Санкт-Петербургский Политехнический Университет
Петра Великого

Институт прикладной математики и механики

Кафедра "Прикладная математика"

Отчёт
Лабораторная работа №6
по дисциплине
"Математическая статистика"

Выполнил студент:
Салихов С.Р.
группа: 3630102/70401

Проверил:
к.ф-м.н., доцент
Баженов Александр Николавич

Санкт-Петербург
2020 г.

Содержание

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 описания данных называют простой линейной регрессией, если

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

, i = 1, ... , n, где x_1, ... , x_n — заданные числа (значения фактора); y_1, ... ,y_n — наблюдаемые значения отклика; ϵ_1, ... , ϵ_n — независимые, нормально распределённые N(0, δ) с нулевым математическим ожиданием и одинаковой (неизвестной) дисперсией случайные величины (ненаблюдаемые); β_0, β_1 — неизвестные параметры, подлежащие оцениванию.
В модели 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.

2.2 Метод наименьших квадратов

При оценивании параметров регрессионной модели 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):

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

Задача минимизации квадратичного критерия (10) носит название задачи метода наименьших квадратов (МНК), а оценки $\hat{\beta}_0$, $\hat{\beta}_1$ параметров β_0, β_1, реализующие минимум критерия, называют МНК-оценками.

2.2.1 Расчётные формулы для МНК-оценок

МНК-оценки параметров $\hat{\beta}_0$ и $\hat{\beta}_1$ находятся из условия обращения функции Q(β_0, β_1) в минимум.
Для нахождения МНК-оценок $\hat{\beta}_0$ и $\hat{\beta}_1$ выпишем необходимые условия экстремума:

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

Далее для упрощения записи сумм будем опускать индекс суммирования. Из системы получим

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

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

Тогда:

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \overline{y} \\ \hat{\beta}_0 \overline{x} + \hat{\beta}_1 \overline{x^2} = \overline{xy} \end{cases}$$

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

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x} * \overline{y}}{\overline{x^2} - \overline{x}^2}$$

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

$$\hat{\beta}_0 = \overline{y} - \overline{x}\hat{\beta}_1$$

2.3 Метод наименьших модулей

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

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

3 Реализация

Для генерации выборки был использован *Python* 3.7 и модуль numpy. Для отрисовки графиков использовался модуль matplotlib. scipy.stats для обработки функций распределений.

4 Результаты

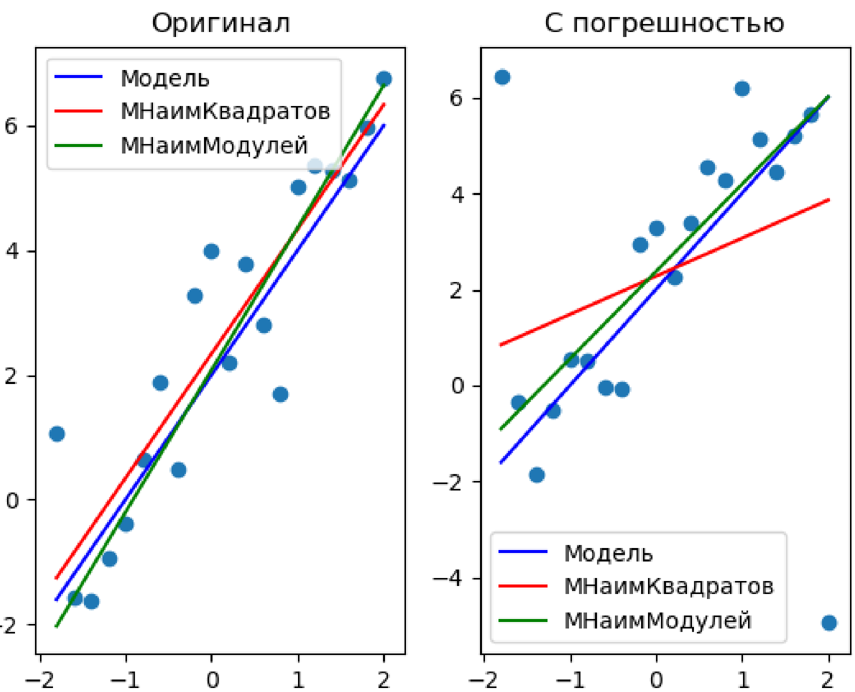

Рис. 1: Графики линейной регрессии при выборке с возмущением и без

4.1 Выборка без возмущений

Критерий наименьших квадратов:

$$\hat{a} \approx 1.93, \hat{b} \approx 2.19$$

Критерий наименьших модулей:

$$\hat{a} \approx 2.24, \hat{b} \approx 1.77$$

4.2 Выборка с возмущниями

Критерий наименьших квадратов:

$$\hat{a} \approx 0.48, \hat{b} \approx 1.76$$

Критерий наименьших модулей:

$$\hat{a} \approx 1.85, \hat{b} \approx 1.39$$

5 Обсуждение

1)МНК оценивает коэффициенты линейной регрессии точнее, на выборке без возмущений.

2)На выборке с возмущениями эффективнее использовать МНМ. Таким образом, метод наименьших модулей устойчив к редким выбросам, в свою очередь МНМ обладает большей сложностью вычислений, чем МНК.

6 Литература

Модуль numpy

Модуль matplotlib

Модуль scipy

Метод наименьших модулей

Метод наименьших квадратов

7 Приложения

Код лабораторной
Код отчёта