МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра математического обеспечения и применения ЭВМ

ОТЧЕТ

по лабораторной работе №5

по дисциплине «Статистические методы обработки экспериментальных данных»

Тема: Элементы регрессионного анализа. Выборочные прямые среднеквадратической регрессии. Корреляционное отношение.

Студентка гр. 7381 _____ Алясова А.Н.

Студент гр. 7381 _____ Кортев Ю.В.

Преподаватель _____ Середа А.-В.И.

Санкт-Петербург

2021

Цель работы.

Ознакомление с основными положениями метода наименьших квадратов (МНК), со статистическими свойствами МНК оценок, с понятием функции регрессии и роли МНК в регрессионном анализе, с корреляционным отношением, как мерой тесноты произвольной (в том числе и линейной) корреляционной связи.

Основные теоретические положения.

Метод наименьших квадратов (МНК) — метод, основанный на поиске минимума суммы квадратов отклонений значений некоторых функций от заданного множества значений. МНК является одним из основных методов регрессионного анализа и применяется для оценки параметров регрессионных моделей на основе выборочных данных.

$$M(X/y) = q_1(y),$$
$$M(Y/x) = q_2(x).$$

Регрессионный анализ – это статистический метод исследования влияния одной или нескольких независимых переменных $X_1, X_2, \dots, X_p$ на зависимую переменную Y.

Линейные функции выборочной среднеквадратической регрессии:

$$y = \overline{y_B} + \overline{r_{xy}}\frac{S_y}{S_x}(x - \overline{x_B}),$$

$$x = \overline{x_B} + \overline{r_{xy}}\frac{S_x}{S_y}(y - \overline{y_B}),$$

г

д

е

yB, xB статистические оценки математических ожиданий выборок X, Y

в

с

р

а

с

Для оценки корреляционной зависимости между случайными величинами в общем, а не только линейной, может быть использовано так называемое корреляционной отношение.

Чтобы рассчитать выборочное корреляционное отношение нужно

Двнгр и межгрупповую *Дмежгр* дисперсии. Оценку общей дисперсии *Х* можно представить, как сумму:

$$D_{\text{общ } yx} = D_{\text{внгр } yx} + D_{\text{межгр } yx}$$

Чтобы рассчитать выборочное корреляционное отношение Y к X нужно рассчитать внутригрупповую и межгрупповую дисперсии.

Внутригрупповая дисперсия вычисляется по формуле:

$$D_{\text{внгр } yx} = \frac{1}{n}\sum_{i=1}^{k_2} n_{x_i} * D_{y \text{ гр } i},$$

где n – объём выборки, k_2 – количество интервалов интервального ряда X, n_{x_i} – абсолютная частота для i-ого интервала интервального ряда X, $D_{y \text{ групп } i}$ – групповая дисперсия элементов выборки Y на i-ом интервале интервального ряда X.

Межгрупповая дисперсия вычисляется по формуле:

$$D_{\text{межгр } yx} = \frac{1}{n}\sum_{i=1}^{k_2} n_{x_i} * (\overline{y_{\text{гр } i}} - \overline{y_B})^2$$

где n – объём выборки, k_2 – количество интервалов интервального ряда X, n_{x_i} – абсолютная частота для i-ого интервала интервального ряда X, $\overline{y_{\text{гр } i}}$ i – групповое математическое ожидание элементов выборки Y на i-ом интервале интервального ряда X, $\overline{y_B}$ – статистическая оценка математического ожидания Y.

В

ы

б

о

р

о

ч

н

о

е

к

о

$$\overline{\eta_{yx}} = \frac{\overline{\sigma_{\overline{y_x}}}}{\overline{\sigma_y}} = \sqrt{\frac{D_{\text{межгр}}}{D_{\text{общ}}}},$$

Где $\overline{\sigma_{\overline{y_x}}} = \sqrt{D_{\text{межгр}}}$, $\overline{\sigma_x} = \sqrt{D_{\text{общ}}}$ – выборочные значения СКВО $\overline{x_y}$ и X соответственно. Аналогично определяется выборочное корреляционное отношение X к Y.

Для расчёта выборочного корреляционного отношения X к Y необходимо рассчитать те же величины по следующим формулам (меняем местами X и Y):

$$D_{\text{внгр } xy} = \frac{1}{n}\sum_{i=1}^{k_1} n_{y_i} * D_{x \text{ гр } i}$$

$$D_{\text{межгр } xy} = \frac{1}{n}\sum_{i=1}^{k_1} n_{y_i} * \left(\overline{x_{\text{гр } \iota}} - \overline{x_B}\right)^2$$

$$D_{\text{общ } xy} = D_{\text{внгр } xy} + D_{\text{межгр } xy}$$

$$\overline{\eta_{xy}} = \frac{\overline{\sigma_{\overline{x_y}}}}{\overline{\sigma_x}} = \sqrt{\frac{D_{\text{межгр}}}{D_{\text{общ}}}},$$

Выборочное уравнение регрессии Y на X:

$$\overline{y_{x_\iota}} = ax^2 + bx + c$$

Значения коэффициентов a, b, c определим с помощью МНК, что приводит к необходимости решать систему линейных уравнений 3го порядка:

$$\begin{cases} a\left(\sum_{i=1}^{m} n_{x_i} x_i^4\right) + b\left(\sum_{i=1}^{m} n_{x_i} x_i^3\right) + c\left(\sum_{i=1}^{m} n_{x_i} x_i^2\right) = \sum_{i=1}^{m} n_{x_i} \overline{y_{x_\iota}} x_i^2 \\ a\left(\sum_{i=1}^{m} n_{x_i} x_i^3\right) + b\left(\sum_{i=1}^{m} n_{x_i} x_i^2\right) + c\left(\sum_{i=1}^{m} n_{x_i} x_i\right) = \sum_{i=1}^{m} n_{x_i} \overline{y_{x_\iota}} x_i \\ a\left(\sum_{i=1}^{m} n_{x_i} x_i^2\right) + b\left(\sum_{i=1}^{m} n_{x_i} x_i\right) + Nc = \sum_{i=1}^{m} n_{x_i} \overline{y_{x_\iota}} \end{cases}$$

Решив данную систему, найдём коэффициенты квадратичной функции выборочной среднеквадратической регрессии.

Постановка задачи.

Для заданной двумерной выборки (X, Y) построить уравнения выборочных прямых среднеквадратической регрессии. Полученные линейные функции регрессии отобразить графически. Найти выборочное корреляционное отношение. Полученные результаты содержательно проинтерпретировать.

Выполнение работы.

1-2) Отобразим двумерную выборку на графике и для заданной выборки построим уравнения средней квадратичной регрессии x на y и y на x соответственно. Построим полученные прямые на множестве выборки.

Линейная функция среднеквадратической регрессии $y(x)$ для заданной выборки:

$$y(x) = \overline{y_B} + \overline{r_{xy}}\,\frac{S_y}{S_x}(x - \overline{x_B});$$

$$y(x) = 0{,}3862 * x - 46{,}214.$$

Линейная функция среднеквадратической регрессии $x(y)$ для заданной выборки:

$$x(y) = \overline{x_B} + \overline{r_{xy}}\,\frac{S_x}{S_y}(y - \overline{y_B});$$

$$x(y) = 2{,}084 * y + 183{,}903.$$

Двумерная выборка и графики линейной функции выборочной среднеквадратической регрессии $y(x)$ и $x(y)$ представлены на рисунке 1.

Найдем оценки остаточной дисперсии для полученных выборочных уравнений регрессии:

$$D_{\text{ост } y} = \frac{1}{n}\sum_{i=1}^{k_1}(y_i - 0{,}3862 * x + 46{,}214) = -4.9977;$$

$$D_{\text{ост } x} = \frac{1}{n}\sum_{i=1}^{k_1}(x_i - 2{,}084 * y - 183{,}903) = 15.7109.$$

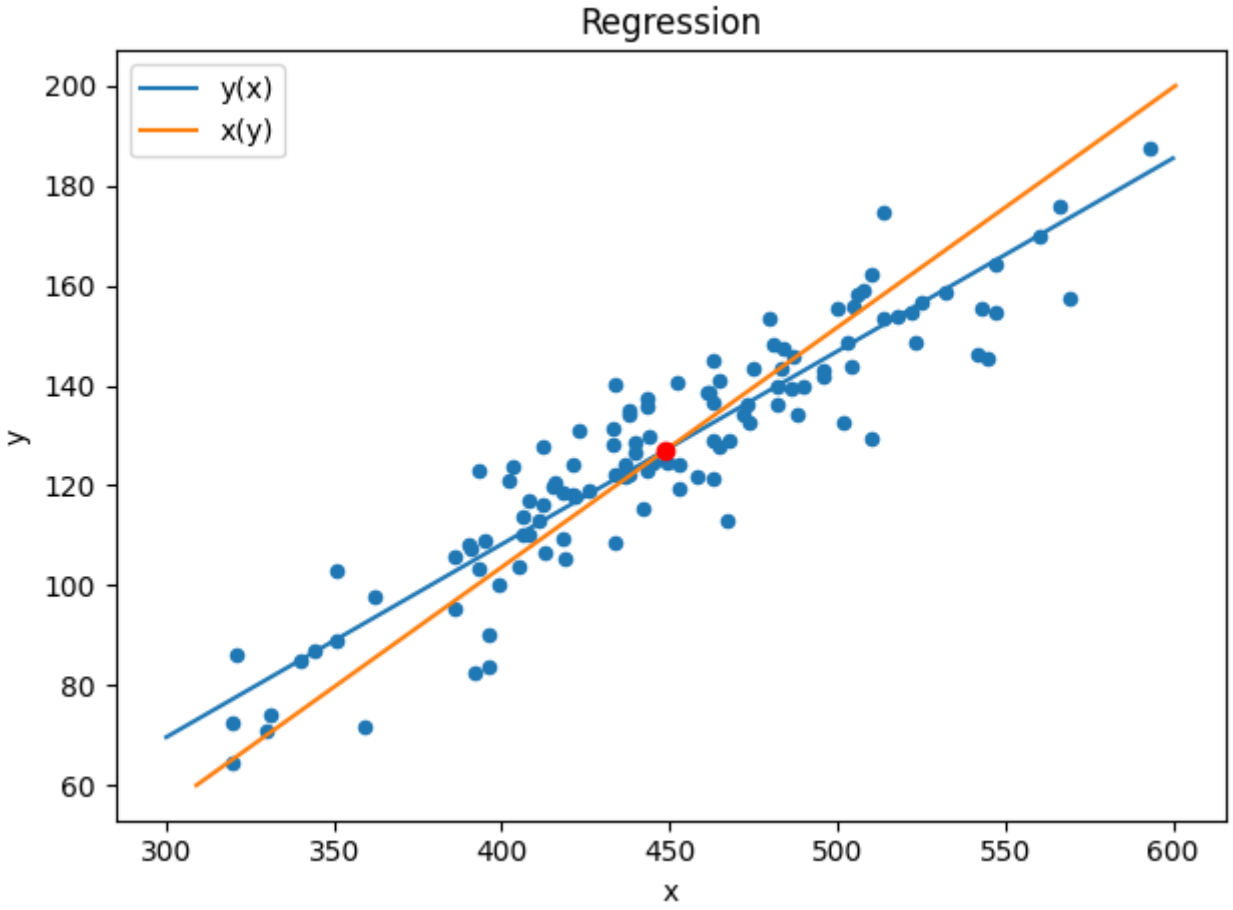

Рисунок 1 - Графики линейной функции выборочной среднеквадратической регрессии $y(x)$ и $x(y)$

3) Составим корреляционную таблицу для нахождения выборочного корреляционного отношения. Убедимся, что неравенства $\eta_{xy} \geq |r_{xy}|$ и $\eta_{yx} \geq |r_{xy}|$ выполняются.

Таблица 1 - Корреляционная таблица

y_i \ x_i	***72***	***87***	***102***	***117***	***132***	***147***	***162***	***177***	***192***	n_y	$\overline{y_{\text{гр}}}$	$D_{y\,\text{гр}\,i}$
337	0	4	4	1	0	0	0	0	0	9	82	100
371	1	0	3	0	0	0	0	0	0	4	94,5	168,750
405	0	3	9	14	1	0	0	0	0	27	109,222	122,840
439	0	0	1	10	12	2	0	0	0	25	126	108
473	0	0	0	3	12	9	0	0	0	24	135,75	98,438

507	0	0	0	0	2	8	6	1	0	17	152,294	130,796
541	0	0	0	0	0	2	5	0	0	7	157,714	45,918
575	0	0	0	0	0	0	1	2	0	3	172	50
609	0	0	0	0	0	0	0	0	1	1	192	0
$\boldsymbol{n_x}$	5	7	14	27	27	21	12	3	1	117		
$\overline{\boldsymbol{x}_{\text{гр}}}$	343,8	366,143	395,286	425,148	457,889	489,190	526,833	552,333	609			
$\boldsymbol{D_{x\ \text{гр}\ i}}$	46,24	23,592	94,367	191,874	228,346	317,179	200,694	513,778	0			

Для того, чтобы посчитать выборочное корреляционное отношение, рассчитаем внутригрупповую, межгрупповую, общую дисперсии.

Расчёт выборочного корреляционного отношения X к Y:

$$D_{\text{внгр}\ xy} = \frac{1}{n}\sum_{i=1}^{k_1} n_{y_i} * D_{x\ \text{гр}\ i} = 185{,}2464$$

$$D_{\text{межгр}\ xy} = \frac{1}{n}\sum_{i=1}^{k_1} n_{y_i} * \left(\overline{x_{\text{гр}\ i}} - \overline{x_B}\right)^2 = 2976{,}4729$$

$$D_{\text{общ}\ xy} = D_{\text{внгр}\ xy} + D_{\text{межгр}\ xy} = 3161{,}7193$$

$$\eta_{xy} = \sqrt{\frac{D_{\text{межгр}\ xy}}{D_{\text{общ}\ xy}}} = 0{,}9702$$

Расчёт выборочного корреляционного отношения Y к X:

$$D_{\text{внгр}\ yx} = \frac{1}{n}\sum_{i=1}^{k_2} n_{x_i} * D_{y\ \text{гр}\ i} = 106{,}1756$$

$$D_{\text{межгр}\ yx} = \frac{1}{n}\sum_{i=1}^{k_2} n_{x_i} * (\overline{y_{\text{гр}\ i}} - \overline{y_B})^2 = 503{,}7590$$

$$D_{\text{общ}\ yx} = D_{\text{внгр}\ yx} + D_{\text{межгр}\ yx} = 609{,}9346$$

$$\eta_{yx} = \sqrt{\frac{D_{\text{межгр}\ yx}}{D_{\text{общ}\ yx}}} = 0{,}9088$$

Проверим выполнение неравенств $\eta_{xy} \geq |r_{xy}|$ и $\eta_{yx} \geq |r_{xy}|$:

$$0{,}9702 = \eta_{xy} \geq r_{xy} = 0.8972$$

$$0{,}9088 = \eta_{yx} \geq |r_{xy}| = 0.8972$$

Неравенства $\eta_{xy} \geq |r_{xy}|$ и $\eta_{yx} \geq |r_{xy}|$ выполняются.

4.1) Для заданной выборки построим корреляционную кривую параболического вида $y = \beta_2 x^2 + \beta_1 x^2 + \beta_0$.

Для определения коэффициентов корреляционной кривой параболического вида $y = ax^2 + bx + c$ была решена следующая система уравнений:

$$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i^2 \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i x_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + cn = \sum_{i=1}^{n} y_i \end{cases}$$

Система была решена с помощью написанной программы на языке Python (код представлен в ПРИЛОЖЕНИИ А). В результате работы программы были получены следующие значения коэффициентов:

$$a = -0{,}000428;$$

$$b = 0{,}763843;$$

$$c = -128{,}094862.$$

Полученное уравнение примет вид:

$$y = -0{,}000428 * x^2 + 0{,}763843 * x - 128{,}094862$$

График квадратичной функции выборочной среднеквадратической регрессии $y(x)$ представлен на рис. 2.

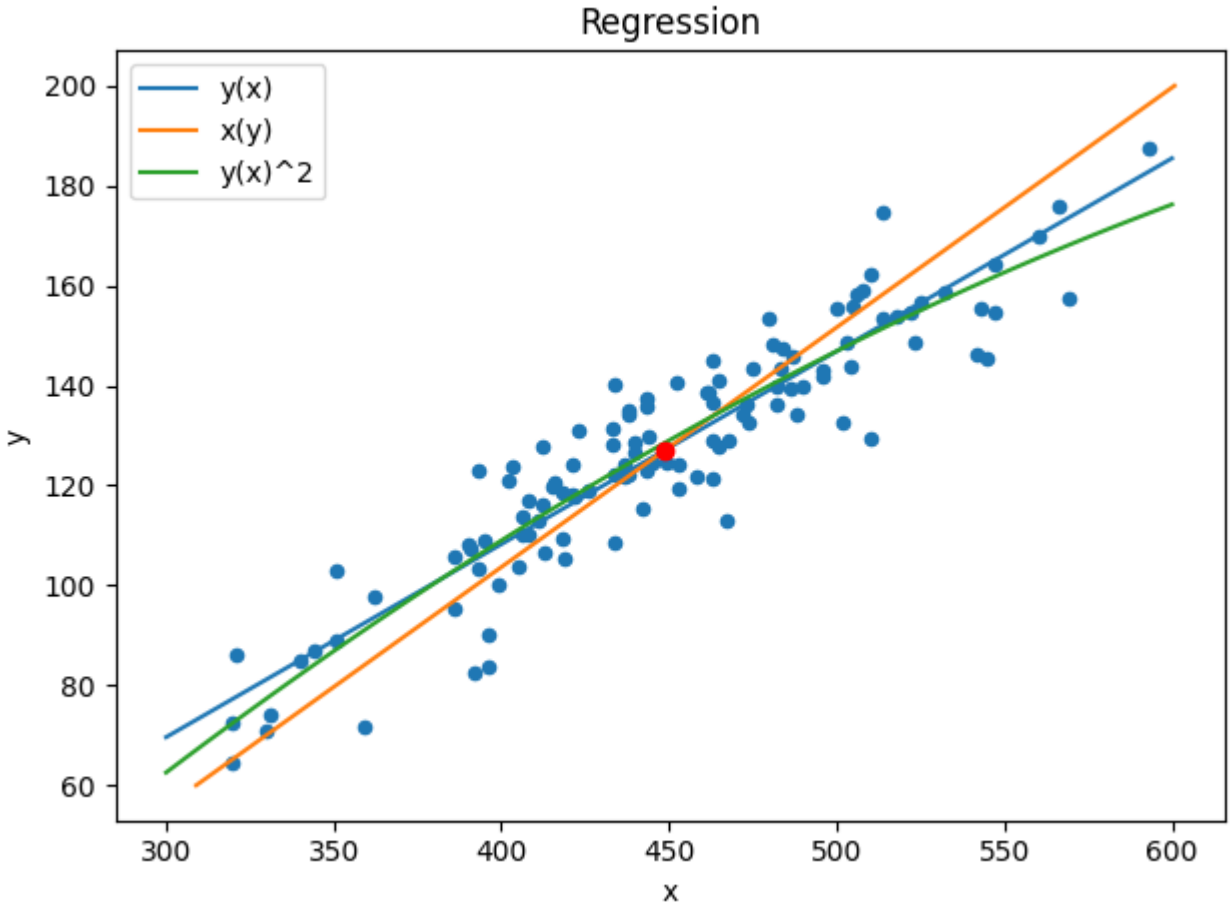

Рисунок 2 - Корреляционная кривая параболического вида

4.2) Построим корреляционную кривую экспоненциальной функции $y = \beta_0 exp(\beta_1 x)$.

Запишем выборочное уравнение в виде $y = a * exp(bx)$.

Найдем коэффициенты a и b с помощью написанной программы на языке Python (код представлен в ПРИЛОЖЕНИИ А), получим следующие коэффициенты:

$$a = 29{,}7009;$$

$$b = 0{,}0032.$$

Корреляционная кривая экспоненциального вида имеет следующий вид:

$$y = 29{,}7009 * exp(0{,}0032 * x).$$

График полученной кривой на множестве выборки представлен на рис. 3.

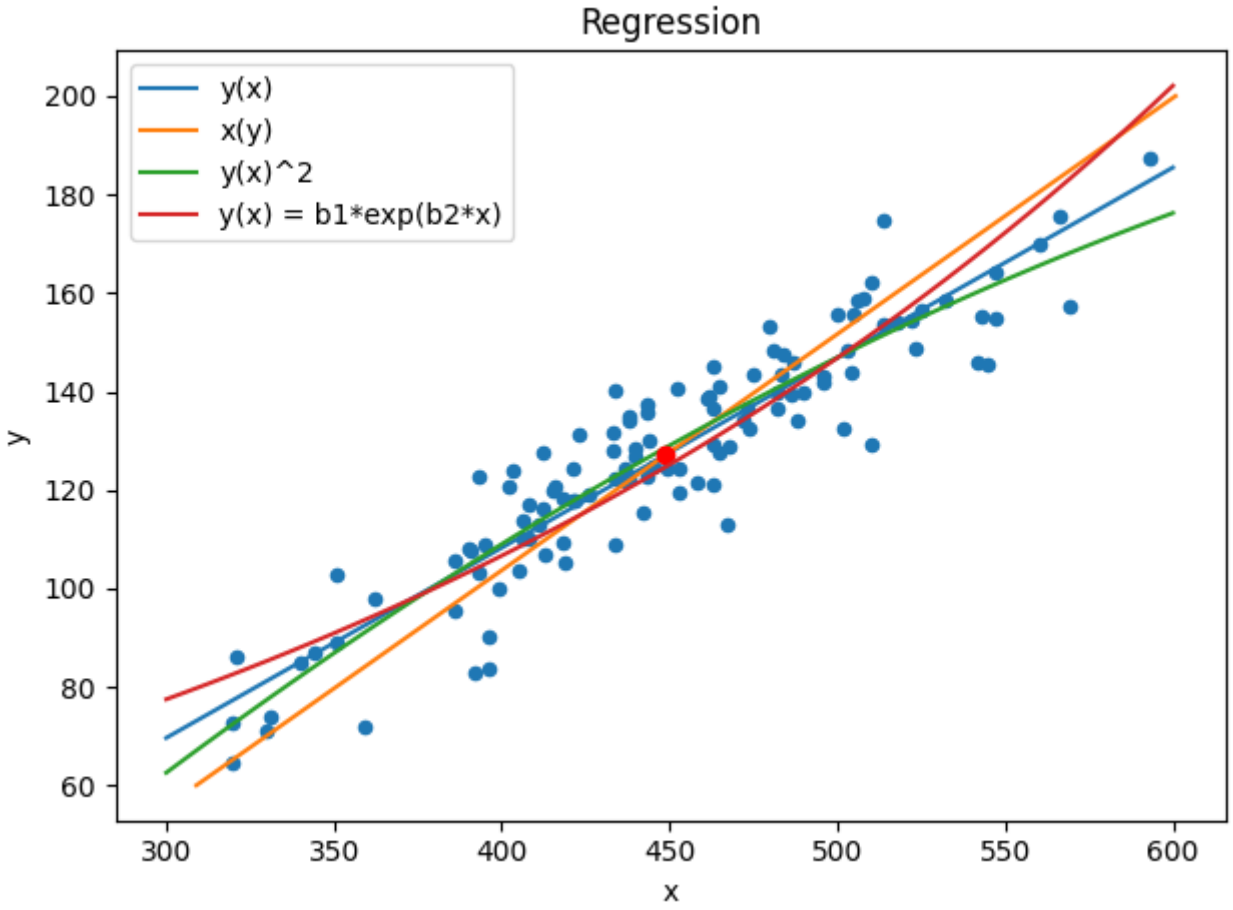
Regression
y(x)
x(y)
y(x)^2
y(x) = b1*exp(b2*x)
200
180
160
140
120
100
80
60
y
300
350
400
450
500
550
600
x

Рисунок 3

Выводы.

Таким образом, были получены уравнения прямых среднеквадратической регрессии.

Были найдены корреляционные соотношения $\eta_{xy} = 0{,}9702$ и $\eta_{yx} = 0{,}9088$. Эти значения близки к 1, что говорит о том, что между X и Y есть сильная зависимость.

Было найдено и построено уравнение выборочных кривых для параболической среднеквадратической регрессии $y = -0{,}000428 * x^2 + 0{,}763843 * x - 128{,}094862$. Была построена корреляционной кривая экспоненциального вида $y = 29{,}7009 * exp(0{,}0032 * x)$.

и

x= 0,3862∗x−46,214 и *xy= 2,084∗y+183,903*.

ПРИЛОЖЕНИЯ А

КОД ПРОГРАММЫ

```
from lab4 import df, cor, s_x, s_y, mean_x, mean_y, means_x, means_y, k,
cor_table, nums_x, nums_y
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np


def mean_sq_regression(s1, s2, mean1, mean2, cor, pref='y(x)='):
    def inner_foo(x, pr=False):
        if pr:
            print(pref + '{} * x'.format(cor * s1 / s2), '{0:+}'.for-
mat(mean1 - cor * s1 / s2 * mean2))
        a1 = cor * s1 / s2
        a2 = mean1 - cor * s1 / s2 * mean2
        return a2 + a1 * x

    return inner_foo


msr_x = mean_sq_regression(s_x, s_y, mean_x, mean_y, cor, 'x(y)=')
msr_y = mean_sq_regression(s_y, s_x, mean_y, mean_x, cor)

ax = df.plot.scatter(x=0, y=1)
y1 = np.array([60, 200])
x2 = np.array([300, 600])

ax.plot(x2, msr_y(x2, pr=True), label='y(x)')
ax.plot(msr_x(y1, pr=True), y1, label='x(y)')

line1 = np.array([[x2[0], msr_y(x2)[0]], [x2[1], msr_y(x2)[1]]])
line2 = np.array([[msr_x(y1)[0], y1[0]], [msr_x(y1)[1], y1[1]]])

t, s = np.linalg.solve(np.array([line1[1] - line1[0], line2[0] -
line2[1]]).T, line2[0] - line1[0])

ax.plot(*((1 - t) * line1[0] + t * line1[1]), 'o', color='red')
ax.set_title('Regression')
ax.set_xlabel('x')
ax.set_ylabel('y')
plt.legend()
plt.show()
```

```
print('Остаточная дисперсия y: ', (np.array(means_y) - msr_y(np.ar-
ray(means_x))).sum() / k)
print('Остаточная дисперсия x: ', (np.array(means_x) - msr_x(np.ar-
ray(means_y))).sum() / k)

cors = cor_table.to_numpy()
rows_y = []
vars_y = []
for row in range(len(cor_table)):
    cols = []
    var = []
    for col in range(len(cor_table.columns)):
        cols.append(cors[row][col] * means_y[col])
    rows_y.append(sum(cols) / nums_x[row])
    for col in range(len(cor_table.columns)):
        var.append(((means_y[col] - rows_y[-1]) ** 2) * cors[row][col])
    vars_y.append(sum(var) / nums_x[row])

cols_x = []
vars_x = []
for col in range(len(cor_table.columns)):
    rows = []
    var = []
    for row in range(len(cor_table)):
        rows.append(cors[row][col] * means_x[row])
    cols_x.append(sum(rows) / nums_y[col])
    for row in range(len(cor_table)):
        var.append(((means_x[col] - cols_x[-1]) ** 2) * cors[row][col])
    vars_x.append(sum(var) / nums_y[col])

cor_table['n_y'] = nums_x
cor_table['mean_y_gr'] = rows_y
cor_table['D_y_gr'] = vars_y
cor_table.to_csv('Table1.csv')
pd.DataFrame({'n_x': nums_y, 'x_mean_gr': cols_x, 'D_x_gr':
vars_x}).T.to_csv('Table1_last_rows.csv')

x_t = pd.DataFrame({'n_x': nums_y, 'x_mean_gr': cols_x, 'D_x_gr':
vars_x})

D_ingr_yx = (cor_table['D_y_gr'].to_numpy() *
x_t['n_x'].to_numpy()).sum() / len(df)
```

```
D_ingr_xy = (x_t['D_x_gr'].to_numpy() * cor_ta-
ble['n_y'].to_numpy()).sum() / len(df)
print('D внутригр xy = {}'.format(D_ingr_xy))
print('D внутригр yx = {}'.format(D_ingr_yx))

D_betwgr_yx = (((cor_table['mean_y_gr'] - mean_y) ** 2).to_numpy() *
x_t['n_x'].to_numpy()).sum() / len(df)
D_betwgr_xy = (((x_t['x_mean_gr'] - mean_x) ** 2).to_numpy() * cor_ta-
ble['n_y'].to_numpy()).sum() / len(df)

print('D межгр xy = {}'.format(D_betwgr_xy))
print('D межгр yx = {}'.format(D_betwgr_yx))

D_gen_xy = D_ingr_xy + D_betwgr_xy
D_gen_yx = D_ingr_yx + D_betwgr_yx

print('D общая xy = {}'.format(D_gen_xy))
print('D общая yx = {}'.format(D_gen_yx))

n_xy = np.sqrt(D_betwgr_xy / D_gen_xy)
n_yx = np.sqrt(D_betwgr_yx / D_gen_yx)

print('n xy = {}'.format(n_xy))
print('n yx = {}'.format(n_yx))

x = df.iloc[:, 0]
y = df.iloc[:, 1]

system = []
b = []
for i in range(3):
    line = []
    for j in range(3):
        line.append((x ** (4 - i - j)).sum())
    system.append(line)
    b.append((y * (x ** (2 - i))).sum())

res = np.linalg.solve(np.array(system), np.array(b))

print('a={}, b={}, c={}'.format(*res))


def sq_regr(x):
    return res[0] * x ** 2 + res[1] * x + res[2]
```

```
ax = df.plot.scatter(x=0, y=1)
y1 = np.array([60, 200])
x2 = np.array([300, 600])

ax.plot(x2, msr_y(x2), label='y(x)')
ax.plot(msr_x(y1), y1, label='x(y)')

x3 = np.linspace(300, 600)
ax.plot(x3, sq_regr(x3), label='y(x)^2')

ax.plot(*((1 - t) * line1[0] + t * line1[1]), 'o', color='red')
ax.set_title('Regression')
ax.set_xlabel('x')
ax.set_ylabel('y')


y = df.iloc[:, 1]
x = df.iloc[:, 0]
z = np.log(y)
a1 = (len(df) * (x*z).sum() - x.sum() * z.sum())/(len(df)*(x*x).sum()-
x.sum()**2)
a0 = z.mean() - a1 * x.mean()

b = a1
a = np.exp(a0)

ax.plot(x3, a * np.exp(b*x3), label='y(x) = b1*exp(b2*x)')
plt.legend()
plt.show()
```