Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Высшая школа прикладной математики и вычислительной физики

Математическая статистика

Отчёт по лабораторной работе №5

Работу выполнил:
А. Н. Баженов
Группа:
5030102/10101
Преподаватель:
П. П. Филиппов

Санкт-Петербург
2024

Содержание

1. Постановка задачи

1. Сгенерировать двумерные выборки размерами 20, 60, 100 для нормального двумерного распределения $N(x, y, 0, 0, 1, 1, \rho)$. Коэффициент корреляции ρ взять равным 0, 0.5, 0.9. Каждая выборка генерируется 1000 раз и для неё вычисляются: среднее значение, среднее значение квадрата и дисперсия коэффициентов корреляции Пирсона, Спирмена и квадрантного коэффициента корреляции. Повторить все вычисления для смеси нормальных распределений:

$$f(x, y) = 0.9N(x, y, 0, 0, 1, 0.9) + 0.1N(x, y, 0, 0, 10, 10, -0.9)$$

   Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

2. Найти оценки коэффициентов линейной регрессии $y_i = a + bx_i + e_i$, используя 20 точек на отрезке [-1.8, 2] с равномерным шагом равным 0.2. Ошибку e_i считать нормально распределённой с параметрами (0, 1). В качестве эталонной зависимости взять $y_i = 2 + 2x_i + e_i$. При построении оценок коэффициентов использовать два критерия: критерий наименьших квадратов и критерий наименьших модулей. Проделать то же самое для выборки, у которой в значения y_1 и $y + 20$ вносятся возмущения 10 и -10.

2. Теоретическая информация

2.1. Коэффициент корреляции

2.1.1. Двумерное нормальное распределение

Двумерная случайная величина (X, Y) называется распределённой нормально (или просто нормальной), если её плотность вероятности определена формулой

$$N(x, y, \bar{x}, \bar{y}, \sigma_x, \sigma_y, \rho) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \times$$
$$\times \exp(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\bar{x})^2}{{\sigma_x}^2} - 2\rho\frac{(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{(y-\bar{y})^2}{{\sigma_y}^2}\right])$$

Компоненты X, Y двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями $\bar{x}$, $\bar{y}$ и средними квадратическими отклонениями σ_x, σ_y соответственно.
Параметр ρ называется коэффициентом корреляции.

2.1.2. Корреляционный момент (ковариация) и коэффициент корреляции

Корреляционный момент, иначе ковариация, двух случайных величин X, Y:

$$K = cov(X, Y) = M[(X - \bar{x})(Y - \bar{y})]$$

Коэффициент корреляции ρ двух случайных величин X, Y:

$$\rho = \frac{K}{\sigma_x\sigma_y}$$

2.1.3. Выборочные коэффициенты корреляции

1. **Выборочный коэффициент корреляции Пирсона:**

$$r = \frac{\frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n^2}\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}} = \frac{K}{s_X s_Y}$$

где K, ${s_X}^2$, ${s_Y}^2$ − выборочные ковариация и дисперсии случайных величин X, Y

2. **Выборочный квадрантный коэффициент корреляции:**

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n}$$

где n_1, n_2, n_3 и n_4 − количества точке с координатами $(x_i,\ y_i)$), попавшими соответственно в I, II, III и IV квадранты декартовой системы с осями $x^{'} = x - medx$, $y^{'} = y - medy$ и с центром

3. **Выборочный коэффициент ранговой корреляции Спирмена** Обозначим ранги, соотвествующие значениям переменной X, через u, а ранги, соотвествующие значениям переменной Y, — через v.
Выборочный коэффициент ранговой корреляции Спирмена:

$$r_S = \frac{\frac{1}{n}\sum(u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\frac{1}{n^2}\sum(u_i - \bar{u})^2(v_i - \bar{v})^2}}$$

где $\bar{u} = \bar{v} = \frac{1+2+\cdots+n}{n} = \frac{n+1}{2}-$ среднее значение рангов

2.1.4. Эллипсы рассеивания

Уравнение проекции эллипса рассеивания на плоскость xOy:

$$\frac{(x - \bar{x})^2}{{\sigma_x}^2} - 2\rho\frac{(x - \bar{x})(y - \bar{y})}{\sigma_x\sigma_y} + \frac{(y - \bar{y})^2}{{\sigma_y}^2} = const$$

Центр эллипса находится в точке с координатами $(\bar{x},\ \bar{y})$; оси симметрии эллипса составляют с осью Ox углы, определяемые уравнением

$$\tan 2\alpha = \frac{2\rho\sigma_x\sigma_y}{{\sigma_x}^2 - {\sigma_y}^2}$$

2.2. Простая линейная регрессия

2.2.1. Модель простой линейной регрессии

Регрессионную модель описания данных называют простой линейной регрессией, если

$$y_i =_0 +_1 x_i + \varepsilon_i, i = 1, 2, \ldots, n$$

где $x_1, x_2, \ldots, x_n-$ заданные числа (значения фактора); $y_1, y_2, \ldots, y_n-$ наблюдаемые значения отклика; $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n-$независимые, нормально распределённые $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые);β_0, β_1-неизвестные параметры, подлежащие оцениванию. В модели (19) отклик y зависит зависит от одного фактора x, и весь разброс экспериментальных точек объясняется только погрешностями наблюдений (результатов измерений) отклика y. Погрешности результатов измерений x в этой модели полагают существенно меньшими погрешностей результатов измерений y, так что ими можно пренебречь

2.2.2. Метод наименьших квадратов

При оценивании параметров регрессионной модели используют различные методы. Один из наиболее распрстранённых подходов заключается в следующем: вводится мера (критерий) рассогласования отклика и регрессионной функции, и оценки параметров регрессии определяются так, чтобы сделать это рассогласование наименьшим. Достаточно простые расчётные формулы для оценок получают при выборе критерия в виде суммы квадратов отклонений значений отклика от значений регрессионной функции (сумма квадратов остатков):

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \to min_{\beta_0, \beta_1}$$

Задача минимизации квадратичного критерия носит название задачи метода наименьших квадратов (МНК), а оценки $\hat{\beta}_0$, $\hat{\beta}_1$ параметров β_0, β_1, реализующие минимум критерия, называют МНК-оценками

2.2.3. Расчётные формулы для МНК-оценок

МНК-оценки параметров $\hat{\beta}_0$ и $\hat{\beta}_1$ находятся из условия обращения функции $Q(\beta_0, \beta_1)$ в минимум. Для нахождения МНК-оценок $\hat{\beta}_0$ и $\hat{\beta}_1$ выпишем необходимые условия экстремума:

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases}$$

Далее для упрощения записи сумм будем опускать индекс суммирования. Из этой системы получим

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum {x_i}^2 = \sum x_i y_i \end{cases}$$

Разделим оба уравнения на n:

$$\begin{cases} n\hat{\beta}_0 + (\frac{1}{n} \sum x_i)\hat{\beta}_1 = \frac{1}{n} \sum y_i \\ (\frac{1}{n} \sum x_i)\hat{\beta}_0 \sum x_i + (\frac{1}{n} \sum {x_i}^2)\hat{\beta}_1 = \frac{1}{n} \sum x_i y_i \end{cases}$$

и, используя известные статистические обозначения для выборочных первых и вторых начальных моментов

$$\bar{x} = \frac{1}{n} \sum x_i, \bar{y} = \frac{1}{n} \sum y_i, \bar{xy} = \frac{1}{n} \sum x_i y_i,$$

получим

$$\begin{cases} n\hat{\beta}_0 + \bar{x}\hat{\beta}_1 = \bar{y} \\ \bar{x}\hat{\beta}_0 + \bar{x^2}\hat{\beta}_1 = \bar{xy}, \end{cases}$$

откуда МНК-оценку $\hat{\beta}_1$ наклона прямой регрессии находим по формуле Крамера

$$\hat{\beta}_1 = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\bar{x^2} - \bar{x}^2}$$

а МНК-оценку $\hat{\beta}_0$ определяем непосредственно из первого уравнения системы:

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

Заметим, что определитель системы

$$\bar{x^2} - \bar{x}^2 = n^{-1}\sum(x_i - \bar{x})^2 = {s_x}^2 > 0,$$

если среди значений $x_1, x_2, \ldots, x_n$ есть различные, что и будем предполагать.
Доказательство минимальности функции $Q(\beta_0, \beta_1)$ в стационарной точке проведём с помощью известного достаточного признака экстремума функции двух переменных. Имеем:

$$\frac{\partial^2 Q}{\partial {\beta_0}^2} = 2n, \frac{\partial^2 Q}{\partial {\beta_0}^2} = 2\sum {x_i}^2 = 2n\bar{x^2}, \frac{\partial^2 Q}{\partial\beta_0\partial\beta_1} = 2\sum x_i = 2n\bar{x}.$$

$$\Delta = \frac{\partial^2 Q}{\partial {\beta_0}^2}\cdot\frac{\partial^2 Q}{\partial {\beta_0}^2} - (\frac{\partial^2 Q}{\partial\beta_0\partial\beta_1})^2 = 4n^2\bar{x^2} - 4n^2\bar{x}^2 = 4n^2\left[\bar{x^2} - \bar{x}^2\right] = 4n^2\left[\frac{1}{n}\sum(x_i - \bar{x})^2\right] = 4n^2{s_x}^2 > 0$$

Этот результат вместе с условием $\frac{\partial^2 Q}{\partial {\beta_0}^2} = 2n > 0$ означает, что в стационарной точке функция Q имеет минимум

2.2.4. Робастные оценки коэффициентов линейной регрессии

Робастность оценок коэффициентов линейной регрессии (т.е. их устойчивость по отношению к наличию в данных редких, но больших по величине выбросов) может быть обеспечена различными способами. Одним из них является использование метода наименьших модулей вместо метода наименьших квадратов:

$$\sum_{i=1}^{n}|y_i - \beta_0 - \beta_1 x_i| \to min_{\beta_0,\beta_1}$$

Напомним, что использование метода наименьших модулей в задаче оценивания параметра сдвига распределений приводит к оценке в виде выборочной медианы, обладающей робастными свойствами. В отличие от этого случая и от задач метода наименьших квадратов, на практике задача решается численно. Соответствующие процедуры представлены в некоторых современных пакетах программ по статистическому анализу. Здесь мы рассмотрим простейшую в вычислительном отношении робастную альтернативу оценкам коэффициентов линейной регрессии по МНК. Для этого сначала запишем выражения для оценок в другом виде:

$$\hat{\beta}_1 = \frac{\bar{xy} - \bar{x}\cdot\bar{y}}{\bar{x^2} - \bar{x}^2} = \frac{k_{xy}}{{s_x}^2} = \frac{k_{xy}}{s_x s_y}\cdot\frac{s_y}{s_x} = r_{xy}\frac{s_y}{s_x}$$

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1$$

В формулах заменим выборочные средние $\bar{x}$ и $\bar{y}$ соответственно на робастные выборочные медианы $medx$ и $medy$, среднеквадратические отклонения s_x и s_y на робастные нормированные интерквартильные широты ${q_x}^*$ и ${q_y}^{**}$, выборочный коэффициент корре-

ляции r_{xy}−на знаковый коэффициент корреляции r_Q :

$$\hat{\beta_{1R}} = r_Q \frac{q_y^*}{q_x^*}$$

$$\hat{\beta_{0R}} = medy - \hat{\beta_{1R}} medx,$$

$$r_Q = \frac{1}{n} \sum_{i=1}^{n} sign(x_i - medx) sign(y_i - medy),$$

$$q_y^* = \frac{y_j - y_l}{k_q(n)}, q_x^* = \frac{x_j - x_l}{k_q(n)},$$

$$l = \begin{cases} [\frac{n}{4}] + 1 & \text{,при } \frac{n}{4} \text{ дробном} \\ \frac{n}{4} & \text{при } \frac{n}{4} \text{ целом} \end{cases}$$

$$j = n - l + 1,$$

$$signz = \begin{cases} 1 & ,z > 0 \\ 0 & ,z = 0 \\ -1 & ,z < 0 \end{cases}$$

Уравнение регрессии здесь имеет вид

$$y = \hat{\beta_{0R}} + \hat{\beta_{1R}} x.$$

Статистики выборочной медианы и интерквартильной широты обладают робастными свойствами в силу того, что основаны на центральных порядковых статистиках, малочувствительных к большим по величине выбросам в данных. Статистика выборочного знакового коэффициента корреляции робастна, так как знаковая функция sign z чувствительна не к величине аргумента, а только к его знаку. Отсюда оценка прямой регрессии обладает очевидными робастными свойствами устойчивости к выбросам по координате y, но она довольно груба.

3. Результаты

3.1. Характеристики распределения

	r	r_S	r_Q
$\rho = 0$			
$E(z)$	-0.0067	0.0015	0.0
$E(z^2)$	0.0264	0.0269	0.04
$D(z)$	0.0544	0.0562	0.0573
$\rho = 0.5$			
$E(z)$	0.5018	0.4654	0.4
$E(z^2)$	0.2518	0.2166	0.16
$D(z)$	0.0341	0.0373	0.0516
$\rho = 0.9$			
$E(z)$	0.9057	0.8812	0.8
$E(z^2)$	0.8203	0.7765	0.64
$D(z)$	0.0027	0.0049	0.0275

Таблица 3.1

Таблица характеристик распределения для n = 20

	r	r_S	r_Q
$\rho = 0$			
$E(z)$	-0.0017	-0.0011	0.0
$E(z^2)$	0.0084	0.0081	0.0044
$D(z)$	0.0172	0.0178	0.0184
$\rho = 0.5$			
$E(z)$	0.502	0.4798	0.3333
$E(z^2)$	0.252	0.2302	0.1111
$D(z)$	0.0103	0.011	0.015
$\rho = 0.9$			
$E(z)$	0.9015	0.8861	0.7333
$E(z^2)$	0.8128	0.7851	0.5378
$D(z)$	0.0007	0.0011	0.0081

Таблица 3.2

Таблица характеристик распределения для n = 60

	r	r_S	r_Q
$\rho = 0$			
$E(z)$	0.0077	0.0065	0.0
$E(z^2)$	0.0047	0.0047	0.0064
$D(z)$	0.01	0.0097	0.0101
$\rho = 0.5$			
$E(z)$	0.5039	0.4828	0.32
$E(z^2)$	0.254	0.2331	0.1024
$D(z)$	0.0054	0.0063	0.0085
$\rho = 0.9$			
$E(z)$	0.9005	0.8883	0.72
$E(z^2)$	0.811	0.7891	0.5184
$D(z)$	0.0004	0.0006	0.005

Таблица 3.3

Таблица характеристик распределения для n = 100

	r	r_S	r_Q
$n = 20$			
$E(z)$	0.799	0.7654	0.6
$E(z^2)$	0.6384	0.5859	0.36
$D(z)$	0.0083	0.0118	0.0322
$n = 60$			
$E(z)$	0.7925	0.7728	0.6
$E(z^2)$	0.6281	0.5973	0.36
$D(z)$	0.0023	0.003	0.0102
$n = 100$			
$E(z)$	0.7934	0.7754	0.6
$E(z^2)$	0.6294	0.6012	0.36
$D(z)$	0.0015	0.0021	0.0067

Таблица 3.4

Таблица характеристик для смеси нормальных распределений

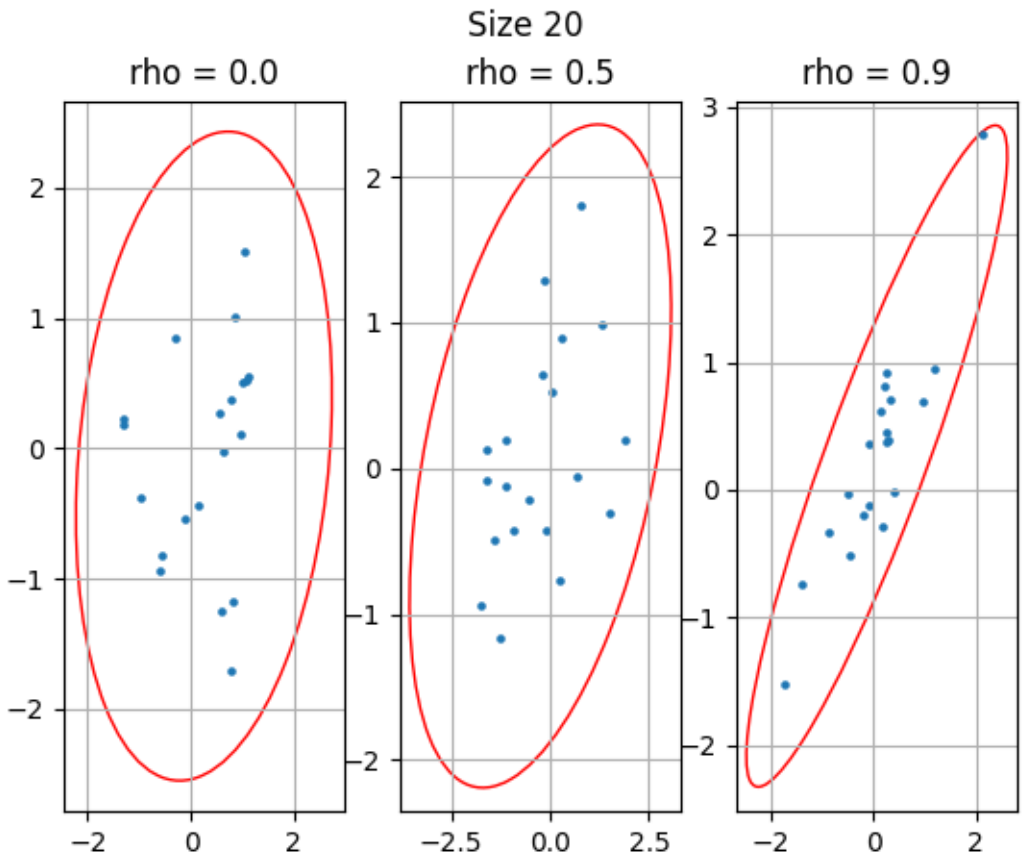

Рисунок 3.1. Эллипс рассеивания для 20 элементов

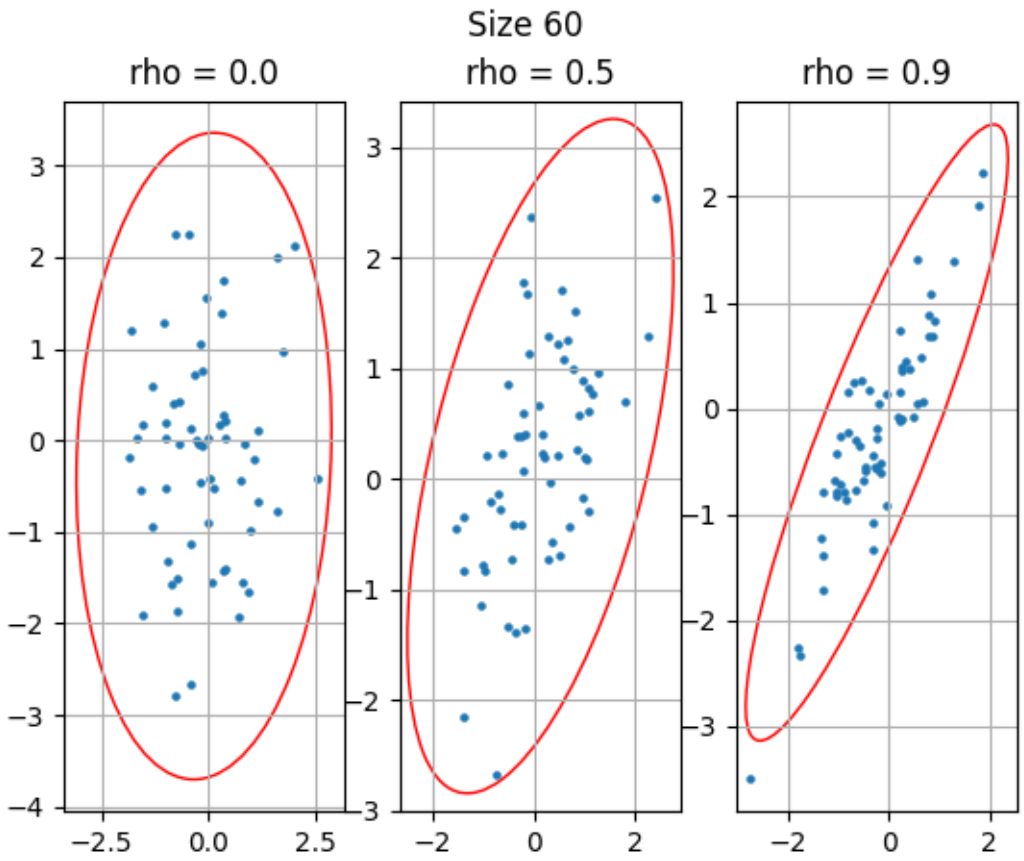

Рисунок 3.2. Эллипс рассеивания для 60 элементов

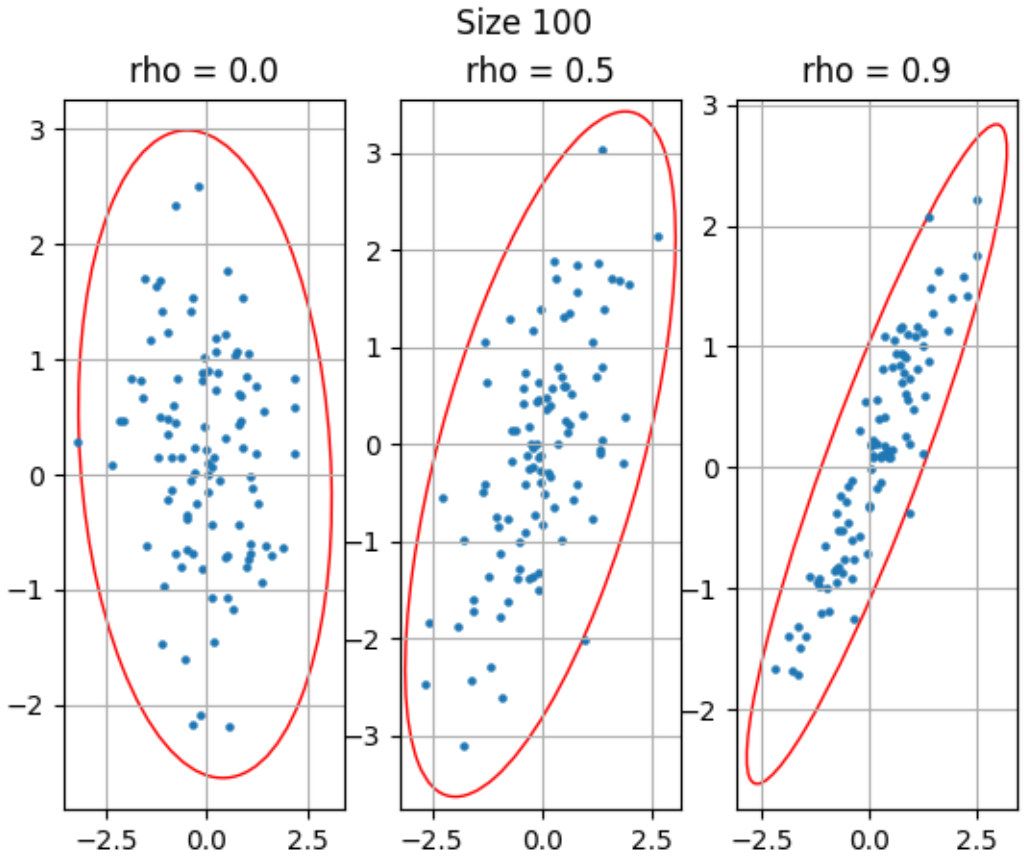

Рисунок 3.3. Эллипс рассеивания для 100 элементов

3.2. Оценки коэффициентов линейной регрессии

$d = \sum_{i=0}^{n} (y_m[i] - y_r[i])^2$

1. **Без возмущений**

   (a) Критерий наименьших квадратов
   $\hat{a} \approx 2.11$, $\hat{b} \approx 2.31$

   (b) Критерий наименьших модулей
   $\hat{a} \approx 2.11$, $\hat{b} \approx 2.31$
   МНК $d = 16.18$
   МНМ $d = 14.90$

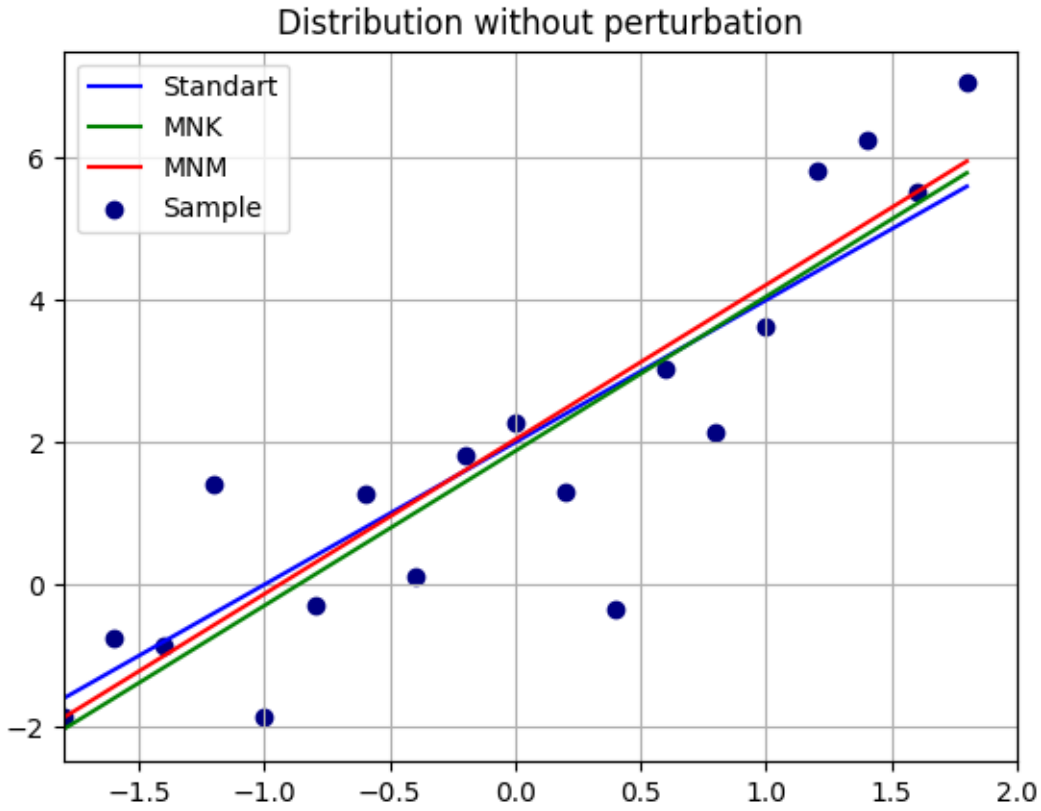

Рисунок 3.4. Выборка без возмущений

2. **С возмущениями**

   (a) Критерий наименьших квадратов
   $\hat{a} \approx 2.07$, $\hat{b} \approx 0.56$

   (b) Критерий наименьших модулей
   $\hat{a} \approx 1.82$, $\hat{b} \approx 1.94$
   МНК $d = 176.38$
   МНМ $d = 30.19$

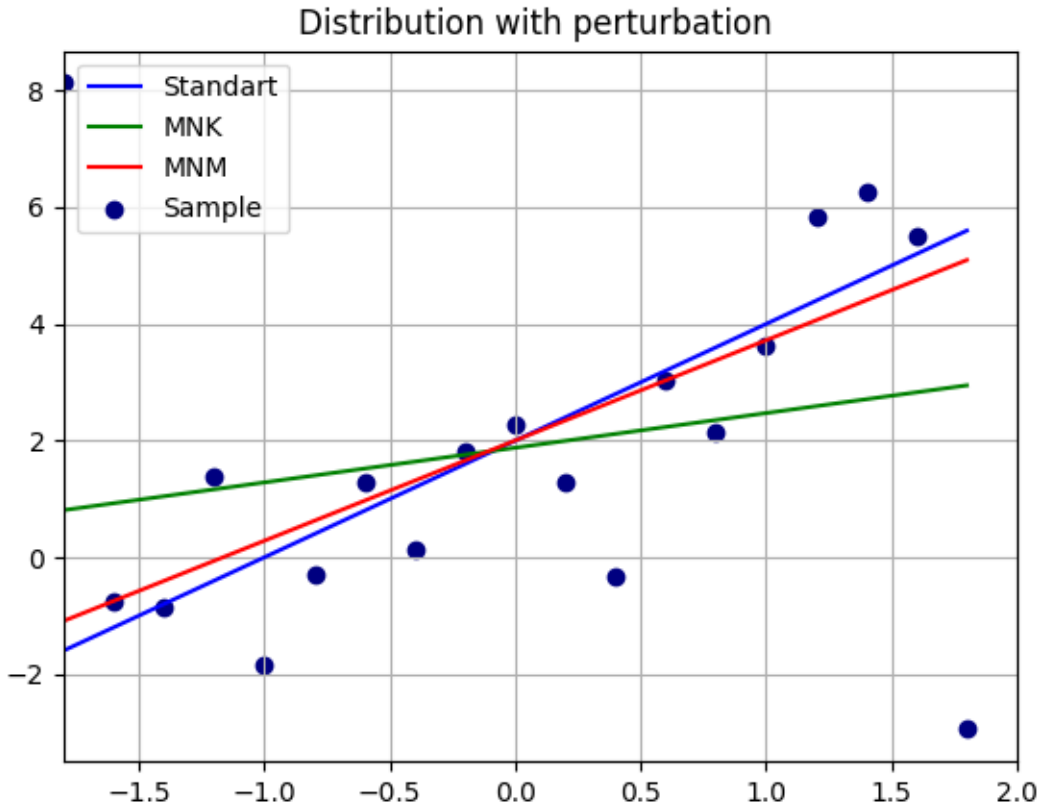

Рисунок 3.5. Выборка с возмущениями