Санкт-Петербургский политехнический университет Петра Великого

Физико-механический институт

Кафедра прикладной математики и информатики

# **Математическая статистика**

Отчет по лабораторной работе №4

Выполнил студент гр. 5030102/20202 Тишковец С.Е.

Преподаватель Баженов А.Н.

Санкт-Петербург

2025

# Оглавление

# 1. Постановка задачи

Найти оценки коэффициентов линейной регрессии $y_i = a + bx_i + \varepsilon_i$, используя 20 точек на отрезке [-1.8, 2] с равномерным шагом равным 0.2. Ошибку $\varepsilon_i$ считать нормально распределённой с параметрами (0, 1). В качестве эталонной зависимости взять $y_i = 2 + 2x_i + \varepsilon_i$.

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. Простая линейная регрессия

Регрессионную модель описания данных называют простой линейной регрессией, если

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1,2,\dots,n$$

где $x_1, x_2, \dots, x_n$ - заданные числа (значения фактора); $y_1, y_2, \dots, y_n$- наблюдаемые значения отклика; $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$- независимые, нормально распределённые $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); $\beta_0, \beta_1$ - неизвестные параметры, подлежащие оцениванию.

В модели отклик y 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. 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):

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

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

## 2.3. Расчётные формулы для МНК-оценок

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

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

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

$$\begin{cases} n\widehat{\beta_0} + \widehat{\beta_1} \sum x_i = \sum y_i \\ \widehat{\beta_0} \sum x_i + \widehat{\beta_1} \sum {x_i}^2 = \sum x_i y_i \end{cases}$$

Разделим оба уравнения на n:

$$\begin{cases} n\widehat{\beta_0} + \left(\dfrac{1}{n}\sum x_i\right)\widehat{\beta_1} = \dfrac{1}{n}\sum y_i \\ \left(\dfrac{1}{n}\sum x_i\right)\widehat{\beta_0} \sum x_i + \left(\dfrac{1}{n}\sum {x_i}^2\right)\widehat{\beta_1} = \dfrac{1}{n}\sum x_i y_i \end{cases}$$

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

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

получим

$$\begin{cases} n\widehat{\beta_0} + \bar{x}\widehat{\beta_1} = \bar{y} \\ \bar{x}\widehat{\beta_0} + \overline{x^2}\widehat{\beta_1} = \overline{xy} \end{cases}$$

откуда МНК-оценку $\widehat{\beta_1}$ наклона прямой регрессии находим по формуле Крамера

$$\widehat{\beta_1} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \overline{x^2}}$$

а МНК-оценку $\widehat{\beta_0}$ определяем непосредственно из первого уравнения системы:

$$\widehat{\beta_0} = \bar{y} - \bar{x}\widehat{\beta_1}$$

Доказательство минимальности функции $Q(\beta_0, \beta_1)$ в стационарной точке проведём с помощью известного достаточного признака экстремума функции двух переменных. Имеем:

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

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

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

## 2.4. 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:

$$\widehat{\beta_1} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \overline{x}^2} = \frac{k_{xy}}{{s_x}^2} = \frac{k_{xy}}{s_x s_y} \cdot \frac{s_y}{s_x} = r_{xy} \frac{s_y}{s_x}$$

$$\widehat{\beta_0} = \bar{y} - \overline{x\widehat{\beta_1}}$$

В формулах заменим выборочные средние $\bar{x}$ и $\bar{y}$ соответственно на робастные выборочные медианы $med\ x$ и $med\ y$, среднеквадратические отклонения $s_x$ и $s_y$ на робастные нормированные интерквартильные широты ${q_x}^*$ и ${q_y}^*$ $*$, выборочный коэффициент корреляции $r_{xy}$ - на знаковый коэффициент корреляции $r_Q$:

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

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

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

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

$$j = n - l + 1$$

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

Уравнение регрессии здесь имеет вид: $y = \widehat{\beta_{0R}} + \widehat{\beta_{1R}} x$.

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 исследования

Оценки коэффициентов линейной регрессии

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

Таблица 1. Результаты для невозмущенной выборки

| Метод | $\hat{a}$ | $\hat{a}/a$ | $\hat{b}$ | $\hat{b}$/b |
|---|---|---|---|---|
| МНК | 1.79 | 0.88 | 2.27 | 1.15 |
| МНМ | 1.85 | 0.92 | 2.64 | 1.33 |

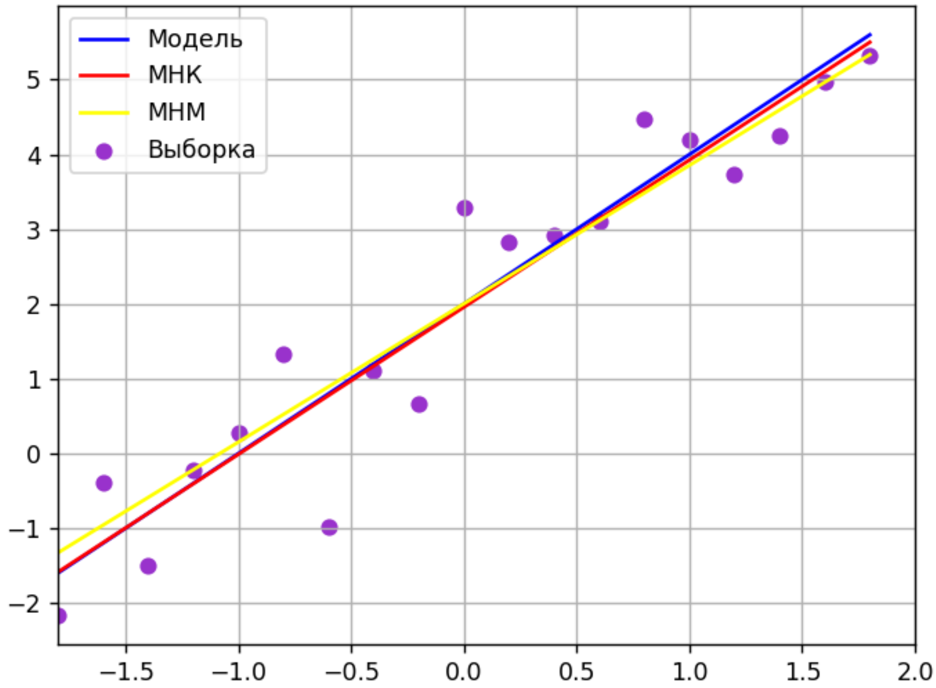


Рис. 1. Выборка без возмущений

МНК d=26.18

МНМ d=28.89

Таблица 2. Результаты для возмущенной выборки

| Метод | $\hat{a}$ | $\hat{a}/a$ | $\hat{b}$ | $\hat{b}/b$ |
|---|---|---|---|---|
| МНК | 1.93 | 0.98 | 0.75 | 0.35 |
| МНМ | 1.85 | 0.72 | 1.84 | 0.94 |

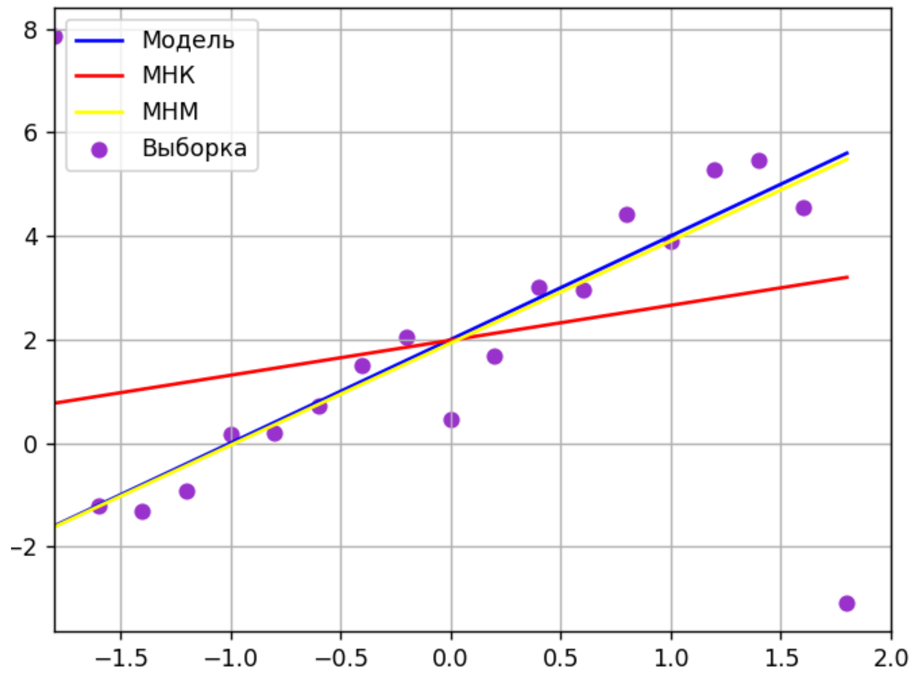


Рис. 2. Выборка с возмущениями

МНК d=146.18

МНМ d=178.82

## 4. Выводы

По полученным результатам можно сделать вывод о том, что используя критерий наименьших квадратов, получится более точно оценить коэффициенты линейной регрессии для выборки без возмущений.

В случае, когда возмущения присутствуют, более подходящим является критерий наименьших модулей