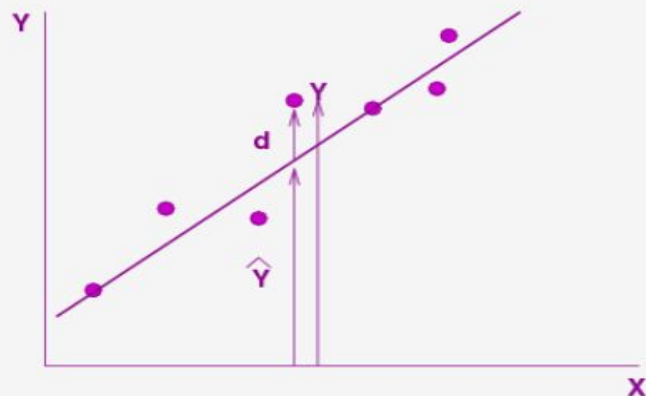


Linear Regression

Regression

- Given a set of data points $x_i; y_i$, what is the relationship between them?
- One kind of question is to ask: are these linearly related in some manner? That is, can we draw a straight line that describes reasonably well the relationship between X and Y ? Remember, the correlation coefficient can tell us if there is a case for such a relationship
- In real life, even if such a relationship held, it will be unreasonable to expect all pairs $x_i; y_i$ to lie precisely on a straight line. Instead, we can probably draw some reasonably well-fitting line. But which one?

Linear Relationship Between 2 Variables I



- ▶ GOAL: fit a line whose equation is of the form $\hat{Y} = a + bX$
- ▶ HOW: minimise $\sum_i d_i^2 = \sum_i (Y_i - \hat{Y}_i)^2$ (the “least squares estimator”)

Linear Relationship Between 2 Variables II

- ▶ The calculation for b is given by:

$$b = \frac{\text{cov}(x, y)}{\text{var}(x)}$$

where $\text{cov}(x, y)$ is the covariance of x and y , given by $\sum_i (x_i - \bar{x})(y_i - \bar{y})$ (see slides on Mathematical Basics)

- ▶ This can be simplified to:

$$b = \sum(xy) / \sum x^2$$

where $x = (X_i - \bar{X})$ and $y = (Y_i - \bar{Y})$

- ▶ $a = \bar{Y} - b\bar{X}$

Do MSE Derivation for getting unbiased estimator for linear regression

