Санкт-Петербургский Политехнический Университет
им. Петра Великого

Институт прикладной математики и механики

Кафедра прикладной математики

Отчёт по лабораторной работе №6 по дисциплине “Математическая статистика”

Простая линейная регрессия

Выполнил студент:

Мишутин Д. В.

Группа:

3630102/70301

Проверил:

К.ф.-м.н., доцент

Баженов Александр Николаевич

Санкт-Петербург

2020 г.

Оглавление

Список иллюстраций и таблиц

1 Постановка задачи

Найти оценки коэффициентов линейной регрессии $y_i = a x_i + b + e_i$, 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 $y_i = 2 x_i + 2 + e_i$. 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

Стандартное нормальное распределение:

$$N(x, 0, 1) = \frac{1}{\sqrt{2\pi}} e^{\frac{-x^2}{2}}$$

2.1 Простая линейная регрессия

Точная формула:

$$y_i = a x_i + b + e_i \; (i = \overline{1, n}),$$

где x_i – заданные числа, y_i – наблюдаемые значения, e_i – независимые, одинаково распределённые значения ошибок с параметрами $E[e_i] = 0$ и $D[e_i] = \sigma^2$, a и b – неизвестные параметры, подлежащие оценке.

2.2 Метод наименьших квадратов (МНК)

Критерий – минимизация RSS-функции (***R**esidual **S**um of **S**quares*):

$$RSS = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(\hat{a} x_i + \hat{b} - y_i\right)^2 \to min$$

В случае линейной регрессии оценочные коэффициенты $\hat{a}$ и $\hat{b}$ можно вычислить по формулам:

$$\begin{cases} \hat{a} = \dfrac{\sigma_y}{\sigma_x} r_{xy} \\ \hat{b} = \overline{y} - \hat{a}\,\overline{x} \end{cases}$$

МНК является несмещённой оценкой. Чувствителен к выбросам (т. к. в вычислениях используется выборочное среднее, крайне неустойчивое к редким, но больших по величине выбросам).

2.3 Метод наименьших модулей (МНМ)

Критерий – минимизация LAD-функции (***L**east **A**bsolute **D**eviations*):

$$LAD = \sum_{i=1}^{n} \left|\hat{a} x_i + \hat{b} - y_i\right| \to min$$

Коэффициенты так же можно вычислить по формулам:

$$\begin{cases} \hat{a} = \dfrac{\widehat{y_2} - \widehat{y_1}}{\widehat{x_2} - \widehat{x_1}} \\ \hat{b} = -med(\hat{a}\,x - \hat{y}) \end{cases}$$

МНМ-оценки обладают свойством робастности. Но на практике решение реализуется только численно.

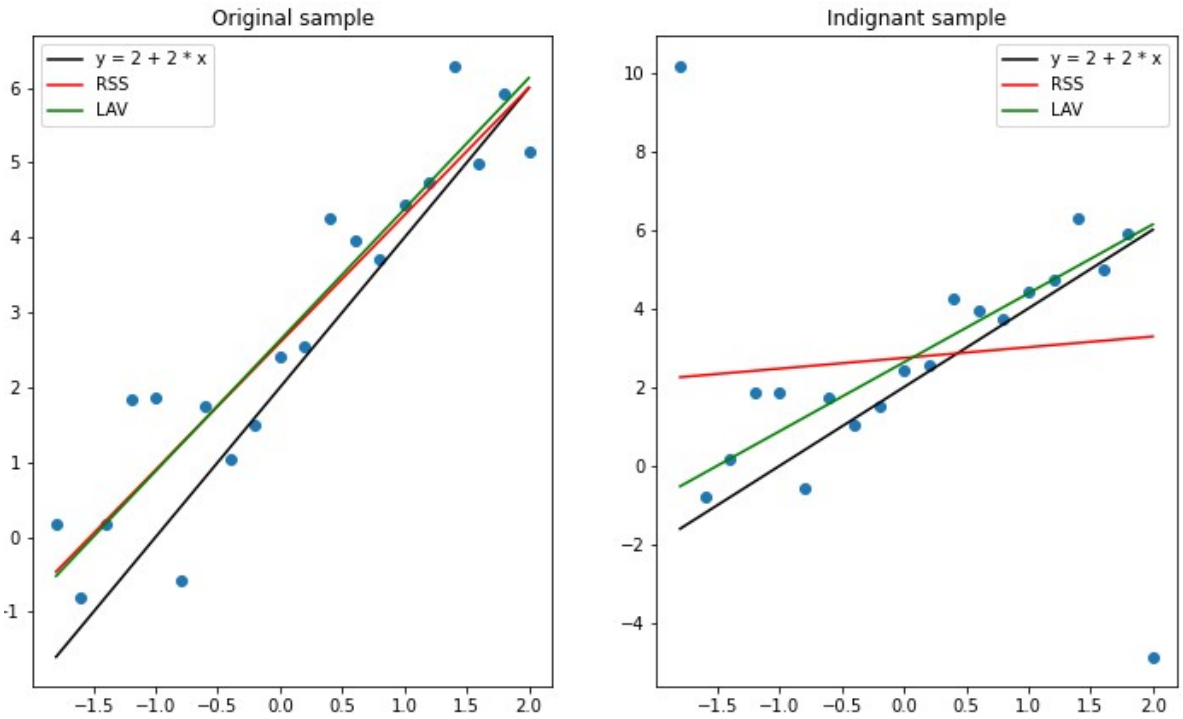
3 Реализация

Был использован язык *Python 3.8.2*: модуль *numpy* для генерации выборок на основе стандартного нормального распределения и вычисления описательных статистик, функция *pearsonr* из модуля *scipy.stats* для расчёта коэффициента корреляции Пирсона, функция *minimize* из модуля *scipy.optimize* для минимизации LAD-функции, модуль *matplotlib* для построения и отображения графиков, модуль *pandas* для оптимального хранения статистических данных и функция *display* из модуля *IPython.display* для их корректного отображения в таблицах.

4 Результаты

4.1 Графики

Рис. 2 График получившейся линейной регрессии



4.2 Оценки коэффициентов линейной регрессии

Таблица 3 Коэффициенты при выборке без возмущений

Исходная выборка	$\hat{a}$	$\hat{b}$
RSS	1.70073	2.599262

LAD	1.75065	2.631271

Таблица 4 Коэффициенты при выборке с возмущениями

Выборка с возмущениями	$\hat{a}$	$\hat{b}$
RSS	0.272159	2.742119
LAD	1.751878	2.631394

5 Выводы

По графикам видно, что оба метода дают хорошую оценку, если нет выбросов. Однако выбросы сильно влияют на оценки по МНК.

Выбросы слабо влияют на оценку по МНМ, но ценой за это является бóльшая вычислительная сложность.

6 Литература

Основы работы с *numpy* (отдельная глава курса)

Документация по *scipy*

Pandas обзор

7 Приложения

Код лабораторной