

Санкт-Петербургский  
Политехнический университет Петра Великого

**Отчет по лабораторным работам №1-6  
по дисциплине  
"Математическая статистика"**

Студент:	Скворцов Владимир Сергеевич
Преподаватель:	Баженов Александр Николаевич
Группа:	5030102/10201

Санкт-Петербург  
2024

# Содержание

<b>1</b>	<b>Постановка задачи</b>	<b>2</b>
1.1	Коэффициент корреляции . . . . .	2
1.2	Простая линейная регрессия . . . . .	2
<b>2</b>	<b>Теоретическое обоснование</b>	<b>2</b>
2.1	Двумерное нормальное распределение . . . . .	2
2.2	Корреляционный момент (ковариация) и коэффициент корреляции . . . . .	2
2.3	Выборочный коэффициент корреляции Пирсона . . . . .	3
2.4	Выборочный квадрантный коэффициент корреляции . . . . .	3
2.5	Выборочный коэффициент ранговой корреляции Спирмена . . . . .	3
2.6	Эллипсы рассеивания . . . . .	3
2.7	Метод наименьших квадратов . . . . .	3
2.8	Метод наименьших модулей . . . . .	4
<b>3</b>	<b>Описание работы</b>	<b>4</b>
<b>4</b>	<b>Результаты</b>	<b>4</b>
4.1	Коэффициент корреляции . . . . .	4
4.2	Простая линейная регрессия . . . . .	8
<b>5</b>	<b>Выводы</b>	<b>13</b>
<b>6</b>	<b>Постановка задачи</b>	<b>13</b>
6.1	Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат . . . . .	13
6.2	Проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей . . . . .	13
<b>7</b>	<b>Теоретическое обоснование</b>	<b>14</b>
7.1	Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат . . . . .	14
7.1.1	Правило проверки гипотезы о законе распределения по методу $\chi^2$ . . . . .	14
7.2	Проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей . . . . .	14
7.2.1	Тест Фишера . . . . .	14

# 1 Постановка задачи

## 1.1 Коэффициент корреляции

Сгенерировать двумерные выборки размерами 20, 60, 100 для нормального двумерного распределения  $N(x, y, 0, 0, 1, 1, \rho)$ . Коэффициент корреляции  $\rho$  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 квадратного коэффициента корреляции. Повторить все вычисления для смеси нормальных распределений:

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

Изобразить сгенерированные точки на плоскости и нарисовать эллипс равновероятности.

## 1.2 Простая линейная регрессия

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. Прodelать то же самое для выборки, у которой в значения  $y_1$  и  $y_{20}$  вносятся возмущения 10 и -10.

# 2 Теоретическое обоснование

## 2.1 Двумерное нормальное распределение

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

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

Компоненты  $X, Y$  двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями  $\bar{x}, \bar{y}$  и средними квадратическими отклонениями  $\sigma_x, \sigma_y$  соответственно.

Параметр  $\rho$  называется коэффициентом корреляции.

## 2.2 Корреляционный момент (ковариация) и коэффициент корреляции

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

$$K = \text{cov}(X, Y) = \mathbf{M}[(X - \bar{x})(Y - \bar{y})] \quad (2)$$

Коэффициент корреляции  $\rho$  двух случайных величин  $X$  и  $Y$ :

$$\rho = \frac{K}{\sigma_x\sigma_y} \quad (3)$$

## 2.3 Выборочный коэффициент корреляции Пирсона

Выборочный коэффициент корреляции Пирсона:

$$r = \frac{\frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum (x_i - \bar{x})^2 \frac{1}{n} \sum (y_i - \bar{y})^2}} = \frac{K}{s_X s_Y}, \quad (4)$$

где  $K$ ,  $s_X^2$ ,  $s_Y^2$  — выборочные ковариации и дисперсии случайных величин  $X$  и  $Y$ .

## 2.4 Выборочный квадрантный коэффициент корреляции

Выборочный квадрантный коэффициент корреляции

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n}, \quad (5)$$

где  $n_1, n_2, n_3, n_4$  — количество точек с координатами  $(x_i, y_i)$ , попавшими, соответственно, в I, II, III, IV квадранты декартовой системы с осями  $x' = x - \text{med}x$ ,  $y' = y - \text{med}y$ .

## 2.5 Выборочный коэффициент ранговой корреляции Спирмена

Обозначим ранги, соответствующие значениям переменной  $X$ , через  $u$ , а ранги, соответствующие значениям переменной  $Y$ , — через  $v$ .

Выборочный коэффициент ранговой корреляции Спирмена:

$$r_S = \frac{\frac{1}{n} \sum (u_i - \bar{u})(v_i - \bar{v})}{\sqrt{\frac{1}{n} \sum (u_i - \bar{u})^2 \frac{1}{n} \sum (v_i - \bar{v})^2}}, \quad (6)$$

где  $\bar{u} = \bar{v} = \frac{1+2+\dots+n}{n} = \frac{n+1}{2}$  — среднее значение рангов.

## 2.6 Эллипсы рассеивания

Уравнение проекции эллипса рассеивания на плоскость  $xOy$ :

$$\frac{(x - \bar{x})^2}{\sigma_x^2} - 2\rho \frac{(x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y} + \frac{(y - \bar{y})^2}{\sigma_y^2} = \text{const}. \quad (7)$$

Центр эллипса 8 находится в точке с координатами  $(x, y)$ ; оси симметрии эллипса составляют с осью  $Ox$  углы, определяемые уравнением

$$\text{tg } 2\alpha = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}. \quad (8)$$

## 2.7 Метод наименьших квадратов

При оценивании параметров регрессионной модели используют различные методы. Один из наиболее распространённых 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):

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \rightarrow \min_{\beta_0, \beta_1}. \quad (9)$$

Задача минимизации квадратичного критерия (9) носит название задачи метода наименьших квадратов (МНК), а оценки  $\beta_0$ ,  $\beta_1$  параметров  $\beta_0$ ,  $\beta_1$ , реализующие минимум критерия (9), называют МНК-оценками.

## 2.8 Метод наименьших модулей

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:

$$\sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i| \rightarrow \min_{\beta_0, \beta_1}. \quad (10)$$

## 3 Описание работы

Лабораторные работы выполнены с использованием Python и его сторонних библиотек `numpy`, `pandas`, `matplotlib`, `seaborn` были построены гистограммы распределений и посчитаны характеристики положения.

Ссылка на GitHub репозиторий: <https://github.com/vladimir-skvortsov/spbstu-mathematical-statistics>

## 4 Результаты

### 4.1 Коэффициент корреляции

$n = 20, \rho = 0$			
	$r$ (4)	$r_S$ (6)	$r_Q$ (5)
Среднее	$8.051 \times 10^{-3}$	$8.633 \times 10^{-3}$	$1.216 \times 10^{-2}$
Среднее квадратов	$5.501 \times 10^{-2}$	$5.418 \times 10^{-2}$	$1.033 \times 10^{-1}$
Дисперсия	$5.495 \times 10^{-2}$	$5.410 \times 10^{-2}$	$1.031 \times 10^{-1}$
$n = 20, \rho = 0.5$			
	$r$ (4)	$r_S$ (6)	$r_Q$ (5)
Среднее	$4.933 \times 10^{-1}$	$4.674 \times 10^{-1}$	$4.644 \times 10^{-1}$
Среднее квадратов	$2.743 \times 10^{-1}$	$2.534 \times 10^{-1}$	$3.139 \times 10^{-1}$
Дисперсия	$3.093 \times 10^{-2}$	$3.496 \times 10^{-2}$	$9.823 \times 10^{-2}$
$n = 20, \rho = 0.9$			
	$r$ (4)	$r_S$ (6)	$r_Q$ (5)
Среднее	$8.938 \times 10^{-1}$	$8.646 \times 10^{-1}$	$9.837 \times 10^{-1}$
Среднее квадратов	$8.014 \times 10^{-1}$	$7.527 \times 10^{-1}$	1.026
Дисперсия	$2.454 \times 10^{-3}$	$5.209 \times 10^{-3}$	$5.804 \times 10^{-2}$
$n = 60, \rho = 0$			
	$r$ (4)	$r_S$ (6)	$r_Q$ (5)
Среднее	$8.143 \times 10^{-3}$	$8.747 \times 10^{-3}$	$8.485 \times 10^{-3}$
Среднее квадратов	$1.709 \times 10^{-2}$	$1.689 \times 10^{-2}$	$3.111 \times 10^{-2}$
Дисперсия	$1.703 \times 10^{-2}$	$1.682 \times 10^{-2}$	$3.104 \times 10^{-2}$
$n = 60, \rho = 0.5$			
	$r$ (4)	$r_S$ (6)	$r_Q$ (5)
Среднее	$4.985 \times 10^{-1}$	$4.757 \times 10^{-1}$	$4.668 \times 10^{-1}$
Среднее квадратов	$2.585 \times 10^{-1}$	$2.373 \times 10^{-1}$	$2.504 \times 10^{-1}$
Дисперсия	$1.000 \times 10^{-2}$	$1.094 \times 10^{-2}$	$3.256 \times 10^{-2}$
$n = 60, \rho = 0.9$			
	$r$ (4)	$r_S$ (6)	$r_Q$ (5)
Среднее	$8.979 \times 10^{-1}$	$8.810 \times 10^{-1}$	$9.937 \times 10^{-1}$
Среднее квадратов	$8.069 \times 10^{-1}$	$7.774 \times 10^{-1}$	1.004
Дисперсия	$7.297 \times 10^{-4}$	$1.202 \times 10^{-3}$	$1.700 \times 10^{-2}$
$n = 100, \rho = 0$			
	$r$ (4)	$r_S$ (6)	$r_Q$ (5)
Среднее	$1.396 \times 10^{-3}$	$8.326 \times 10^{-5}$	$1.584 \times 10^{-3}$
Среднее квадратов	$9.856 \times 10^{-3}$	$9.848 \times 10^{-3}$	$1.972 \times 10^{-2}$
Дисперсия	$9.854 \times 10^{-3}$	$9.848 \times 10^{-3}$	$1.972 \times 10^{-2}$
$n = 100, \rho = 0.5$			
	$r$ (4)	$r_S$ (6)	$r_Q$ (5)
Среднее	$5.013 \times 10^{-1}$	$4.812 \times 10^{-1}$	$4.723 \times 10^{-1}$
Среднее квадратов	$2.568 \times 10^{-1}$	$2.375 \times 10^{-1}$	$2.407 \times 10^{-1}$
Дисперсия	$5.481 \times 10^{-3}$	$6.013 \times 10^{-3}$	$1.762 \times 10^{-2}$
$n = 100, \rho = 0.9$			
	$r$ (4)	$r_S$ (6)	$r_Q$ (5)
Среднее	$8.999 \times 10^{-1}$	$8.866 \times 10^{-1}$	1.003
Среднее квадратов	$8.103 \times 10^{-1}$	$7.868 \times 10^{-1}$	1.017
Дисперсия	$4.017 \times 10^{-4}$	$6.665 \times 10^{-4}$	$1.049 \times 10^{-2}$

Таблица 1: Характеристики нормального двумерного распределения

$n = 20$			
	$r$ (4)	$r_S$ (6)	$r_Q$ (5)
Среднее	$-7.987 \times 10^{-2}$	$-7.020 \times 10^{-2}$	$-6.336 \times 10^{-2}$
Среднее квадратов	$5.968 \times 10^{-2}$	$5.944 \times 10^{-2}$	$1.112 \times 10^{-1}$
Дисперсия	$5.330 \times 10^{-2}$	$5.451 \times 10^{-2}$	$1.072 \times 10^{-1}$
$n = 60$			
	$r$ (4)	$r_S$ (6)	$r_Q$ (5)
Среднее	$9.290 \times 10^{-2}$	$-8.988 \times 10^{-2}$	$-8.730 \times 10^{-2}$
Среднее квадратов	$2.606 \times 10^{-2}$	$2.553 \times 10^{-2}$	$4.290 \times 10^{-2}$
Дисперсия	$1.743 \times 10^{-2}$	$1.745 \times 10^{-2}$	$3.528 \times 10^{-2}$
$n = 100$			
	$r$ (4)	$r_S$ (6)	$r_Q$ (5)
Среднее	$-1.013 \times 10^{-1}$	$-9.639 \times 10^{-2}$	$-9.011 \times 10^{-2}$
Среднее квадратов	$2.047 \times 10^{-2}$	$1.984 \times 10^{-2}$	$2.968 \times 10^{-2}$
Дисперсия	$1.021 \times 10^{-2}$	$1.054 \times 10^{-2}$	$2.156 \times 10^{-2}$

Таблица 2: Характеристики смеси нормальных распределений

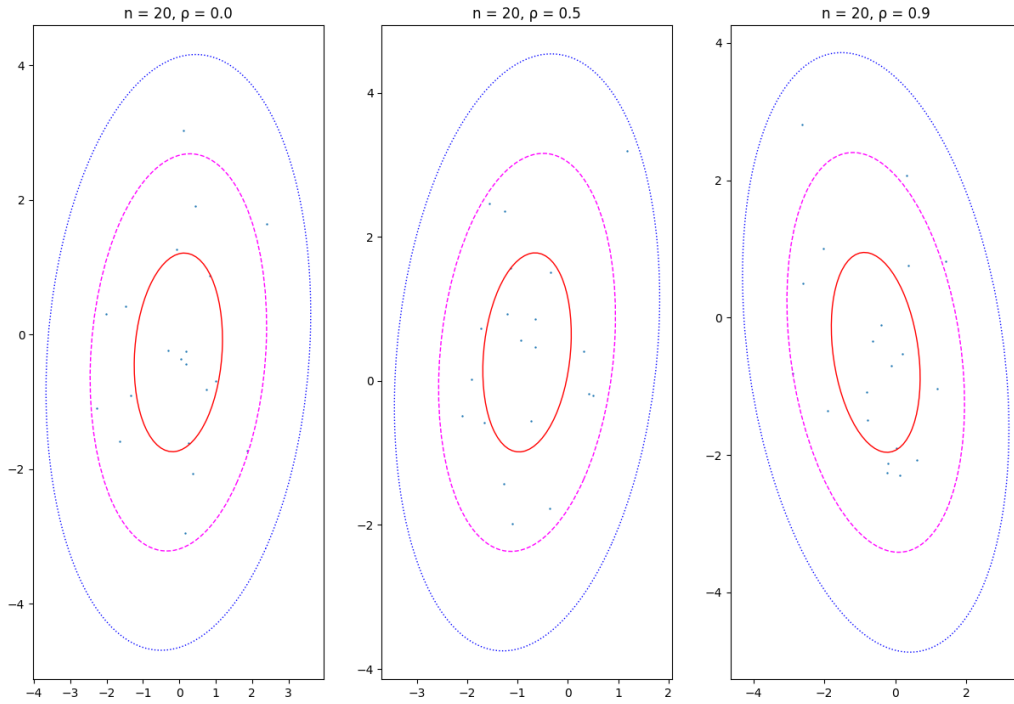


Рис. 1: Смесь нормальных распределений и эллипсы равновероятности (  $n = 20$  )

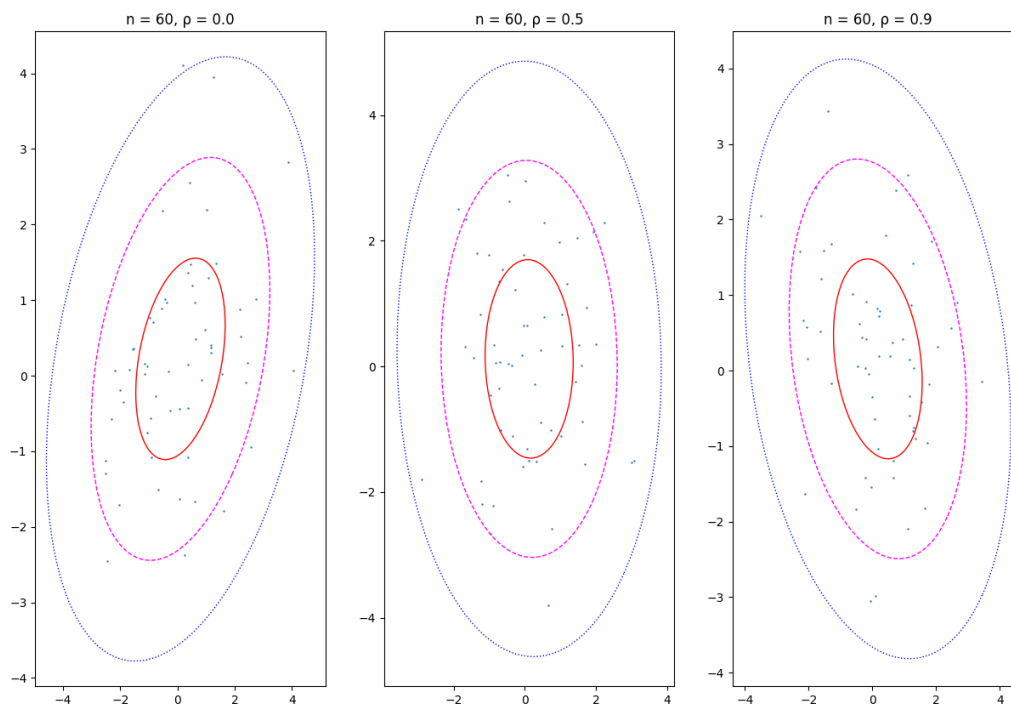


Рис. 2: Смесь нормальных распределений и эллипсы равновероятности (  $n = 60$  )

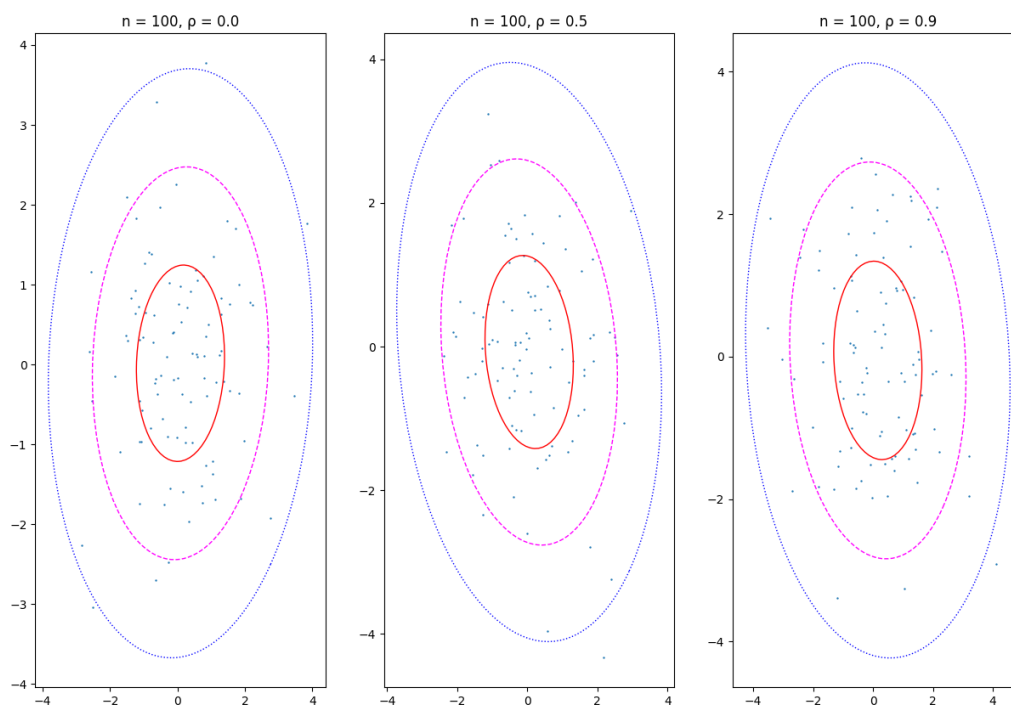


Рис. 3: Смесь нормальных распределений и эллипсы равновероятности (  $n = 100$  )



## 4.2 Простая линейная регрессия

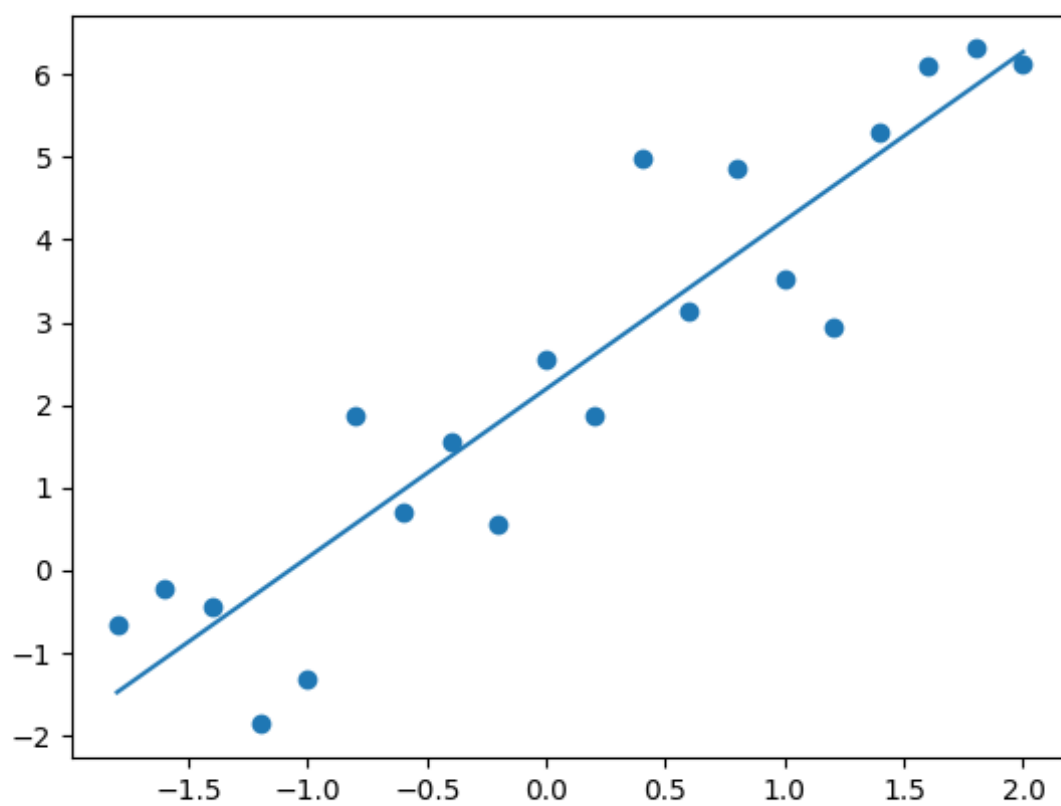


Рис. 4: Метод наименьших квадратов (2.0388, 2.1955)

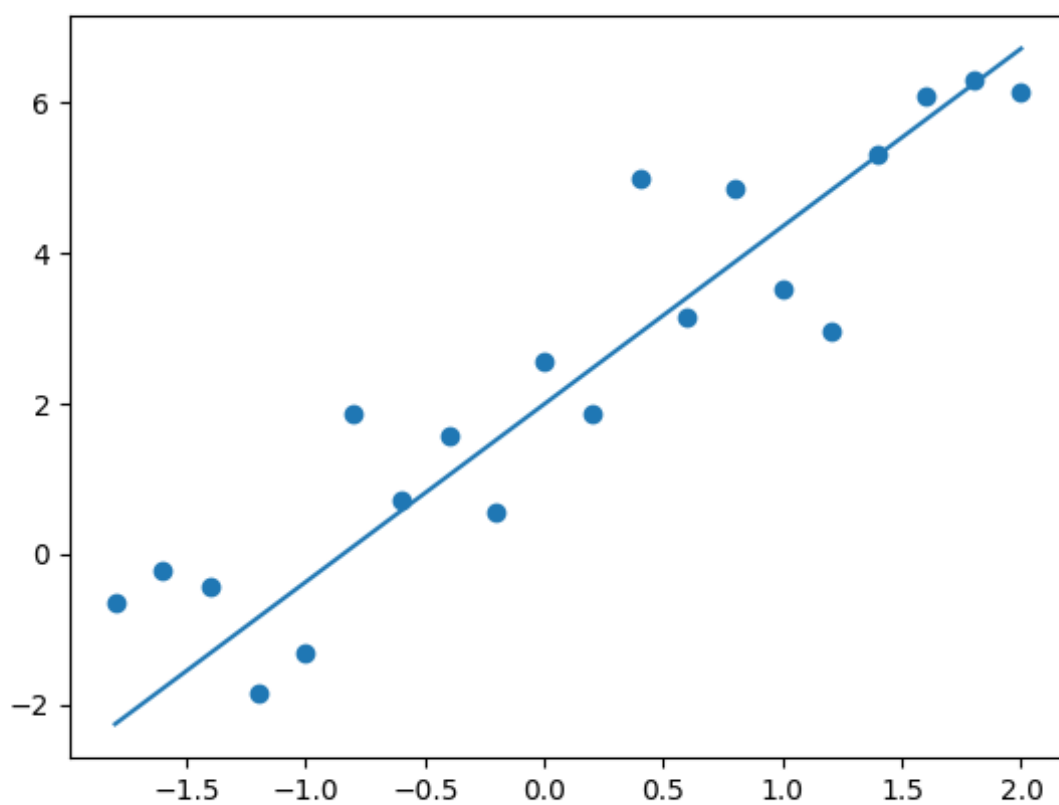


Рис. 5: Метод наименьших модулей (2.3634, 1.9945)

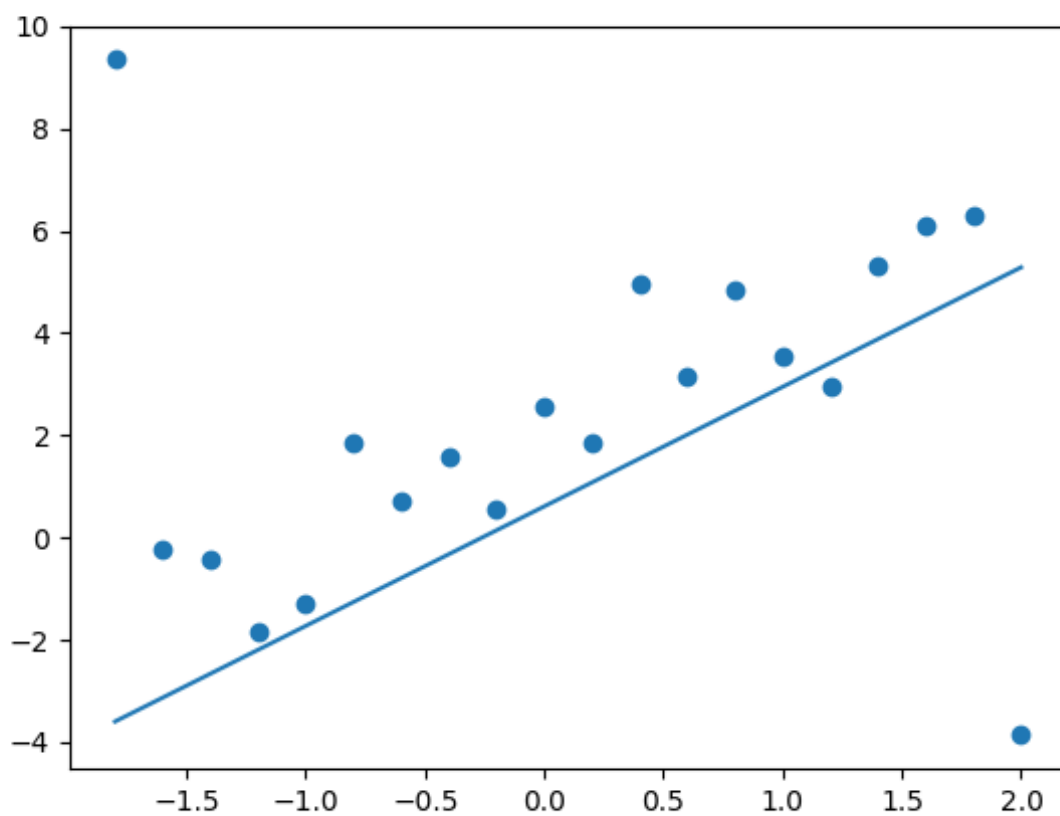


Рис. 6: Метод наименьших квадратов с возмущениями (2.3383, 0.6102)

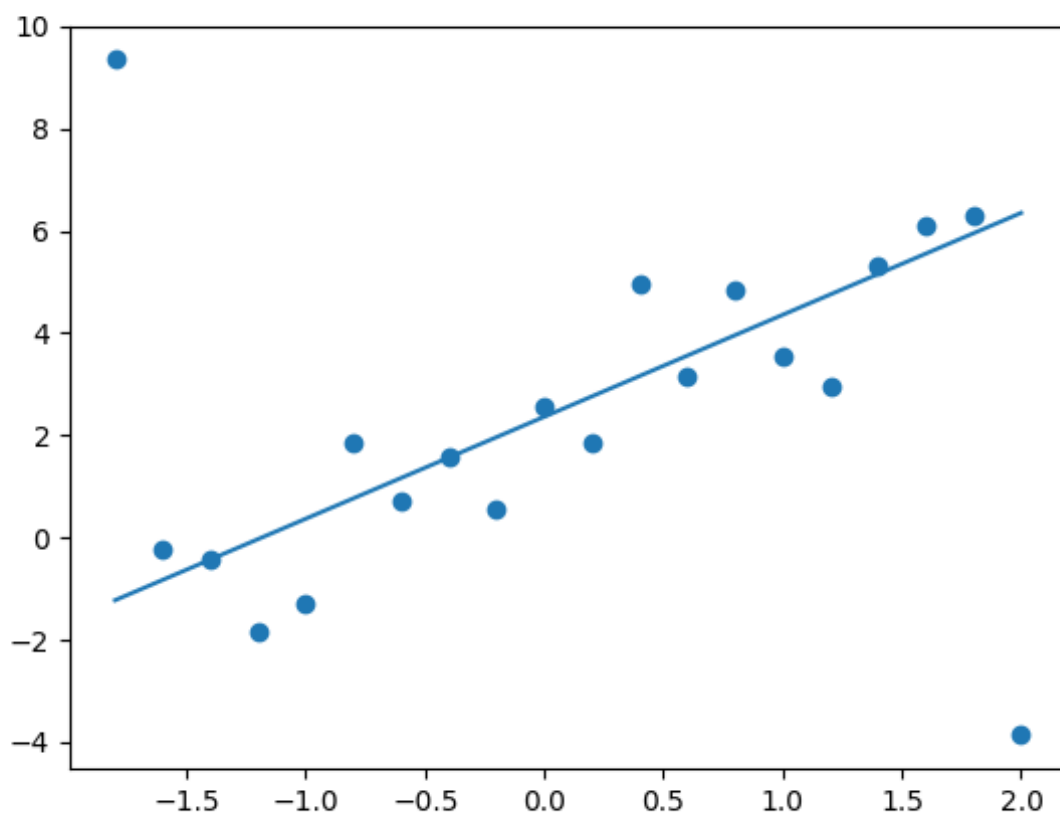


Рис. 7: Метод наименьших модулей с возмущениями (1.9944, 2.3634)

	$a$	$a'$	$b$	$b'$	$\Delta a$	$\Delta b$
МНК	2.0388	2.3383	2.1955	0.6102	1.3383	0.7221
МНМ	2.3634	1.9944	1.9945	2.3634	0.1561	0.1849

Таблица 3: Таблица коэффициентов

Здесь:

- $y = ax + b$  — уравнение линейной регрессии для методов МНК и МНМ без выбросов
- $y = a'x + b'$  — уравнение линейной регрессии для методов МНК и МНМ с выбросами
- $\Delta a = \frac{|a-a'|}{a}$
- $\Delta b = \frac{|b-b'|}{b}$

## 5 Выводы

На основе полученных характеристик (включая среднее значение, среднее значение квадрата и дисперсию) для различных коэффициентов корреляции и размеров выборки, можно сделать следующие наблюдения:

1. При увеличении размера выборки повышается точность оценок, что видно по уменьшению дисперсий коэффициентов корреляции. Это соответствует принципам центральной предельной теоремы и закона больших чисел.
2. При увеличении коэффициента корреляции  $\rho$ , средние значения коэффициентов Пирсона, Спирмена и квадратичного коэффициента корреляции тоже увеличиваются. Это указывает на прямую связь между  $\rho$  и другими коэффициентами корреляции.

Из результатов оценок коэффициентов линейной регрессии при использовании двух критериев (критерий наименьших квадратов и критерий наименьших модулей) можно сделать следующие выводы:

1. Метод наименьших квадратов показал себя эффективно в случае, когда нет значительных выбросов в данных, в то время как метод наименьших модулей проявил себя лучше в присутствии значительных возмущений.
2. Важно выбирать метод, исходя из особенностей данных. Если в данных присутствуют выбросы, метод наименьших модулей будет предпочтительнее из-за его устойчивости к выбросам.

## 6 Постановка задачи

### 6.1 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

Создать распределения согласно нормальному, распределению Стьюдента и равномерному распределению с мощностями выборки  $n = 20, 100$ .

Провести исследование по методу  $\chi^2$ .

### 6.2 Проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей

Мощность нормального распределения  $N = 100$ .

1. Выбрать две выборки мощностью 20 и 40
2. Выбрать две выборки мощностью 20 и 100

Провести исследование по методу теста Фишера для случаев 1 и 2.

## 7 Теоретическое обоснование

### 7.1 Проверка гипотезы о законе распределения генеральной совокупности. Метод хи-квадрат

#### 7.1.1 Правило проверки гипотезы о законе распределения по методу $\chi^2$

1. Выбираем уровень значимости  $\alpha$ ,
2. По таблице находим квантиль  $\chi^2_{1-\alpha}(k-1)$  распределения хи-квадрат с  $k-1$  степенями свободы порядка  $1-\alpha$ ,
3. С помощью гипотетической функции распределения  $F(x)$  вычисляем вероятности  $p_i = P(X \in \Delta_i)$ ,  $i \in \overline{1, k}$ ,
4. Находим частоты  $n_i$  попадания элементов выборки в подмножества  $\Delta_i$ ,  $i \in \overline{1, k}$ .
5. Вычисляем выборочное значение статистики критерия  $\chi^2$ :

$$x_B^2 = \sum_{i=1}^k \frac{(n_i - np_i)^2}{np_i}$$

6. Сравниваем  $\chi_B^2$  и квантиль  $\chi^2_{1-\alpha}(k-1)$ .
  - (а) Если  $\chi_B^2 < \chi^2_{1-\alpha}(k-1)$ , то гипотеза  $H_0$  на данном этапе проверки принимается.
  - (б) Если  $\chi_B^2 \geq \chi^2_{1-\alpha}(k-1)$ , то гипотеза  $H_0$  отвергается, выбирается одно из альтернативных распределений, и процедура проверки повторяется.

### 7.2 Проверка гипотезы о равенстве дисперсий двух нормальных генеральных совокупностей

Несмещенные оценки дисперсий:

$$s_X^2 = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})^2; \quad s_Y^2 = \frac{1}{n-1} \sum_{i=1}^m (y_i - \bar{y})^2; \quad (11)$$

Статистика критерия Фишера:

$$F = s_X^2 / s_Y^2 \quad (12)$$

#### 7.2.1 Тест Фишера

1. Вычисляем несмещенные оценки дисперсий (11),
2. Выбираем статистику критерия (12),
3. Выбираем уровень значимости  $\alpha$ ,
4. По таблице квантиль  $F_{1-\frac{\alpha}{2}}(k_1, k_2)$  распределения Фишера.
5. Вычисляем выборочное значение  $F_V$  статистики критерия.
6. Сравниваем  $F_B$  и  $F_{1-\frac{\alpha}{2}}(k_1, k_2)$ . Если  $F_B < F_{1-\frac{\alpha}{2}}(k_1, k_2)$ , то гипотеза  $H_0$  на выбранном уровне значимости  $\alpha$  принимается. В противном случае — отвергается.