Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Московский государственный технический университет имени Н.Э. Баумана»
(МГТУ им. Н.Э.Баумана)
Мытищинский филиал

ФАКУЛЬТЕТ КОСМИЧЕСКИЙ

КАФЕДРА К-1 САУ

ЛАБОРАТОРНАЯ РАБОТА №8

ПО ДИСЦИПЛИНЕ
“МАТЕМАТИЧЕСКАЯ СТАТИСТИКА”

НА ТЕМУ:

«КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ»

Студент К1-41Б (Группа) **29.04.23** (Подпись, дата) **К. А. Тимофеев** (ФИО)

Руководитель (Подпись, дата) **О. М. Полещук** (ФИО)

2023 г.

Цель:

При помощи регрессионного анализа найти коэффициенты линейного уравнения регрессии $Y = a \ \ X + b$.

Решение:

xi	yi
-4,132	-4,539
-3,204	-3,306
-2,142	-2,018
-1,285	-1,223
-0,506	-0,739
1,013	-0,823
1,964	-1,873
3,167	1,055
4,658	0,989
5,243	0,116
6,296	6,213
7,259	0,856
8,275	0,743
9,202	5,137
9,687	2,209

Решим задачу, написав программу для построения графика линейной регрессии на языке Python с использованием библиотек numpy, pandas, matplotlib, а также импортируем модуль Linear Regression для регрессионного анализа из библиотеки sklearn.linear_model.

Имеется набор данных из исходных значений (x_i и y_i). Построим модель на основе полученных данных и обучим её:

```
model = LinearRegression()
X = pd.DataFrame(data)
Y = pd.DataFrame(pred_data)
model.fit(X,Y)
```

Далее функциями intercept_ и coef_ найдем значения параметров a и b:

```
a = model.intercept_
print("A = ",float(a))

b = model.coef_
print("B = ", float(b))
```

Сразу же можем получить информацию о том, насколько достоверна наша модель. Переменная R_sqr хранит в себе значение R^2 - коэффициент детерминации, равный $R^2 = 1 - \frac{\sum_{i=1}^{n} (\tilde{y}_i - y_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$, где $\tilde{y}$ - это модельные значения величины y, а $\bar{y}$ - среднее выборочное по исходным данным. При этом известно, что чем ближе коэффициент детерминации к 1, тем лучше построена модель и тем больше доля общей дисперсии объясняется уравнением.

```
predict_Y = model.predict(X)
```

Модельные значения в свою очередь можно получить с помощью функции predict(), которая принимает значение исходных данных X.

Проверим гипотезу $H_0 : b = b_0 = 0$, конкурентом которой выступает $H_1 : b \quad b_0$.
Тогда расчёт проверки основной гипотезы приведён ниже:

$n := \mathrm{cols}(X) = 15$

$$average(A) := \frac{\sum_{i=1}^{\mathrm{cols}(A)} A_i}{\mathrm{cols}(A)} \quad \sigma(A) := \sqrt{\frac{\sum_{i=1}^{\mathrm{cols}(A)} \left(A_i - average(A)\right)^2}{\mathrm{cols}(A)}} \qquad b0 := 0$$

Оценки неизвестных параметров модели

$$b`` := \frac{n \cdot \sum_{i=1}^{n} X_i \cdot Y_i - \left(\sum_{i=1}^{n} X_i\right) \cdot \left(\sum_{i=1}^{n} Y_i\right)}{n \cdot \sum_{i=1}^{n} {X_i}^2 - \left(\sum_{i=1}^{n} X_i\right)^2} = 0,5056$$

$$a`` := \frac{1}{n} \cdot \sum_{i=1}^{n} Y_i - \frac{b``}{n} \cdot \sum_{i=1}^{n} X_i = -1,3472$$

$$E(A; B) := \left| \begin{array}{l} \text{for } i \in [1..n] \\ \quad M_i := B_i - a`` - b`` \cdot A_i \\ M^T \end{array} \right.$$

$$e\_ := E(X; Y) = [\,-1,1025 \;\; -0,3388 \;\; 0,4122 \;\; 0,7739 \;\; 0,864 \;\; 0,0119 \;\; -1,5189 \;\cdots]$$

$$S\_sqr := \frac{1}{n-2} \cdot \sum_{i=1}^{n} {e\_}_i^{\,2} = 2,9272 \qquad \sigma\_b := \sqrt{\frac{(\sigma(Y))^2}{\sum_{i=1}^{n}\left(\left(X_i\right)^2\right) - n \cdot \left(average(X)^2\right)}} = 0,1598$$

$$b`` = 0,5056 \qquad S\_b := \frac{\sigma\_b \cdot \sqrt{S\_sqr}}{\sigma(Y)} = 0,0989 \qquad \begin{array}{ll} k := n - 2 = 13 & a := 0,05 \\ t_{k,a} := 2,1604 & \end{array}$$

$$b`` - t_{k,a} \cdot S\_b \le b0 = 0 \qquad b`` + t_{k,a} \cdot S\_b = 0,7193$$

$$b`` + t_{k,a} \cdot S\_b \ge b0 = 1 \qquad b`` - t_{k,a} \cdot S\_b = 0,292 \qquad t_{k,a} \cdot S\_b = 0,2137$$

Т.е. $b0$ не принадлежит отрезку $\left[\left(b`` - t_{k,a} \cdot S\_b\right)..\left(b`` + t_{k,a} \cdot S\_b\right)\right]$, а значит,
принимается гипотеза H1: $b \ne \left(b_0 := 0\right)$

Взяв для построения интервала статистику $\dfrac{Y - (a + b \;\; X)}{S_b} \quad t_{n-2}$, получим что,

доверительный интервал для Y выглядит как:

$$a + b \;\; X - t_{k,\alpha} \;\; S_b \quad Y \quad a + b \;\; X + t_{k,\alpha} \;\; S_b$$

Далее находим параметры для оценки регрессии:

```
#-------------------------Параметры регрессии ------------------------------
print("\***----------ПАРАМЕТРЫ РЕГРЕССИИ----------***/")
n_regression = 1 #степени свободы регрессии
SS_regression = sum(np.square(np.subtract(np.array(Y), average_Y))) #сумма квадратов отклонений
MS_regression = SS_regression/n_regression #средний квадрат отклонений
F = sum(np.square(np.subtract(np.array(predict_Y), average_Y)))/sum(np.square(np.subtract(np.array(Y), np.array(predict_Y))))
print("\*** Сумма квадратов отклонений SS =", float(SS_regression))
print("\*** Средний квадрат отклонений MS =", float(MS_regression))
print("\*** Отношение факторной к остаточной дисперсий F =", float(F))
print("\*** -------------------------------------- ***/")
#--------------------------------------------------------------------------
```

Здесь ключевой является переменная F, которая также показывает успешность модели, а именно, чем больше F, тем больше факторная дисперсия, а соответственно лучше построена модель, остаточная дисперсия же указывает на ошибки модели и в идеале должна стремиться к нулю.

	df	SS	MS	F	Значимость F
Регрессия	1	114,5694	114,5694	2,011	2,1604
Остаток	13	38,053	2,93		
Итого	14	115,0909			

здесь

df - степени свободы статистики

SS - сумма квадратов отклонений

MS - средний квадрат отклонений

Значимость F - квантиль t-критерия Стьюдента $t_{n-2,a=0,05} = 2{,}1604$, находят из таблицы значений критерия Стьюдента.

Выведем необходимые значения и график регрессионной модели:

```
PS C:\Users\Админ> & C:/Users/Админ/AppData/Local/Programs/Python/Pyth
 Среднее выборочное X = [3.033]
 Среднее выборочное Y = [0.18646667]
\*** Получим коэффициенты регрессии ***/
A =  -1.3471529512675944
B =  0.5056444503574881
Коэфициент детерминации R^2 =  0.6678604306903887
\*** Тогда уравнение линейной регрессии примет вид:
--------------------------------------------------
 Y[i] =  -1.3471529512675944 * X[i] + 0.5056444503574881 + e[i]
 где e[i] - ошибки регрессии
--------------------------------------------------
\***----------ПАРАМЕТРЫ РЕГРЕССИИ----------***/
\*** Сумма квадратов отклонений SS = 114.56944773333333
\*** Средний квадрат отклонений MS = 114.56944773333333
\*** Отношение факторной к остаточной дисперсий F = 2.01078249146469
\*** ---------------------------------------- ***/
\*** Ошибки регрессии ***/
\*** ----------------------- ***/
[[-1.10252418]
 [-0.33876223]
 [ 0.41224336]
 [ 0.77390607]
 [ 0.86400904]
 [ 0.01193512]
 [-1.51893275]
 [ 0.80077698]
 [-0.0191389 ]
 [-1.1879409 ]
 [ 4.37661549]
 [-1.46732011]
 [-2.09405488]
 [ 1.83121272]
 [-1.34202484]]
\*** ----------------------- ***/
```

```
\*** ------------------------ ***/
\***----------ПАРАМЕТРЫ ОСТАТКОВ----------***/
\*** Сумма квадратов отклонений для остатков SS = 38.053047026189354
\*** Средний квадрат отклонений для остатков MS = 2.9271574635530273
\*** ---------------------------------------- ***/
\***----------        ИТОГО        ----------***/
\*** Сумма квадратов отклонений для графы итога SS = 115.09099500000013
\*** Средний квадрат отклонений для остатков MS = 2.718074787584954
\*** ---------------------------------------- ***/
PS C:\Users\Админ>
```

Получили что:

Коэффициент детерминации $R^2 = 0.67$, что говорит о неплохом уровне модели, однако это не предел ожиданий.

Линейное уравнение регрессии получилось в виде $Y = 0.51 - 1.34 \ X$ с коэффициентами $a = -1.34$ и $b = 0.51$.

Зная коэффициенты, можем построить доверительный интервал для Y:

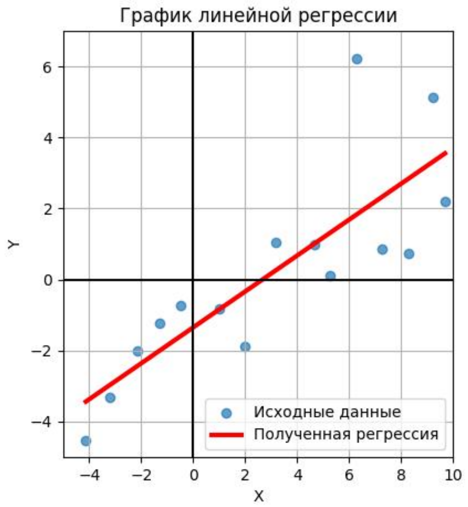
$$a + b \ X - t_{k,\alpha} \ S_b \quad Y \quad a + b \ X + t_{k,\alpha} \ S_b$$

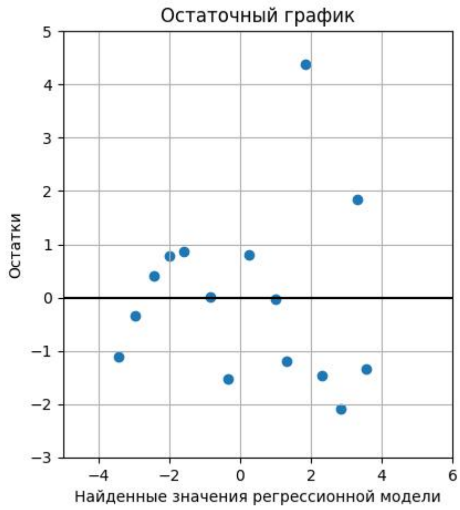
$$-1{,}34 + 0.51 \ X - 0{,}2137 \quad Y \quad -1{,}34 + 0.51 \ X + 0{,}2137$$

$$-1{,}5537 + 0{,}51 \ X \quad Y \quad 1{,}1263 + 0.51 \ X$$

Также нашли отношение факторной к остаточной дисперсий $F = 2.011$. Однако это значение значительно меньше ожидаемого, что также не говорит о хорошей работе регрессионной модели.

Наконец-то рассмотрим полученные графики.





На графике линейной регрессии наблюдаем одновременно точки, координатами которых являются исходные x_i и y_i, и прямую образованную модельными значениями, найденными в ходе работы. Именно красную прямую описывает найденное уравнение линейной регрессии, иначе говоря, именно эта прямая из всех других возможных по мнению нашей модели имеет наименьшую ошибку или остаток для относительно всех точек, облако которых мы построили с помощью исходной информации.

На графике остатков (остаточный график) мы можем отчетливее пронаблюдать отклонение от прямой $Y = 0.51 - 1.34\ X$. Сумма этих отклонений стремится к нулю.

Вывод:

Произвели рассчет с помощью метода регрессионного анализа, нашли коэффициенты линейного уравнения регрессии, и построили его график. Поскольку коэффициент детерминации оказался мал, можем сделать вывод о том, что исходные данные не ложатся на линейную регрессию, а значит имеет смысл использовать нелинейный регрессионный анализ.