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. Корреляционные отношения.

Студент гр. 8383 _____ Бабенко Н.С.
Студент гр. 8383 _____ Сахаров В.М.
Преподаватель _____ Середа А.-В.И.

Санкт-Петербург
2022

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 — 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.

$$M(X/y) = q_1(y);\ M(Y/x) = q_2(x)$$

Пусть имеется двумерная случайная величина $\{X, Y\}$, где X и Y зависимые случайные величины. Функцию $g(x)$ называют линейной функцией среднеквадратической регрессии Y на X.

$$g(x) = m(Y/x) = m(Y) + r_{xy}\frac{\sigma_y}{\sigma_x}[x - m(X)]$$

В случае, когда известны только выборочные данные – двумерная выборка значений случайных величин X и Y, возможно построение только выборочных прямых среднеквадратической регрессии. Уравнения выборочных прямых среднеквадратической регрессии:

$$\overline{y_x} = \overline{y}_{\text{в}} + \overline{r_{xy}}\frac{S_y}{S_x}(x - \overline{x}_{\text{в}});\ \overline{x_y} = \overline{x}_{\text{в}} + \overline{r_{xy}}\frac{S_x}{S_y}(y - \overline{y}_{\text{в}})$$

Оценку общей дисперсии можно представить, как сумму внутригрупповой и межгрупповой дисперсии:

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

Внутригрупповая дисперсия вычисляется, как взвешенная по объемам групп средняя арифметическая групповых дисперсий, межгрупповая – как дисперсия условных средних $\overline{x_{y_i}}$ относительно выборочной средней $\overline{x}_{\text{в}}$.

Выборочное корреляционное отношение Y к X определяется в соответствии с выражением:

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

Запишем выборочное уравнение регрессии Y на X в параболическом виде:

$$\overline{y_x} = ax^2 + bx + c$$

Значения коэффициентов a, b и c можно определить с помощью МНК, что приводит к необходимости решать систему линейных уравнений третьего порядка:

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

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

- Двумерная выборка и прямые регрессии

Для заданной двумерной выборки были построены уравнения средней квадратичной регрессии x на y и y на x. Далее полученные прямые были отображены на множестве выборки.

Выборочные прямые средней квадратичной регрессии x на y и y на x:

$$\overline{x_y} = \bar{x}_в + \overline{r_{xy}} \frac{S_x}{S_y}(y - \bar{y}_в)$$

$$\overline{y_x} = \bar{y}_в + \overline{r_{xy}} \frac{S_y}{S_x}(x - \bar{x}_в)$$

$$x(y) = 2.331 * y + 157.371$$

$$y(x) = 0.3122 * x - 13.1442$$

Полученные прямые, отображенные на множестве выборки представлены на рис. 1.

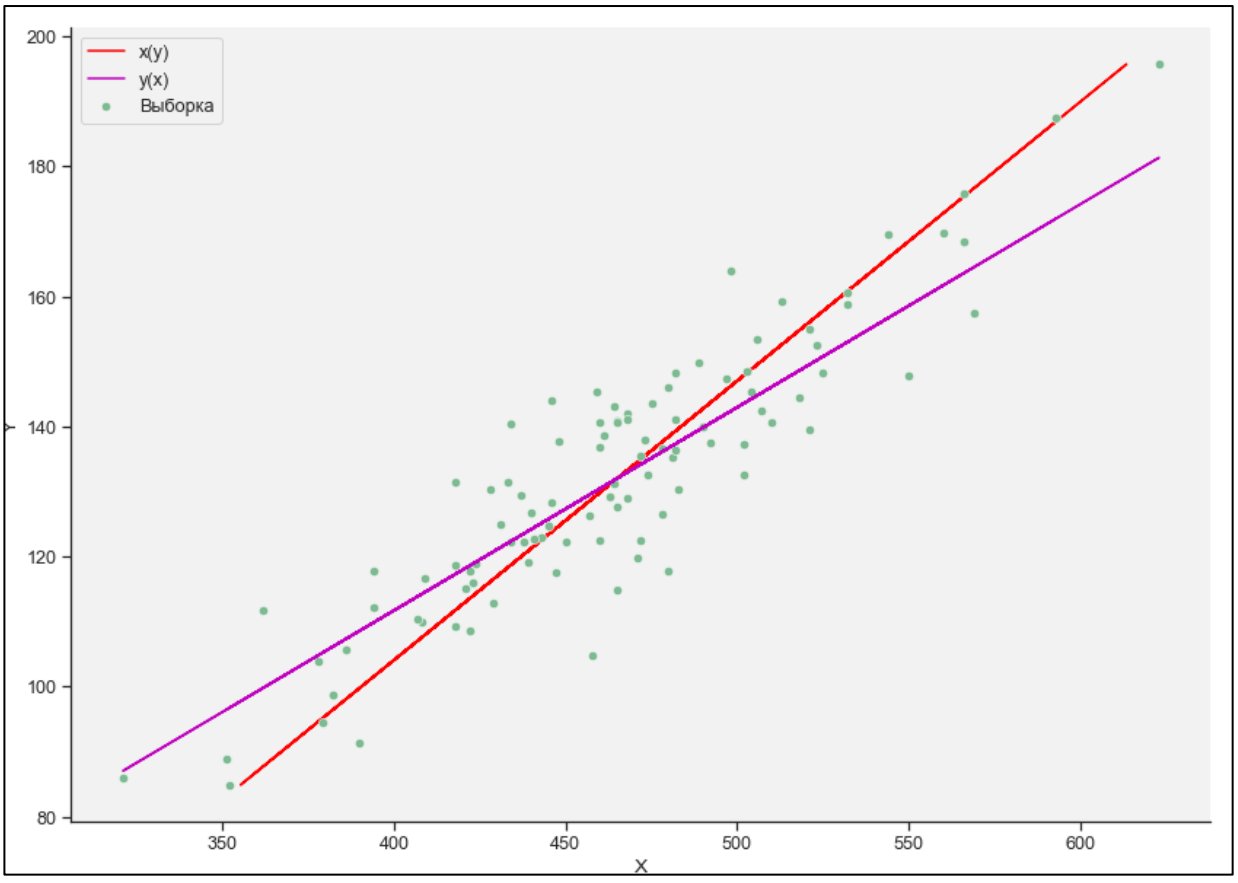

Рисунок 1 - Выборочные прямые средней квадратичной регрессии x на y и y на x

Были найдены статистические оценки остаточной дисперсии для полученных выборочных прямых средней квадратичной регрессии x на y и y на x:

$$D_{\text{ост } x} = S_x^2\left(1 - \overline{r_{xy}}^2\right) = 811.149$$

$$D_{\text{ост } y} = S_y^2\left(1 - \overline{r_{xy}}^2\right) = \ 108.63$$

- Нахождение выборочного корреляционного отношения

Была составлена корреляционная таблица для нахождения выборочного корреляционного отношения, которая представлена в таблице 1. Были посчитаны условные выборочные средние и дисперсии.

Таблица 1 - Корреляционная таблица

Y	**X**									
	343	**387**	**431**	**475**	**519**	**563**	**604**	$\mathbf{n_y}$	$\bar{\mathbf{x}}_{\text{гр}}$	$\mathbf{D}_{\mathbf{x}_{\text{гр}}}$
92.9	3	3	0	0	0	0	0	6	365	484
108.9	1	5	6	2	0	0	0	14	415.29	1270.64
124.9	0	1	18	12	1	0	0	32	448.88	704.5
140.9	0	0	3	20	9	1	0	33	485.67	821.65
156.9	0	0	0	1	7	1	0	9	519	430.22
172.9	0	0	0	0	0	4	0	4	563	0
188.3	0	0	0	0	0	0	2	2	604	0
$\mathbf{n}_x$	4	9	27	35	17	6	2	100		
$\bar{\mathbf{y}}_{\text{гр}}$	96.9	105.34	123.12	134.04	146.55	164.9	188.3			
$\mathbf{D}_{\mathbf{y}_{\text{гр}}}$	48	102.07	82.72	107.35	87.72	149.33	0			

Выборочное корреляционное отношение Y к X определяется в соответствии с выражением:

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

Аналогично для выборочного корреляционного отношения X к Y.

Для этого были рассчитаны внутригрупповая, межгрупповая и общая дисперсии. Выборочное корреляционное отношение Y к X:

$$D_{\text{внгр } yx} = \frac{1}{n}\sum_{i=1}^{k1} \mathrm{D}_{\mathrm{y}_{\text{гр } i}} n_{x_i} = 94.8854$$

$$D_{\text{межгр } yx} = \frac{1}{n}\sum_{i=1}^{k1} \left(\bar{\mathrm{y}}_{\text{гр i}} - \bar{y_{\text{в}}}\right)^2 n_{x_i} = 300.316$$

$$D_{\text{общ } yx} = D_{\text{внгр } yx} + D_{\text{межгр } yx} = 395.2014$$

$$\overline{\eta_{yx}} = \sqrt{\frac{D_{\text{межгр } yx}}{D_{\text{общ } yx}}} = 0.8717$$

Выборочное корреляционное отношение X к Y:

$$D_{\text{внгр } xy} = \frac{1}{n}\sum_{i=1}^{k2} \mathrm{D}_{\mathrm{x}_{\text{гр i}}} n_{y_i} = 742.2339$$

$$D_{\text{межгр } xy} = \frac{1}{n}\sum_{i=1}^{k2} \left(\bar{\mathrm{x}}_{\text{гр i}} - \bar{x_{\text{в}}}\right)^2 n_{y_i} = 2203.048$$

$$D_{\text{общ } xy} = D_{\text{внгр } xy} + D_{\text{межгр } xy} = 2945.2819$$

$$\overline{\eta_{xy}} = \sqrt{\frac{D_{\text{межгр } xy}}{D_{\text{общ } xy}}} = 0.8649$$

Убедимся, что неравенства $\overline{\eta_{xy}} \geq \left|\overline{r_{xy}}\right|$ и $\overline{\eta_{yx}} \geq \left|\overline{r_{xy}}\right|$ выполняются:

$$\overline{r_{xy}} = 0.853$$

$$\overline{\eta_{xy}} = 0.8649$$

$$\overline{\eta_{yx}} = 0.8717$$

Неравенство $\overline{\eta_{xy}} \geq \left|\overline{r_{xy}}\right|$ выполняется, так же, как и неравенство $\overline{\eta_{yx}} \geq \left|\overline{r_{xy}}\right|$.

- Построение корреляционных кривых
  1. Параболический вид

Для заданной выборки была построена корреляционная кривая параболического вида $y = \beta_2 x^2 + \beta_1 x^2 + \beta_0$.

Запишем выборочное уравнение регрессии Y на X в параболическом виде:

$$\overline{y_x} = ax^2 + bx + c$$

Значения коэффициентов определим с помощью МНК. Была решена следующая система уравнений:

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

Чтобы удобно рассчитать приведенные суммы была построена таблица 2.

Таблица 2 – Таблица сумм МНК

$\mathbf{x}$	$\mathbf{n_x}$	$\mathbf{\bar{y}_x}$	$\mathbf{n_x x}$	$\mathbf{n_x x^2}$	$\mathbf{n_x x^3}$	$\mathbf{n_x x^4}$	$\mathbf{n_x \bar{y}_x}$	$\mathbf{n_x \bar{y}_x x}$	$\mathbf{n_x \bar{y}_x x^2}$
343	4	96.9	1372	470596	161414428	55365148804	387.6	132946.8	45600752.4
387	9	105.34	3483	1347921	521645427	201876780249	948	366899.22	141989998.14
431	27	123.12	11637	5015547	2161700757	931693026267	3324.24	1432747.44	617514146.64
475	35	134.04	16625	7896875	3751015625	1781732421875	4691.4	2228415	1058497125
519	17	146.55	8823	4579137	2376572103	1233440921457	2491.35	1293010.65	671072527.35
563	6	164.9	3378	1901814	1070721282	602816081766	989.4	557032.2	313609128.6
604	2	188.3	1208	729632	440697728	266181427712	376.6	227466.4	137389705.6
$\sum$	100		46526	21941522	10483767350	5073105808130	13208.65	6238517.7	2985673383.73

Система была решена с помощью написанной программы. В результате были получены следующие значения коэффициентов:

$$a = 0.0003$$

$$b = 0.0021$$

$$c = 57.4223$$

Тогда, выборочное уравнение регрессии Y на X:

$$y(x) = 0.0003 * x^2 + 0.0021 * x + 57.4223$$

Корреляционная кривая параболического вида на множестве выборки представлена на рис. 2.

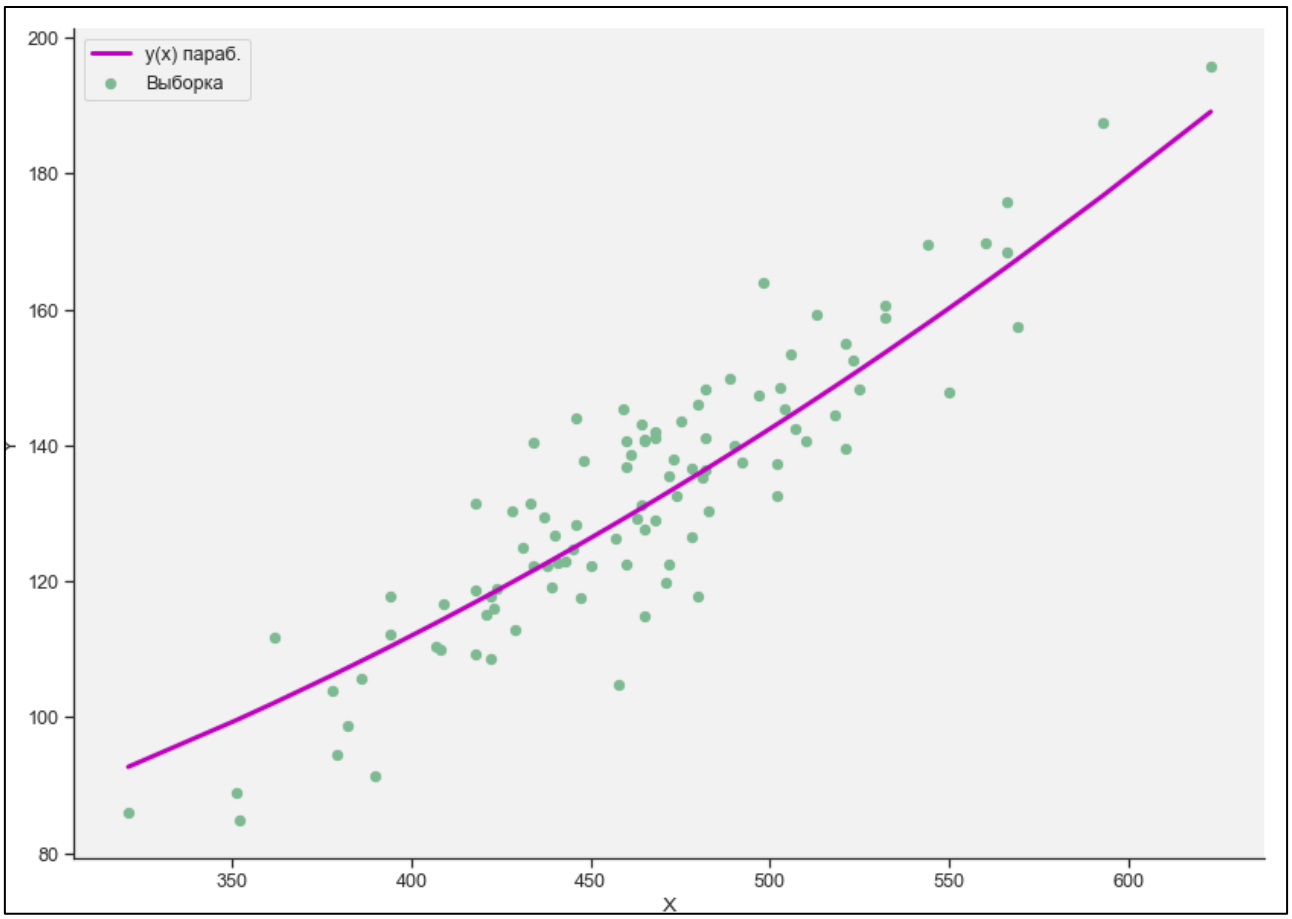

Рисунок 2 – Корреляционная кривая параболического вида

2. Логарифмическая функция

Для заданной выборки построим корреляционную кривую логарифмической функции $y = \beta_1 \ln x + \beta_0$. Выборочное уравнение регрессии Y на X:

$$\overline{y_x} = a + b \ln x$$

Применяя МНК, можно получить формулы для расчета значений коэффициентов a и b:

$$\begin{cases} b = \dfrac{n \sum_{i=1}^{K} (\overline{y_{x_i}} * \ln x_i) - \sum_{i=1}^{K} \ln \overline{y_{x_i}} * \sum_{i=1}^{K} \ln x_i}{n \sum_{i=1}^{K} (\ln x_i)^2 - (\sum_{i=1}^{K} \ln x_i)^2} \\ a = \dfrac{\sum_{i=1}^{K} \overline{y_{x_i}} - \left(\sum_{i=1}^{K} \ln x_i\right) b}{n} \end{cases}$$

С помощью написанной программы на языке Python были найдены данные коэффициенты:

$$a = -845.15$$

$$b = 159.4$$

Тогда, выборочное уравнение регрессии Y на X:

$$y(x) = -845.15 + 159.4 * \ln x$$

Корреляционная кривая логарифмической функции представлена на рисунке 3.

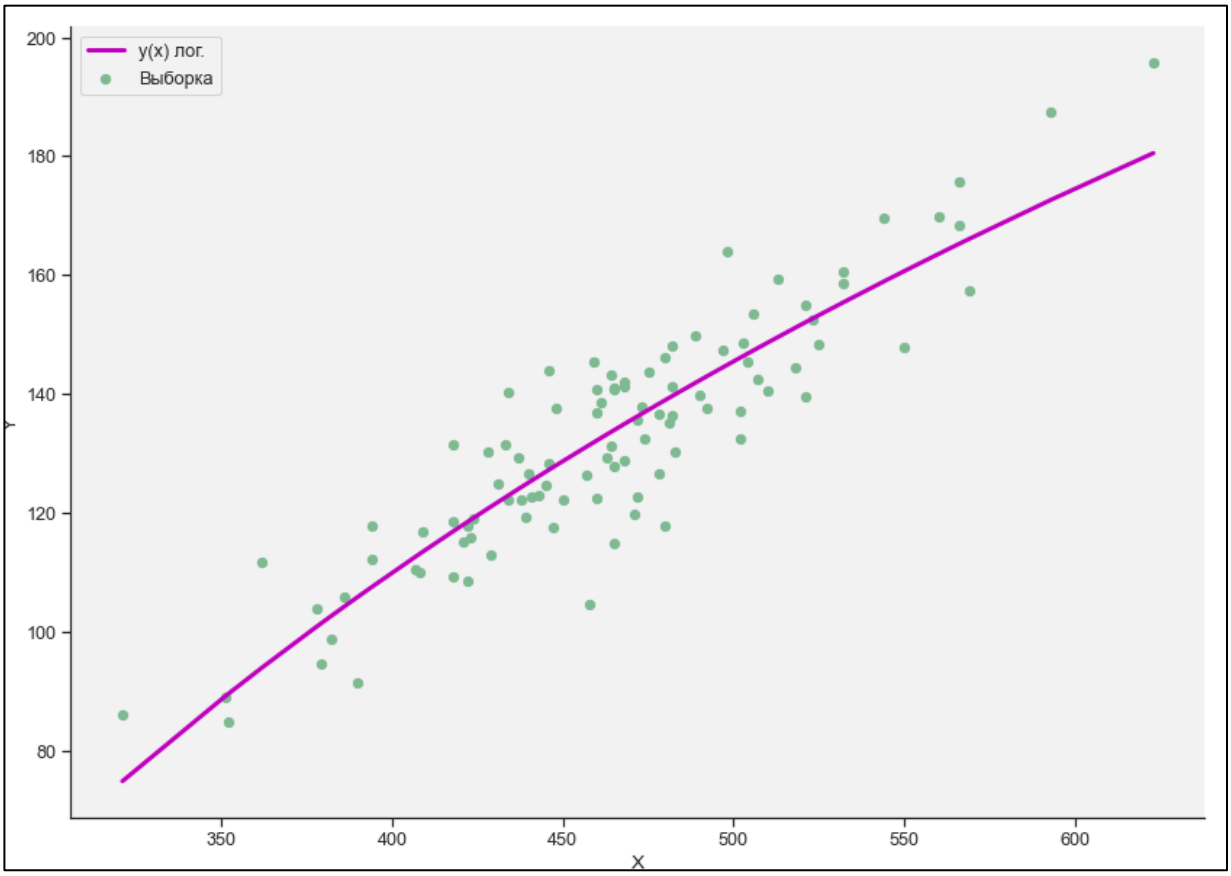

Рисунок 3 – Корреляционная кривая логарифмической функции

Выводы

Для заданной двумерной выборки были получены выборочные прямые средней квадратичной регрессии x на y и y на x. Данные прямые были построены на множестве выборки.

Были найдены условные выборочные средние и дисперсии и посчитаны внутригрупповая, межгрупповая и общая дисперсии для расчёта выборочного корреляционного отношения x к y и y к x.

Найдены выборочные корреляционные отношения $\overline{\eta_{xy}} = 0.8649$ и $\overline{\eta_{yx}} = 0.8717$. Выяснено, что выполняются неравенства $\overline{\eta_{xy}} \geq \left|\overline{r_{xy}}\right|$ и $\overline{\eta_{yx}} \geq \left|\overline{r_{xy}}\right|$. В результате, на основании полученных значений выборочного корреляционного отношения было выдвинуто предположение о корреляционной зависимости признаков, однако зависимость не линейная корреляционная и не функциональная.

Были построены корреляционные кривые параболического и логарифмического вида. Коэффициенты уравнений были найдены с помощью МНК. Исходя из построенных графиков, можно увидеть, что корреляционная зависимость может быть выражена обеими функциями.

ПРИЛОЖЕНИЕ А
ИСХОДНЫЙ КОД

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('data/main_data.csv')
X = df['nu']
Y = df['E']
int_rowX = pd.read_csv('data/interval.csv')
int_rowY = pd.read_csv('data/interval2.csv')
kor = pd.read_csv('data/kor.csv')

sns.set_theme(palette='crest', font_scale=1.15)
sns.set_style("ticks", {"axes.facecolor": ".95"})
ax = sns.relplot(data=df, x='nu', y='E', kind='scatter', height=8.27, as-
pect=11.7/8.27)
ax.set_axis_labels('X', 'Y')
plt.savefig('pics/1.png')

N = 100
xv, yv = 465.26, 132.09
sx, sy = 54.57, 19.97
r = 0.853

regr_xy = lambda y: xv + r*(sx/sy)*(y-yv)

ost_var_xy = (sx**2)*(1-r**2)

regr_yx = lambda x: yv + r*(sy/sx)*(x-xv)

ost_var_yx = (sy**2)*(1-r**2)
```

```
ax = sns.relplot(data=df, x='nu', y='E', kind='scatter', height=8.27,
                 aspect=11.7/8.27, s=40, label='Выборка')
plt.plot(regr_xy(df['E']), df['E'], label='x(y)', zorder=0, c='r')
plt.plot(df['nu'], regr_yx(df['nu']), label='y(x)', zorder=1, c='m')
ax.set_axis_labels('X', 'Y')
plt.legend()
plt.savefig('pics/2.png')

ost_var_xy
ost_var_yx

kor.loc[1:7,'Xi'] = [np.sum(kor.iloc[i,1:8]) for i in range(1,8)]
kor.iloc[8,1:8] = [np.sum(kor.iloc[1:8,i]) for i in range(1,8)]
kor.iloc[8,8] = 100

kor.loc[1:7,'yX']
=[(np.dot(kor.iloc[0,1:8],kor.iloc[i,1:8])/kor.loc[i,'Xi']).round(2)  for
i in range(1,8)]

kor.iloc[9,1:8]
=[(np.dot(kor.iloc[1:8,0],kor.iloc[1:8,i])/kor.iloc[8,i]).round(2)  for  i
in range(1,8)]

kor['D_grX'] = np.NaN
for i in range(1,8):
    x0_arg_kv = kor.iloc[0,1:8]**2
    dt = np.dot(x0_arg_kv,kor.iloc[i,1:8])/kor.loc[i,'Xi']
    dt -= kor.loc[i,'yX']**2
    kor.loc[i,'D_grX'] =(dt).round(2)


kor = kor.append(pd.Series(dtype='float64'), ignore_index=True)
for i in range(1,8):
    y0_arg_kv = kor.iloc[1:8,0]**2
```

```
    dt2 = np.dot(y0_arg_kv,kor.iloc[1:8,i])/kor.iloc[8,i]
    dt2 -= kor.iloc[9,i]**2
    kor.iloc[10,i] =(dt2).round(2)

D_vngr_xy = np.dot(kor.loc[1:7,'Xi'],kor.loc[1:7,'D_grX'])/kor.iloc[8,8]
D_vngr_xy.round(4)

kv_mezh_xy = (kor.loc[1:7,'yX']-xv)**2
D_mezh_xy = np.dot(kor.loc[1:7,'Xi'],kv_mezh_xy)/kor.iloc[8,8]
D_mezh_xy.round(4)

D_obsh_xy = D_vngr_xy + D_mezh_xy
D_obsh_xy.round(4)

eta_xy = np.sqrt(D_mezh_xy/D_obsh_xy)
eta_xy.round(4)
r

D_vngr_yx = np.dot(kor.iloc[8,1:8],kor.iloc[10,1:8])/kor.iloc[8,8]
D_vngr_yx

kv_mezh_yx = (kor.iloc[9,1:8]-yv)**2
D_mezh_yx = np.dot(kor.iloc[8,1:8],kv_mezh_yx)/kor.iloc[8,8]
D_mezh_yx.round(4)

D_obsh_yx = D_vngr_yx + D_mezh_yx
D_obsh_yx.round(4)

eta_yx = np.sqrt(D_mezh_yx/D_obsh_yx)
eta_yx.round(4)
r
kor
```

```
df_prbl_x = pd.DataFrame({'x': kor.iloc[0,1:8], 'n': kor.iloc[8,1:8], 'y': kor.iloc[9,1:8]})

for i in range(1,5):
    df_prbl_x[f'nx{i}'] = df_prbl_x['n']*(df_prbl_x['x']**i)
df_prbl_x['ny'] = df_prbl_x['n']*df_prbl_x['y']
df_prbl_x['nyx1'] = df_prbl_x['nx1']*df_prbl_x['y']
df_prbl_x['nyx2'] = df_prbl_x['nx2']*df_prbl_x['y']

df_prbl_xf = df_prbl_x.append(df_prbl_x.sum(), ignore_index=True)
df_prbl_xf.iloc[-1,[0,2]] = 0
df_prbl_xf.to_csv('data/parabolxy.csv', index=False)
df_prbl_xf

M1 = np.array([[df_prbl_xf.loc[7,'nx4'],df_prbl_xf.loc[7,'nx3'],df_prbl_xf.loc[7,'nx2']],

[df_prbl_xf.loc[7,'nx3'],df_prbl_xf.loc[7,'nx2'],df_prbl_xf.loc[7,'nx1']],

[df_prbl_xf.loc[7,'nx2'],df_prbl_xf.loc[7,'nx1'],df_prbl_xf.loc[7,'n']]])
v1 = np.array([df_prbl_xf.loc[7,'nyx2'],df_prbl_xf.loc[7,'nyx1'],df_prbl_xf.loc[7,'ny']])
a, b, c = np.linalg.solve(M1, v1)
parab_regr = lambda x: a*x*x+b*x+c
a.round(4), b.round(4), c.round(4)

ax = sns.relplot(data=df, x='nu', y=parab_regr(df['nu']), kind='line', linewidth=3,
                 height=8.27, aspect=11.7/8.27, label='y(x) параб.', color='m')
plt.scatter(df['nu'], df['E'], s=40, label='Выборка')
```

```
ax.set_axis_labels('X', 'Y')
plt.legend()
plt.savefig('pics/3.png')

df_step_x = pd.DataFrame({'x': kor.iloc[0,1:8], 'n': kor.iloc[8,1:8], 'y':
kor.iloc[9,1:8]})
df_step_x

df_step_x['ln(x)'] = np.log(df_step_x['x'])
df_step_x['ln2(x)'] = (np.log(df_step_x['x']))**2
df_step_x['ln(x)y'] = df_step_x['ln(x)']*df_step_x['y']

df_step_xf = df_step_x.append(df_step_x.sum(), ignore_index=True)
df_step_xf.iloc[-1,[0]] = np.NaN
df_step_xf.round(3).to_csv('data/logxy.csv', index=False)
df_step_xf

b2         =         ((df_step_xf.loc[7,'n']*df_step_xf.loc[7,'ln(x)y'])-
(df_step_xf.loc[7,'y']*df_step_xf.loc[7,'ln(x)']))/(((df_step_xf.loc[7,'n
']*df_step_xf.loc[7,'ln2(x)'])-(df_step_xf.loc[7,'ln(x)'])**2))
a2                         =                         (df_step_xf.loc[7,'y']-
(df_step_xf.loc[7,'ln(x)']*b2))/df_step_xf.loc[7,'n']
a2.round(4), b2.round(4)
log_regr = lambda x: a2+b2*np.log(x)

X_new_ln = np.hstack((np.ones((N,1)),np.expand_dims(np.log(X),1)))
beta_curr_hat         =         np.matmul(np.matmul(np.linalg.inv(np.mat-
mul(X_new_ln.T,X_new_ln)),X_new_ln.T),Y)
plt.scatter(X,Y)
plt.xlabel("radius")
plt.ylabel("perimeter")
plt.plot(X,beta_curr_hat[0] + np.log(X) * beta_curr_hat[1],"-r")
plt.show()
a2 = beta_curr_hat[0]
```

```
b2 = beta_curr_hat[1]
log_regr = lambda x: a2+b2*np.log(x)
a2, b2

ax = sns.relplot(data=df, x='nu', y=log_regr(df['nu']), kind='line', lin-
ewidth=3,
                 height=8.27,   aspect=11.7/8.27,   label='y(x)   лог.',
color='m')
plt.scatter(df['nu'], df['E'], s=40, label='Выборка')
ax.set_axis_labels('X', 'Y')
plt.legend()
plt.savefig('pics/4.png')

dfst = df.copy()
dfst['1'] = parab_regr(dfst['nu'])
dfst['2'] = log_regr(dfst['nu'])
dfstm = dfst.melt(id_vars='nu', value_vars=['1','2'])

ax   =   sns.relplot(data=dfstm,   x='nu',   y='value',   hue='variable',
kind='line', linewidth=2.5,
                 height=8.27, aspect=11.7/8.27)
plt.scatter(df['nu'], df['E'], s=50, label='Выборка')
ax.set_axis_labels('nu', 'E')
plt.legend()
```