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 группы 3630102/70401

Мельникова Анна Николаевна

Проверил
к. ф.-м. н., доцент

Баженов Александр Николаевич

Санкт-Петербург
2020

Содержание

Список таблиц

Список иллюстраций

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. Проделать 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

2.1.1 Модель простой линейной регрессии

Регрессионную модель (гипотезу, которая должна быть подвергнута статистической проверке) описания данных называют *простой линейной регрессией*, если

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = \overline{1, n}, \tag{1}$$

где

- $x_1, ..., x_n$ — заданные числа (значения фактора);
- $y_1, ..., y_n$ — наблюдаемые значения отклика;
- $\varepsilon_1, ..., \varepsilon_n$ — независимые, нормально распределённые $\sim N(0, \sigma)$: 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 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].

2.1.2 Метод наименьших квадратов

При оценивании параметров регрессионной модели (β_0, β_1) используют различные методы. Один из наиболее распрстранённых 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):

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

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

2.1.3 Расчётные формулы для МНК-оценок

МНК-оценки параметров $\widehat{\beta}_0, \widehat{\beta}_1$ находятся из условия обращения функции $Q(\beta_0, \beta_1)$ в минимум.
Для нахождения МНК-оценок $\widehat{\beta}_0, \widehat{\beta}_1$ выпишем необходимые условия экстремума

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

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

$$\begin{cases} n\widehat{\beta}_0 + \widehat{\beta}_1 \sum x_i = \sum y_i \\ \widehat{\beta}_0 \sum x_i + \widehat{\beta}_1 \sum x_i^2 = \sum x_i y_i \end{cases}$$

Разделим оба уравнения на n:

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

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

$$\overline{x} = \frac{1}{n} \sum x_i, \overline{y} = \frac{1}{n} \sum y_i, \overline{x^2} = \frac{1}{n} \sum x_i^2, \overline{xy} = \frac{1}{n} \sum x_i y_i,$$

получим

$$\begin{cases} \widehat{\beta}_0 + \widehat{\beta}_1 \overline{x} = \overline{y}, \\ \widehat{\beta}_0 \overline{x} + \widehat{\beta}_1 \overline{x^2} = \overline{xy}, \end{cases} \tag{4}$$

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

$$\widehat{\beta}_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} \tag{5}$$

а МНК-оценку $\widehat{\beta}_0$ определяем непосредственно из первого уравнения системы (4):

$$\widehat{\beta}_0 = \overline{y} - \overline{x}\widehat{\beta}_1 \tag{6}$$

Заметим, что определитель системы (4):

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

если среди значений $x_1, ..., x_n$ 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:

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n, \frac{\partial^2 Q}{\partial \beta_1^2} = 2\sum x_i^2 = 2n\overline{x^2}, \frac{\partial^2 Q}{\partial \beta_1 \partial \beta_0} = 2\sum x_i = 2n\overline{x}$$

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

Этот результат вместе с условием $\frac{\partial^2 Q}{\partial \beta_0^2} = 2n > 0$ означает, что в стационарной точке функция Q имеет минимум [1, с. 508-511].

2.2 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:

$$\sum_{i=1}^{n} |y_i - \beta_0 - \beta_1 x_i| \to \min_{\beta_0, \beta_1} \qquad (7)$$

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 (7) решается численно. Соответствующие процедуры представлены в некоторых современных пакетах программ по статистическому анализу.

Здесь мы рассмотрим простейшую в вычистлительном отношении робастную альтернативу оценкам коэффициентов линейной регрессии по МНК. Для этого сначала запишем выражения для оценок (5) и (6) в другом виде:

$$\widehat{\beta}_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - (\overline{x})^2} = \frac{k_{xy}}{s_x^2} = \frac{k_{xy}}{s_x s_y} \cdot \frac{s_y}{s_x} = r_{xy}\frac{s_y}{s_x}, \qquad \widehat{\beta}_0 = \overline{y} - \overline{x}\widehat{\beta}_1 \qquad (8)$$

В формулах (8) заменим выборочные средние $\bar{x}$ и $\bar{y}$ соответственно на робастные выборочные медианы $medx$ и $medy$, 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:

$$\widehat{\beta}_{1R} = r_Q \frac{q_y^*}{q_x^*},$$

$$\widehat{\beta}_{0R} = medy - \widehat{\beta}_{1R} medx,$$

$$r_Q = \frac{1}{n} \sum_{i=1}^{n} sgn(x_i - medx) sgn(y_i - medy),$$

$$q_y^* = \frac{y_{(j)} - y_{(l)}}{k_q(n)}, q_x^* = \frac{x_{(j)} - x_{(l)}}{k_q(n)},$$

$$\begin{cases} [\frac{n}{4}] + 1 \text{ при } \frac{n}{4} \text{ дробном,} \\ \frac{n}{4} \text{ при } \frac{n}{4} \text{ целом.} \end{cases}$$

$$j = n - l + 1$$

$$sgn(z) = \begin{cases} 1 \text{ при } z > 0 \\ 0 \text{ при } z = 0 \\ -1 \text{ при } z < 0 \end{cases}$$

Уравнение регрессии здесь имеет вид

$$y = \hat{\beta}_{0R} + \hat{\beta}_{1R} x \tag{9}$$

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 $sgn(z)$ чувствительна не к величине аргумента, а только к его знаку. Отсюда оценка прямой регрессии (9) обладает очевидными робастными свойствами устойчивости к выбросам по координате y, но она довольно груба [1, с. 518-519].

3 Реализация

Лабораторная работа выполнена с помощью встроенных средств языка программирования Python в среде разработки PyCharm (для визуализации использовался пакет matplotlib, для вычислений - numpy). Исходный код лабораторной работы и tex-файлы отчета приведёны в приложении в виде ссылки на репозиторий GitHub.

4 Результаты

4.1 Оценки коэффициентов линейной регрессии

4.1.1 Выборка без возмущений

- Критерий наименьших квадратов:

$$\hat{a}_{ls} \approx 1.86, \hat{b}_{ls} \approx 2.18$$

- Критерий наименьших квадратов:

$$\hat{a}_{lm} \approx 1.9, \hat{b}_{lm} \approx 2.02$$

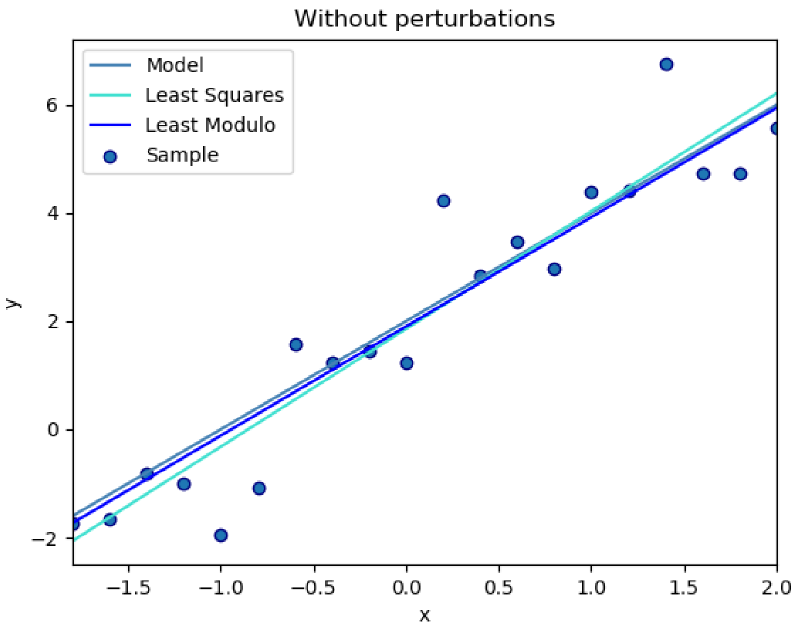

Рис. 1: Выборка без возмущений

4.1.2 Выборка с возмущениями

- Критерий наименьших квадратов:

$$\hat{a}_{ls} \approx 2.0, \hat{b}_{ls} \approx 0.77$$

- Критерий наименьших квадратов:

$$\hat{a}_{lm} \approx 1.91, \hat{b}_{lm} \approx 1.95$$

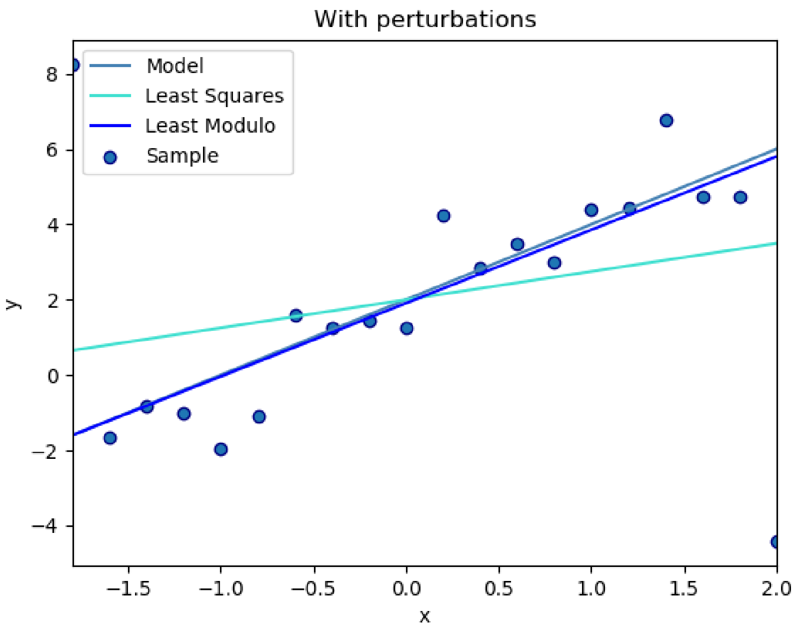

Рис. 2: Выборка с возмущениями

5 Обсуждение

Оценим, какие коэффициенты, полученные каким методом лучше аппроксимируют модельную зависимость: для МНК и МНМ для выборки с возмущениями и без них посчитаем сумму по всем $x \in [-1.8, 2]$, взятым с шагом 0.2:

$$Distance_{ls(lm)} = \sum (y_{model} - y_{ls(lm)})^2.$$

$$y_{model} = 2 + 2 \cdot x$$

$$y_{ls} = \hat{a}_{ls} + \hat{b}_{ls} \cdot x$$

$$y_{ls} = \hat{a}_{lm} + \hat{b}_{lm} \cdot x$$

Для данных выборок получаем:

- Выборка без возмущений:

  $$Distance_{ls} < Distance_{lm}$$

  $$0.563 < 1.637$$

  Критерий наименьших квадратов точнее оценивает коэффициенты линейной регрессии на выборке без возмущений.

- Выборка без возмущений:

  $$Distance_{lm} < Distance_{ls}$$

$$1.393 < 43.805$$

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, что соответствует ожиданиям, ведь он обладает робастными свойствами.

6 Приложения

Код программы - GitHub URL: https://github.com/kaustika/Statistics2020

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.