**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 |  | Киреев К.А. |
| Студент гр. 8383 |  | Муковский Д.В. |
| 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 (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.

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:

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

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

Внутригрупповая дисперсия вычисляется, как взвешенная по объемам групп средняя арифметическая групповых дисперсий.

Межгрупповая дисперсия вычисляется, как дисперсия условных (групповых) средних относительно выборочной средней .

Выборочное корреляционное отношение к определяется в соответствии с выражением:

где , – выборочные значения СКВО и соответственно. Аналогично определяется выборочное корреляционное отношение к .

Выборочное уравнение регрессии на параболического вида:

Значения коэффициентов и определим 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***

Двумерная выборка показана на рис. 1.

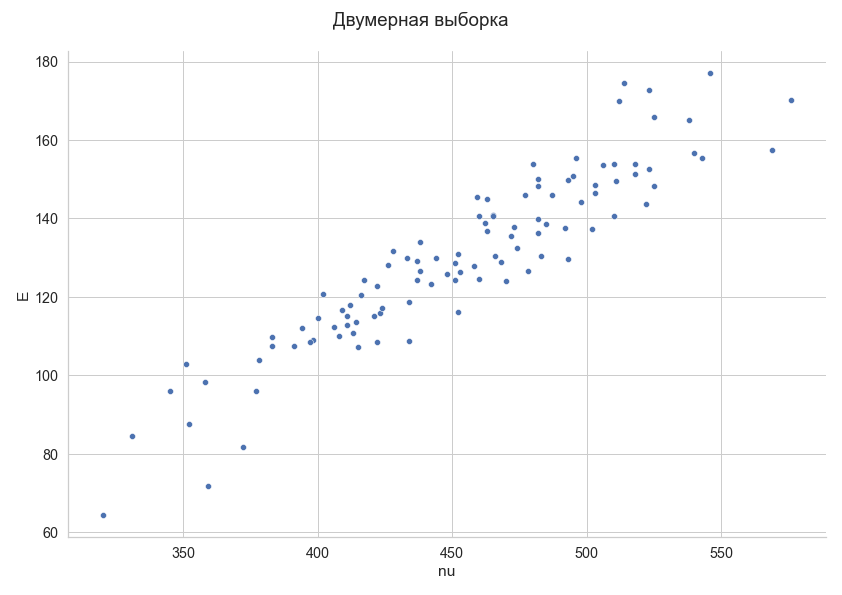


Рисунок 1 – Двумерная выборка

***Уравнения средней квадратичной регрессии***

Для заданной выборки построим уравнения средней квадратичной регрессии на и на и отобразим полученные прямые на множестве выборки.

Выборочная прямая средней квадратичной регрессии на :

Выборочная прямая средней квадратичной регрессии на :

Двумерная выборка и выборочные прямые средней квадратичной регрессии представлены на рис. 2.

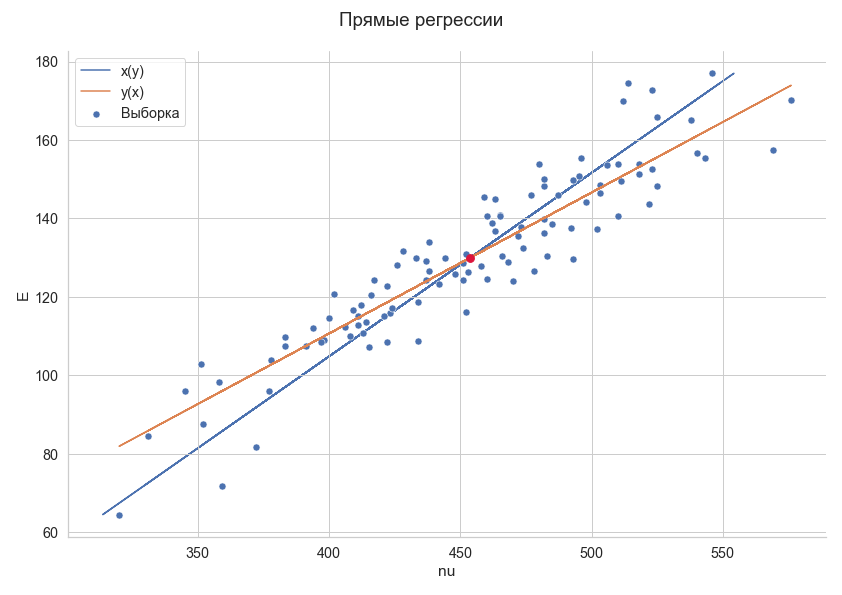
****

Рисунок 2 - Выборочные прямые средней квадратичной регрессии

Можно заметить, что пересечение выборочных прямых средней квадратичной регрессии находится в точке с координатами выборочного среднего для каждого из признаков.

Статистические оценки остаточной дисперсии для полученных выборочных прямых регрессии:

***Выборочное корреляционное отношение***

Корреляционная таблица для нахождения выборочного корреляционного отношения представлена в таблице 1. В данной таблице рассчитаны групповые выборочные средние и групповые выборочные дисперсии.

*Таблица 1 - Корреляционная таблица*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | | | | | | |
| *338.5* | *375.5* | *412.5* | *449.5* | *486.5* | *523.5* | *559* |  |  |  |
| 72.55 | 1 | 1 | - | - | - | - | - | 2 | 357 | 342.25 |
| 88.65 | 3 | 2 | - | - | - | - | - | 5 | 353.3 | 328.56 |
| 104.75 | 1 | 5 | 8 | 1 | - | - | - | 15 | 397.7 | 693.63 |
| 120.85 | - | - | 14 | 11 | 2 | - | - | 27 | 433.06 | 536.99 |
| 136.95 | - | - | 1 | 12 | 12 | 2 | - | 27 | 479.06 | 638.07 |
| 153.05 | - | - | - | 1 | 10 | 8 | 2 | 21 | 505.74 | 715.28 |
| 169.05 | - | - | - | - | - | 5 | 2 | 7 | 533.64 | 260.24 |
|  | 5 | 8 | 23 | 25 | 24 | 15 | 4 |  | - | - |
|  | 88.65 | 96.7 | 115.95 | 129.22 | 142.32 | 156.24 | 161.05 | - | - | - |
|  | 103.68 | 129.6 | 77.42 | 106.69 | 99.86 | 108.7 | 64 | - | - | - |

Выборочное корреляционное отношение к рассчитывается как отношение выборочных значений СКО и соответсвенно. Для этого были вычислены внутригрупповая, межгрупповая и общая дисперсии.

Неравенство выполняется.

Выборочное корреляционное отношение к :

Неравенство выполняется.

***Корреляционные кривые***

Для заданной выборки построим корреляционную кривую параболического вида . Выборочное уравнение регрессии на :

Значения коэффициентов определим с помощью МНК, решив систему уравнений:

Для вычисления сумм была построена таблица 2.

*Таблица 2 – Таблица сумм МНК*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
| 338.5 | 5.0 | 88.65 | 1692.5 | 572911.25 | 193930458.125 | 65645460075.3125 | 443.25 | 150040.125 | 50788582.3125 |
| 375.5 | 8.0 | 96.7 | 3004.0 | 1128002.0 | 423564751.0 | 159048564000.5 | 773.6 | 290486.8 | 109077793.4 |
| 412.5 | 23.0 | 115.95 | 9487.5 | 3913593.75 | 1614357421.875 | 665922436523.4375 | 2666.85 | 1100075.625 | 453781195.3125 |
| 449.5 | 25.0 | 129.22 | 11237.5 | 5051256.25 | 2270539684.375 | 1020607588126.5625 | 3230.5 | 1452109.75 | 652723332.625 |
| 486.5 | 24.0 | 142.32 | 11676.0 | 5680374.0 | 2763501951.0 | 1344443699161.5 | 3415.68 | 1661728.3199999998 | 808430827.68 |
| 523.5 | 15.0 | 156.24 | 7852.5 | 4110783.75 | 2151995293.125 | 1126569535950.9375 | 2343.6 | 1226874.6 | 642268853.1 |
| 559.0 | 4.0 | 161.05 | 2236.0 | 1249924.0 | 698707516.0 | 390577501444.0 | 644.2 | 360107.80000000005 | 201300260.20000002 |
|  | 104.0 | - | 47186.0 | 21706845.0 | 10116597075.5 | 4772814785282.25 | 13517.68 | 6241423.019999999 | 2918370844.6299996 |

В результате решения системы были получены следующие значения коэффициентов:

Выборочное уравнение регрессии на :

Корреляционная кривая параболического вида представлена на рис. 3.

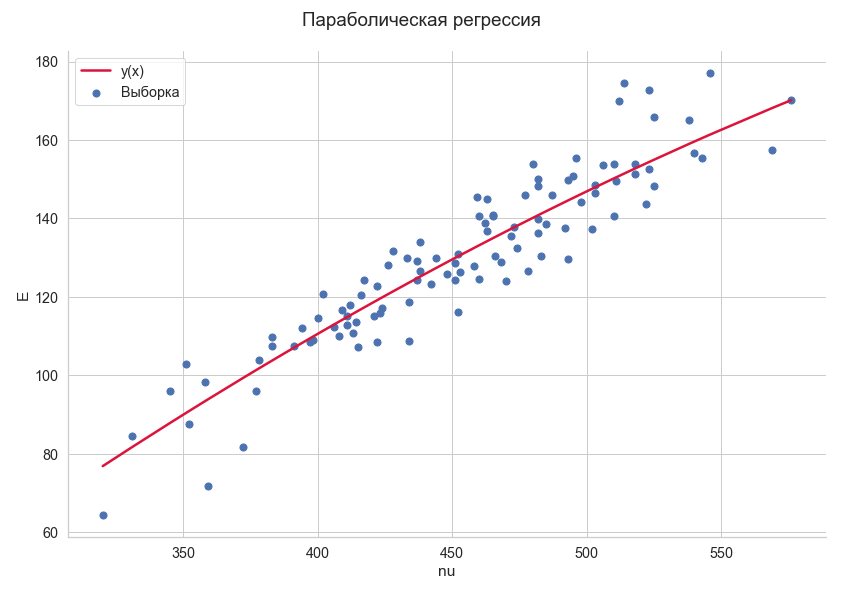


Рисунок 3 – Корреляционная кривая параболического вида

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

Значения коэффициентов определим с помощью МНК, решив систему уравнений:

Для вычисления сумм была построена таблица 3.

*Таблица 3 – Таблица сумм МНК*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| 338.5 | 5.0 | 88.65 | 5.825 | 33.925 | 4.485 | 26.121 |
| 375.5 | 8.0 | 96.7 | 5.928 | 35.144 | 4.572 | 27.102 |
| 412.5 | 23.0 | 115.95 | 6.022 | 36.267 | 4.753 | 28.625 |
| 449.5 | 25.0 | 129.22 | 6.108 | 37.309 | 4.862 | 29.695 |
| 486.5 | 24.0 | 142.32 | 6.187 | 38.282 | 4.958 | 30.677 |
| 523.5 | 15.0 | 156.24 | 6.261 | 39.194 | 5.051 | 31.624 |
| 559.0 | 4.0 | 161.05 | 6.326 | 40.02 | 5.082 | 32.148 |
|  | 104.0 | - | 42.657 | 260.142 | 33.762 | 205.991 |

В результате решения системы были получены следующие значения коэффициентов:

Выборочное уравнение регрессии на :

Корреляционная кривая степенной функции представлена на рис. 4.

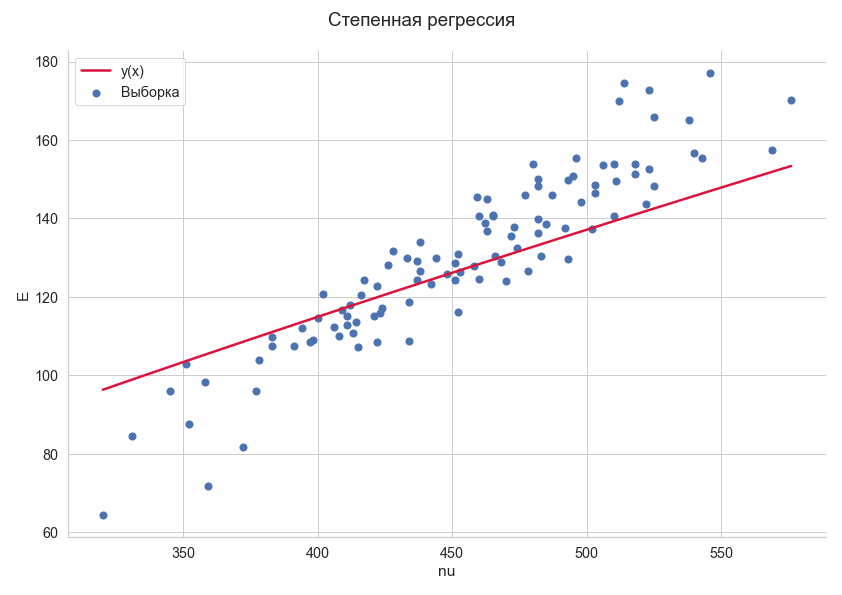


Рисунок 4 – Корреляционная кривая степенной функции

**Выводы**

Для заданной выборки были получены выборочные прямые средней квадратичной регрессии на и на .

Прямые были построены на множестве выборки.

Найдены выборочные корреляционные отношения и . Определено, что выполняются неравенства и . На основе полученных значений выборочного корреляционного отношения можно предположить, что и связаны корреляционной зависимостью, но не линейной корреляционной зависимостью и не функциональной зависимостью. Характер корреляционной зависимости не определен.

Были построены корреляционные кривые параболического и степенного вида. Визуально можно сделать вывод о том, что корреляционная зависимость признаков может быть выражена параболической функцией, но в меньшей мере степенной.

**ПРИЛОЖЕНИЕ А**

**ИСХОДНЫЙ КОД**

#!/usr/bin/env python

# coding: utf-8

# In[1]:

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

from IPython.core.interactiveshell import InteractiveShell

InteractiveShell.ast\_node\_interactivity = "all"

# pd.set\_option('display.max\_columns', None)

# pd.set\_option('display.max\_rows', None)

# ## Выборка

# In[2]:

df = pd.read\_csv('c:/Users/gandh/dev/unv/smoed/me/data/main\_data.csv')

X = df['nu']

Y = df['E']

int\_rowX = pd.read\_csv('c:/Users/gandh/dev/unv/smoed/me/data/interval.csv')

int\_rowY = pd.read\_csv('c:/Users/gandh/dev/unv/smoed/me/data/interval2.csv')

kor = pd.read\_csv('c:/Users/gandh/dev/unv/smoed/me/data/kor.csv')

# In[3]:

sns.set\_theme(style="whitegrid", palette='deep', context='notebook', font\_scale=1.3)

ax = sns.relplot(data=df, x='nu', y='E', kind='scatter', height=8.27, aspect=11.7/8.27)

ax.set\_axis\_labels('nu', 'E')

ax.fig.suptitle('Двумерная выборка')

plt.tight\_layout()

plt.savefig('pics/1.png')

# ## Прямые регрессии

# In[4]:

N = 104

xv, yv = 453.71, 129.98

sx, sy = 53.79, 22.06

r = 0.8765

# ### Прямая x на y

# In[5]:

regr\_xy = lambda y: xv + r\*(sx/sy)\*(y-yv)

# In[6]:

ost\_var\_xy = (sx\*\*2)\*(1-r\*\*2)

# ### Прямая y на x

# In[7]:

regr\_yx = lambda x: yv + r\*(sy/sx)\*(x-xv)

# In[8]:

ost\_var\_yx = (sy\*\*2)\*(1-r\*\*2)

# ### График

# In[9]:

# Регрессия Y на X

# ax = sns.lmplot(data=df, x='nu', y='E', height=8.27, aspect=11.7/8.27)

# In[10]:

ax = sns.relplot(data=df, x='nu', y='E', kind='scatter', height=8.27,

aspect=11.7/8.27, s=50, label='Выборка')

plt.plot(regr\_xy(df['E']), df['E'], label='x(y)', zorder=0)

plt.plot(df['nu'], regr\_yx(df['nu']), label='y(x)', zorder=1)

plt.scatter(xv, yv, s=60, c='crimson', zorder=2)

ax.set\_axis\_labels('nu', 'E')

ax.fig.suptitle('Прямые регрессии')

plt.legend()

plt.tight\_layout()

plt.savefig('pics/2.png')

# In[11]:

ost\_var\_xy

ost\_var\_yx

# ## Выборочное корреляционное отношение

# ### Таблица

# In[12]:

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] = 104

# #### Средний x для данного y (условный выборочный x)

# In[13]:

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)]

# #### Средний y для данного x (условный выборочный y)

# In[14]:

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)]

# #### Групповая выборочная дисперсия X

# In[15]:

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)

# #### Групповая выборочная дисперсия Y

# In[16]:

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)

# In[17]:

kor

# ### Дисперсии X к Y

# #### Внутригрупповая дисперсия X к Y

# In[18]:

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)

# #### Межгрупповая дисперсия X к Y

# In[19]:

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)

# #### Общая дисперсия X к Y

# In[20]:

D\_obsh\_xy = D\_vngr\_xy + D\_mezh\_xy

D\_obsh\_xy.round(4)

# #### Выборочное корреляционное отношение X к Y

# In[21]:

eta\_xy = np.sqrt(D\_mezh\_xy/D\_obsh\_xy)

eta\_xy.round(4)

r

# ### Дисперсии Y к X

# #### Внутригрупповая дисперсия Y к X

# In[22]:

D\_vngr\_yx = np.dot(kor.iloc[8,1:8],kor.iloc[10,1:8])/kor.iloc[8,8]

D\_vngr\_yx

# #### Межгрупповая дисперсия Y к X

# In[23]:

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)

# #### Общая дисперсия Y к X

# In[24]:

D\_obsh\_yx = D\_vngr\_yx + D\_mezh\_yx

D\_obsh\_yx.round(4)

# #### Выборочное корреляционное отношение Y к X

# In[25]:

eta\_yx = np.sqrt(D\_mezh\_yx/D\_obsh\_yx)

eta\_yx.round(4)

r

# ## Корреляционные кривые

# In[26]:

kor

# ### Параболическая регерессия Y на X

# In[27]:

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

# In[28]:

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]] = np.NaN

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, b, c

ax = sns.relplot(data=df, x='nu', y=parab\_regr(df['nu']), kind='line', linewidth=2.5,

height=8.27, aspect=11.7/8.27, label='y(x)', color='crimson')

plt.scatter(df['nu'], df['E'], s=50, label='Выборка')

ax.set\_axis\_labels('nu', 'E')

ax.fig.suptitle('Параболическая регрессия')

plt.legend()

plt.tight\_layout()

plt.savefig('pics/3.png')

# ### Степенная регерессия Y на X

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['log\_x'] = np.log(df\_step\_x['x'])

df\_step\_x['log\_x2'] = np.log(df\_step\_x['x'])\*\*2

df\_step\_x['log\_y'] = np.log(df\_step\_x['y'])

df\_step\_x['log\_x\_log\_y'] = df\_step\_x['log\_x']\*df\_step\_x['log\_y']

df\_step\_xf = df\_step\_x.append(df\_step\_x.sum(), ignore\_index=True)

df\_step\_xf.iloc[-1,[0,2]] = np.NaN

df\_step\_xf.round(3).to\_csv('data/stepxy.csv', index=False)

df\_step\_xf

M1 = np.array([[df\_step\_xf.loc[7,'n'],df\_step\_xf.loc[7,'log\_x']],

[df\_step\_xf.loc[7,'log\_x'],df\_step\_xf.loc[7,'log\_x2']]])

v1 = np.array([df\_step\_xf.loc[7,'log\_y'],df\_step\_xf.loc[7,'log\_x\_log\_y']])

a2, b2 = np.linalg.solve(M1, v1)

step\_regr = lambda x: np.exp(a2)\*(x\*\*b2)

np.exp(a2), b2, a2

dfst = df.copy()

dfst['1'] = parab\_regr(dfst['nu'])

dfst['2'] = step\_regr(dfst['nu'])

dfstm = dfst.melt(id\_vars='nu', value\_vars=['1','2'])

dfstm

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()

ax = sns.relplot(data=df, x='nu', y=step\_regr(df['nu']), kind='line', linewidth=2.5,

height=8.27, aspect=11.7/8.27, label='y(x)', color='crimson')

plt.scatter(df['nu'], df['E'], s=50, label='Выборка')

ax.set\_axis\_labels('nu', 'E')

ax.fig.suptitle('Степенная регрессия')

plt.legend()

plt.tight\_layout()

plt.savefig('pics/4.png')