Санкт-Петербургский политехнический университет
Петра Великого

Институт прикладной математики и механики
Кафедра «Прикладная математика»

Отчет по лабораторной работе № 6
по дисциплине: Математическая статика.

Выполнила студентка:
Заболотских Екатерина Дмитриевна
группа: 3630102/70301

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

Санкт-Петербург
2020 г.

Оглавление

Список иллюстраций

Постановка задачи

Найти оценки коэффициентов a, b 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_2 вносятся возмущения 10 и -10.

Теория

Простая линейная регрессия

Регрессионную модель описания данных называют простой линейной регрессией, если

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \tag{1}$$

, где $\{x_i\}_{i=1}^n$ - значения фактора, $\{y_i\}_{i=1}^n$ - наблюдаемые значения отклика, а $\{\varepsilon_i\}_{i=1}^n$ - независимые, нормально распределенные по закону $\mathcal{N}(0, \sigma)$ случайные величины, а β_0, β_1 - оцениваемые параметры. Для оценки применяются различные методы, в данной работе рассмотрен следующий подход: вводится критерий рассогласования отклика и регрессионной функции, после чего оценки параметров регрессии выводятся из задачи минимизации критерия.

Критерий наименьших квадратов

Достаточно простые расчетные формулы для оценок получают при выборе критерия в виде суммы квадратов отклонений значений отклика от значений регрессионной функции:

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

Оценки $\widehat{\beta_0}, \widehat{\beta_1}$ параметров β_0, β_1, реализующие минимум критерия (2), называют МНК-оценками:

$$\widehat{\beta_1} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} \quad \widehat{\beta_0} = \bar{y} - \bar{x} \cdot \widehat{\beta_1} \tag{3}$$

Критерий наименьших модулей

Робастность оценок коэффициентов линейной регрессии (т.е. их устойчивость по отношению к наличию данных редких, но больших по величине выбросов) может быть обеспечена различными способами. Одним из них является метод нами меньших модулей вместо метода МНК:

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

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 задач метода МНК, на практике задача (4) решается численно.

В данной работе был использован метод Нелдера-Мида, применимый к негладким функциям (в том числе к $M(\beta_0, \beta_1)$).

Реализация

Код программы, реализующий данную задачу, был написан на языке Python в интегрированной среде разработке PyCharm.

Были использованы библиотеки:

- **Numpy** – библиотека для работы с данными.
- **Matplotlib** – вывод графиков.
- **SciPy** – модуль “stats” для генерации данных по эталонной зависимости, оценок МНК, модуль “optimize” для метода Нелдера-Мида.

Результаты

Выборка без выбросов

- Критерий наименьших квадратов

  $\widehat{\beta_0}$ (3) = 2.16

  $\widehat{\beta_1}$ (3) = 2.11

  Q (2) = 6.9024

  M (4) = 9.8735

- Критерий наименьших модулей

  $\widehat{\beta_0}$ (3) = 1.93

  $\widehat{\beta_1}$ (3) = 2.03

  Q (2) = 8.0054

  M (4) = 9.3798

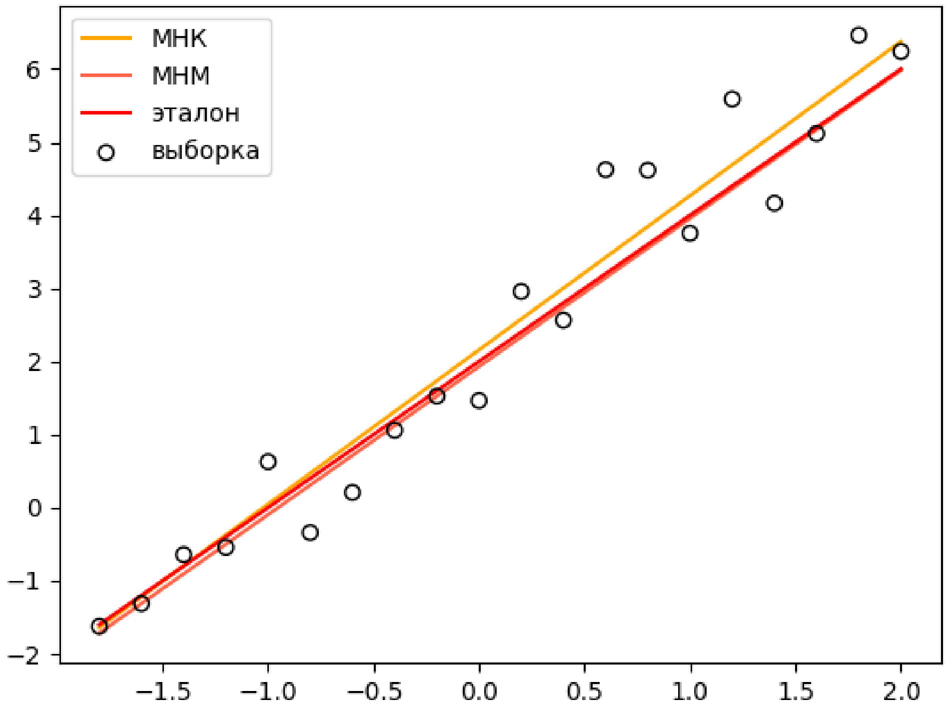

Рисунок 1: без выбросов

Выборка с выбросами

- Критерий наименьших квадратов

  $\widehat{\beta_0}$ (3) = 0.73

  $\widehat{\beta_1}$ (3) = 2.25

  Q (2) = 258.5346

  M (4) = 45.4088

- Критерий наименьших модулей

  $\widehat{\beta_0}$ (3) = 1.9

  $\widehat{\beta_1}$ (3) = 2.01

  Q (2) = 203.7506

  M (4) = 28.7808

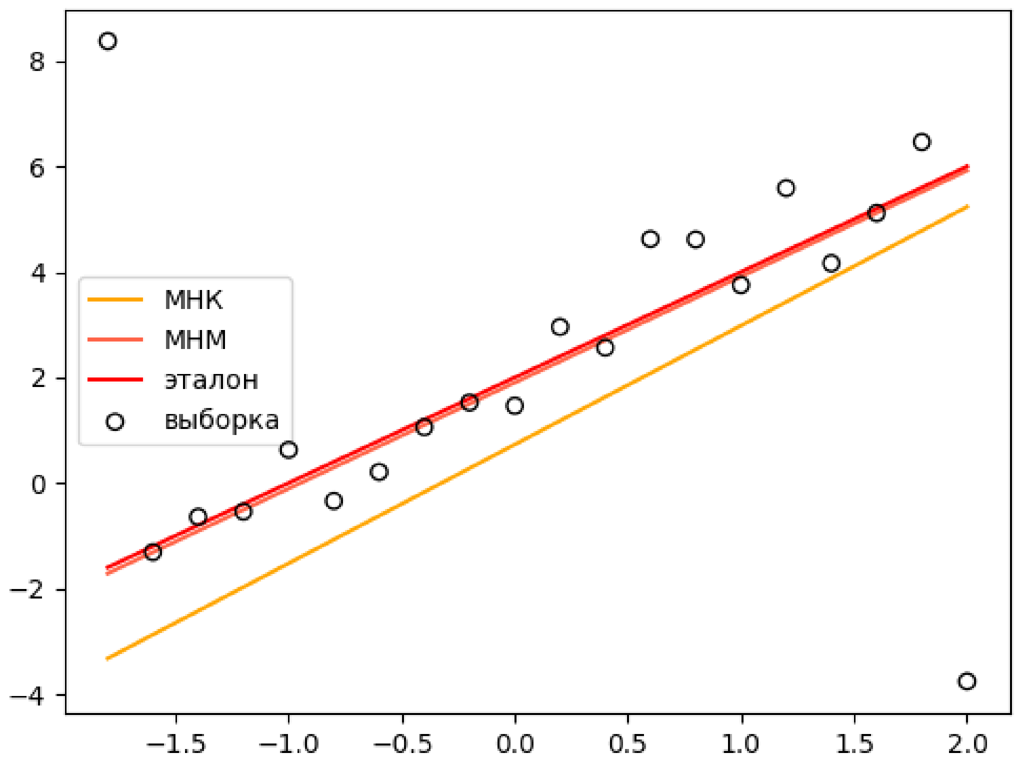

Рисунок 2: с выбросами

Обсуждение

Графики показали, что оценка по критерию наименьших модулей значительно лучше приближает эталонную зависимость при наличии выбросов.
С другой стороны, критерий наименьших квадратов лучше в случае отсутствия выбросов.

Полученные значения M, Q упорядочены, для оценки МНК значение Q меньше, чем для любой другой, аналогично для оценки МНМ и значения M.

Список литературы

1. Конспекты лекции
2. Википедия: https://ru.wikipedia.org/wiki

Ссылка на github: https://github.com/KateZabolotskih/MathStatLabs