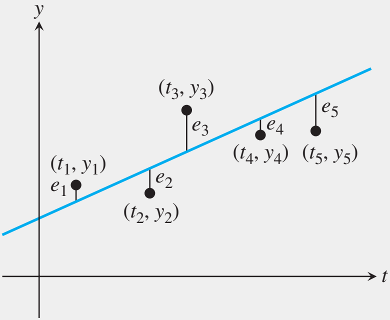
Метод наименьших квадратов (Least Squares)

Вычислительная математика. Лекция 6

Исупов К.С.

November 8, 2021



Введение

Экспериментальный анализ

Пусть изучаемый процесс характеризуется двумя величинами:

- x — независимая переменная
- y — зависимая переменная

В результате проведения эксперимента значениям $X = (x_1, x_2, \dots, x_n)$ поставлены в соответствие значения $Y = (y_1, y_2, \dots, y_n)$:

x_i	x_1	x_2	x_3	x_4	$\dots$	x_i	$\dots$	x_n
y_i	y_1	y_2	y_3	y_4	$\dots$	y_i	$\dots$	y_n

Требуется определить форму связи между исследуемыми величинами, выявить *существенные* факторы.

Экспериментальный анализ

Пусть изучаемый процесс характеризуется двумя величинами:

- x — независимая переменная
- y — зависимая переменная

В результате проведения эксперимента значениям $X = (x_1, x_2, \ldots, x_n)$ поставлены в соответствие значения $Y = (y_1, y_2, \ldots, y_n)$:

x_i	x_1	x_2	x_3	x_4	$\ldots$	x_i	$\ldots$	x_n
y_i	y_1	y_2	y_3	y_4	$\ldots$	y_i	$\ldots$	y_n

Требуется определить форму связи между исследуемыми величинами, выявить *существенные* факторы.

Особенности

- Число данных в таблице, n, может быть очень большим.
- Значения y_i получены с погрешностями.

Экспериментальный анализ

Пусть изучаемый процесс характеризуется двумя величинами:

- x — независимая переменная
- y — зависимая переменная

В результате проведения эксперимента значениям $X = (x_1, x_2, \dots, x_n)$ поставлены в соответствие значения $Y = (y_1, y_2, \dots, y_n)$:

x_i	x_1	x_2	x_3	x_4	$\dots$	x_i	$\dots$	x_n
y_i	y_1	y_2	y_3	y_4	$\dots$	y_i	$\dots$	y_n

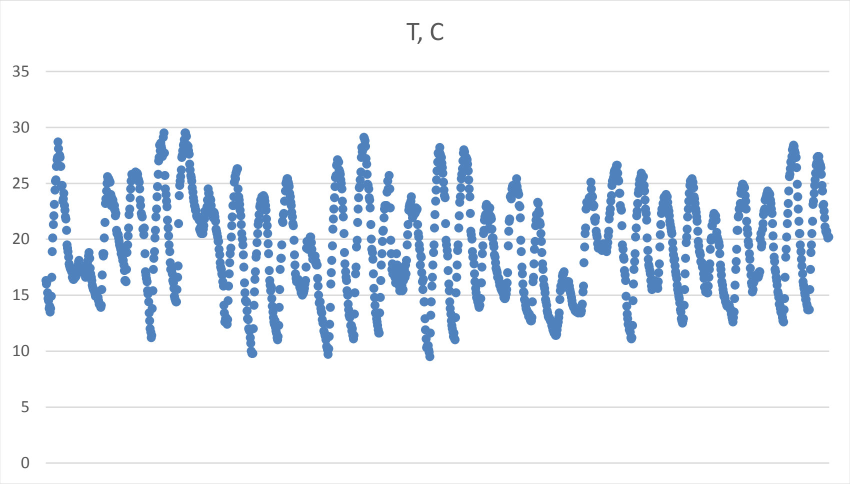
Требуется определить форму связи между исследуемыми величинами, выявить *существенные* факторы.

Особенности

- Число данных в таблице, n, может быть очень большим.
- Значения y_i получены с погрешностями.

Применение интерполяции для обработки таких данных неэффективно!

Пример: данные о температуре за месяц с промежутком 30 минут.



Задача

Задача

В результате проведения эксперимента значениям $X = (x_1, x_2, \ldots, x_n)$ поставлены в соответствие значения $Y = (y_1, y_2, \ldots, y_n)$:

x_i	x_1	x_2	x_3	x_4	$\ldots$	x_i	$\ldots$	x_n
y_i	y_1	y_2	y_3	y_4	$\ldots$	y_i	$\ldots$	y_n

Требуется определить *достаточно простую* (с небольшим числом параметров) зависимость, которая лучшим возможным образом связывает переменные x и y. Такая зависимость называется *эмпирической*.

Задача

В результате проведения эксперимента значениям $X = (x_1, x_2, \ldots, x_n)$ поставлены в соответствие значения $Y = (y_1, y_2, \ldots, y_n)$:

x_i	x_1	x_2	x_3	x_4	$\ldots$	x_i	$\ldots$	x_n
y_i	y_1	y_2	y_3	y_4	$\ldots$	y_i	$\ldots$	y_n

Требуется определить *достаточно простую* (с небольшим числом параметров) зависимость, которая лучшим возможным образом связывает переменные x и y. Такая зависимость называется *эмпирической*.

Этапы решения задачи:

1. Выбор вида эмпирической зависимости (формулы).
2. Определение наилучших значений параметров выбранной зависимости.

Выбор вида эмпирической зависимости

- Не имеет четкой математической постановки.
- Базируется на опыте и интуиции исследователя.
- Стремятся, чтобы формула имела достаточно простую структуру и небольшое количество параметров, подлежащих определению.

Выбор вида эмпирической зависимости

- Не имеет четкой математической постановки.
- Базируется на опыте и интуиции исследователя.
- Стремятся, чтобы формула имела достаточно простую структуру и небольшое количество параметров, подлежащих определению.

После того как *вид* зависимости выбран, выполняется второй этап — определение *наилучших значений параметров* выбранной зависимости. На данном этапе применяют метод наименьших квадратов.

Применение метода наименьших квадратов при обработке экспериментальных данных

Метод наименьших квадратов

Концепция

За наилучшие значения параметров зависимости $f(x)$ принимают такие, для которых **сумма квадратов отклонений** экспериментальных значений y_i (исходные данные в таблице) от вычисленных по эмпирической формуле $y_i^T = f(x_i)$ имеет наименьшее значение:

$$F(a, b, c, \dots) = \sum_{i=1}^{n} (y_i - f(x_i))^2 \to \min . \tag{1}$$

Метод наименьших квадратов

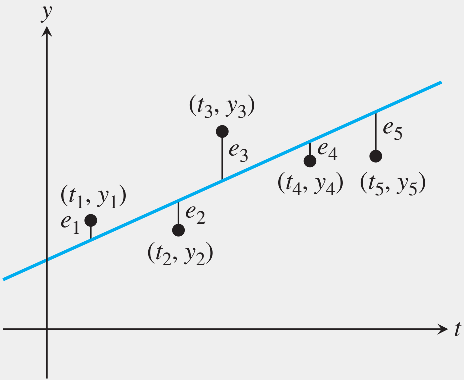
Концепция

За наилучшие значения параметров зависимости $f(x)$ принимают такие, для которых **сумма квадратов отклонений** экспериментальных значений y_i (исходные данные в таблице) от вычисленных по эмпирической формуле $y_i^T = f(x_i)$ имеет наименьшее значение:

$$F(a, b, c, \dots) = \sum_{i=1}^{n} (y_i - f(x_i))^2 \to \min . \tag{1}$$

Таким образом, основная цель метода наименьших квадратов — решение задачи минимизации (1).

Метод наименьших квадратов



Лучшей среди зависимостей вида $y = at + b$ является та, при которой квадрат ошибки $e_1^2 + e_2^2 + \cdots + e_5^2$ является наименьшим.

Метод наименьших квадратов

Сжатие данных

Метод наименьших квадратов — классический пример сжатия данных. Входные данные состоят из набора точек, а выходными данными является модель (эмпирическая зависимость), которая при относительно небольшом числе параметров максимально соответствуют данным. Обычно причиной использования наименьших квадратов является замена зашумленных данных на правдоподобную базовую модель, которая используется в дальнейшем для прогнозирования сигнала или в целях классификации.

Метод наименьших квадратов для линейной зависимости

Пусть эмпирическая зависимость является линейной:

$$y = ax + b$$

Метод наименьших квадратов для линейной зависимости

Пусть эмпирическая зависимость является линейной:

$$y = ax + b$$

Для нее минимизируется функция:

$$F(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \to \min \tag{2}$$

Метод наименьших квадратов для линейной зависимости

Пусть эмпирическая зависимость является линейной:

$$y = ax + b$$

Для нее минимизируется функция:

$$F(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \to \min \tag{2}$$

Условия экстремума функции (2):

$$\frac{\partial F}{\partial a} = 0, \quad \frac{\partial F}{\partial b} = 0.$$

Метод наименьших квадратов для линейной зависимости

Пусть эмпирическая зависимость является линейной:

$$y = ax + b$$

Для нее минимизируется функция:

$$F(a, b) = \sum_{i=1}^{n} (y_i - ax_i - b)^2 \to \min \tag{2}$$

Условия экстремума функции (2):

$$\frac{\partial F}{\partial a} = 0, \quad \frac{\partial F}{\partial b} = 0.$$

Легко видеть, что:

$$\frac{\partial F}{\partial a} = \sum_{i=1}^{n} (y_i - ax_i - b)x_i, \quad \frac{\partial F}{\partial b} = \sum_{i=1}^{n} (y_i - ax_i - b).$$

Метод наименьших квадратов для линейной зависимости

То есть, для нахождения наилучших значений параметров a и b линейной зависимости нужно решить следующую систему уравнений:

$$\begin{cases} \frac{\partial F}{\partial a} = \sum_{i=1}^{n}(y_i - ax_i - b)x_i = 0 \\ \frac{\partial F}{\partial b} = \sum_{i=1}^{n}(y_i - ax_i - b) = 0 \end{cases} \tag{3}$$

Метод наименьших квадратов для линейной зависимости

То есть, для нахождения наилучших значений параметров a и b линейной зависимости нужно решить следующую систему уравнений:

$$\begin{cases} \frac{\partial F}{\partial a} = \sum_{i=1}^{n}(y_i - ax_i - b)x_i = 0 \\ \frac{\partial F}{\partial b} = \sum_{i=1}^{n}(y_i - ax_i - b) = 0 \end{cases} \tag{3}$$

Для упрощения, преобразуем ее к следующему виду:

$$\begin{cases} a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \end{cases} \tag{4}$$

Решение этой системы позволяет найти наилучшие значения параметров a и b линейной зависимости.

Метод наименьших квадратов для квадратичной зависимости

Пусть эмпирическая зависимость является квадратичной параболой:

$$y = ax^2 + bx + c$$

Метод наименьших квадратов для квадратичной зависимости

Пусть эмпирическая зависимость является квадратичной параболой:

$$y = ax^2 + bx + c$$

Для нее минимизируется функция:

$$F(a, b, c) = \sum_{i=1}^{n} (y_i - ax_i^2 - bx_i - c)^2 \to \min \tag{5}$$

Метод наименьших квадратов для квадратичной зависимости

Система уравнений для нахождения параметров a, b, c:

$$\begin{cases} \frac{\partial F}{\partial a} = \sum_{i=1}^{n}(y_i - ax_i^2 - bx^i - c)x_i^2 = 0 \\ \frac{\partial F}{\partial b} = \sum_{i=1}^{n}(y_i - ax_i^2 - bx_i - c)x_i = 0 \\ \frac{\partial F}{\partial c} = \sum_{i=1}^{n}(y_i - ax_i^2 - bx_i - c) = 0 \end{cases} \tag{6}$$

Метод наименьших квадратов для квадратичной зависимости

Система уравнений для нахождения параметров a, b, c:

$$\begin{cases} \frac{\partial F}{\partial a} = \sum_{i=1}^{n} (y_i - ax_i^2 - bx^i - c)x_i^2 = 0 \\ \frac{\partial F}{\partial b} = \sum_{i=1}^{n} (y_i - ax_i^2 - bx_i - c)x_i = 0 \\ \frac{\partial F}{\partial c} = \sum_{i=1}^{n} (y_i - ax_i^2 - bx_i - c) = 0 \end{cases} \tag{6}$$

Преобразуем ее к виду:

$$\begin{cases} a\sum_{i=1}^{n} x_i^4 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i \\ a\sum_{i=1}^{n} x_i^3 + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \\ a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i + cn = \sum_{i=1}^{n} y_i \end{cases} \tag{7}$$

Решение этой системы позволяет найти наилучшие значения параметров *a, b* и *c* линейной зависимости.

Замена переменных

Множество аналитических зависимостей могут быть сведены к линейной путем замены переменных:

- Показательная зависимость:

$$y = ab^x \to \ln y = Y, \ln a = A, \ln b = B \to Y = A + Bx$$

Замена переменных

Множество аналитических зависимостей могут быть сведены к линейной путем замены переменных:

- Показательная зависимость:

$$y = ab^x \to \ln y = Y, \ln a = A, \ln b = B \to Y = A + Bx$$

- Степенная зависимость:

$$y = ax^b \to \ln y = Y, \ln a = A, \ln x = X \to Y = A + bX$$

Замена переменных

Множество аналитических зависимостей могут быть сведены к линейной путем замены переменных:

- Показательная зависимость:

$$y = ab^x \to \ln y = Y, \ln a = A, \ln b = B \to Y = A + Bx$$

- Степенная зависимость:

$$y = ax^b \to \ln y = Y, \ln a = A, \ln x = X \to Y = A + bX$$

- Гиперболическая зависимость:

$$y = a + b/x \to 1/x = X \to y = a + bX$$

Замена переменных

Множество аналитических зависимостей могут быть сведены к линейной путем замены переменных:

- Показательная зависимость:

$$y = ab^x \to \ln y = Y, \ln a = A, \ln b = B \to Y = A + Bx$$

- Степенная зависимость:

$$y = ax^b \to \ln y = Y, \ln a = A, \ln x = X \to Y = A + bX$$

- Гиперболическая зависимость:

$$y = a + b/x \to 1/x = X \to y = a + bX$$

- Дробно-рациональная зависимость:

$$y = x/(ax + b) \to 1/y = Y, 1/x = X \to Y = a + bX$$

Замена переменных

Множество аналитических зависимостей могут быть сведены к линейной путем замены переменных:

- Показательная зависимость:

$$y = ab^x \to \ln y = Y, \ln a = A, \ln b = B \to Y = A + Bx$$

- Степенная зависимость:

$$y = ax^b \to \ln y = Y, \ln a = A, \ln x = X \to Y = A + bX$$

- Гиперболическая зависимость:

$$y = a + b/x \to 1/x = X \to y = a + bX$$

- Дробно-рациональная зависимость:

$$y = x/(ax + b) \to 1/y = Y, 1/x = X \to Y = a + bX$$

- Логарифмическая зависимость:

$$y = a \ln x + b) \to \ln x = X \to y = aX + b$$

Пример (T. Sauer, Numerical Analysis, 2012)

Измеренные значения температуры в г. Вашингтон, округ Колумбия, 1 января 2001 года:

time of day	t	temp (C)
12 mid.	0	-2.2
3 am	$\frac{1}{8}$	-2.8
6 am	$\frac{1}{4}$	-6.1
9 am	$\frac{3}{8}$	-3.9
12 noon	$\frac{1}{2}$	0.0
3 pm	$\frac{5}{8}$	1.1
6 pm	$\frac{3}{4}$	-0.6
9 pm	$\frac{7}{8}$	-1.1

- Выбранная зависимость (модель), которая учитывает цикличность изменения температуры в краткосрочном периоде:

$$y = c_1 + c_2 \cos 2\pi t + c_3 \sin 2\pi t$$

Пример (T. Sauer, Numerical Analysis, 2012)

Подстановка исходных данных в модель дает переопределенную систему линейных уравнений:

$$
\begin{aligned}
c_1 + c_2 \cos 2\pi(0) + c_3 \sin 2\pi(0) &= -2.2 \\
c_1 + c_2 \cos 2\pi\left(\frac{1}{8}\right) + c_3 \sin 2\pi\left(\frac{1}{8}\right) &= -2.8 \\
c_1 + c_2 \cos 2\pi\left(\frac{1}{4}\right) + c_3 \sin 2\pi\left(\frac{1}{4}\right) &= -6.1 \\
c_1 + c_2 \cos 2\pi\left(\frac{3}{8}\right) + c_3 \sin 2\pi\left(\frac{3}{8}\right) &= -3.9 \\
c_1 + c_2 \cos 2\pi\left(\frac{1}{2}\right) + c_3 \sin 2\pi\left(\frac{1}{2}\right) &= 0.0 \\
c_1 + c_2 \cos 2\pi\left(\frac{5}{8}\right) + c_3 \sin 2\pi\left(\frac{5}{8}\right) &= 1.1 \\
c_1 + c_2 \cos 2\pi\left(\frac{3}{4}\right) + c_3 \sin 2\pi\left(\frac{3}{4}\right) &= -0.6 \\
c_1 + c_2 \cos 2\pi\left(\frac{7}{8}\right) + c_3 \sin 2\pi\left(\frac{7}{8}\right) &= -1.1
\end{aligned}
$$

Пример (T. Sauer, Numerical Analysis, 2012)

Применение наименьших квадратов позволяет получить следующие наилучшие значения параметров модели (выбранной зависимости):

$$c_1 = -1.95, \quad c_2 = -0.7445 \quad c_3 = -2.5594.$$

Т.е., лучшей версией модели в смысле наименьших квадратов является:

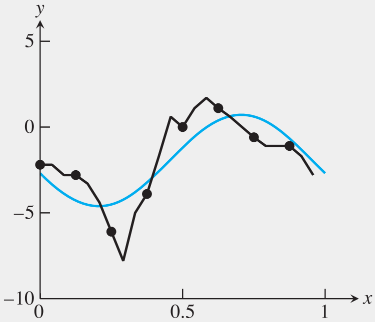
$$y = -1.95 - 0.7445 \cos 2\pi t - 2.5594 \sin 2\pi t \tag{8}$$

Применение наименьших квадратов позволяет получить следующие наилучшие значения параметров модели (выбранной зависимости):

$$c_1 = -1.95, \quad c_2 = -0.7445 \quad c_3 = -2.5594.$$

Т.е., лучшей версией модели в смысле наименьших квадратов является:

$$y = -1.95 - 0.7445 \cos 2\pi t - 2.5594 \sin 2\pi t \tag{8}$$



Пример (T. Sauer, Numerical Analysis, 2012)

- Выберем уточненную модель:

$$y = c_1 + c_2 \cos 2\pi t + c_3 \sin 2\pi t + c_4 \cos 4\pi t$$

Подставим исходные данные в модель:

$$\begin{aligned}
c_1 + c_2 \cos 2\pi(0) + c_3 \sin 2\pi(0) + c_4 \cos 4\pi(0) &= -2.2 \\
c_1 + c_2 \cos 2\pi\left(\frac{1}{8}\right) + c_3 \sin 2\pi\left(\frac{1}{8}\right) + c_4 \cos 4\pi\left(\frac{1}{8}\right) &= -2.8 \\
c_1 + c_2 \cos 2\pi\left(\frac{1}{4}\right) + c_3 \sin 2\pi\left(\frac{1}{4}\right) + c_4 \cos 4\pi\left(\frac{1}{4}\right) &= -6.1 \\
c_1 + c_2 \cos 2\pi\left(\frac{3}{8}\right) + c_3 \sin 2\pi\left(\frac{3}{8}\right) + c_4 \cos 4\pi\left(\frac{3}{8}\right) &= -3.9 \\
c_1 + c_2 \cos 2\pi\left(\frac{1}{2}\right) + c_3 \sin 2\pi\left(\frac{1}{2}\right) + c_4 \cos 4\pi\left(\frac{1}{2}\right) &= 0.0 \\
c_1 + c_2 \cos 2\pi\left(\frac{5}{8}\right) + c_3 \sin 2\pi\left(\frac{5}{8}\right) + c_4 \cos 4\pi\left(\frac{5}{8}\right) &= 1.1 \\
c_1 + c_2 \cos 2\pi\left(\frac{3}{4}\right) + c_3 \sin 2\pi\left(\frac{3}{4}\right) + c_4 \cos 4\pi\left(\frac{3}{4}\right) &= -0.6 \\
c_1 + c_2 \cos 2\pi\left(\frac{7}{8}\right) + c_3 \sin 2\pi\left(\frac{7}{8}\right) + c_4 \cos 4\pi\left(\frac{7}{8}\right) &= -1.1,
\end{aligned}$$

Пример (T. Sauer, Numerical Analysis, 2012)

Применение наименьших квадратов позволяет получить следующие наилучшие значения параметров модели:

$$c_1 = -1.95, \quad c_2 = -0.7445 \quad c_3 = -2.5594 \quad c_4 = 1.125$$

Т.е., лучшей версией модели в смысле наименьших квадратов является:

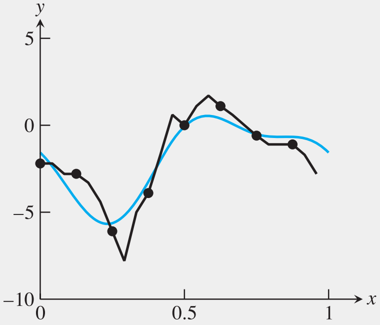
$$y = -1.95 - 0.7445 \cos 2\pi t - 2.5594 \sin 2\pi t + 1.125 \cos 4\pi t. \qquad (9)$$

Применение наименьших квадратов позволяет получить следующие наилучшие значения параметров модели:

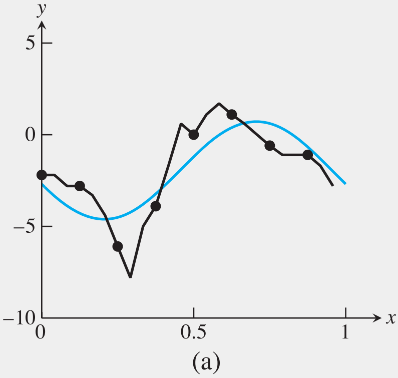
$$c_1 = -1.95, \quad c_2 = -0.7445 \quad c_3 = -2.5594 \quad c_4 = 1.125$$

Т.е., лучшей версией модели в смысле наименьших квадратов является:

$$y = -1.95 - 0.7445 \cos 2\pi t - 2.5594 \sin 2\pi t + 1.125 \cos 4\pi t. \qquad (9)$$



Сравнение двух моделей:



end