

Санкт-Петербургский политехнический университет Петра Великого

Институт прикладной математики и механики

Кафедра «Прикладная математика»

Отчёт

по лабораторной работе №6

по дисциплине

«Математическая статистика»

Выполнил студент

В. А. Рыженко

Проверил:

к.ф.-м.н., доцент

Баженов Александр Николаевич

Санкт-Петербург, 2020 г.

Содержание

1. Постановка задачи	3
2. Теория	3
2.1. Простая линейная регрессия	3
2.2. Двумерное нормальное распределение	3
2.2.1. Метод наименьших квадратов	3
2.2.2. Расчётные формулы для МНК-оценок	4
2.3. Робастные оценки коэффициентов линейной регрессии	5
2.4. Оценка качества	6
3. Реализация	6
4. Результаты	7
4.1. Оценки коэффициентов линейной регрессии	7
4.1.1. Выборка без возмущений	7
4.1.2. Выборка с возмущениями	7
5. Обсуждение	8
6. Приложения	8

Список иллюстраций

1	Выборка без возмущений	7
2	Выборка с возмущениями	8

1. Постановка задачи

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 $y_i = 2 + 2x_i + e_i$. При построении оценок коэффициентов использовать два критерия: критерий наименьших квадратов и критерий наименьших модулей. Прodelать то же самое для выборки, у которой в значения y_1 и y_{20} вносятся возмущения 10 и -10.

2. Теория

2.1. Простая линейная регрессия

2.2. Двумерное нормальное распределение

Регрессионную модель описания данных называют простой линейной регрессией, если

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, i = 1..n \quad (1)$$

где x_1, \dots, x_n — заданные числа (значения фактора); y_1, \dots, y_n — наблюдаемые значения отклика; $\epsilon_1, \dots, \epsilon_n$ — независимые, нормально распределённые $N(0, \sigma)$ с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); β_0, β_1 — неизвестные параметры, подлежащие оцениванию. В модели (1) отклик y зависит от одного фактора x , и весь разброс экспериментальных точек объясняется только погрешностями наблюдений (результатов измерений) отклика y . Погрешности результатов измерений x в этой модели полагают существенно меньшими погрешностей результатов измерений y , так что ими можно пренебречь [1, с. 507].

2.2.1. Метод наименьших квадратов

При оценивании параметров регрессионной модели используют различные методы. Один из наиболее распространённых подходов заключается в следующем: вводится мера (критерий) рассогласования отклика и регрессионной функции, и оценки параметров регрессии определяются так, чтобы сделать это рассогласование наименьшим. Достаточно простые расчётные формулы для оценок получают при выборе критерия в виде суммы квадратов отклонений значений отклика от значений регрессионной функции (сумма квадратов остатков):

$$Q(\beta_0, \beta_1) = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \rightarrow \min_{\beta_0, \beta_1} \quad (2)$$

Задача минимизации квадратичного критерия (2) носит название задачи метода наименьших квадратов (МНК), а оценки $\hat{\beta}_0, \hat{\beta}_1$ параметров β_0, β_1 , реализующие минимум критерия (2), называют МНК-оценками [1, с. 508].

2.2.2. Расчётные формулы для МНК-оценок

МНК-оценки параметров $\hat{\beta}_0, \hat{\beta}_1$ находятся из условия обращения функции $Q(\beta_0, \beta_1)$ в минимум.

Для нахождения МНК-оценок $\hat{\beta}_0, \hat{\beta}_1$ выпишем необходимые условия экстремума

$$\begin{cases} \frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0 \\ \frac{\partial Q}{\partial \beta_1} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \end{cases} \quad (3)$$

Далее для упрощения записи сумм будем опускать индекс суммирования. Из системы (3) получим:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum x_i = \sum y_i \\ \hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i \end{cases} \quad (4)$$

Разделим оба уравнения на n :

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1 (\frac{1}{n} \sum x_i) = \frac{1}{n} \sum y_i \\ \hat{\beta}_0 (\frac{1}{n} \sum x_i) + \hat{\beta}_1 (\frac{1}{n} \sum x_i^2) = \frac{1}{n} \sum x_i y_i \end{cases} \quad (5)$$

и, используя известные статистические обозначения для выборочных первых и вторых начальных моментов

$$\bar{x} = \frac{1}{n} \sum x_i, \bar{y} = \frac{1}{n} \sum y_i, \bar{x}^2 = \frac{1}{n} \sum x_i^2, \bar{x}\bar{y} = \frac{1}{n} \sum x_i y_i, \quad (6)$$

получим

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = \bar{y} \\ \hat{\beta}_0 \bar{x} + \hat{\beta}_1 \bar{x}^2 = \bar{x}\bar{y}, \end{cases} \quad (7)$$

откуда МНК-оценку $\hat{\beta}_1$ наклона прямой регрессии находим по формуле Крамера

$$\hat{\beta}_1 = \frac{\bar{x}\bar{y} - \bar{x} \cdot \bar{y}}{\bar{x}^2 - (\bar{x})^2} \quad (8)$$

а МНК-оценку $\hat{\beta}_0$ определяем непосредственно из первого уравнения системы (7):

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \quad (9)$$

Заметим, что определитель системы (7):

$$\bar{x}^2 - (\bar{x})^2 = \frac{1}{n} \sum (x_i - \bar{x})^2 = s_x^2 > 0, \quad (10)$$

если среди значений x_1, \dots, x_n есть различные, что и будем предполагать.

Доказательство минимальности функции $Q(\beta_0, \beta_1)$ в стационарной точке проведём

с помощью известного достаточного признака экстремума функции двух переменных. Имеем:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n, \frac{\partial^2 Q}{\partial \beta_1^2} = 2 \sum x_i^2 = 2n\bar{x}^2, \frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0} = 2 \sum x_i = 2n\bar{x} \quad (11)$$

$$\Delta = \frac{\partial^2 Q}{\partial \beta_0^2} \cdot \frac{\partial^2 Q}{\partial \beta_1^2} - \left(\frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0} \right)^2 = 4n^2 \bar{x}^2 - 4n^2 (\bar{x})^2 = 4n^2 [\bar{x}^2 - (\bar{x})^2] = 4n^2 \left[\frac{1}{n} \sum (x_i - \bar{x})^2 \right] = 4n^2 s_x^2 > 0. \quad (12)$$

Этот результат вместе с условием $\frac{\partial^2 Q}{\partial \beta_0^2} = 2n > 0$ означает, что в стационарной точке функция Q имеет минимум.

2.3. Робастные оценки коэффициентов линейной регрессии

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

$$\sum_{i=1}^n |y_i - \beta_0 - \beta_1 x_i| \rightarrow \min_{\beta_0, \beta_1} \quad (13)$$

Напомним, что 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 от задач метода наименьших квадратов, на практике задача (13) решается численно. Соответствующие процедуры представлены в некоторых современных пакетах программ по статистическому анализу.

Здесь мы рассмотрим простейшую в вычислительном отношении робастную альтернативу оценкам коэффициентов линейной регрессии по МНК. Для этого сначала запишем выражения для оценок (9) и (8) в другом виде:

$$\hat{\beta}_1 = \frac{\bar{x}\bar{y} - \bar{x} \cdot \bar{y}}{\bar{x}^2 - (\bar{x})^2} = \frac{k_{xy}}{s_x^2} = \frac{k_{xy}}{s_x s_y} \cdot \frac{s_y}{s_x} = r_{xy} \frac{s_y}{s_x}, \hat{\beta}_0 = \bar{y} - \bar{x} \hat{\beta}_1 \quad (14)$$

В формулах (14) заменим выборочные средние \bar{x} и \bar{y} соответственно на робастные выборочные медианы $\text{med } x$ и $\text{med } y$, среднеквадратические отклонения s_x и s_y на робастные нормированные интерквартильные широты q_x^* и q_y^* , выборочный коэффициент корреляции r_{xy} — на знаковый коэффициент корреляции r_Q :

$$\hat{\beta}_{1R} = r_Q \frac{q_y^*}{q_x^*}, \quad (15)$$

$$\hat{\beta}_{0R} = \text{med } y - \hat{\beta}_{1R} \text{med } x, \quad (16)$$

$$r_Q = \frac{1}{n} \sum_{i=1}^n \text{sgn}(x_i - \text{med}x) \text{sgn}(y_i - \text{med}y), \quad (17)$$

$$\begin{aligned} q_y^* &= \frac{y_{(j)} - y_{(l)}}{k_q(n)}, q_x^* = \frac{x_{(j)} - x_{(l)}}{k_q(n)}, \\ &\begin{cases} \lfloor \frac{n}{4} \rfloor + 1 \text{ при } \frac{n}{4} \text{ дробном,} \\ \frac{n}{4} \text{ при } \frac{n}{4} \text{ целом.} \end{cases} \\ &j = n - l + 1 \\ \text{sgn}(z) &= \begin{cases} 1 \text{ при } z > 0 \\ 0 \text{ при } z = 0 \\ -1 \text{ при } z < 0 \end{cases} \end{aligned} \quad (18)$$

Уравнение регрессии здесь имеет вид

$$y = \hat{\beta}_{0R} + \hat{\beta}_{1R}x \quad (19)$$

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 $\text{sgn} z$ чувствительна не к величине аргумента, а только к его знаку. Отсюда оценка прямой регрессии (19) обладает очевидными робастными свойствами устойчивости к выбросам по координате y , но она довольно груба

2.4. Оценка качества

Проводить оценку качества работы метод будем проводить при помощи метода наименьших квадратов 13

3. Реализация

Лабораторная работа выполнена с помощью встроенных средств языка программирования Python в среде разработки Visual Code. Исходный код лабораторной работы приведён в приложении.

4. Результаты

4.1. Оценки коэффициентов линейной регрессии

4.1.1. Выборка без возмущений

- Критерий наименьших квадратов: $\hat{a} \approx 2.36$, $\hat{b} \approx 2.06$, quality ≈ 12.63
- Критерий наименьших модулей: $\hat{a} \approx 2.65$, $\hat{b} \approx 2.14$, quality ≈ 13.79

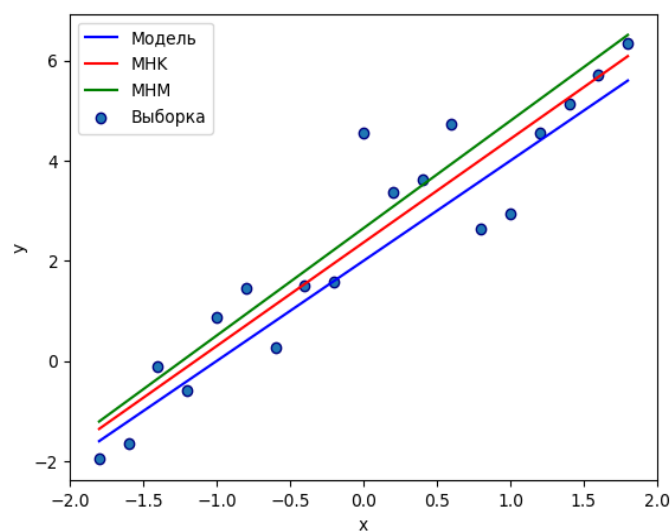


Рис. 1. Выборка без возмущений

4.1.2. Выборка с возмущениями

- Критерий наименьших квадратов: $\hat{a} \approx 2.36$, $\hat{b} \approx 0.48$, quality ≈ 38.23
- Критерий наименьших модулей: $\hat{a} \approx 2.65$, $\hat{b} \approx 1.69$, quality ≈ 30.76

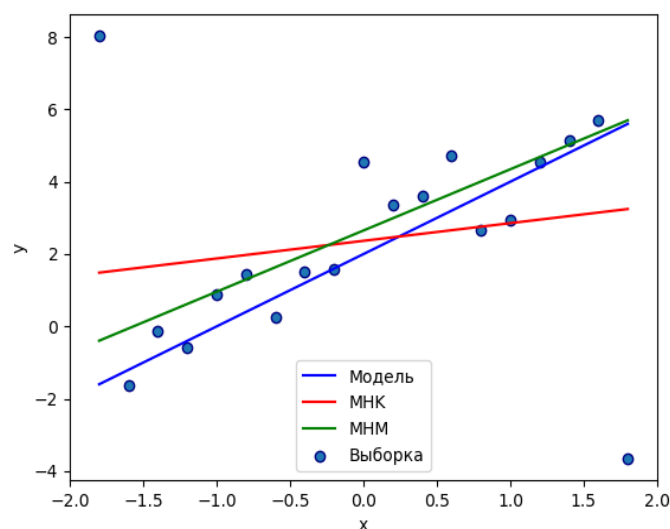


Рис. 2. Выборка с возмущениями

5. Обсуждение

- Критерий наименьших квадратов точнее оценивает коэффициенты линейной регрессии на выборке без возмущений.
- Для выборки с возмущениями результат получается точнее при оценке критерием наименьших модулей.
- Таким образом, критерий наименьших модулей устойчив к редким выбросам, в отличие от критерия наименьших квадратов.

6. Приложения

Репозиторий на GitHub с релизацией: github.com.

Список литературы

- [1] Вероятностные разделы математики. Учебник для бакалавров технических направлений. // Под ред. Максимова Ю.Д. — Спб.: «Иван Федоров», 2001. — 592 с., илл.
- [2] Вентцель Е.С. Теория вероятностей: Учеб. для вузов. — 6-е изд. стер. — М.: Высш. шк., 1999. — 576 с.