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:

Кондратьев Д. А.
группа: 3630102/70301

Проверил:

к.ф.-м.н., доцент
Баженов Александр Николаевич

Санкт-Петербург
2020 г.

Содержание

Список иллюстраций

Список таблиц

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

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

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, σ) 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].

2.2. 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):

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

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

2.2.1. Расчётные формулы для МНК-оценок

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

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

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

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

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

Разделим оба уравнения на n 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

$$\begin{cases} \hat{\beta}_0 + \hat{\beta}_1 \overline{x} = \overline{y} \\ \hat{\beta}_0 \overline{x} + \hat{\beta}_1 \overline{x^2} = \overline{xy} \end{cases} \tag{5}$$

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

$$\hat{\beta}_1 = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2} \tag{6}$$

а МНК-оценку $\hat{\beta}_0$ определяем непосредственно из первого уравнения системы (5):

$$\hat{\beta}_0 = \overline{y} - \overline{x} \hat{\beta}_1 \tag{7}$$

2.3. Метод наименьших модулей

Критерий наименьших модулей — заключается в минимизации следующей функции:

$$M(a,b) = \sum_{i=1}^{n} |y_i - ax_i - b| \to \min \tag{8}$$

3. Реализация

Лабораторная работа выполнена на программном языке *Python 3.8* в среде разработки *Jupyter Notebook 6.0.3*. В работе использовались следующие пакеты языка *Python*:

- *numpy* — для генерации выборки и работы с массивами;
- *matplotlib.pyplot* и *seaborn* — для построения графиков;
- *scipy.optimize* — для решения задач оптимизации;

Ссылка на исходный код лабораторной работы приведена в приложении.

4. Результаты

4.1. Оценки коэффициентов линейной регрессии

Выборка без возмущений

Критерий	a	b
МНК	1.502	1.878
МНМ	1.709	1.563

Выборка с возмущениями

Критерий	a	b
МНК	0.394	1.895
МНМ	1.457	1.564

Таблица 1. Оценки коэффициентов линейной регрессии

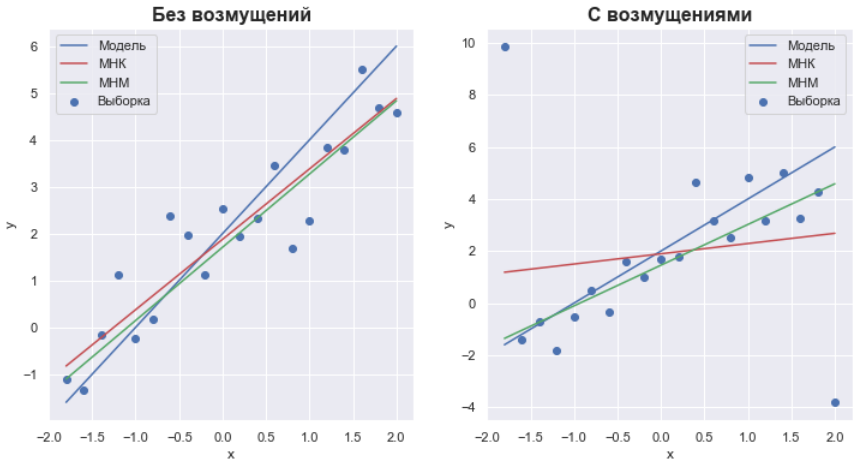

Рис. 1. Линейная регрессия

Введем следующую метрику:

$$\rho = \sum (y_i^* - y_i)^2$$

где y_i — значение эталонной функции в точке x_i, y_i^* — значение функции в точке x_i, полученное путем оценки.

Теперь посчитаем данные метрики для МНК и МНМ:

- выборка без возмущений: $\rho_{\text{МНК}} = 7.18,\ \rho_{\text{МНМ}} = 7.33$;
- выборка с возмущениями:$\rho_{\text{МНК}} = 70.06,\ \rho_{\text{МНМ}} = 11.94$.

5. Обсуждение

Исходя из полученных результатов можно сделать следующие выводы:

- Критерий наименьших квадратов точнее оценивает коэффициенты линейной регрессии на выборке без возмущений, так как $\rho_{\text{МНК}} < \rho_{\text{МНМ}}$.
- Критерий наименьших модулей точнее оценивает коэффициенты линейной регрессии на выборке с возмущениями, так как $\rho_{\text{МНМ}} < \rho_{\text{МНК}}$.

- Критерий наименьших модулей устойчив к редким выбросам по сравнению с критерием наименьших квадратов. Но при этом обладает большей вычислительной сложностью из-за необходимости решения задачи минимизации.

6. Литература

1) **Вероятностные разделы математики.** Учебник для бакалавров технических направлений.//Под ред. Максимова Ю.Д. — Спб.: «Иван Федоров», 2001. — 592 с., илл.

2) Least squares. URL: `https://en.wikipedia.org/wiki/Least_squares`

3) Least absolute deviations. URL: `https://en.wikipedia.org/wiki/Least_absolute_deviations`

7. Приложение

1) Код лабораторной. URL: `https://github.com/DmitriiKondratev/MatStat/blob/master/Lab_6/Lab_6.ipynb`

2) Код отчёта. URL: `https://github.com/DmitriiKondratev/MatStat/blob/master/Lab_6/Lab_report_6.tex`