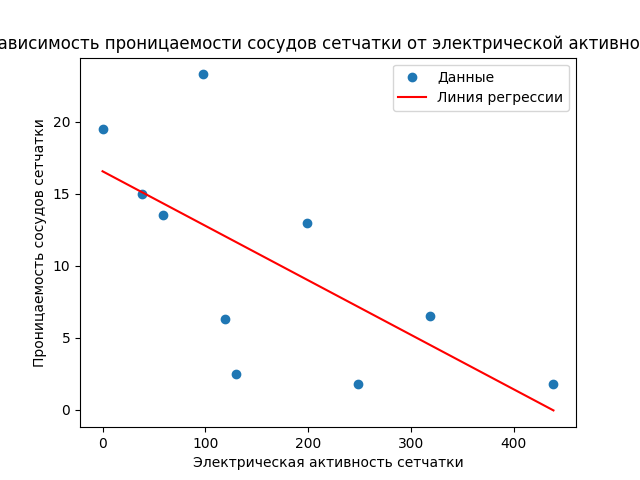
plt.xlabel('Электрическая активность сетчатки')

plt.ylabel('Проницаемость сосудов сетчатки')

plt.title('Зависимость проницаемости сосудов сетчатки от электрической активности')

plt.legend()

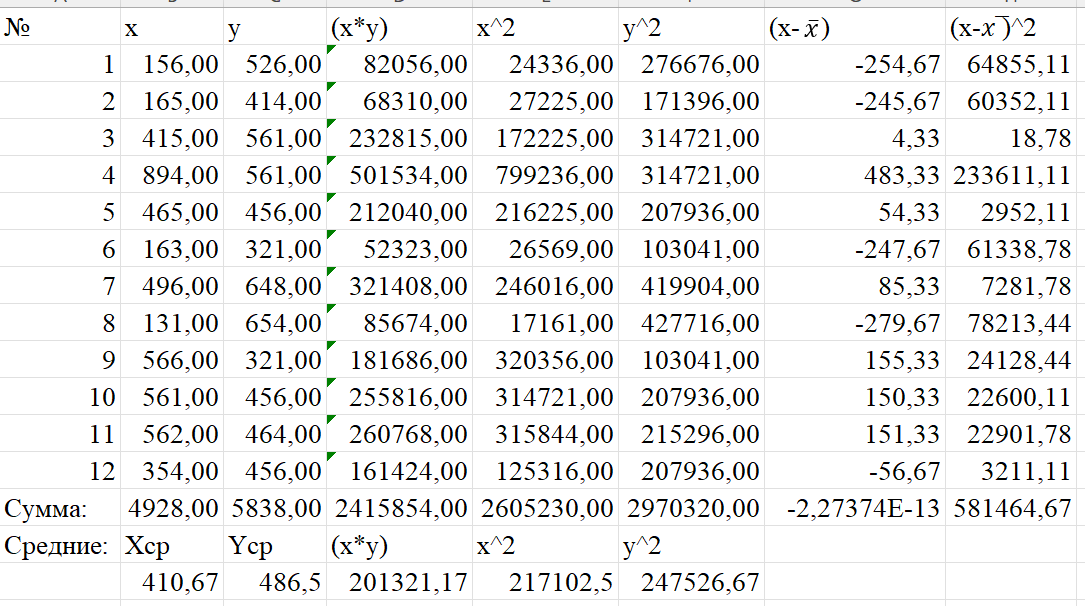
plt.show()



Задание 2:

1. Для нахождения коэффициентов уравнения регрессии выполнить все табличные расчеты.
2. Составить уравнение регрессии.
3. Выяснить, являются ли полученные оценки коэффициентов уравнения регрессии эффективными и состоятельными.
4. Построить на одном графике корреляционное поле, чтобы сделать предварительный вывод является ли регрессия линейная, а так же построить линию тренда.
5. Получить выводы о проделанной работе.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| n | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| X | 156 | 165 | 415 | 894 | 465 | 163 | 496 | 131 | 566 | 561 | 562 | 354 |
| Y | 526 | 414 | 561 | 561 | 456 | 321 | 648 | 654 | 321 | 456 | 464 | 456 |



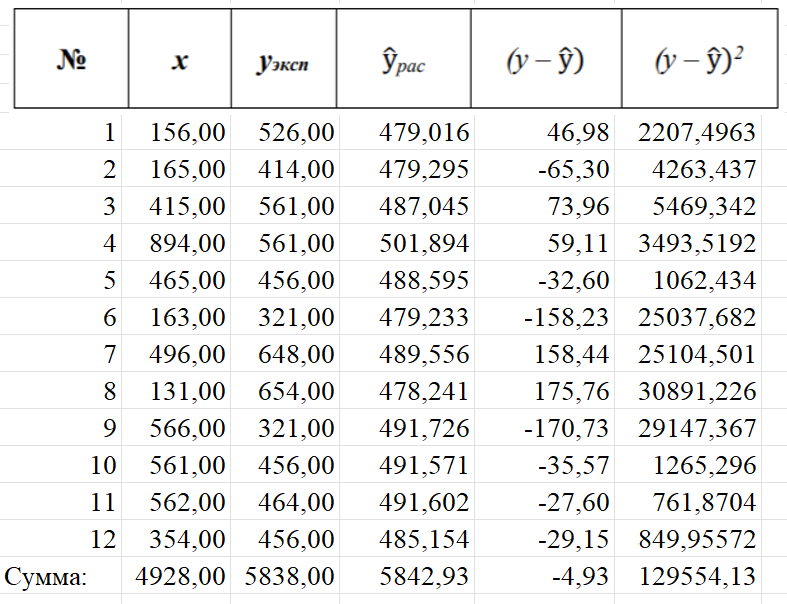
= 217102,5 – (410,67) ^ 2 = 48452,65

= 247526,67 – (486,5) ^ 2 = 10844,42

b = (201321,17 – 410,67 \* 486,5) / (48452,65) = 0,031

a = 486,5 – 0,03 \* 410,67 = 474,18

= 474,18 + 0,031x



Интервал, в котором находится рассчитанный коэффициент определяется по шкале Чеддока. Коэффициент находится в интервале от 0,1 до 0,3, связь слабая, прямая.

Найдем t – критерий Стьюдента для обоих параметров

*0,000634*

*import numpy as np*

*import matplotlib.pyplot as plt*

*from scipy import stats*

*n = np.array([1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12])*

*X = np.array([156, 165, 415, 894, 465, 163, 496, 131, 566, 561, 562, 354])*

*Y = np.array([526, 414, 561, 561, 456, 321, 648, 654, 321, 456, 464, 456])*

*mean\_X = np.mean(X)*

*mean\_Y = np.mean(Y)*

*numerator = np.sum((X - mean\_X) \* (Y - mean\_Y))*

*denominator = np.sum((X - mean\_X) \*\* 2)*

*b1 = numerator / denominator*

*b0 = mean\_Y - b1 \* mean\_X*

*print("Коэффициенты уравнения регрессии:")*

*print("b0 =", b0)*

*print("b1 =", b1)*

*plt.scatter(X, Y, label='Данные')*

*plt.plot(X, b0 + b1\*X, color='red', label='Линия тренда')*

*plt.xlabel('X')*

*plt.ylabel('Y')*

*plt.title('Корреляционное поле и линия тренда')*

*plt.legend()*

*plt.show()*

*n = len(X)*

*SSR = np.sum((b0 + b1 \* X - Y) \*\* 2)*

*SE\_b0 = np.sqrt(SSR / (n - 2) / np.sum((X - mean\_X) \*\* 2))*

*SE\_b1 = np.sqrt(SSR / (n - 2) / np.sum((X - mean\_X) \*\* 2))*

*t\_b0 = b0 / SE\_b0*

*t\_b1 = b1 / SE\_b1*

*alpha = 0.05*

*t\_critical = stats.t.ppf(1 - alpha / 2, n - 2)*

*print("t-статистика для b0:", t\_b0)*

*print("t-статистика для b1:", t\_b1)*

*print("Критическое значение t-статистики для alpha =", alpha, "и степеней свободы", n - 2, ":", t\_critical)*

*if abs(t\_b0) > t\_critical:*

*print("b0 является значимым коэффициентом")*

*else:*

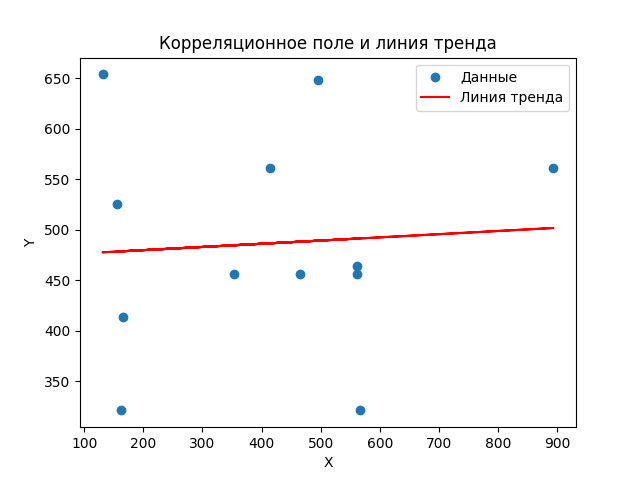
*print("b0 не является значимым коэффициентом")*

*if abs(t\_b1) > t\_critical:*

*print("b1 является значимым коэффициентом")*

*else:*

*print("b1 не является значимым коэффициентом")*

**

Контрольные вопросы:

1. Цели изучения зависимости между переменными могут быть разнообразными, включая понимание взаимосвязей между явлениями, прогнозирование будущих значений, выявление факторов, влияющих на исследуемый процесс, оптимизацию решений и многое другое.
2. Виды связей между переменными:

* Прямая (положительная) связь: значения одной переменной возрастают при увеличении значений другой переменной.
* Обратная (отрицательная) связь: значения одной переменной уменьшаются при увеличении значений другой переменной.
* Нет связи: изменения в одной переменной не влияют на значения другой переменной.

1. Функциональная зависимость означает, что значения одной переменной (зависимой) определяются значениями другой переменной (независимой) согласно определенной функции или закону.
2. Корреляционная связь указывает на степень взаимосвязи между переменными, но не обязательно на причинно-следственные отношения.
3. Примеры корреляционной связи:

* Положительная корреляция между уровнем образования и доходом.
* Обратная корреляция между количеством затраченного времени на подготовку к экзамену и результатом экзамена.

1. Коэффициент корреляции Пирсона измеряет степень линейной взаимосвязи между двумя переменными. Он принимает значения от -1 до 1, где 1 указывает на положительную линейную корреляцию, -1 - на отрицательную, а 0 - на отсутствие корреляции.
2. Примеры графиков зависимостей:

* График с положительным коэффициентом корреляции будет иметь наклон вверх отлево направо.
* График с отрицательным коэффициентом корреляции будет иметь наклон вниз отлево направо.
* График без корреляции будет представлять собой случайное распределение точек без определенной направленности.

1. Принцип регрессионного анализа заключается в изучении взаимосвязи между зависимыми и независимыми переменными и создании математических моделей для прогнозирования значений зависимой переменной на основе значений независимых переменных.
2. Уравнение регрессии представляет собой математическую модель, описывающую функциональную зависимость между зависимой и независимыми переменными. Линия регрессии на графике представляет собой линию, которая наилучшим образом соответствует точкам данных и используется для предсказания значений зависимой переменной на основе независимых переменных.
3. Уровень значимости корреляции (обычно обозначается как alpha) определяет вероятность ошибки первого рода при отклонении нулевой гипотезы о отсутствии корреляции между переменными. Обычно используют значения alpha равные 0.05 или 0.01.