Linear Regression

- **Straight-line linear regression**:
  - involves a response variable y and a single predictor variable x

    $$\mathbf{y = w_0 + w_1\, x}$$

  - w_0 : y-intercept
  - w_1 : slope
  - w_0 & w_1 are **regression coefficients**

Linear regression

- **Method of least squares**: estimates the best-fitting straight line as the one that minimizes the error between the actual data and the estimate of the line.

$$w_1 = \frac{\sum_{i=1}^{|D|}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{|D|}(x_i - \bar{x})^2} \qquad w_0 = \bar{y} - w_1\bar{x}$$

- *D:* a training set
- *x*: values of predictor variable
- *y*: values of response variable
- *|D|* : data points of the form(*x*1, *y*1), (*x*2, *y*2),…, (*x|D|*, *y|D|*).
- $\bar{x}$ *:* the mean value of *x*1, *x*2, … , *x|D|*
- $\bar{y}$ *:* the mean value of *y*1, *y*2, … , *y|D|*

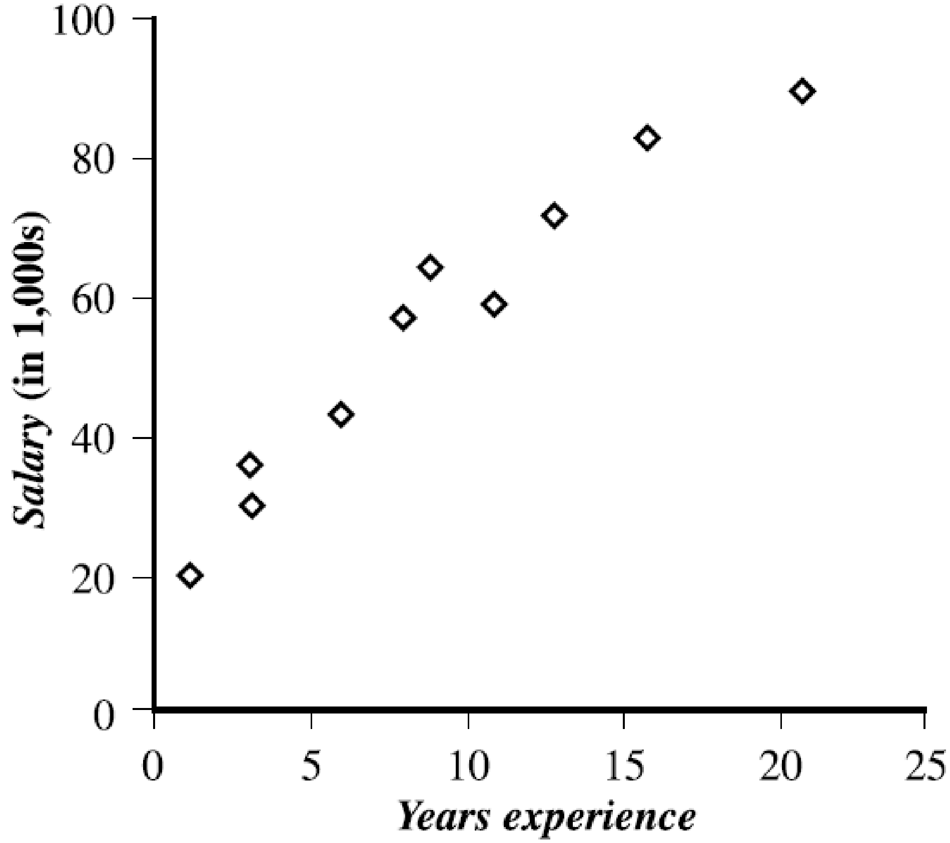
Example: Salary problem

- The table shows a set of paired data where *x* is the number of years of work experience of a college graduate and *y* is the corresponding salary of the graduate.

x years experience	y salary (in \$1000s)
3	30
8	57
9	64
13	72
3	36
6	43
11	59
21	90
1	20
16	83

Linear Regression

- The 2-D data can be graphed on a **scatter plot**.
- The plot suggests a linear relationship between the two variables, *x and y.*



Example: Salary data

- Given the above data, we compute

$$\bar{x} = 9.1 \text{ and } \bar{y} = 55.4$$

- we get

$$w_1 = \frac{(3-9.1)(30-55.4)+(8-9.1)(57-55.4)+\cdots+(16-9.1)(83-55.4)}{(3-9.1)^2+(8-9.1)^2+\cdots+(16-9.1)^2} = 3.5$$

$$w_0 = 55.4-(3.5)(9.1) = 23.6$$

- The equation of the least squares line is estimated by

$$y = 23.6 + 3.5x$$

Example: Salary data