Санкт-Петербургский политехнический университет
Петра Великого

Институт прикладной математики и механики

Кафедра «Прикладная математика»

Отчет по лабораторной работе №6

по дисциплине
"Математическая статистика"

Выполнил студент
Группы 3630102/80101 Шао Цзяци

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

Санкт-Петербург
2021 г.

Содержание

Список иллюстраций

1. Постановка задачи

Дано двухмерное нормальное распределение $N(x, y, 0, 0, 1, 1, \rho)$, требуется:

- 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 $y_i = 2 + x_i + e_i$.
- 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_1и $y_2 0$ 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

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = 1, 2, ..., n \tag{1}$$

где, x_i - заданные числа(значения фактора), y_i - наблюдаемые значения отклика, ε_i - независимые, нормально распределенные $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой(независимой) дисперсией случайные величины (не наблюдаемые), β_0, β_1 - неизвестные параметры, подлежащие оцениванию.

2.1.2. Метод наименьших квадратов

Метод наименьших квадратов(МНК) — математический метод, применяемый для решения различных задач, основанный на минимизации суммы квадратов отклонений некоторых функций от искомых переменных. Он может использоваться для аппроксимации точечных значений некоторой функции. МНК является одним из базовых методов регрессионного анализа для оценки неизвестных параметров регрессионных моделей по выборочным данным.

Расчетная формула для МНК-оценок:

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

Расчетные формулы для МНК-оценок

$$\beta_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \tag{3}$$

$$\beta_0 = \overline{y} - \overline{x}\beta_1 \tag{4}$$

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: *метод наименьших модулей*:

$$\sum_{i=1}^{n} |\varepsilon_i| = \sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \xrightarrow{\beta_0,\beta_1} min \tag{5}$$

Робастная альтернатива оценкам коэффициентов линейной регрессии по МНК:

$$\begin{cases} \beta_{1R} = r_Q \frac{q_y^*}{q_x^*} \\ \beta_{0R} = medy - \beta_{1R} medx \end{cases} \tag{6}$$

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

$$\begin{cases} r_Q = \frac{1}{n} \sum_{i=1}^{n} sgn(x_i - medx) sgn(y_i - medy) \\ q_x^* = \frac{x_{(j)} - x_{(l)}}{k_q(n)} \\ q_y^* = \frac{y_{(j)} - y_{(l)}}{k_q(n)} \\ l = \begin{cases} [n/4] + 1, n/4 - \text{дробь} \\ n/4, n/4 - \text{целое} \end{cases} \\ j = n - l + 1 \end{cases} \tag{7}$$

3. Реализация

Лабораторная работа выполнена с помощью встроенных средств языка программирования python в среде разработки Pycharm с дополнительными библиотеками.

- scipy
- numpy
- matplotlib
- math

Исходный код лабораторной работы размещен в Github-репозитории.
URL: `https://github.com/Shaots/shaoMathStatistic/tree/master/Lab6`

4. Результаты

4.1. Выборка без возмущения

Оценка коэффициентов по методу наименьших квадратов(2):

$$\begin{cases} \beta_0 = 1.9357957 \\ \beta_1 = 1.7281780 \end{cases} \tag{8}$$

Удаленность по мере в пространстве l^2: 2.1324297

Оценка коэффициентов по методу наименьших модулей(5):

$$\begin{cases} \beta_{0R} = 2.2563289 \\ \beta_{1R} = 1.8532691 \end{cases} \tag{9}$$

Удаленность по мере в пространстве l^1: 4.922955

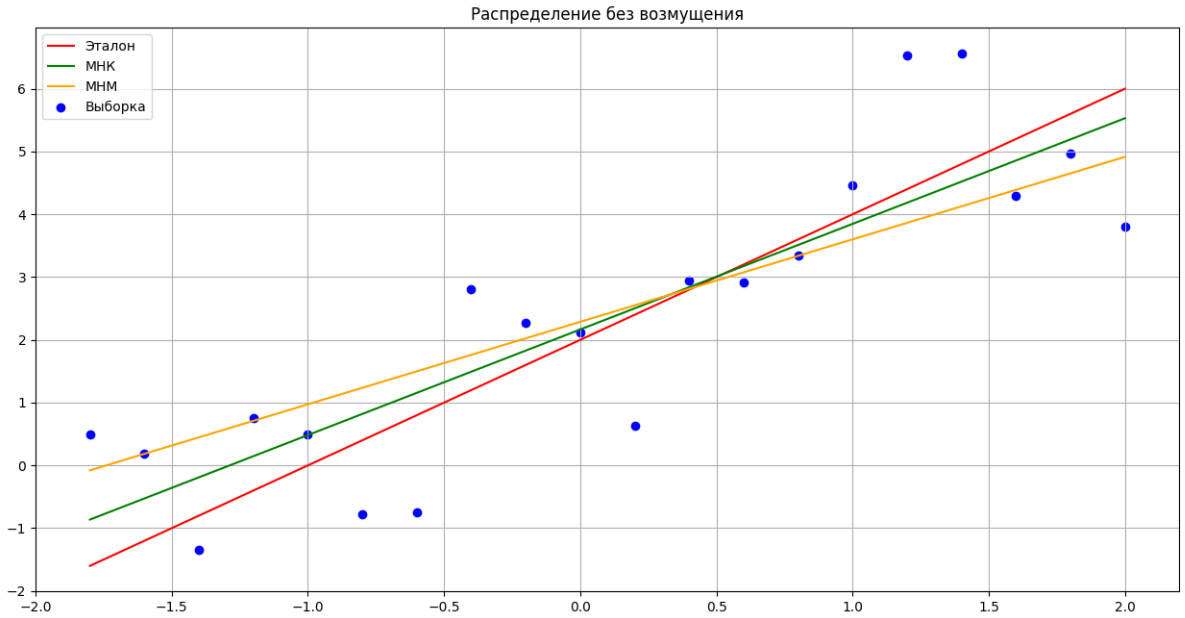

Рис. 1. Выборка из 20 элементов без возмущения

4.2. Выборка с возмущением

Оценка коэффициентов по методу наименьших квадратов(2):

$$\begin{cases} \beta_0 = 2.0786528 \\ \beta_1 = 0.2996066 \end{cases} \tag{10}$$

Удаленность по мере в пространстве l^2: 77.076617.

Оценка коэффициентов по методу наименьших модулей(5):

$$\begin{cases} \beta_{0R} = 2.0786528 \\ \beta_{1R} = 1.2342563 \end{cases} \tag{11}$$

Удаленность по мере в пространстве l^1: 15.314874.

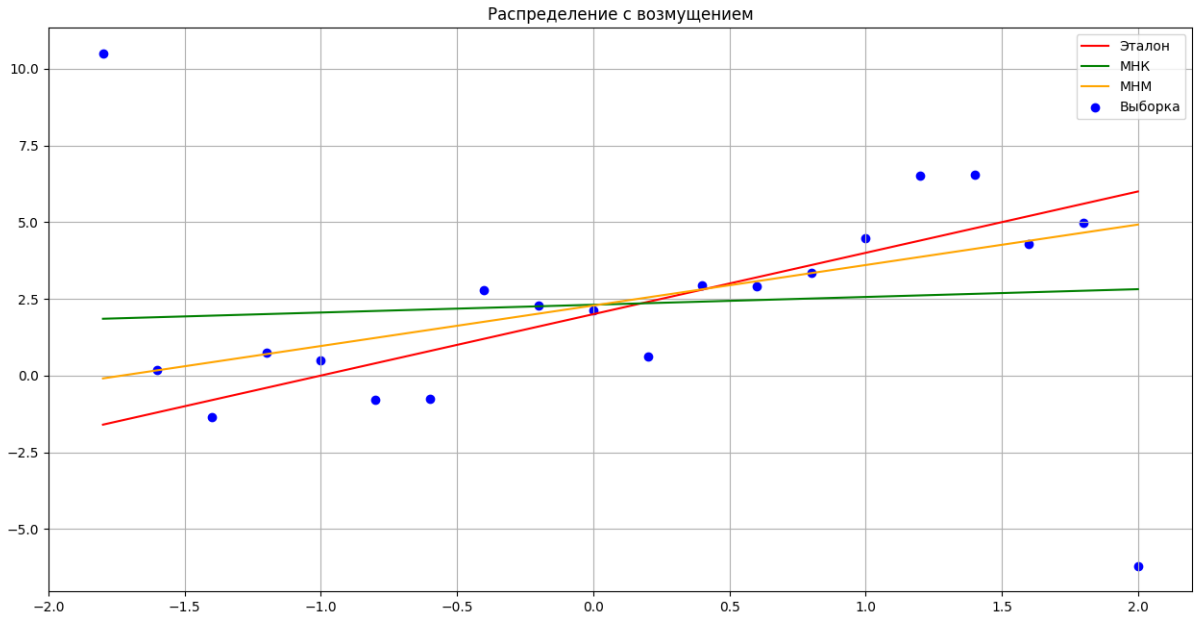

Рис. 2. Выборка из 20 элементов с возмущением

5. Обсуждение

По полученным результатам (8)(10) можно сказать, при выборке без возмущений используя критерий наименьших квадратов удастся точнее оценить коэффициенты линейной регрессии. Если редкие возмущения присутствуют, тогда лучше использовать критерий наименьших модулей.