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. 8382 _____ Звегинцева Е.Н.

Студент гр. 8382 _____ Мирончик П.Д.

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.

$$\sum_i (y_i - f_i(x_1, \ldots, x_n))^2 \to \min_x$$

Т.е. необходимо найти такие значения переменных, чтобы значения функций f_i были близки к y_i.

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

Линейные функции выборочной среднеквадратической регрессии:

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

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

где $\bar{x}, \bar{y}$ - выборочное среднее X, Y соответственно, $\overline{r_{xy}}$ – выборочный коэффициент корреляции, S_x и S_y – статистические оценки среднеквадратических отклонений для выборок X, Y.

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 рассчитать внутригрупповую D_{y_x} и межгрупповую $D_{\bar{y}}$ дисперсии. Оценку общей дисперсии X можно представить, как сумму:

$$D_y = D_{y_x} + D_{\bar{y}}$$

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

Внутригрупповая дисперсия вычисляется по формуле (среднее дисперсий внутри каждой группы):

$$D_{y_x} = \frac{1}{N}\sum_{i=1}^{k_2} n_{x_i}\left(\frac{1}{n_{x_i}}\sum_{i=1}^{k_2}(y_{x,i} - \overline{y_x})^2\right),$$

Межгрупповая дисперсия вычисляется по формуле (дисперсии средних значений $\overline{y_x}$внутри каждой группы относительно $\bar{y}$):

$$D_{\bar{y}} = \frac{\sum_{i=1}^{k_2} n_{x_i} * (\overline{y_x} - \bar{y})^2}{N}$$

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

$$\overline{\eta_{yx}} = \frac{\sigma_{\bar{y}}}{\sigma_y} = \sqrt{\frac{D_{\bar{y}}}{D_{y_x}}},$$

И определяется следующими свойствами:

- $0 \leq \bar{\eta} \leq 1$
- $\bar{\eta} = 0$ – нет оснований предположить, что между величинами X и Y существует корреляционная зависимость
- $\bar{\eta} = 1$ – выборочные данные дают основание предположить, что между величинами X и Y существует функциональная зависимость
- $\bar{\eta} \geq |\bar{r}_{xy}|$

- $\bar{\eta} = |\bar{r}_{xy}|$ – выборочные данные согласованы с предположением, что случайные величины X и Y связаны линейной корреляционной зависимостью.

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.

Для выполнения данной работы была использована выборка, сформированная в первой лабораторной работе. Из лабораторных работ №2 и 4 берем выборочные параметры распределения величин E и v (выборочное среднее, дисперсия, исправленная оценка среднеквадратического отклонения) и выборочный коэффициент корреляции, что представлены в таблице на рис.1.

	mean	variance	sd
u	453.280702	2541.960680	50.640457
v	129.843860	439.691059	21.061390

Рисунок 1 – Выборочные параметры из предыдущих лабораторных работ

Зависимость между двум величинами можно наблюдать на диаграмме рассеяния, представленной на рис.2.

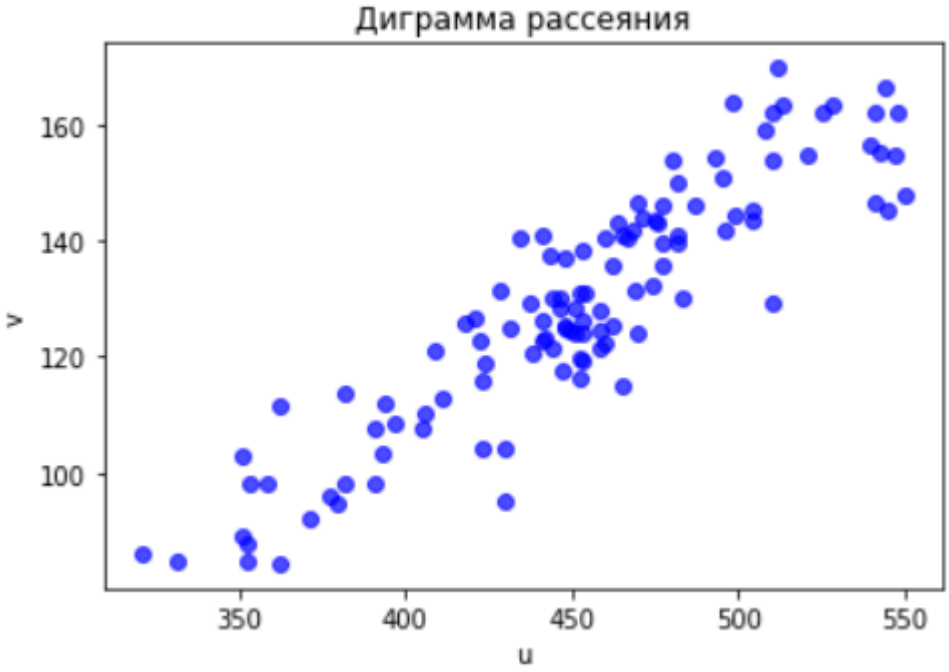

Рисунок 2 – Диаграмма рассеяния

Далее построим уравнение выборочной прямой среднеквадратичной регрессии, а также уравнение среднеквадратичной регрессии для величин E и v:

$$E_v = \bar{E} + \rho_{E,v}(v - \bar{v})$$

$$v_E = \bar{v} + \rho_{v,E}(E - \bar{E})$$

$$\rho_{E,v} = \bar{r}_{E,v} \frac{S_E}{S_v}$$

$$\rho_{v,E} = \bar{r}_{v,E} \frac{S_v}{S_E}$$

Рассчитаем эти значения:

	mean	variance	sd	reg_coeff
u	453.280702	2541.960680	50.640457	2.141776
v	129.843860	439.691059	21.061390	0.370470

Данные регрессионные прямые можно отобразить на графике, представленном на рис.3. Точкой отмечена координата, $(\bar{\nu}, \bar{E})$ которая является пересечением наших прямых.

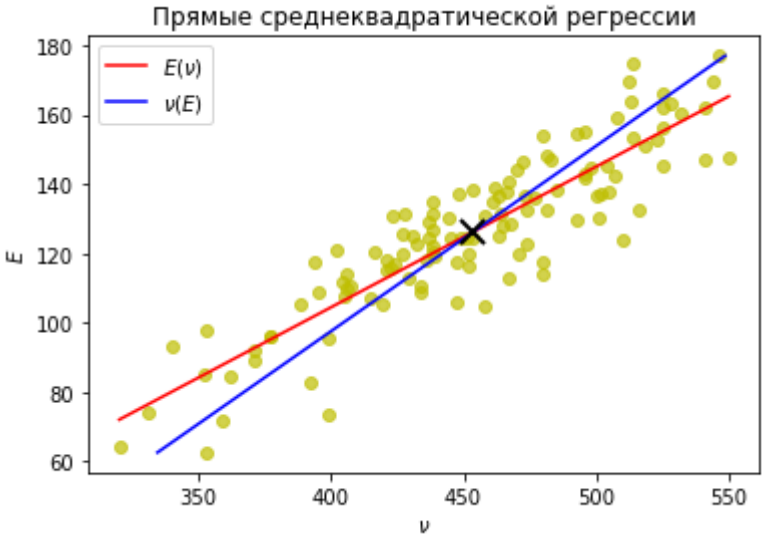

Рисунок 3 – график прямых среднеквадратичной регрессии

Далее можно найти дисперсию отклонений для величин E и v, данная остаточная дисперсия равна:

$$D_{\varepsilon E} = S_E^2(1 - \bar{r}_{xy}^2)$$

$$D_{\varepsilon v} = S_v^2(1 - \bar{r}_{xy}^2)$$

Вычислим данные значения:

$$D_{\varepsilon E} = 91.616004$$

$$D_{\varepsilon v} = 529.654346$$

Можно наблюдать, что остаточная дисперсия величины E сильно меньше величины v, что можно наблюдать на рис.3, где по оси E точки меньше отступают от прямой $E(v)$, чем у второй величины.

Составим корреляционную таблицу для нахождения выборочного корреляционного отношения. Для фиксированных значений величин рассматриваем наблюдения (E, v) и (v, E) как группу и вычисляем средние значения величины в этой группе. Представим это в таблице.

v u	89.8	100.8	111.8	122.8	133.8	144.8	155.8	166.8	N_u	mean_u
335.5	2.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	89.8
364.5	5.0	4.0	1.0	0.0	0.0	0.0	0.0	0.0	10.0	96.4
393.5	1.0	3.0	6.0	0.0	0.0	0.0	0.0	0.0	10.0	106.3
422.5	0.0	3.0	2.0	6.0	1.0	1.0	0.0	0.0	13.0	118.6
451.5	0.0	0.0	2.0	19.0	11.0	4.0	0.0	0.0	36.0	128.0
480.5	0.0	0.0	0.0	1.0	4.0	12.0	2.0	0.0	19.0	142.5
509.5	0.0	0.0	0.0	0.0	1.0	4.0	4.0	4.0	13.0	154.1
538.5	0.0	0.0	0.0	0.0	0.0	3.0	3.0	5.0	11.0	157.8
N_v	8.0	10.0	11.0	26.0	17.0	24.0	9.0	9.0	NaN	NaN
mean_v	360.9	390.6	406.7	445.9	460.0	485.3	512.7	525.6	NaN	NaN

Для каждой величины вычислим внутрегрупповую D_{y_x} и межгрупповую $D_{\bar{y}}$ дисперсии, а также их сумму:

	within groups	across groups	within + across	sample variance
v	80.067729	359.623330	439.691059	439.691059
u	461.182198	2080.778482	2541.960680	2541.960680

Как мы можем наблюдать, сумма равна выборочной дисперсии, следовательно вычисления произведены верно.

Далее вычислим корреляционные отношения E к v, v к E.

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

$$\overline{\eta_{yx}} = \frac{\sigma_{\bar{y}}}{\sigma_y} = \sqrt{\frac{D_{\bar{y}}}{D_{y_x}}},$$

```
0.9043782558968885
0.9047498291801275
```

$\bar{\eta}_{E,v}$= 0.904378

$\bar{\eta}_{E,v}$= 0.904750

Выборочное корреляционное отношение должно быть не меньше модуля выборочного коэффициента корреляции, который в нашем случае равен:

$\bar{r}_{v,E} = 0.8907$

Как мы видим, значения корреляционных отношений соответствуют ожиданиям и близки к 1, следовательно между величинами присутствует сильная корреляционная зависимость.

Далее необходимо построить параболическую корреляционную кривую, определив значения коэффициентов с помощью МНК.

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 матричным способом. В результате работы программы были получены следующие значения коэффициентов:

	a	b	c
v(u)	-0.000224	0.561539	-78.352315
u(v)	-0.004316	3.359413	92.115520

В связи с тем, что коэффициент *a* достаточно мал, то кривые будут близки к прямым, что можно увидеть на рис. 4.

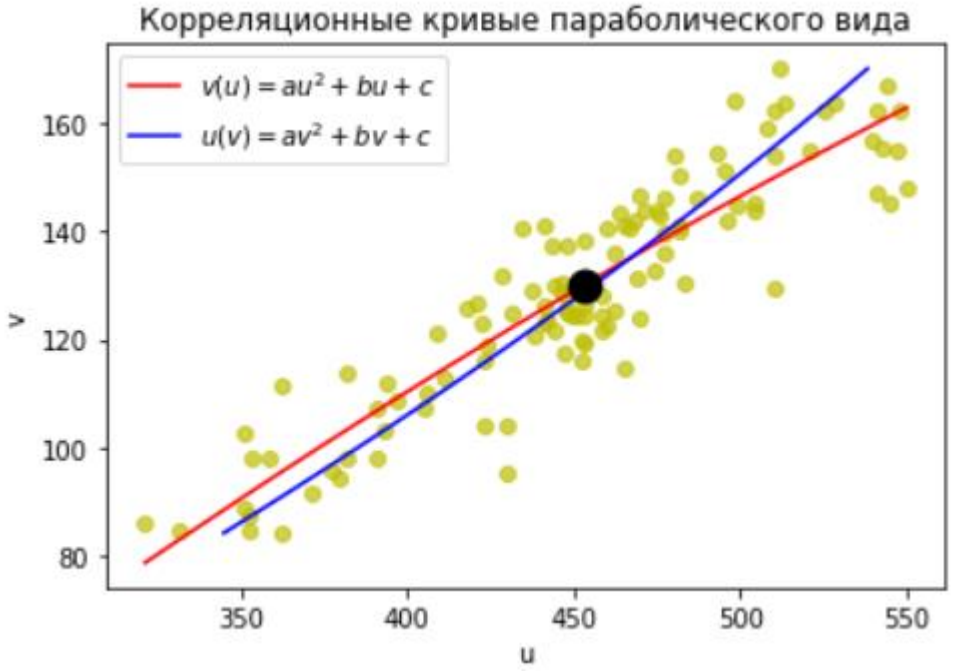

Рисунок 4 - Корреляционные кривые параболического вида

Согласно варианту, построим корреляционную кривую экспоненциальной функции $y = \frac{\beta_1}{x} + \beta_0$.

Для этого воспользуемся методом МНК и выразим с его помощью искомые переменные. После расчета на исходных данных получим следующие значения коэффициентов кривой обратной функции:

	beta_0	beta_1
v(u)	-131.123735	-940.559295
u(v)	-459.263810	916.301059

И построим график для найденных коэффициентов:

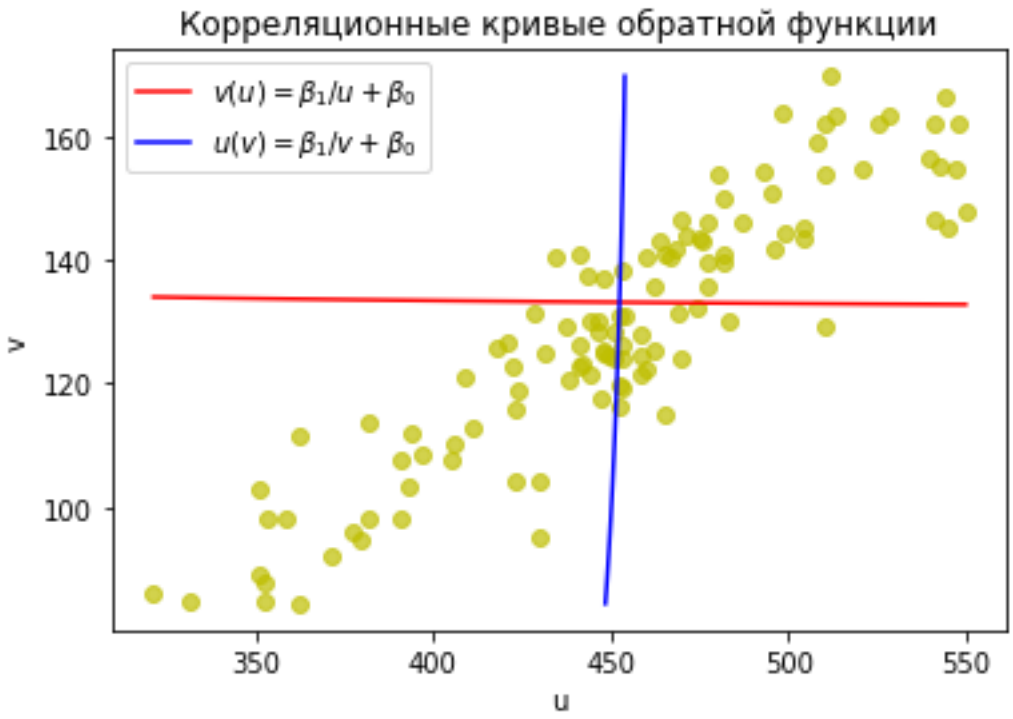

Рисунок 5 – Полученная кривая на множестве выборки

Видно, что полученные прямые очень плохо приближают значения распределения – вероятно, это связано с тем, что используемая функция не подходит для рассматриваемых данных, т.к. между распределениями величин E и v наблюдается прямая зависимость.

Выводы.

Была построена выборочная прямая линейной регрессии, коэффициенты которой вычислены с помощью метода наименьших квадратов. Было вычислено корреляционное отношение величины E к v и величины v к E, что подтвердило сильную корреляционную зависимость величин. Были построены корреляционные кривые параболического вида и обратной функции.