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

Отчёт по лабораторной работе №4

по дисциплине
«Математическая статистика»

Выполнила студентка
группы 5030101/20202 — Чинь Тхи Тху Хоай

Проверил
Преподаватель — Баженов Александр Николаевич

Санкт-Петербург
2025

Содержание

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 ε_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 Теория

2.1 Простая линейная регрессия

2.1.1 Модель простой линейной регрессии

Регрессионную модель описания данных называют простой линейной регрессией, если

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

где $x_1, ..., x_n$– заданные числа (значения фактора); $y_1, ...y_n$– наблюдаемые значения отклика; $\varepsilon_1, ..., \varepsilon_n$– независимые, нормально распределенные $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); β_0, β_1– неизвестные параметры, подлежащие оцениванию.
В модели (1) 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.

2.1.2 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):

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

Задача минимизации квадратичного критерия $Q(\beta_0, \beta_1)$ носит название задачи метода наименьших квадратов (МНК), а оценки $\hat{\beta}_0, \hat{\beta}_1$ параметров β_0, β_1, реализующие минимум критерия $Q(\beta_0, \beta_1)$, называют МНК-оценками.

2.1.3 Расчётные формулы для МНК-оценок

МНК-оценки параметров $\hat{\beta}_0, \hat{\beta}_1$ находятся из условия обращения функции $Q(\beta_0, \beta_1)$ в минимум.

Для нахождения МНК-оценок $\hat{\beta}_0, \hat{\beta}_1$ выпишем необходимые условия экстремума

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i = 0 \end{cases} \tag{3}$$

Далее для упрощения записи сумм будем опускать индекс суммирования. Из системы (3) получим:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i \end{cases} \tag{4}$$

Разделим оба уравнения на n:

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1(\frac{1}{n}\sum x_i) = \frac{1}{n}\sum y_i \\ \hat{\beta}_0(\frac{1}{n}\sum x_i) + \hat{\beta}_1(\frac{1}{n}\sum x_i^2) = \frac{1}{n}\sum x_i y_i \end{cases} \tag{5}$$

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

$$\bar{x} = \frac{1}{n}\sum x_i, \bar{y} = \frac{1}{n}\sum y_i, \bar{x^2} = \frac{1}{n}\sum x_i^2, \bar{xy} = \frac{1}{n}\sum x_i y_i, \tag{6}$$

получим

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \\ \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \bar{x^2} = \bar{xy}, \end{cases} \tag{7}$$

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

$$\hat{\beta}_1 = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\bar{x^2} - (\bar{x})^2} \tag{8}$$

а МНК-оценку $\hat{\beta}_0$ определяем непосредственно из первого уравнения системы (7):

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \tag{9}$$

Заметим, что определитель системы (7):

$$\bar{x^2} - (\bar{x})^2 = \frac{1}{n}\sum (x_i - \bar{x})^2 = s_x^2 > 0, \tag{10}$$

если среди значений $x_1, ..., x_n$ 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:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n, \frac{\partial^2 Q}{\partial \beta_1^2} = 2\sum x_i^2 = 2n\bar{x^2}, \frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0} = 2\sum x_i = 2n\bar{x} \tag{11}$$

$$\triangle = \frac{\partial^2 Q}{\partial \beta_0^2} \cdot \frac{\partial^2 Q}{\partial \beta_1^2} - (\frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0})^2 = 4n^2\bar{x^2} - 4n^2(\bar{x})^2 = 4n^2\left[\bar{x^2} - (\bar{x})^2\right] = 4n^2\left[\frac{1}{n}\sum(x_i - \bar{x})\right] = 4n^2 s_x^2 > 0. \tag{12}$$

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 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:

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \to \min_{\beta_0, \beta_1} \tag{13}$$

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 (13) 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 (9) и (8) в другом виде:

$$\begin{cases} \hat{\beta}_1 = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\bar{x^2} - (\bar{x})^2} = \frac{k_{xy}}{s_x^2} = \frac{k_{xy}}{s_x s_y} \cdot \frac{s_y}{s_x} = r_{xy}\frac{s_y}{s_x} \\ \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \end{cases} \tag{14}$$

В формулах (14) 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 корреляции r_{xy} — на знаковый коэффициент корреляции r_Q:

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

$$\hat{\beta}_{0R} = medy - \hat{\beta}_{1R} medx, \tag{16}$$

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

$$q_y^* = \frac{y_{(j)} - y_{(l)}}{k_q(n)}, \quad q_x^* = \frac{x_{(j)} - x_{(l)}}{k_q(n)},$$
$$\begin{cases} [\frac{n}{4}] + 1 \text{ при } \frac{n}{4} \text{ дробном,} \\ \frac{n}{4} \text{ при } \frac{n}{4} \text{ целом.} \end{cases}$$
$$j = n - l + 1$$
$$sgn(z) = \begin{cases} 1 \text{ при } z > 0 \\ 0 \text{ при } z = 0 \\ -1 \text{ при } z < 0 \end{cases} \tag{18}$$

Уравнение регрессии здесь имеет вид

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

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 $sgnz$ чувствительна не к величине аргумента, а только к его знаку. Отсюда оценка прямой регрессии (19) обладает очевидными робастными свойствами устойчивости к выбросам по координате y, но она довольно груба.

3 Программная реализация

Лабораторная работа выполнена на языке Python 3.12.6 в среде разработки Visual Studio Code. Использовались дополнительные библиотеки:

1. scipy (встроенные функции оптимизации)
2. matplotlib (визуализация)
3. numpy (вычисление ряда числовых характеристик)

В приложении находится ссылка на GitHub репозиторий с исходныи кодом.

4 Результаты

4.1 Оценки коэффициентов линейной регрессии

Метрика удаленности: $distance = \sum_{i=0}^{n}(y_{model}[i] - y_{regr}[i])^2$

4.1.1 Выборка без возмущений

Таблица 1: Результаты для невозмущенной выборки

Метод	$\hat{a}$	$\frac{\hat{a}}{a}$	$\hat{b}$	$\frac{\hat{b}}{b}$
МНК	1.89	0.95	2.18	1.09
МНМ	1.94	0.97	2.48	1.24

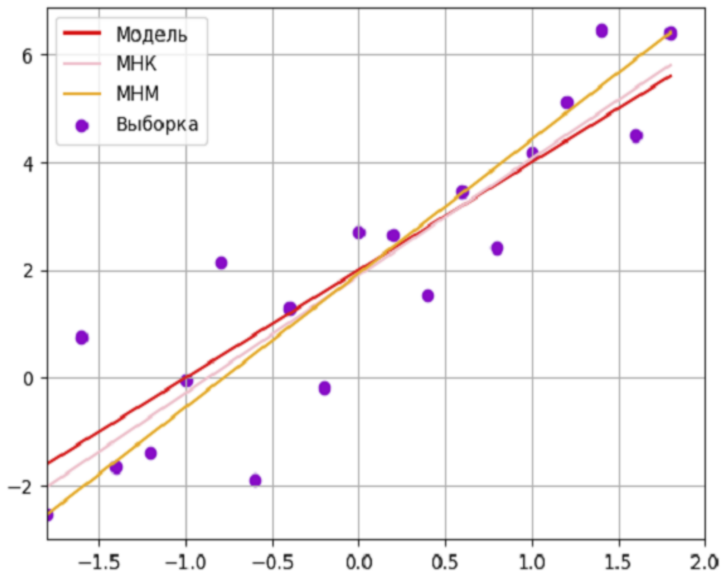

Рис. 1: Выборка без возмущений

$$\text{МНК } distance = 27.06$$
$$\text{МНМ } distance = 29.27$$

4.1.2 Выборка с возмущениями

Таблица 2: Результаты для невозмущенной выборки

Метод	$\hat{a}$	$\frac{\hat{a}}{a}$	$\hat{b}$	$\frac{\hat{b}}{b}$
МНК	2.09	1.05	0.58	0.29
МНМ	1.81	0.91	1.72	0.86

1. Критерий наименьших квадратов: $\hat{a} \approx 2.09$, $\hat{b} \approx 0.58$

2. Критерий наименьших модулей: $\hat{a} \approx 1.81$, $\hat{b} \approx 1.72$

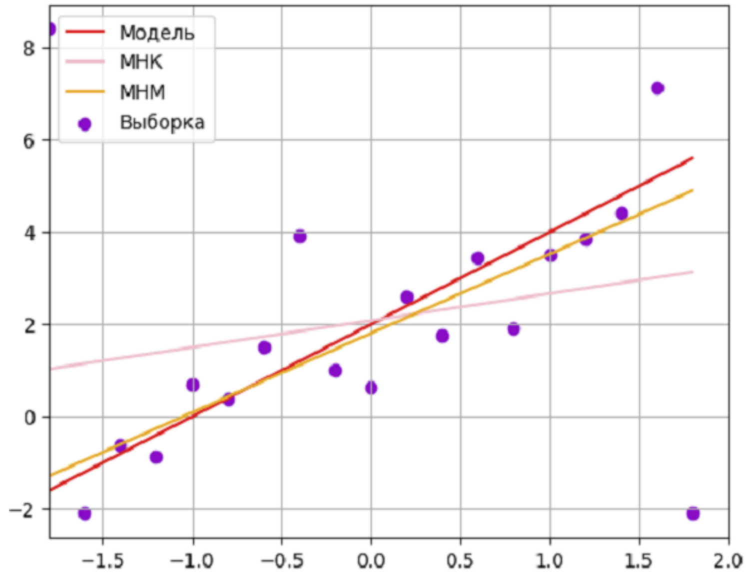

Рис. 2: Выборка с возмущениями

$$\text{МНК } distance = 133.17$$
$$\text{МНМ } distance = 164.08$$

5 Обсуждение

5.1 Оценки коэффициентов линейной регрессии

По полученным результатам (см. метрику удаленности модельной прямой от теоретической - *distance*) можно сказать, что используя критерий наименьших квадратов удастся точнее оценить коэффициенты линейной регрессии для выборки без возмущений. Если же редкие возмущения присутствуют, тогда лучше использовать критерий наименьших модулей.

6 Приложение

Код программы GitHub URL:

https://github.com/Akira1707/Math-Statistic/tree/main/Lab4