Санкт-Петербургский
Политехнический университет Петра Великого

Отчет по лабораторным работам №5-6
по дисциплине
"Математическая статистика"

Построение гистограм различных вероятностных распределений и получение оценок положения и оценок рассеяния соответсвующих распределений

Студент: Белоус Фёдор Васильевич
Преподаватель: Баженов Александр Николаевич

Группа: 5030102/10101

Санкт-Петербург
2024

Содержание

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 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 корреля- ции Пирсона, Спирмена и квадрантного коэффициента корреляции. Повторить все вычисления для смеси нормальных распределений:

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

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. Проделать 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 x, y и средними квадратическими отклонениями σ_x, σ_y соответственно.

Параметр ρ называется коэффициентом корреляции.

2.2 Корреляционный момент (ковариация) и коэффициент корреляции

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

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

Коэффициент корреляции ρ 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:

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

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

2.4 Выборочный квадрантный коэффициент корреляции

Выборочный квадрантный коэффициент корреляции

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

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 $x' = x - \mathbf{med}x$, $y' = y - \mathbf{med}y$.

2.5 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:

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

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

2.6 Эллипсы рассеивания

Рассмотрим поверхность распределения, изображающую функцию (1). Она имеет вид холма, вершина которого находится над точкой $(\bar{x}, \bar{y})$.
В сечении поверхности распределения плоскостями, параллельными оси $N(x, y, \bar{x}, \bar{y}, \sigma_x, \sigma_y, \rho)$, получаются кривые, подобные нормальным кривым распределения. В сечении поверхности распределения плоскостями, параллельными плоскости xOy, получаются эллипсы. Напишем уравнение проекции такого эллипса на плоскость xOy:

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

Уравнение эллипса 7 можно проанализировать обычными методами аналитической геометрии. Применяя их, убеждаемся, что центр эллипса 7 находится в точке с координатами $(\bar{x}, \bar{y})$; что касается направления осей симметрии эллипса, то они составляют с осью Ox углы, определяемые уравнением

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

Это уравнение дает два значения углов: α и α_1, различающиеся на $\frac{\pi}{2}$.
Таким образом, ориентация эллипса 7 относительно координатных осей находится в

прямой зависимости от коэффициента корреляции ρ системы (X, Y); если величины не коррелированны (т.е. в данном случае и независимы), то оси симметрии эллипса параллельны координатным осям; в противном случае они составляют с координатными осями некоторый угол.
Пересекая поверхность распределения плоскостями, параллельными плоскости xOy, и проектируя сечения на плоскость xOy мы получим целое семейство подобных и одинаково расположенных эллипсов с общим центром $(\bar{x}, \bar{y})$. Во всех точках каждого из таких эллипсов плотность распределения $N(x, y, \bar{x}, \bar{y}, \sigma_x, \sigma_y, \rho)$ постоянна. Поэтому такие эллипсы называются эллипсами равной плотности или, короче эллипсами рассеивания. Общие оси всех эллипсов рассеивания называются главными осями рассеивания.

2.7 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):

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

Задача минимизации квадратичного критерия (9) носит название задачи метода наименьших квадратов (МНК), а оценки β_0, β_1 параметров β_0, β_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:

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \to \min_{\beta_0, \beta_1}. \tag{10}$$

3 Описание работы

Лабораторные работы выполнены с использованием Python и сторонних библиотек `numpy`, `pandas`, `matplotlib`, `seaborn`.

Ссылка на GitHub репозиторий:
https://github.com/feodorrussia/Mathematical-statistics/tree/master/Lab_3

4 Результаты

4.1 Коэффициент корреляции

$n = 20,\ \rho = 0$	r (4)	r_S (6)	r_Q (5)
Среднее	$6.16 * 10^{-3}$	$4.84 * 10^{-3}$	$5.66 * 10^{-4}$
Среднее квадратов	$4.98 * 10^{-2}$	$4.97 * 10^{-2}$	$1.01 * 10^{-1}$
Дисперсия	$4.98 * 10^{-2}$	$4.97 * 10^{-2}$	$1.01 * 10^{-1}$
$n = 20,\ \rho = 0.5$	r (4)	r_S (6)	r_Q (5)
Среднее	$4.92 * 10^{-1}$	$4.64 * 10^{-1}$	$4.66 * 10^{-1}$
Среднее квадратов	$2.72 * 10^{-1}$	$2.49 * 10^{-1}$	$3.15 * 10^{-1}$
Дисперсия	$2.99 * 10^{-2}$	$3.33 * 10^{-2}$	$9.72 * 10^{-2}$
$n = 20,\ \rho = 0.9$	r (4)	r_S (6)	r_Q (5)
Среднее	$8.95 * 10^{-1}$	$8.66 * 10^{-1}$	$9.77 * 10^{-1}$
Среднее квадратов	$8.04 * 10^{-1}$	$7.54 * 10^{-1}$	$1.026 * 10^{0}$
Дисперсия	$2.45 * 10^{-3}$	$5.21 * 10^{-3}$	$5.80 * 10^{-2}$
$n = 60,\ \rho = 0$	r (4)	r_S (6)	r_Q (5)
Среднее	$7.00 * 10^{-4}$	$2.00 * 10^{-4}$	$2.83 * 10^{-4}$
Среднее квадратов	$1.79 * 10^{-2}$	$1.73 * 10^{-2}$	$3.37 * 10^{-2}$
Дисперсия	$1.79 * 10^{-2}$	$1.73 * 10^{-2}$	$3.37 * 10^{-2}$
$n = 60,\ \rho = 0.5$	r (4)	r_S (6)	r_Q (5)
Среднее	$4.91 * 10^{-1}$	$4.69 * 10^{-1}$	$4.62 * 10^{-1}$
Среднее квадратов	$2.59 * 10^{-1}$	$2.37 * 10^{-1}$	$2.51 * 10^{-1}$
Дисперсия	$1.00 * 10^{-2}$	$1.09 * 10^{-2}$	$3.26 * 10^{-2}$
$n = 60,\ \rho = 0.9$	r (4)	r_S (6)	r_Q (5)
Среднее	$8.98 * 10^{-1}$	$8.81 * 10^{-1}$	$9.94 * 10^{-1}$
Среднее квадратов	$8.07 * 10^{-1}$	$7.77 * 10^{-1}$	$1.004 * 10^{0}$
Дисперсия	$7.30 * 10^{-4}$	$1.20 * 10^{-3}$	$1.70 * 10^{-2}$
$n = 100,\ \rho = 0$	r (4)	r_S (6)	r_Q (5)
Среднее	$5.12 * 10^{-3}$	$3.51 * 10^{-3}$	$3.677 * 10^{-3}$
Среднее квадратов	$1.04 * 10^{-2}$	$1.03 * 10^{-2}$	$2.08 * 10^{-2}$
Дисперсия	$1.04 * 10^{-2}$	$1.03 * 10^{-2}$	$2.08 * 10^{-2}$
$n = 100,\ \rho = 0.5$	r (4)	r_S (6)	r_Q (5)
Среднее	$5.01 * 10^{-1}$	$4.81 * 10^{-1}$	$4.72 * 10^{-1}$
Среднее квадратов	$2.57 * 10^{-1}$	$2.38 * 10^{-1}$	$2.407 * 10^{-1}$
Дисперсия	$5.48 * 10^{-3}$	$6.01 * 10^{-3}$	$1.76 * 10^{-2}$
$n = 100,\ \rho = 0.9$	r (4)	r_S (6)	r_Q (5)
Среднее	$9.00 * 10^{-1}$	$8.87 * 10^{-1}$	$1.00 * 10^{0}$
Среднее квадратов	$8.10 * 10^{-1}$	$7.87 * 10^{-1}$	$1.02 * 10^{0}$
Дисперсия	$4.02 * 10^{-4}$	$6.67 * 10^{-4}$	$1.05 * 10^{-2}$

Таблица 1: Характеристики нормального двумерного распределения

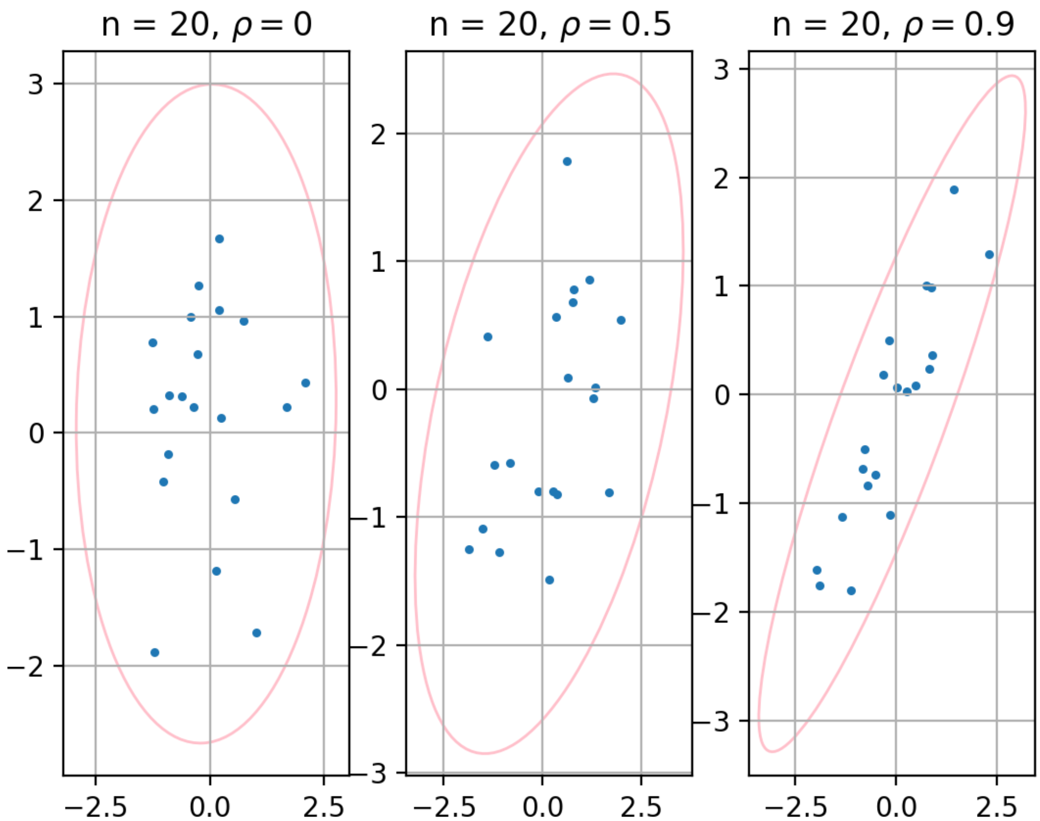

Рис. 1: Эллипсы равновероятности для выборки размером 20

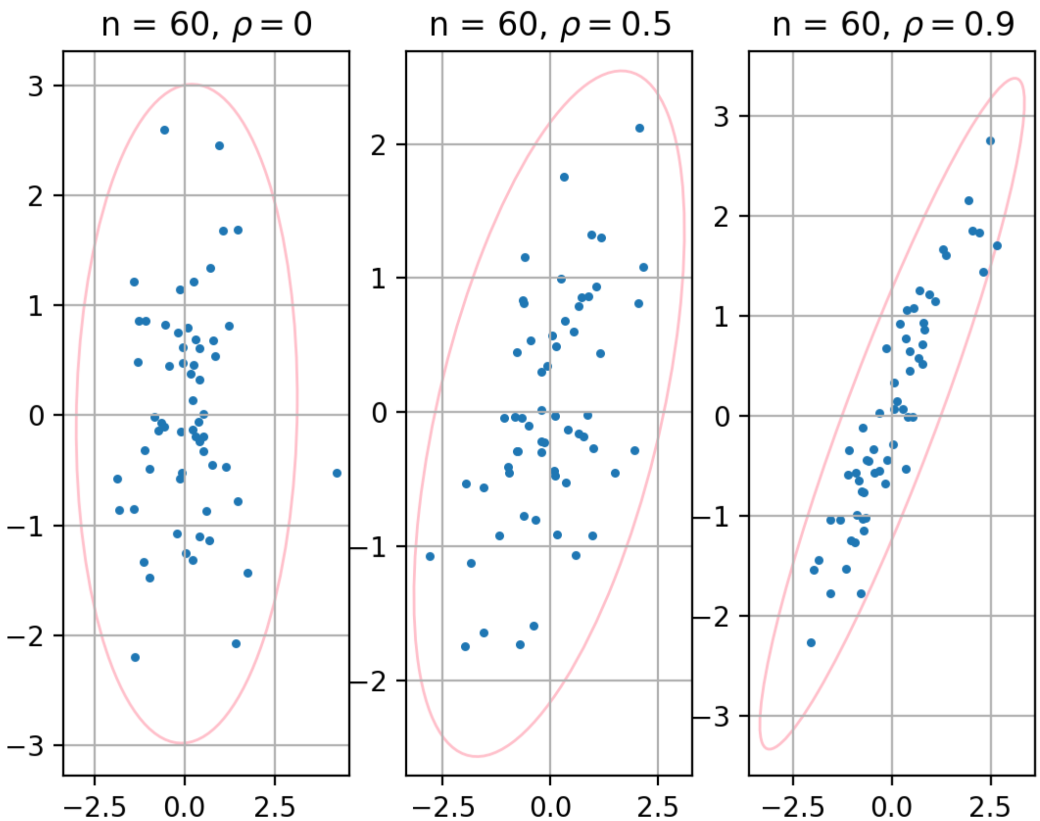

Рис. 2: Эллипсы равновероятности для выборки размером 60

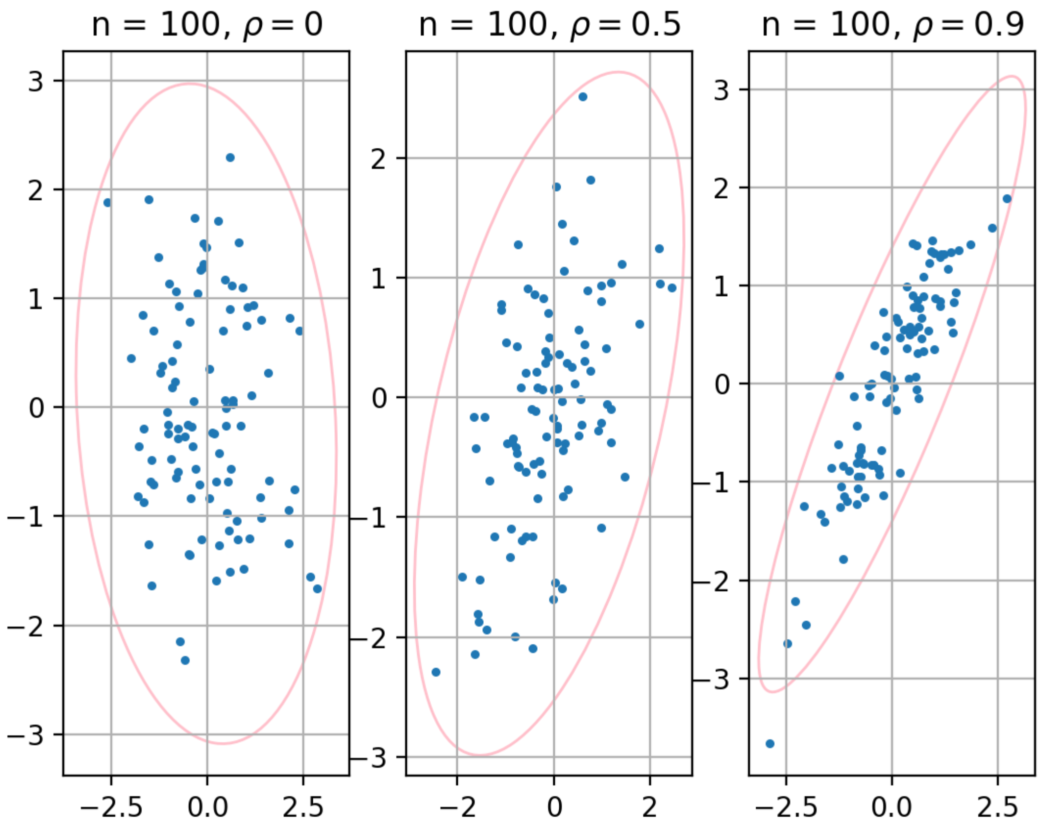

Рис. 3: Эллипсы равновероятности для выборки размером 100

$n = 20$	r (4)	r_S (6)	r_Q (5)
Среднее	$-9.16 * 10^{-2}$	$-8.83 * 10^{-2}$	$-9.42 * 10^{-2}$
Среднее квадратов	$6.12 * 10^{-2}$	$6.15 * 10^{-2}$	$1.22 * 10^{-1}$
Дисперсия	$5.29 * 10^{-2}$	$5.37 * 10^{-2}$	$1.13 * 10^{-1}$
$n = 60$	r (4)	r_S (6)	r_Q (5)
Среднее	$-8.40 * 10^{-2}$	$-7.92 * 10^{-2}$	$-7.81 * 10^{-2}$
Среднее квадратов	$2.35 * 10^{-2}$	$2.27 * 10^{-2}$	$3.97 * 10^{-2}$
Дисперсия	$1.64 * 10^{-2}$	$1.65 * 10^{-2}$	$3.36 * 10^{-2}$
$n = 100$	r (4)	r_S (6)	r_Q (5)
Среднее	$-1.01 * 10^{-1}$	$-9.53 * 10^{-2}$	$-8.67 * 10^{-2}$
Среднее квадратов	$2.10 * 10^{-2}$	$1.98 * 10^{-2}$	$2.85 * 10^{-2}$
Дисперсия	$1.08 * 10^{-2}$	$1.08 * 10^{-2}$	$2.10 * 10^{-2}$

Таблица 2: Характеристики смеси нормальных распределений

4.2 Простая линейная регрессия

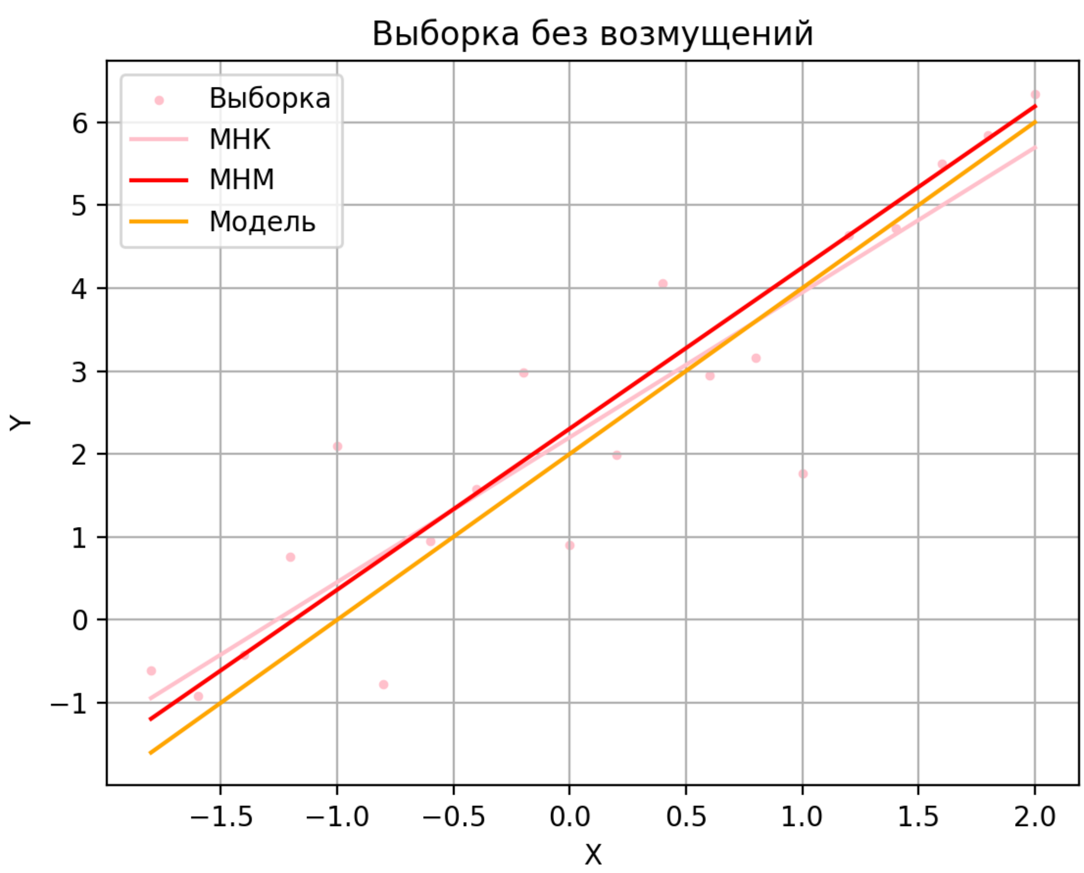

Рис. 4: МНК и МНМ без возмущений

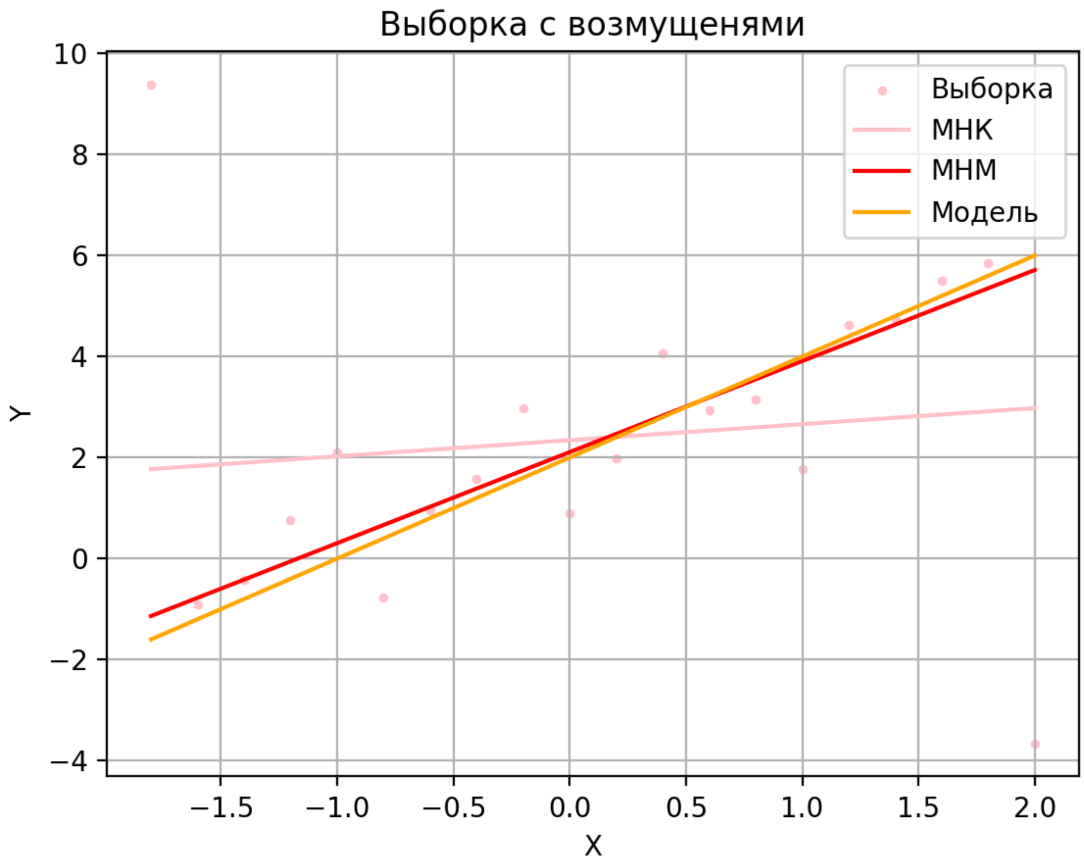

Рис. 5: МНК и МНМ с возмущениями

5 Выводы

При увеличении размера выборки наблюдается повышение точности оценок, что проявляется в уменьшении дисперсий коэффициентов корреляции. Этот эффект соответствует основным принципам центральной предельной теоремы и закона больших чисел. Увеличение коэффициента корреляции ρ сопровождается ростом средних значений коэффициентов Пирсона, Спирмена и квадратичного коэффициента корреляции вследствие прямой зависимости между ρ и другими коэффициентами корреляции.
Метод наименьших квадратов проявляет эффективность в условиях, когда данные не содержат значительных выбросов, в то время как метод наименьших модулей демонстрирует превосходство в случае наличия значительных возмущений. При выборе метода следует учитывать особенности данных: при наличии выбросов предпочтительнее использовать метод наименьших модулей из-за его устойчивости к выбросам.