

САНКТ-ПЕТЕРБУРГСКИЙ ПОЛИТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ
ИМ. ПЕТРА ВЕЛИКОГО

ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ И МЕХАНИКИ
КАФЕДРА "ПРИКЛАДНАЯ МАТЕМАТИКА"

ОТЧЁТ
ЛАБОРАТОРНАЯ РАБОТА № 6
ПО ДИСЦИПЛИНЕ
"МАТЕМАТИЧЕСКАЯ СТАТИСТИКА"

ВЫПОЛНИЛ СТУДЕНТ:
ШАРАПОВ СЕРГЕЙ АНДРЕЕВИЧ
ГРУППА: 3630102/70401

ПРОВЕРИЛ:
К.Ф-М.Н., ДОЦЕНТ
БАЖЕНОВ АЛЕКСАНДР НИКОЛАЕВИЧ

САНКТ-ПЕТЕРБУРГ
2020 год

Содержание

1. Список иллюстраций	3
2. Список таблиц	3
3. Постановка задачи	4
4. Теория	4
4.1. Метод наименьших квадратов	4
4.2. Метод наименьших модулей	4
5. Реализация	5
6. Результаты	5
7. Выводы	6
8. Литература	6
9. Приложения	6

1 Список иллюстраций

1	Графики линейной регрессии	5
---	--------------------------------------	---

2 Список таблиц

1	Таблица оценок коэффициентов линейной регрессии без возмущений . . .	6
2	Таблица оценок коэффициентов линейной регрессии с возмущениями . . .	6

3 Постановка задачи

Необходимо найти оценки линейной регрессии $y_i = a + bx_i + e_i$, используя 20 точек отрезка $[-1.8; 2]$ с равномерным шагом 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.

Прodelать то же самое для выборки, у которой в значении y_1 и y_{20} вносятся возмущения 10 и -10 соответственно.

4 Теория

Простая линейная регрессия [?]:

$$y_i = ax_i + b + e_i, \quad i = \overline{1, n}, \quad (1)$$

где x_i – заданные числа, y_i – наблюдаемые значения отклика, e_i – независимые, нормально распределённые с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые), a и b – неизвестные параметры, подлежащие оцениванию.

4.1 Метод наименьших квадратов

Критерий – минимизация функции [?]:

$$Q(a, b) = \sum_{i=1}^n (y_i - ax_i - b)^2 \rightarrow \min \quad (2)$$

Оценка \hat{a} и \hat{b} параметров a и b , в которых достигается минимум $Q(a, b)$, называются МНК-оценками. Формулы для их вычисления:

$$\begin{cases} \hat{a} = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2} \\ \hat{b} = \bar{y} - \hat{a}\bar{x} \end{cases} \quad (3)$$

Оценка по методу наименьших квадратов является несмещённой оценкой.

МНК чувствителен к выбросам (т.к. в вычислении используется выборочное среднее, значение которого крайне неустойчиво к большим по относительной величине выбросам)

4.2 Метод наименьших модулей

Критерий наименьших модулей – заключается в минимизации следующей функции [?]:

$$M(a, b) = \sum_{i=1}^n |y_i - ax_i - b| \rightarrow \min \quad (4)$$

Формулы для вычисления робастных параметров:

$$\begin{cases} \hat{a}_R = r_Q \frac{q_y^*}{q_x^*} \\ \hat{b}_R = med y - \hat{a}_R med x \end{cases} \quad (5)$$

, где

$$r_Q = \frac{1}{n} \sum sgn(x_i - med x) sgn(y_i - med y) \quad (6)$$

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 МНМ обладает очевидными робастными свойствами устойчивости к выбросам по координате y , но она довольно груба.

5 Реализация

Работы была выполнена на языке *Python3.8.2* Для генерации выборок использовался модуль *numpy*. Для построения графиков использовалась библиотека *matplotlib*. Регрессионные модели использовались из библиотеки *statsmodels*.

6 Результаты

Рис. 1: Графики линейной регрессии

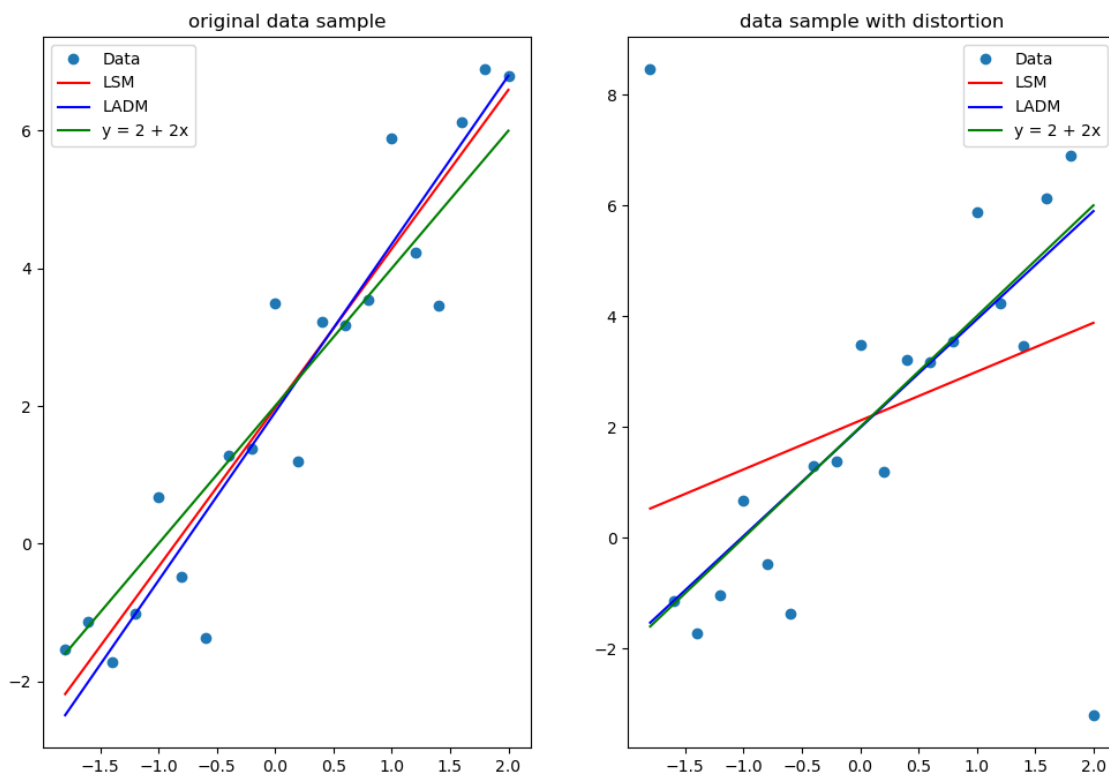


Таблица 1: Таблица оценок коэффициентов линейной регрессии без возмущений

	\hat{a}	\hat{b}
МНК	1.739737	1.924716
МНМ	1.935446	1.742526

Таблица 2: Таблица оценок коэффициентов линейной регрессии с возмущениями

	\hat{a}	\hat{b}
МНК	0.311165	2.067573
МНМ	1.877536	1.707785

7 Выводы

По графику видно, что оба метода дают хорошую оценку коэффициентов линейной регрессии, при отсутствии выбросов. Однако выбросы сильно влияют на оценки по МНК.

Выбросы мало влияют на оценку по МНМ, но ценой за это является бóльшая по сравнению с МНК сложность вычисления. На практике зачастую легче просто отсеять выбросы из выборки.

8 Литература

[Модуль numpy](#)

[Модуль matplotlib](#)

[Модуль scipy](#)

Шевляков Г. Л. Лекции по математической статистике, 2019.

9 Приложения

[Код лабораторной](#)