Lecture 1.8

- Least Squares Method for Linear Regression

Least Squares Method

- **The Least Squares Method** is a mathematical approach used to find the best-fitting line (or curve) to a given set of data points by minimizing the sum of the squares of the differences between the observed values and the values predicted by the model
- It is commonly used in regression analysis to estimate the parameters of a linear or nonlinear model

Idea

- Given a set of data points $(\boldsymbol{x_1}, \boldsymbol{y_1}), (\boldsymbol{x_2}, \boldsymbol{y_2}), \dots, (\boldsymbol{x_n}, \boldsymbol{y_n})$ the goal is to find the function $y = f(x)$ (e.g., a line $\boldsymbol{y = mx + c}$) that **minimizes the residuals** (the differences between observed and predicted values)
- The residuals are squared to avoid canceling out positive and negative differences

Mathematical Formulation

- The residual for each data point is

$$r_i = y_i - (mx_i + c)$$

- The sum of squared residuals is

$$S = \sum_{i=1}^{n} (y_i - (mx_i + c))^2$$

- We minimize S by taking partial derivatives with respect to m and c, and setting them to zero

$$\frac{\partial S}{\partial m} = 0 \text{ and } \frac{\partial S}{\partial c} = 0$$

Derivative with respect to m

$$\frac{\partial S}{\partial m} = \frac{\partial \sum_{i=1}^{n}(y_i - (mx_i + c))^2}{\partial m}$$

$$\Rightarrow \frac{\partial S}{\partial m} = -2\sum_{i=1}^{n} x_i(y_i - (mx_i + c))$$

$$Taking\, \frac{\partial S}{\partial m} = 0$$

$$-2\sum_{i=1}^{n} x_i(y_i - (mx_i + c)) = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = m\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_i$$

Derivative with respect to *c*

$$\frac{\partial S}{\partial c} = \frac{\partial \sum_{i=1}^{n}(y_i - (mx_i + c))^2}{\partial c}$$

$$\Rightarrow \frac{\partial S}{\partial c} = -2\sum_{i=1}^{n}(y_i - (mx_i + c))$$

$$Taking\ \frac{\partial S}{\partial c} = 0$$

$$-2\sum_{i=1}^{n}(y_i - (mx_i + c)) = 0$$

$$\Rightarrow \sum_{i=1}^{n} y_i = m\sum_{i=1}^{n} x_i + nc$$

Solution for m and c

$$m = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

$$c = \frac{\sum y_i - m\sum x_i}{n}$$

Example 1

x_i	y_i
1	1.2
2	1.9
3	3.0
4	4.1

Solution

x_i	y_i	x_i^2	x_iy_i
1	1.2	1	1.2
2	1.9	4	3.8
3	3.0	9	9.0
4	4.1	16	16.4
$\sum x_i = 10$	$\sum y_i = 10.2$	$\sum x_i^2 = 30$	$\sum x_iy_i = 30.4$

$$m = \frac{n\sum x_iy_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{4 \times 30.4 - 10 \times 10.2}{4 \times 30 - 10^2} = 0.98$$

$$c = \frac{\sum y_i - m\sum x_i}{n} = \frac{10.2 - 0.98 \times 10}{4} = 0.1$$