Санкт-Петербургский политехнический университет Петра Великого
Институт прикладной математики и механики
Кафедра прикладной математики

ОТЧЕТ

Тема: ***Линейная регрессия***

Направление: 01.03.02 Прикладная математика и информатика

Выполнил студент гр. 33631/4 Камалетдинова Ю.

Преподаватель Баженов А.

Санкт-Петербург
2019

Содержание

Постановка задачи

Рассматривается линейная модель зависимости данных. Необходимо найти оценки коэффициентов линейной регрессии $y_i = a + bx_i + e_i$, используя $n = 20$ точек на отрезке $[-1.8; 2]$ с равномерным шагом равным 0.2. Ошибку e_i считать нормально распределенной с параметрами (0, 1). В качестве эталонной зависимости взять функцию

$$y_i = 2 + 2x_i + e_i \tag{1}$$

При построении оценок коэффициентов использовать два критерия: критерий наименьших квадратов и критерий наименьших модулей. Требуется проделать описанную работу для выборки, у которой в значения y_1 и y_{20} вносятся возмущения 10 и -10.

Описание алгоритма

Метод наименьших квадратов

Введем обозначение для уравнения прямой, полученного по тому или иному критерию рассогласованности отклика и регрессионной модели

$$\hat{y}_i = \hat{a} + \hat{b}x_i, \tag{2}$$

где $\hat{a}, \hat{b}$ — оценки параметров a, b

Запишем минимизируемое выражение для случая критерия наименьших квадратов (МНК)

$$Q(a, b) = \sum_{i=1}^{n} \epsilon^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2 \to \min_{a,b} \tag{3}$$

Опустим запись необходимых условий экстремума и доказательства минимальности функции (3) в стационарной точке, описанных в [1], и приведем МНК-оценки коэффициентов

$$\hat{b} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} \tag{4}$$

$$\hat{a} = \overline{y} - \overline{x}\hat{b}, \tag{5}$$

где $\overline{x}$, $\overline{x^2}$, $\overline{y}$, $\overline{xy}$ — выборочные первые и вторые начальные моменты

Метод наименьших модулей

Одной из альтернатив МНК является метод наименьших модулей (МНМ)

$$A(a, b) = \sum_{i=1}^{n} |y_i - a - bx_i| \to \min_{a,b} \tag{6}$$

Запишем выражения для оценок (4), (5) в другом виде

$$\hat{b} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{k_{xy}}{s_x s_y} = \frac{k_{xy}}{s_y^2}\frac{s_y}{s_x} = r_{xy}\frac{s_y}{s_x} \tag{7}$$

$$\hat{a} = \overline{y} - \overline{x}\hat{b}, \tag{8}$$

В формулах (7), (8) заменим выборочные средние $\overline{x}$, $\overline{y}$ на выборочные медианы $med\ x$, $med\ y$, а среднеквадратические отклонения s_x, s_y на интерквартильные широты IQR_x, IQR_y; выборочный коэффициент корреляции r_{xy} — на знаковый коэффициент корреляции r_Q

$$\hat{b}_R = r_q \frac{IQR_y}{IQR_x}, \tag{9}$$

$$\hat{a}_R = med\ y - \hat{b}_R\ med\ x, \tag{10}$$

$$r_Q = \frac{1}{n}\sum_{i=1}^{n} sign(x_i - med\ x)\ sign(y_i - med\ y) \tag{11}$$

$$sign\ z = \begin{cases} 1, \ z > 0 \\ 0, \ z = 0 \\ -1, \ z < 0 \end{cases} \tag{12}$$

Формулы (7), (8), (9), (10), (11), (12) указаны в учебнике [1]. Уравнение регрессии примет вид

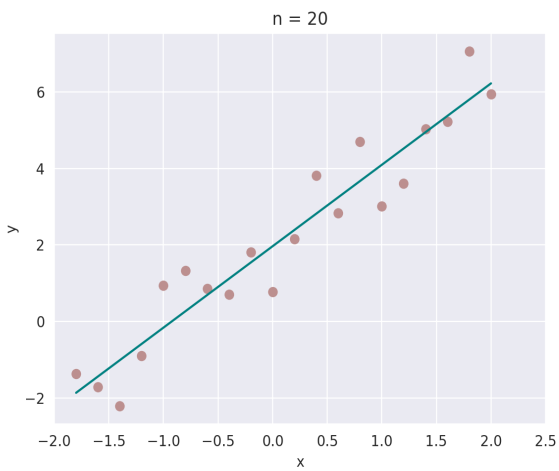
$$y = \hat{a}_R + \hat{b}_R x \tag{13}$$

Реализация

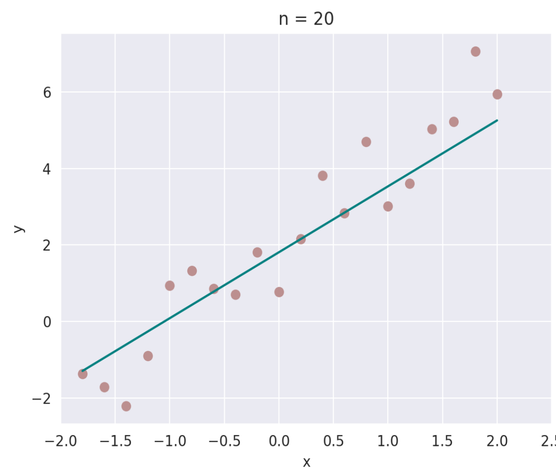
Для выполнения поставленной задачи будем пользоваться библиотеками для языка Python: *numpy, scipy* – расчеты, законы распределения вероятностей; *matplotlib, seaborn* – визуализация результатов. Ход работы:

- Задаем вектор точек $x_n = [-1.8, -1.6, \ldots, 2.0]$ с шагом 0.2, $n = 20$
- Вычисляем вектор значений функции (1)
- Рассчитываем оценки коэффициентов линейной регрессии по формулам (4), (5), (9), (10)
- Вносим возмущения $+10$ и -10 в первое и последнее значения регрессионной функции соответсвтенно и повторяем шаги 2, 3
- Изображаем полученные результаты на графике и сравниваем коэффициенты, рассчитанные по разным критериям

Результат



(a)

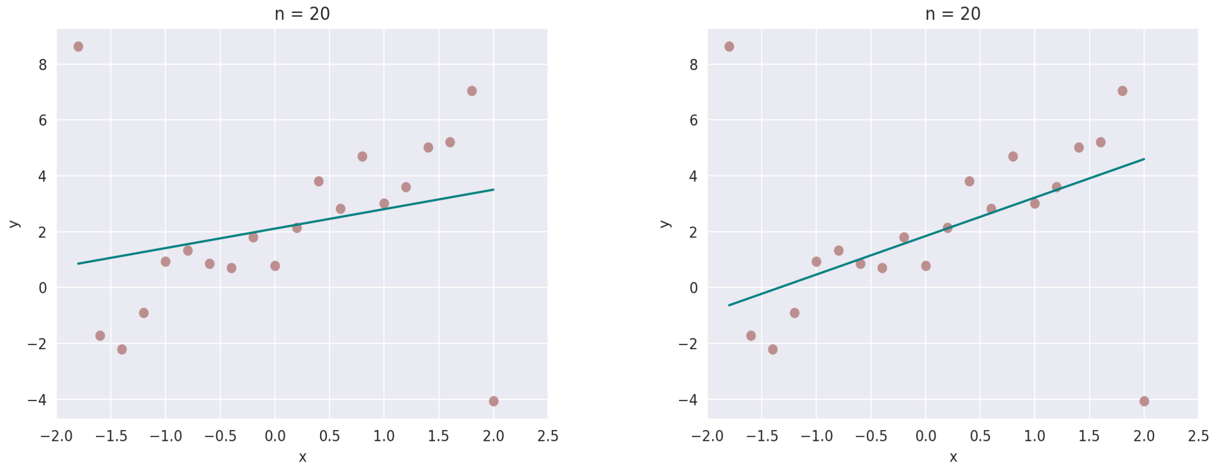


(b)

Рис. 1: График прямой, исходные данные без возмущений (a) МНК; (b) МНМ

1(a) $\hat{a} = 1.9677,\ \hat{b} = 2.1254$

1(b) $\hat{a} = 1.8099,\ \hat{b} = 1.7207$



(a) (b)

Рис. 2: График прямой, исходные данные с возмущениями (a) МНК; (b) МНМ

2(a) $\hat{a} = 2.1106,\ \hat{b} = 0.6968$

2(b) $\hat{a} = 1.8443,\ \hat{b} = 1.3766$

Результаты проведенной работы показывают, что наиболее устойчивым критерием к выбросам является метод наименьших модулей. Выборочная медиана и интерквартильные широты менее чувствительны к выбросам, что и объясняет полученные результаты.

Также можно заметить, что использование метода наименьших квадратов в случае отсутствия наблюдений, не свойственных данной выборке, дает лучшие результаты. Применение МНК при наличии больших по величине выбросов имеет смысл после предварительной отбраковки значений.

Список литературы

[1] *Амосова Н.Н., Куклин Б.А., Макарова С.Б., Максимов Ю.Д., Митрофанова Н.М., Полищук В.И., Шевляков Г.Л.* Вероятностные разделы математики. Учебник для бакалавров технических направлений. — СПб.: Иван Федоров, 2001. — 592 с.: илл. — ISBN 5-81940-050-X.