Linear Regression

We define the line of best fit line as

$\hat{y}_i = ax_i + b$

Such that the best fit line looks to minimize the cost function we named E , n is a number of data points.

$$E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$E = \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

To minimize our cost function, E, we must find where the first derivative of E is equal to 0 with respect to a and b. The closer a and b are to 0, the less the total error for each point is. Let's start with the partial derivative of a first.

$$\frac{dE}{da} = -2\sum_{i=1}^{n} x_i(y_i - ax_i - b) = 0$$

$$\sum_{i=1}^{n} y_i x_i - a\sum_{i=1}^{n} x_i^2 - b\sum_{i=1}^{n} x_i = 0$$

$$a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(i)$$

$$\frac{dE}{db} = -2\sum_{i=1}^{n} (y_i - ax_i - b) = 0$$

$$\sum_{i=1}^{n} y_i + a\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} b = 0$$

$$a\sum_{i=1}^{n} x_i + bn = \sum_{i=1}^{n} y_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots..(ii)$$

Substitute Equation (i) and (ii) ....Equation (i) multiply by **n** and Equation (ii) multiply by $\sum_{i=1}^{n} x_i$

$$na\sum_{i=1}^{n} x_i^2 + nb\sum_{i=1}^{n} x_i = n\sum_{i=1}^{n} x_i y_i$$

$$a(\sum_{i=1}^{n} x_i)^2 + nb\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

(-) (-)

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$a = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \quad where\ Cov(x,y) = \frac{1}{n}\sum(x_i - \bar{x})(y_i - \bar{y}) \quad and\ var(x) = \frac{1}{n}\sum(x_i - \bar{x})^2$$

$$b = \bar{y} - a\bar{x} \quad where\ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \quad and\ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$a = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i y_i - x_i\bar{y} - \bar{x} y_i - \bar{x}\bar{y})}{\sum({x_i}^2 - 2x_i\bar{x} + \bar{x}^2)} = \frac{\sum x_i y_i - \bar{y}\sum x_i - \bar{x}\sum y_i + \sum \bar{x}\bar{y}}{\sum {x_i}^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2}$$

$$= \frac{\sum x_i y_i - \bar{y}n\bar{x} \cancel{- \bar{x}n\bar{y} + n\bar{x}\bar{y}}}{\sum {x_i}^2 - 2n\bar{x}^2 + n\bar{x}^2} = \frac{\sum x_i y_i - n\bar{y}\bar{x}}{\sum {x_i}^2 - n\bar{x}^2} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum {x_i}^2 - \sum x_i^2}$$