Санкт-Петербургский политехнический университет Петра Великого

Институт прикладной математики и механики

Кафедра «Телематика (при ЦНИИ РТК)»

Отчет по лабораторным работам № 5, 6

По дисциплине «Теория вероятностей и Математическая статистика»

Выполнил
Студент гр. 3630201/80101 _____ Печеный Н. А.

Руководитель
к.ф.-м.н., доцент _____ Баженов А. Н.

«___»_____ 2021г.

Санкт-Петербург
2021

Содержание

Список иллюстраций

Список таблиц

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:

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

   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 Теория

2.1 Двумерное нормальное распределение

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

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

Компоненты X, Y двумерной нормальной случайной величины также распределены нормально с математическими ожиданиями $\overline{x}, \overline{y}$ и средними квадратическими отклонениями σ_x, σ_y соответственно [1]. Параметр ρ называется коэффициентом корреляции.

2.2 Корреляционный момент (ковариация) и коэффициент корреляции

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

$$K = cov(X, Y) = M\left[(X - \overline{x})(Y - \overline{y})\right] \quad (3)$$

Коэффициент корреляции ρ двух случайных величин X и Y:

$$\rho = \frac{K}{\sigma_x\sigma_y} \quad (4)$$

2.3 Выборочные коэффициенты корреляции

2.3.1 Выборочный коэффициент корреляции Пирсона

Выборочный коэффициент корреляции Пирсона [1]:

$$r = \frac{\frac{1}{n}\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum(x_i - \overline{x})^2\frac{1}{n}\sum(y_i - \overline{y})^2}} = \frac{K}{s_X s_Y}, \quad (5)$$

где K, s_X^2, s_Y^2 – выборочные ковариация и дисперсии с.в. X и Y.

2.3.2 Выборочный квадрантный коэффициент корреляции

Выборочный квадрантный коэффициент корреляции [1]:

$$r_Q = \frac{(n_1 + n_3) - (n_2 + n_4)}{n}, \quad (6)$$

где n_1, n_2, n_3, n_4 – количества точек с координатами (x_i, y_i), попавшими соответственно в I, II, III и IV квадранты декартовой системы с осями $x' = x - med\ x, y' = y - med\ y$ и с центром в точке с координатами $(med\ x, med\ y)$.

2.3.3 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 [1]:

$$r_S = \frac{\frac{1}{n}\sum(u_i - \overline{u})(v_i - \overline{v})}{\sqrt{\frac{1}{n}\sum(u_i - \overline{u})^2 \frac{1}{n}\sum(v_i - \overline{v})^2}}, \tag{7}$$

где $\overline{u} = \overline{v} = \frac{1+2+\ldots+n}{n} = \frac{n+1}{2}$ – среднее значение рангов.

2.4 Эллипсы рассеивания

Уравнение проекции эллипса рассеивания на плоскость xOy:

$$\frac{(x-\overline{x})^2}{\sigma_x^2} - 2\rho\frac{(x-\overline{x})(y-\overline{y})}{\sigma_x\sigma_y} + \frac{(y-\overline{y})^2}{\sigma_y^2} = const. \tag{8}$$

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

$$tg2\alpha = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}. \tag{9}$$

2.5 Простая линейная регрессия

2.5.1 Модель простой линейной регрессии

Регрессионную модель описания данных называют простой линейной регрессией, если

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{10}$$

где $x_1, \ldots, x_n$ — заданные числа (значения фактора); $y_1, \ldots, y_n$ — наблюдаемые значения отклика; $\varepsilon_1, \ldots, \varepsilon_n$ — 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.

2.5.2 Метод наименьших квадратов

Метод наименьших квадратов (МНК) [1]:

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \to \min_{\beta_0, \beta_1}. \tag{11}$$

2.5.3 Расчётные формулы для МНК-оценок

МНК-оценки параметров β_0 и β_1 [1]:

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} \tag{12}$$

$$\hat{\beta}_0 = \overline{y} - \overline{x}\hat{\beta}_1 \tag{13}$$

2.6 Робастные оценки коэффициентов линейной регрессии

Метод наименьших модулей [1]:

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \to \min_{\beta_0,\beta_1}. \tag{14}$$

$$\hat{\beta}_{1R} = r_Q \frac{q_y^*}{q_x^*}, \tag{15}$$

$$\hat{\beta}_{0R} = med\ y - \hat{\beta}_{1R} med\ x, \tag{16}$$

$$r_Q = \frac{1}{n} \sum_{i=1}^{n} sgn(x_i - med\ x) sgn(y_i - med\ y), \tag{17}$$

$$q_y^* = \frac{y_j - y_l}{k_q(n)}, \qquad q_x^* = \frac{x_j - x_l}{k_q(n)} \tag{18}$$

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

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

$$sgn\ z = \begin{cases} 1 \text{ при } z > 0, \\ 0 \text{ при } z = 0, \\ -1 \text{ при } z < 0. \end{cases}$$

Уравнение регрессии здесь имеет вид

$$y = \hat{\beta}_{0R} + \hat{\beta}_{1R} x. \tag{19}$$

3 Реализация

Расчёты проводились в среде аналитических вычислений Maxima. Для генерации выборок и создания и отрисовки графиков были использованы библиотечные функции среды разработки. Код скрипта представлен в репозитории на GitHub, ссылка на репозиторий находится в **Приложении А**.

4 Результаты

Ниже, в таблицах 1 - 3 представлены выборочные коэффициенты корреляции Пирсона, Спирмена и квадрантный коэффициент корреляции для выборок размером 20, 60 и 100 элементов двумерного нормального распределения $N(x, y, 0, 0, 1, 1, \rho)$ с коэффициентами корреляции $\rho = 0, 0.5, 0.9$.

Таблица 1: Выборочные коэффициенты корреляции для двумерного нормального распределения, $\rho = 0$

$\rho = 0$	r (5)			r_Q (6)			r_S (7)		
	$E(z)$	$E(z^2)$	$D(z)$	$E(z)$	$E(z^2)$	$D(z)$	$E(z)$	$E(z^2)$	$D(z)$
$N = 20$	0.005	0.053	0.053	0.006	0.053	0.053	0.007	0.052	0.052
$N = 60$	0.001	0.018	0.018	0.004	0.017	0.017	0.004	0.018	0.017
$N = 100$	0.003	0.011	0.011	0.001	0.01	0.01	0.004	0.011	0.011

Таблица 2: Выборочные коэффициенты корреляции для двумерного нормального распределения, $\rho = 0.5$

$\rho = 0.5$	r			r_Q			r_S		
	$E(z)$	$E(z^2)$	$D(z)$	$E(z)$	$E(z^2)$	$D(z)$	$E(z)$	$E(z^2)$	$D(z)$
$N = 20$	0.5	0.3	0.03	0.324	0.155	0.05	0.46	0.24	0.035
$N = 60$	0.5	0.26	0.01	0.32	0.12	0.016	0.5	0.24	0.01
$N = 100$	0.5	0.26	0.005	0.33	0.12	0.009	0.5	0.24	0.006

Таблица 3: Выборочные коэффициенты корреляции для двумерного нормального распределения, $\rho = 0.9$

$\rho = 0.9$	r			r_Q			r_S		
	$E(z)$	$E(z^2)$	$D(z)$	$E(z)$	$E(z^2)$	$D(z)$	$E(z)$	$E(z^2)$	$D(z)$
$N = 20$	0.9	0.8	0.003	0.6916	0.5064	0.03	0.9	0.75	0.005
$N = 60$	0.89	0.8	0.0007	0.7	0.5	0.009	0.89	0.78	0.001
$N = 100$	0.9	0.8	0.0004	0.7042	0.5	0.005	0.9	0.8	0.0006

Ниже в таблице 4 представлены выборочные коэффициенты корреляции Пирсона, Спирмена и квадрантный коэффициент корреляции для выборок смеси двумерных нормальных распределений (1) размером 20, 60 и 100 элементов.

Таблица 4: Выборочные коэффициенты корреляции для смеси двумерных нормальных распределений

	r			r_Q			r_S		
	$E(z)$	$E(z^2)$	$D(z)$	$E(z)$	$E(z^2)$	$D(z)$	$E(z)$	$E(z^2)$	$D(z)$
$N = 20$	-0.35	0.57	0.45	0.5332	0.32	0.036	0.47	0.3	0.08
$N = 60$	-0.65	0.5	0.08	0.56	0.32	0.01	0.5	0.25	0.026
$N = 100$	-0.7	0.52	0.03	0.56	0.32	0.007	0.47	0.23	0.02

Ниже на рисунках 1 - 3 представлены эллипсы рассеивания для выборок двумерного нормального распределения размером 20, 60 и 100 элементов, синим цветом обозначены элементы выборок, эллипсы построены согласно формуле (8).

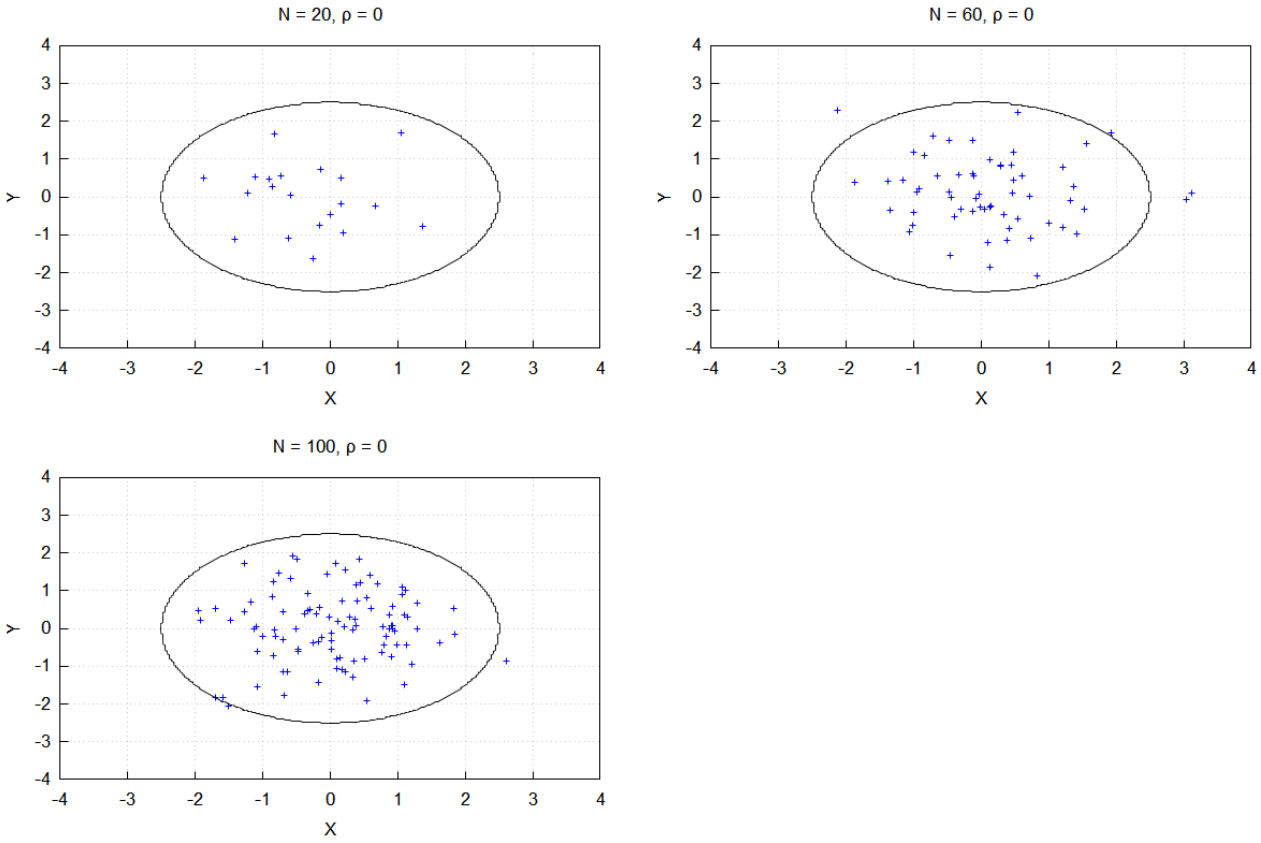

Рис. 1: Эллипсы рассеивания для выборок нормального распределения, $\rho = 0$

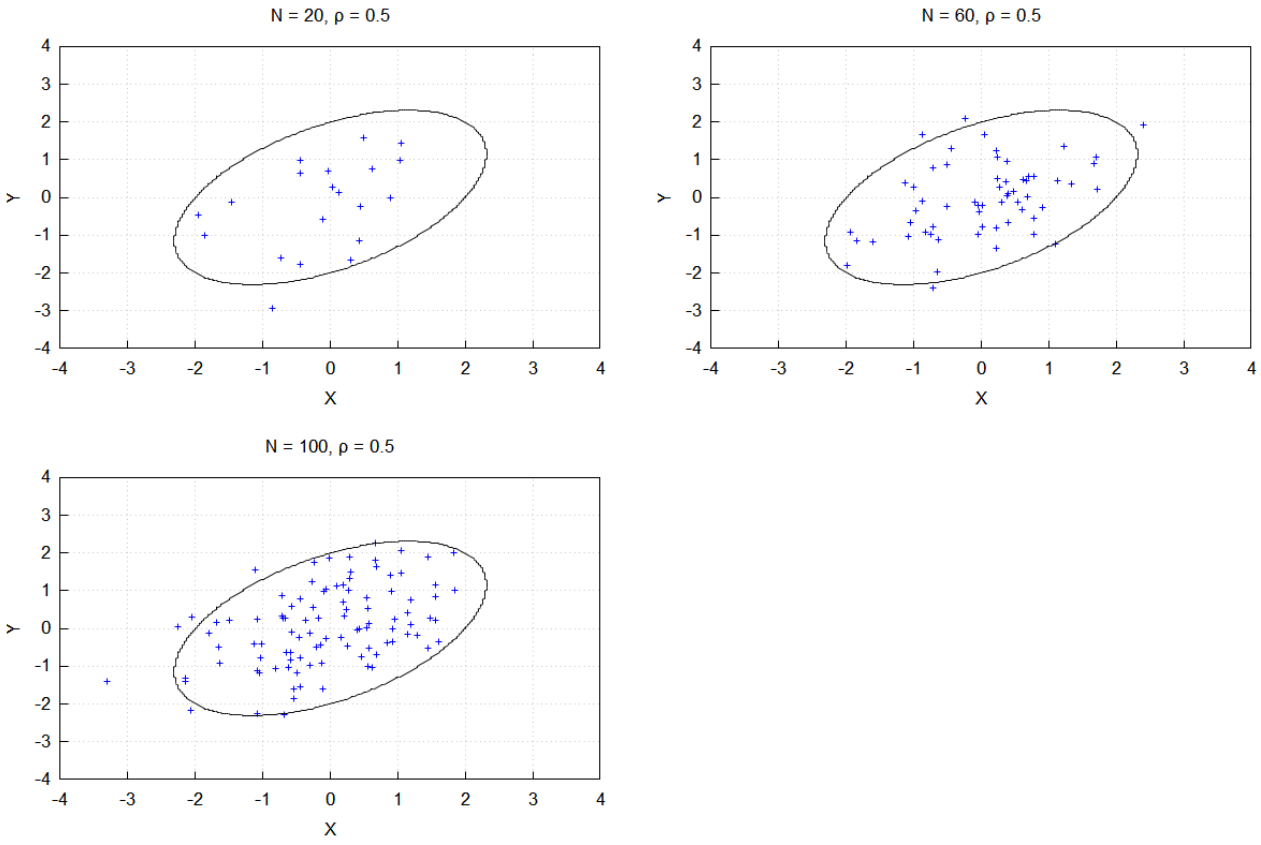

Рис. 2: Эллипсы рассеивания для выборок нормального распределения, $\rho = 0.5$

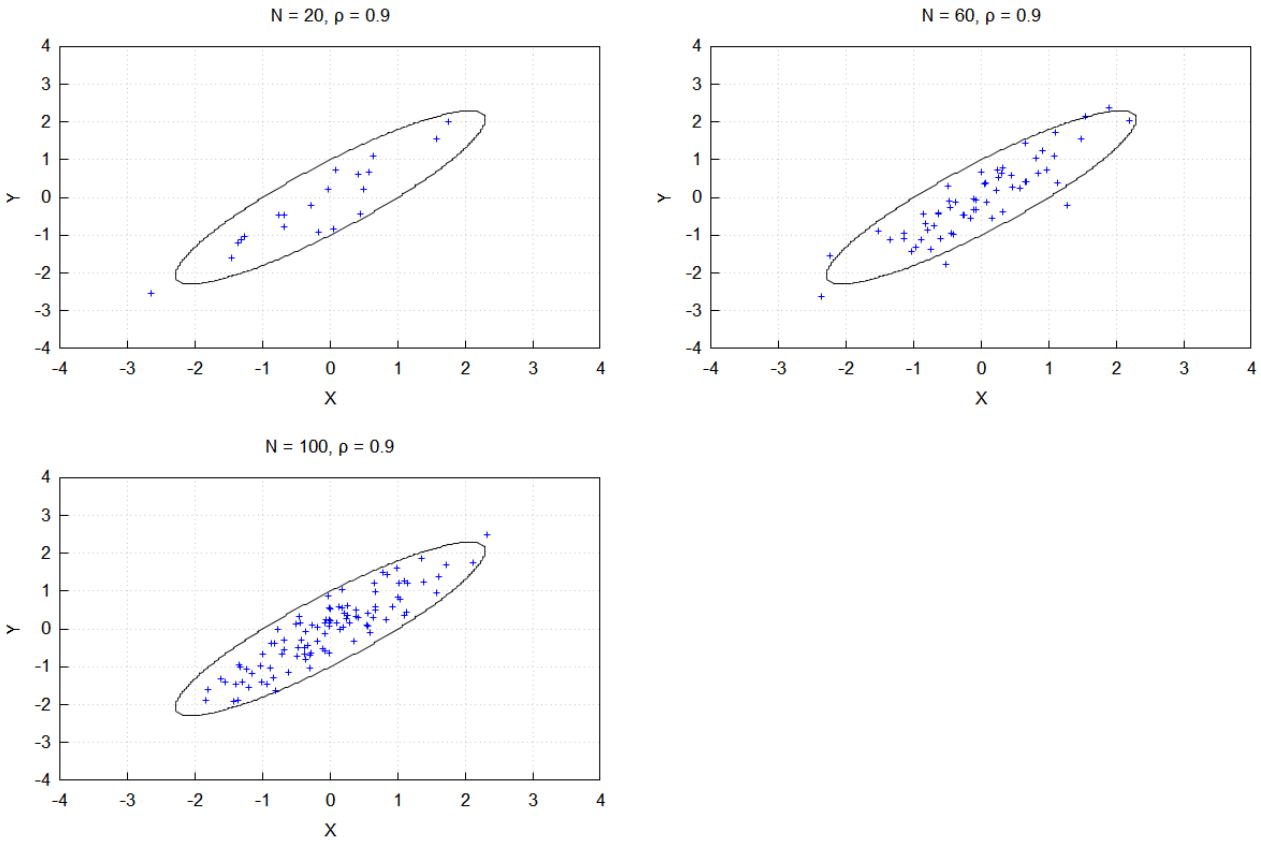

Рис. 3: Эллипсы рассеивания для выборок нормального распределения, $\rho = 0.9$

4.1 Оценки коэффициентов линейной регрессии

4.1.1 Выборка без возмущений

- Критерий наименьших квадратов (12 - 13):

$$\hat{\beta}_0 \approx 1.903, \quad \hat{\beta}_1 \approx 2.037$$

- Критерий наименьших модулей (15 - 16):

$$\hat{\beta}_{0R} \approx 2.239, \quad \hat{\beta}_{1R} \approx 1.612$$

На рисунке 4 представлена выборка без возмущений, график теоретической модели, а также графики, соответствующие линейным регрессиям с коэффициентами, вычисленными согласно методам наименьших квадратов и наименьших модулей.

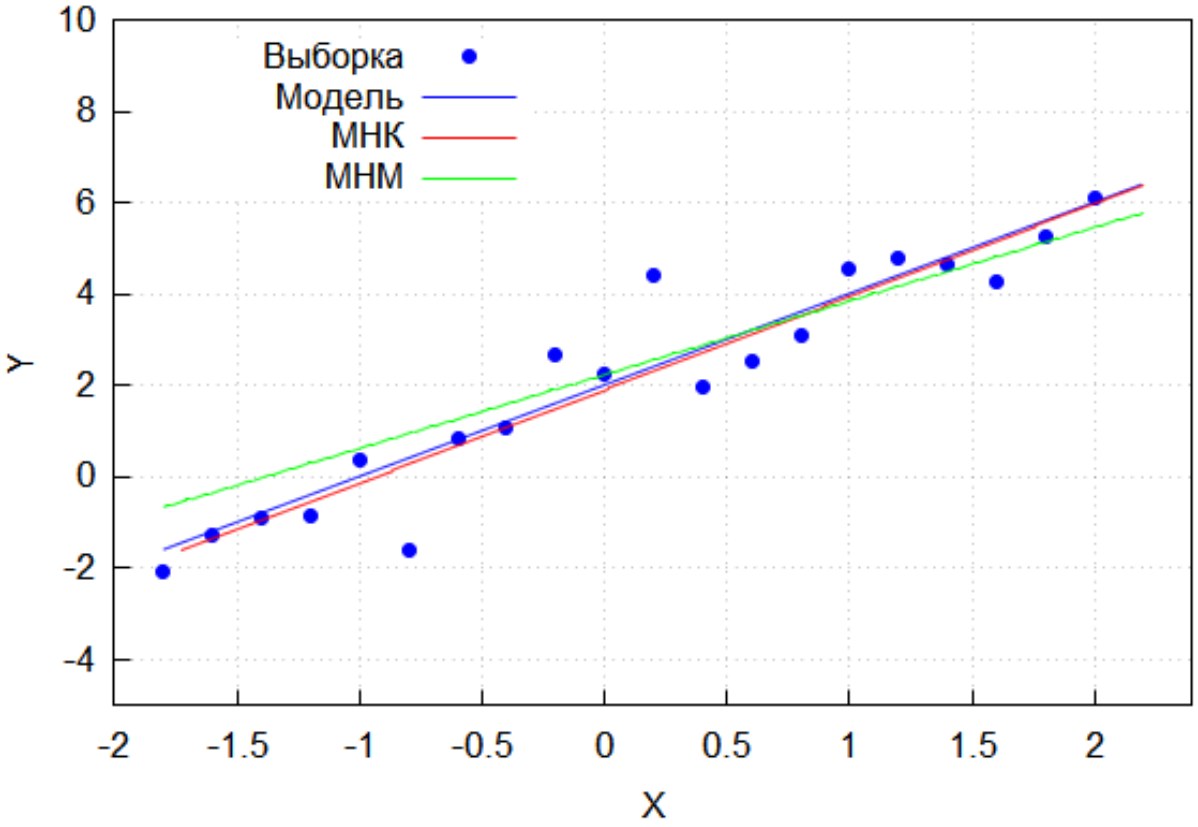

Рис. 4: Выборка без возмущений

4.1.2 Выборка с возмущениями

- Критерий наименьших квадратов (12 - 13):

$$\hat{\beta}_0 \approx 2.046, \quad \hat{\beta}_1 \approx 0.608$$

- Критерий наименьших модулей (15 - 16):

$$\hat{\beta}_{0R} \approx 2.279, \quad \hat{\beta}_{1R} \approx 1.209$$

На рисунке 5 представлена выборка с возмущениями, график теоретической модели, а также графики, соответствующие линейным регрессиям с коэффициентами, вычисленными согласно методам наименьших квадратов и наименьших модулей.

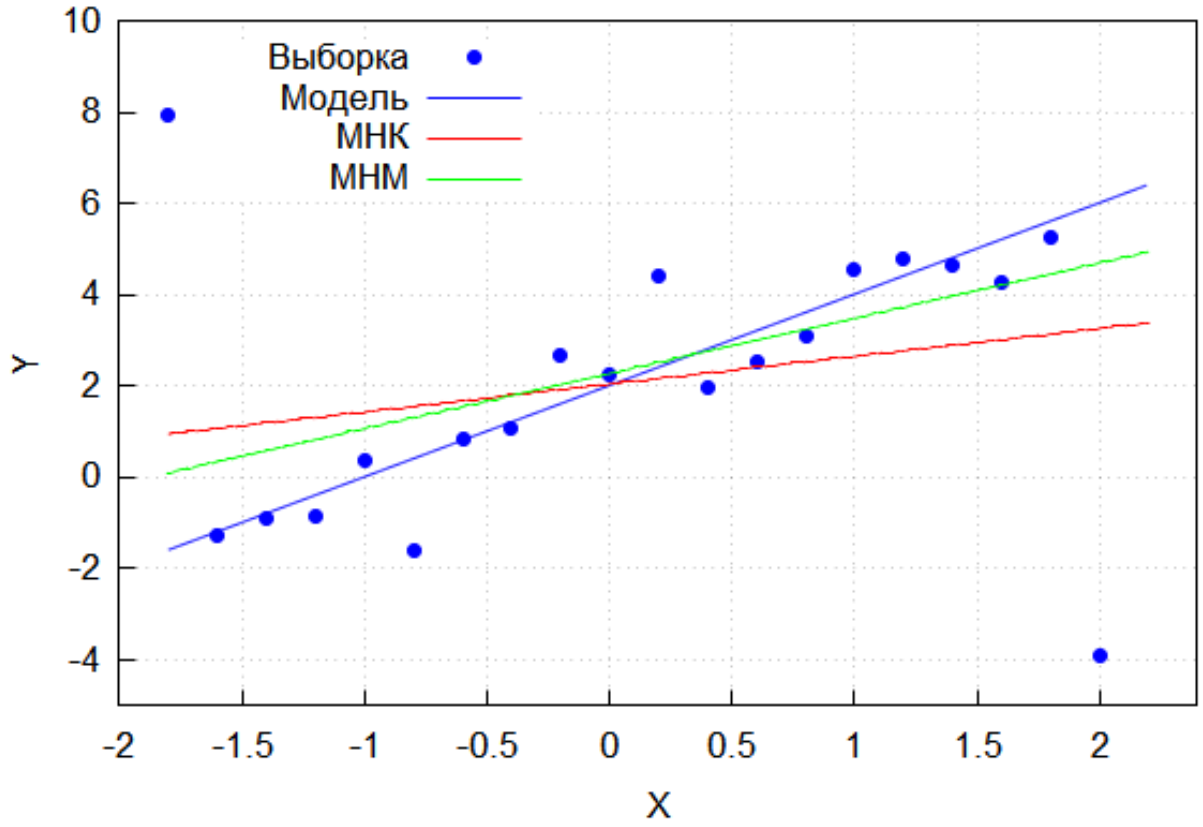

Рис. 5: Выборка с возмущениями

Заключение

В ходе выполнения лабораторной работы были построены выборки двумерного нормального распределения с коэффициентами корреляции $\rho = 0, 0.5, 0.9$ и выборки смеси нормальных распределений. Исходя из оценок выборочных коэффициентов корреляции можно сделать вывод, что квадрантный выборочный коэффициент имеет наибольшее отклонение от теоретического значения коэффициента корреляции, но с увеличением мощности выборки все выборочные коэффициенты корреляции стремятся к своим теоретическим значениям.

Из графиков двумерных нормальных распределений видно, что чем больше коэффициент корреляции, тем больше сужается эллипс рассеивания, в пределе вырождаясь в прямую, при модуле коэффициента корреляции равном единице.

Также в ходе выполнения работы были построены линейные регрессии для нормально распределенной выборки. Коэффициенты регрессии вычислялись с помощью методов наименьших квадратов и наименьших модулей.

Исходя из полученных результатов можно сделать вывод о том, что метод наименьших квадратов позволяет более точно определить коэффициенты регрессии, однако этот метод не устойчив к возмущениям в выборке. С другой стороны метод наименьших модулей даёт менее точную оценку, однако, являясь робастным, он гораздо устойчивее к возмущениям и хорошо подходит для работы с выборками, имеющими сильные отклонения в отдельных точках.

Список литературы

[1] Теоретическое приложение к лабораторным работам №5-8 по дисциплине «Математическая статистика». – СПб.: СПбПУ, 2020. – 22 с

Приложение А. Репозиторий с исходным кодом

Ссылка на репозиторий GitHub с исходным кодом: `https://github.com/pchn/TeorVer`