Санкт-Петербургский политехнический университет
Петра Великого

Институт прикладной математики и механики

Кафедра «Прикладная математика»

Отчет по лабораторной работе №6

по дисциплине
"Математическая статистика"

Выполнила студентка
группы 3630102/80201 Деркаченко Анна Олеговна

Проверил
доцент, к.ф.-м.н. Баженов Александр Николаевич

Санкт-Петербург
2021 г.

Содержание

Список иллюстраций

1. Постановка задачи

Дано нормальное двумерное распределение $N(x, y, 0, 0, 1, 1, \rho)$.

Необходимо:

1) 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 распределенной с параметрами $(0, 1)$
2) В качестве эталонной зависимости взять $y_i = 2 + 2x_i + e_i$
3) При построении оценок коэффициентов использовать два критерия:
   - критерий наименьших квадратов
   - критерий наименьших модулей
4) Проделать те же действия 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

Простая линейная регрессия - регрессионная модель описания данных, такая что:

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

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.

Отклик y зависит зависит от одного фактора x, и весь разброс экспериментальных точек объясняется только погрешностями наблюдений 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

В данном методе вводится мера (критерий) рассогласования отклика и регрессионной функции, и оценки параметров регрессии определяются так, чтобы сделать это рассогласование наименьшим. Используется критерий в виде суммы квадратов отклонений значений отклика от значений регрессионной функции (сумма квадратов остатков):

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

Расчётные формулы для МНК-оценок:

$$\begin{cases} \hat{\beta}_1 = \frac{\bar{xy} - \bar{x} \cdot \bar{y}}{\bar{x^2} - (\bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \end{cases} \tag{3}$$

2.2. Робастные оценки коэффициентов линейной регрессии

Робастность оценок коэффициентов линейной регрессии - их устойчивость по отношению к наличию в данных редких, но больших по величине выбросов. Она может быть обеспечена использованием метода наименьших модулей вместо метода наименьших квадратов:

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \to \min_{\beta_0, \beta_1} \tag{4}$$

Робастная альтернатива оценкам коэффициентов линейной регрессии по МНК:

$$\begin{cases} \hat{\beta}_{1R} = r_Q \frac{q_y^*}{q_x^*} \\ \hat{\beta}_{0R} = medy - \hat{\beta}_{1R} medx \end{cases} \tag{5}$$

где $medx$ и $medy$ - робастные выборочные медианы, q_x^* и q_y^* - робастные нормированные интерквартильные широты, r_Q - знаковый коэффициент корреляции. Причем:

$$\begin{cases} r_Q = \frac{1}{n} \sum_{i=1}^{n} sgn(x_i - medx) sgn(y_i - medy) \\ q_x^* = \frac{x_j - x_l}{k_q(n)} \\ q_y^* = \frac{y_j - y_l}{k_q(n)} \\ l = \begin{cases} [\frac{n}{4}] + 1 \text{ при } \frac{n}{4} \text{ - дробном} \\ \frac{n}{4} \text{ при } \frac{n}{4} \text{ - целом} \end{cases} \\ j = n - l + 1 \end{cases} \tag{6}$$

Уравнение регрессии принимает вид:

$$y = \hat{\beta}_{0R} + \hat{\beta}_{1R} x \tag{7}$$

3. Реализация

Реализация лабораторной работы проводилась на языке Python в среде разработки PyCharm с использованием дополнительных библиотек:

- scipy
- numpy
- matplotlib

Исходный код лабораторной работы размещен в GitHub-репозитории.

URL: https://github.com/derkanw/Mathstat/tree/main/lab6

4. Результаты

4.1. Выборка без возмущения

Оценка коэффициентов по критерию наименьших квадратов:

$$\begin{cases} \hat{\beta}_0 = 2.198856 \\ \hat{\beta}_1 = 1.942401 \end{cases} \tag{8}$$

Удаленность по мере в пространстве l^2 : 0.826971

Оценка коэффициентов по критерию наименьших модулей:

$$\begin{cases} \hat{\beta}_{0R} = 2.154839 \\ \hat{\beta}_{1R} = 1.947938 \end{cases} \tag{9}$$

Удаленность по мере в пространстве l^1 : 2.941948

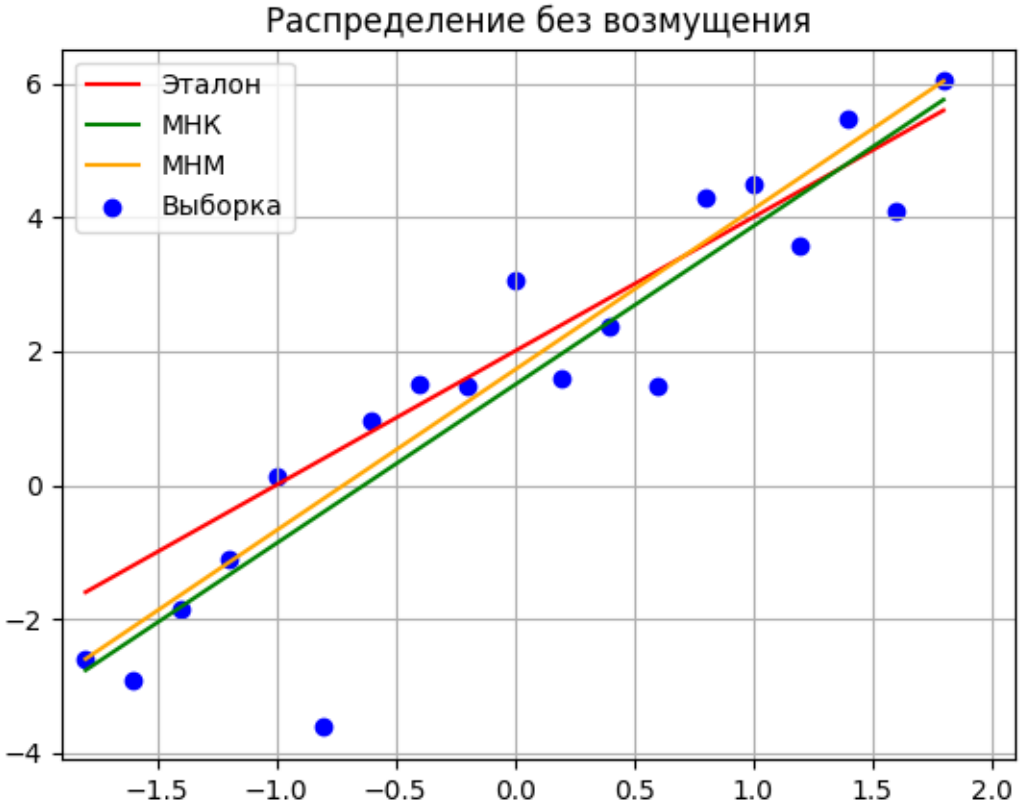

Рис. 1. Выборка из 20 элементов без возмущения

4.2. Выборка с возмущением

Оценка коэффициентов по критерию наименьших квадратов:

$$\begin{cases} \hat{\beta}_0 = 2.198856 \\ \hat{\beta}_1 = 0.363454 \end{cases} \tag{10}$$

Удаленность по мере в пространстве l^2 : 61.816189

Оценка коэффициентов по критерию наименьших модулей:

$$\begin{cases} \hat{\beta}_{0R} = 2.316183 \\ \hat{\beta}_{1R} = 1.440918 \end{cases} \tag{11}$$

Удаленность по мере в пространстве l^1 : 10.97349

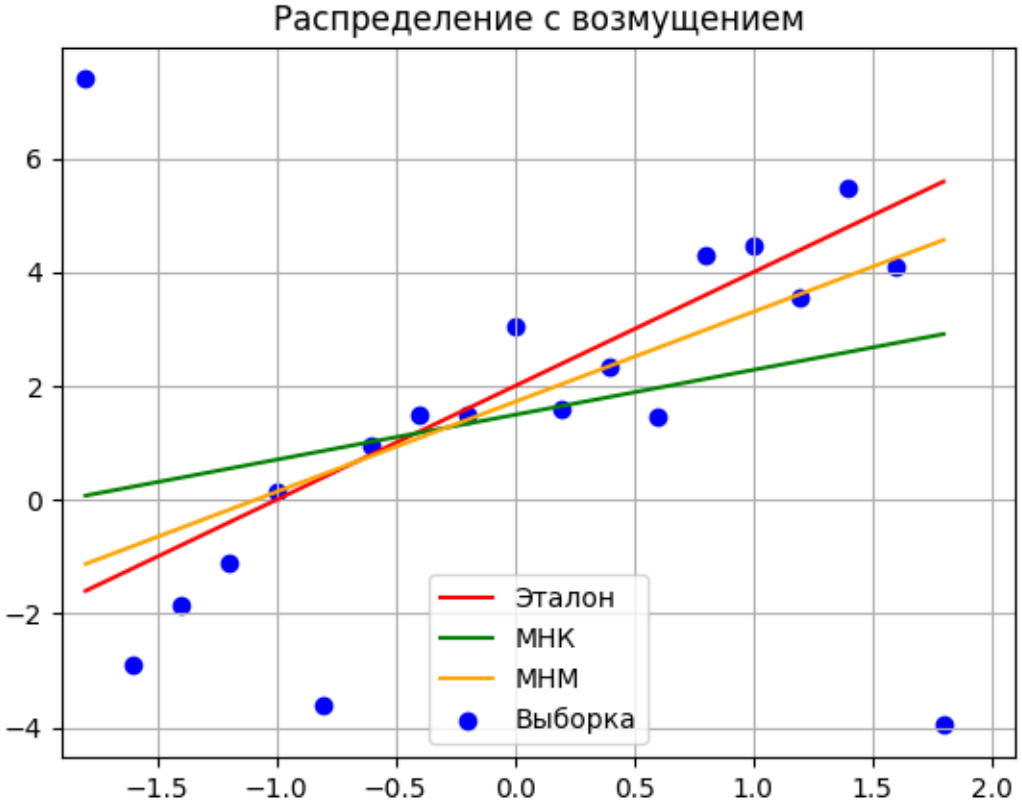

Рис. 2. Выборка из 20 элементов с возмущением

5. Обсуждение

Метод наименьших квадратов и наименьших модулей работают в разный пространствах: в l^2 и в l^1 соответственно. Поэтому удаленность экспериментально построенной прямой от эталонной стоит рассматривать только в контексте каждого из пространств.

В ходе исследования полученный результатов можно сделать вывод, что метод наименьших квадратов дает более точную оценку коэффициентов линейной регрессии, но менее пригоден при выборке с редкими возмущениями достаточной величины. То есть метод наименьших модулей менее точен, но более устойчив, так как использует робастные величины.