Санкт-Петербургский политехнический университет Петра Великого

Институт прикладной математики и механики

Высшая школа прикладной математики

ОТЧЁТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №6

по дисциплине «Математическая статистика»

Выполнила
студентка гр.3630102/80101

А.А. Тимофеева

Руководитель доцент, к.ф.-м.н.

А.Н.Баженов

Санкт-Петербург 2021

СОДЕРЖАНИЕ

СПИСОК ИЛЛЮСТРАЦИЙ

1 ПОСТАНОВКА ЗАДАЧИ

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 Простая линейная регрессия

2.1.1 Модель простой линейной регрессии

Регрессионную модель описания данных называют *простой линейной регрессией*, если

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

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.

2.1.2 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 (14) носит название задачи

метода наименьших квадратов (МНК), а оценки $\widehat{\beta_0}, \widehat{\beta_1}$ параметров β_0, β_1, реализующие минимум критерия (2), называют МНК-оценками.

2.1.3 Расчетные 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:

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

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

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

Разделим оба уравнения на n:

$$\begin{cases} \widehat{\beta_0} + \widehat{\beta_1}\sum(\frac{1}{n}x_i) = \frac{1}{n}\sum y_i \\ \widehat{\beta_0}\sum(\frac{1}{n}x_i) + \beta_1(\frac{1}{n}\widehat{\sum x_i^2)} = \frac{1}{n}\sum x_\iota y_\iota \end{cases}$$

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

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

получим:

$$\begin{cases} \widehat{\beta_0} + \widehat{\beta_1}\bar{x} = \bar{y}, \\ \widehat{\beta_0}\bar{x} + \widehat{\beta_1}\overline{x^2} = \overline{xy}, \end{cases} \tag{4}$$

откуда МНК-оценку β_1 наклона прямой регрессии находим по формуле Крамера:

$$\widehat{\beta_1} = \frac{\overline{xy} - \bar{x} * \bar{y}}{\overline{x^2} - (\bar{x})^2} \tag{5}$$

а МНК-оценку $\beta 0$ определяем непосредственно из первого уравнения системы:

$$\widehat{\beta_0} = \bar{y} - \bar{x}\widehat{\beta_1} \tag{6}$$

Заметим, что определитель системы (4):

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

если среди значений $x_1, \dots, x_n$ 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:

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

$$\Delta = \frac{\partial^2 Q}{\partial \beta_0^2} * \frac{\partial^2 Q}{\partial \beta_1^2} - \left(\frac{\partial^2 Q}{\partial \beta_0^2 \beta_1^2}\right)^2 = 4n^2\overline{x^2} - 4n^2(\bar{x})^2 = 4n^2\left[\overline{x^2} - (\bar{x})^2\right] =$$

$$= 4n^2[\frac{1}{n}\sum (x_i - \bar{x})] \quad = 4n^2 s_x^2 > 0.$$

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

2.2 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:

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \quad \underset{\beta_0, \beta_1}{\to min} \tag{7}$$

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 (7) 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 (5) и (6) в другом виде:

$$\beta_1 = \frac{\overline{xy} - \bar{x} * \bar{y}}{\overline{x^2} - (\overline{x})^2} = \frac{k_{xy}}{s_x^2} = \frac{k_{xy}}{s_x s_y} * \frac{s_y}{s_x} = r_{xy}\frac{s_y}{s_x}, \qquad \beta_0 = \bar{y} - \bar{x}\widehat{\beta_1} \tag{8}$$

В формулах (8) 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:

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

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

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

$$q_y^* = \frac{y_{(j)} - y_{(l)}}{k_q(n)}, \qquad q_x^* = \frac{x_{(j)} - x_{(l)}}{k_q(n)},$$

$$\begin{cases} \left[\frac{n}{4}\right] + 1 \text{ при } \frac{n}{4} \text{ дробном,} \\ \frac{n}{4} \text{ при } \frac{n}{4} \text{ целом.} \end{cases}$$

$$j = n - l + 1$$

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

Уравнение регрессии здесь имеет вид

$$y = \widehat{\beta_{0R}} + \widehat{\beta_{1R}} \tag{9}$$

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 $sgn(z)$ чувствительна не к величине аргумента, а только к его знаку. Отсюда оценка прямой регрессии (9) обладает очевидными робастными свойствами устойчивости к выбросам по координате y, но она довольно груба.

3 РЕАЛИЗАЦИЯ

Лабораторная работа выполнена с помощью встроенных средств языка программирования Python в среде разработки PyCharm. Исходный код лабораторной работы приведён в приложении.

4 РЕЗУЛЬТАТЫ

4.1. Оценки коэффициентов линейной регрессии

4.1.1 Выборка без возмущений

1. Критерий наименьших квадратов: $\hat{a} \approx 1.77, \hat{b} \approx 1.65$
2. Критерий наименьших модулей: $\hat{a} \approx 1.57, \hat{b} \approx 1.73$

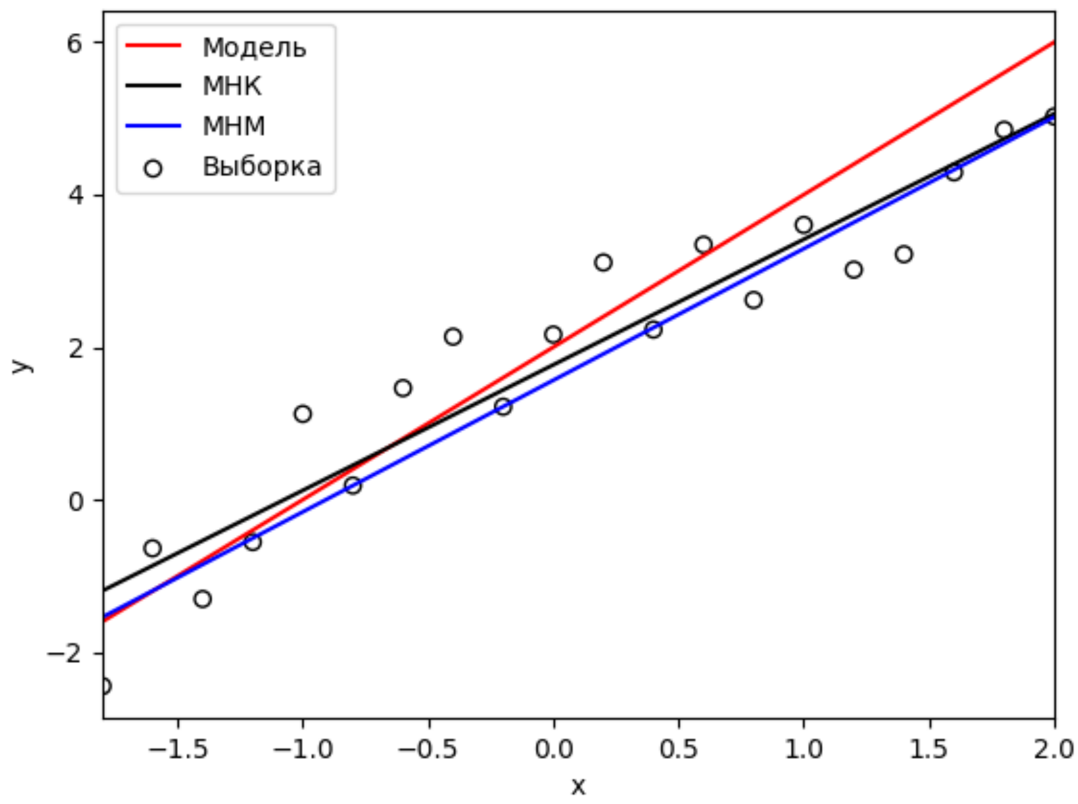

Рисунок 1: Выборка без возмущений

МНК dist = 4.76
МНМ dist = 6.26

4.1.2 Выборка с возмущениями

1. Критерий наименьших квадратов: $\hat{a} \approx 1.91, \hat{b} \approx 0.22$
2. Критерий наименьших модулей: $\hat{a} \approx 2.02, \hat{b} \approx 1.42$

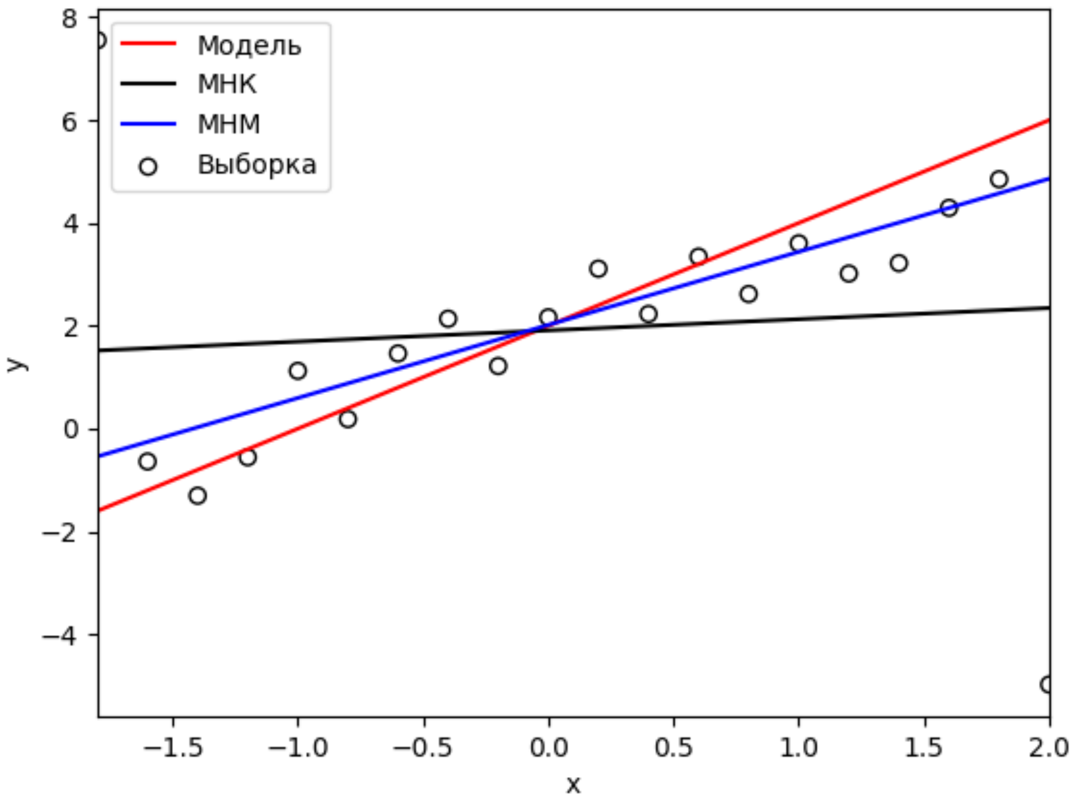

Рисунок 2: Выборка с возмущениями

МНК dist = 85.92
МНМ dist = 8.93

5 ОБСУЖДЕНИЕ

Таким образом, можно сделать вывод, что критерий наименьших квадратов точнее оценивает коэффициенты линейной регрессии на выборке без возмущений, на выборке с возмущениями лучше использовать критерий наименьших модулей. Также можно сказать, что если присутствуют редкие возмущения, то лучше использовать критерий наименьших модулей.

6 ПРИЛОЖЕНИЕ

Код программы URL:https://github.com/tmffv/MathStat/blob/master/lab6/lab6.py