**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.

Пусть имеется двумерная случайная величина , где и зависимые случайные величины. Функцию называют линейной функцией среднеквадратической регрессии на .

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

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

Внутригрупповая дисперсия вычисляется, как взвешенная по объемам групп средняя арифметическая групповых дисперсий, межгрупповая – как дисперсия условных средних относительно выборочной средней .

Выборочное корреляционное отношение к определяется в соответствии с выражением:

Запишем выборочное уравнение регрессии на в параболическом виде:

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

**Постановка задачи**

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

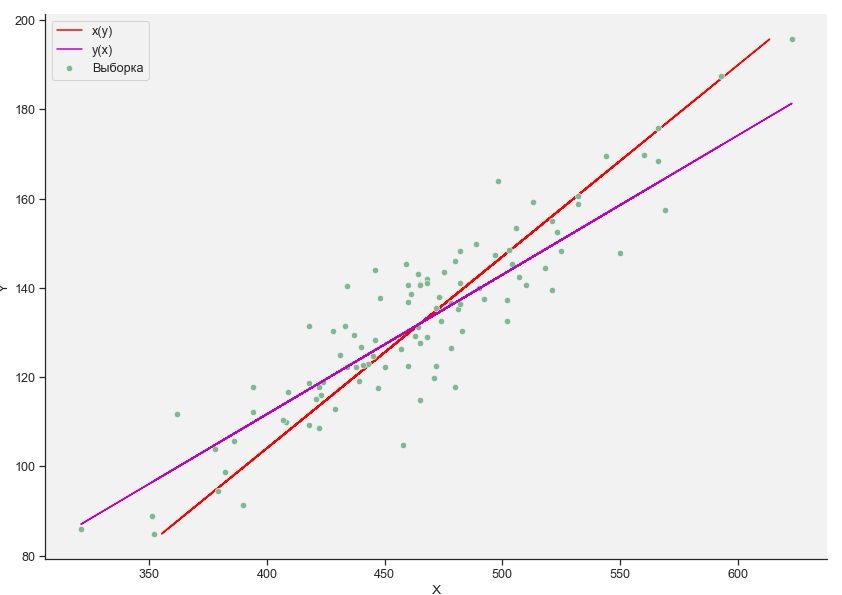
**Выполнение работы**

* Двумерная выборка и прямые регрессии

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

Выборочные прямые средней квадратичной регрессии на и на:

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 на и на:

* 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

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | | | | | | |
| **343** | **387** | **431** | **475** | **519** | **563** | **604** |  |  |  |
| 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 |
|  | 4 | 9 | 27 | 35 | 17 | 6 | 2 |  |  |  |
|  | 96.9 | 105.34 | 123.12 | 134.04 | 146.55 | 164.9 | 188.3 |  |  |  |
|  | 48 | 102.07 | 82.72 | 107.35 | 87.72 | 149.33 | 0 |  |  |  |

Выборочное корреляционное отношение к определяется в соответствии с выражением:

Аналогично для выборочного корреляционного отношения к .

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

Выборочное корреляционное отношение к :

Убедимся, что неравенства и выполняются:

Неравенство выполняется, так же, как и неравенство .

* Построение корреляционных кривых
  1. Параболический вид

Для заданной выборки была построена корреляционная кривая параболического вида .

Запишем выборочное уравнение регрессии на в параболическом виде:

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

Чтобы удобно рассчитать приведенные суммы была построена таблица 2.

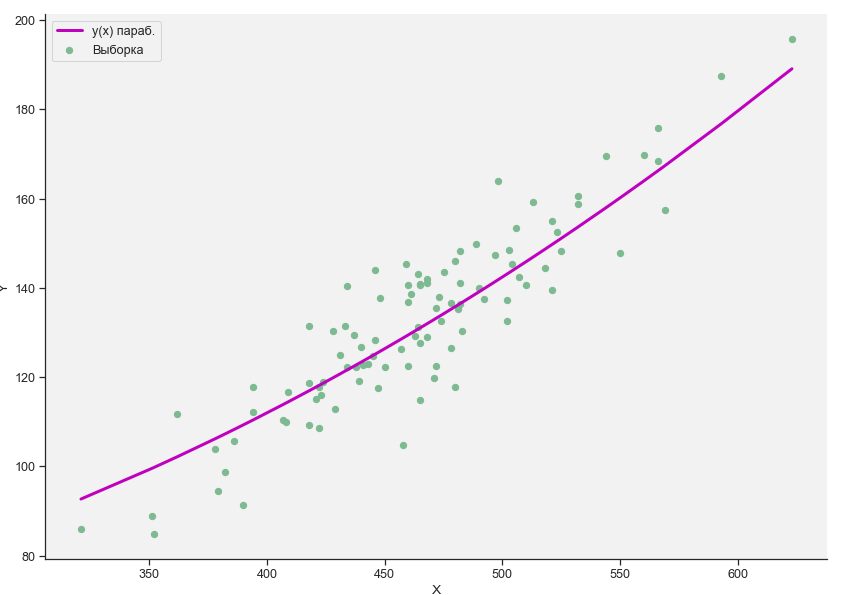
Таблица 2 – Таблица сумм МНК

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
| 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 |
|  | 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:

Тогда, выборочное уравнение регрессии на :

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



*Рисунок 2 – Корреляционная кривая параболического вида*

* 1. Логарифмическая функция

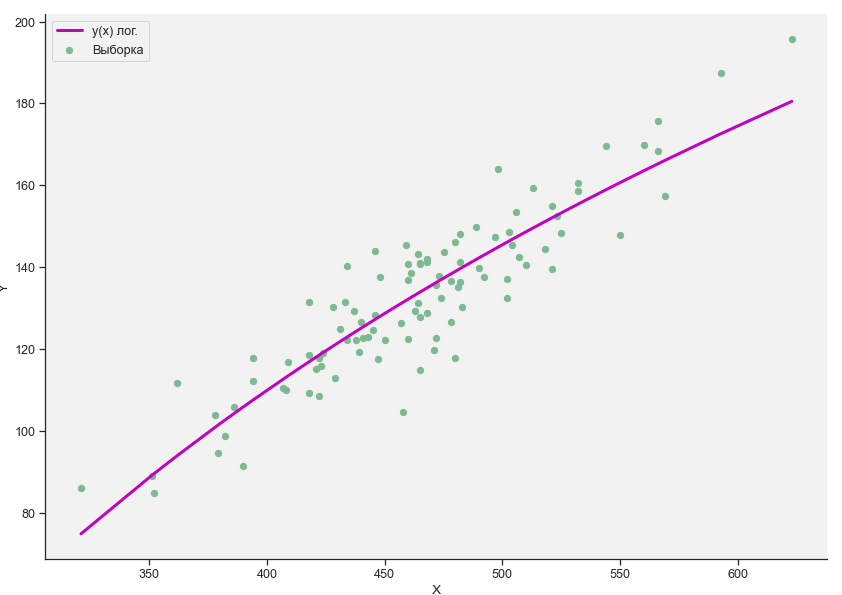
Для заданной выборки построим корреляционную кривую логарифмической функции . Выборочное уравнение регрессии на :

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

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

Тогда, выборочное уравнение регрессии на :

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 на и на . 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 к и к .

Найдены выборочные корреляционные отношения и . Выяснено, что выполняются неравенства и . 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, aspect=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.matmul(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', linewidth=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()