Regression Analysis

Regression is a statistical technique used to model and analyze the relationship between a dependent variable (response) and one or more independent variable (Predictors).

Linear regression

$$y = \beta_0 + \beta_1 x + \epsilon$$

Multiple Linear regression

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \epsilon$$

Polynomial Linear Regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \dots + \beta_n x^n + \epsilon$$

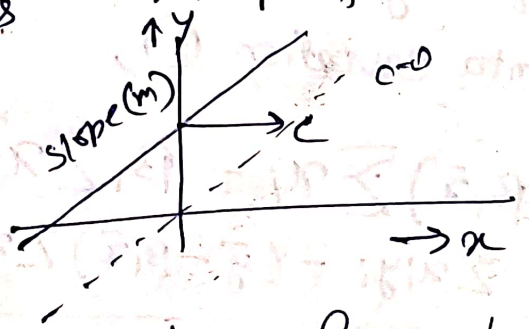
Logistic Regression → Talk about it later/tomorrow.

Slope intercept of line

$y = mx + c$ → y intercept of line (point at which the line crosses the y-axis)

↳ slope of the line (steepness of the line)

variables



$$\text{slope}(m) = \frac{y_2 - y_1}{x_2 - x_1}$$

Unit change in x, how much y-changes.

Simple Linear Regression

$$y_i = \beta_0 + \beta_1 x_i + \epsilon \quad \rightarrow (1)$$

- y_i: dependent variable
- β_0: intercept
- β_1: slope
- x_i: independent variable
- ϵ: random error term

Now we aim to find the values of the parameters β_0, β_1 to reduce/minimize the sum of squared errors.

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

$$\Rightarrow S = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = \sum_{i=1} (y_i - \beta_0 - \beta_1 x_i)$$

To minimize S, we take partial derivative w.r.t β_0, β_1

finding minimum in single variable optimization

$$\frac{\partial S}{\partial \beta_0} = -2 \sum (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\Rightarrow \sum y_i = \beta_0 + \beta_1 \sum x_i \rightarrow (1)$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum x_i (y_i - \beta_0 - \beta_1 x_i) = 0.$$

$$\Rightarrow \sum (x_i y_i - x_i \beta_0 - \beta_1 x_i^2) = 0$$

$$\Rightarrow \sum x_i y_i = \sum x_i \beta_0 + \beta_1 x_i^2 \rightarrow (2)$$

from eqn (1),

$$\Rightarrow \frac{1}{n} \sum y_i = \frac{1}{n} \{ \beta_0 + \beta_1 \sum x_i \}$$

$$\Rightarrow \bar{y} = \beta_0 + \beta_1 \bar{x}$$

$$\Rightarrow \beta_0 = \bar{y} - \beta_1 \bar{x} \rightarrow (3)$$

Substituting eqn (3) into equation (2),

$$\sum x_i y_i = (\bar{y} - \beta_1 \bar{x}) \sum x_i + \beta_1 \sum x_i^2$$

$$\beta_1 = \frac{\sum x_i y_i - (\bar{y} - \beta_1 \bar{x}) \sum x_i}{\sum x_i^2}$$

$$= \frac{\sum y_i - (\bar{y} - \beta_1 \bar{x})}{\sum x_i}$$

$$\Rightarrow \sum x_i y_i = \bar{y} \sum x_i - \beta_1 \bar{x} \sum x_i + \beta_1 \sum x_i^2$$

$$\Rightarrow \sum x_i y_i - \bar{y} \sum x_i = \beta_1 \left(\sum x_i^2 - \bar{x} \sum x_i \right) \rightarrow (4)$$

$$\Rightarrow \beta_1 = \frac{\sum x_i y_i - \bar{y} \sum x_i}{\sum x_i^2 - \bar{x} \sum x_i}$$

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

We know, $\frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x} \Rightarrow \sum x_i = n\bar{x}$

$\therefore \sum x_i^2 - \bar{x}\sum x_i = \sum x_i^2 - n\bar{x}^2$

and $\sum x_i y_i - \bar{y}\sum x_i = \sum x_i y_i - n\bar{x}\bar{y}$

From eqn (4)

$$\sum x_i y_i - n\bar{x}\bar{y} = \beta_1\left(\sum x_i^2 - n\bar{x}^2\right)$$

$$\boxed{\beta_1 = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2}} \rightarrow (5) \Rightarrow \beta_1 = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}$$

$$\sum (x_i-\bar{x})(y_i-\bar{y}) = \sum x_i y_i - n\bar{x}\bar{y}$$

$$= \sum (x_i y_i - x_i\bar{y} - y_i\bar{x} + \bar{x}\bar{y})$$

$$= \sum x_i y_i - \bar{y}\sum x_i - \bar{x}\sum y_i + \sum \bar{x}\bar{y} \quad \Big| \quad \sum x_i = n\bar{x};\ \sum y_i = n\bar{y}$$

and $\bar{x}, \bar{y}$ are const and does not depend on i

$$\sum \bar{x}\bar{y} = n\bar{x}\bar{y}$$

$$= \sum x_i y_i - \bar{y}(n\bar{x}) - \bar{x}(n\bar{y}) + n\bar{x}\bar{y}$$

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

$$= \sum x_i y_i - n\bar{x}\bar{y}$$

$$\boxed{\beta_1 = \frac{\sum (x_i-\bar{x})(y_i-\bar{y})}{\sum (x_i-\bar{x})^2}} =$$

$$\sum (x_i-\bar{x})^2 = \sum (x_i^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2$$

$$= \sum x_i^2 - 2\bar{x}(n\bar{x}) + n\bar{x}^2$$

$$= \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2$$

$$= \sum x_i^2 - n\bar{x}^2$$

All is done to show $\beta_1 = \frac{Cov(x,y)}{Var(x)}$ → tells how 2 variables vary together

+ve cov −ve cov

→ the spread of data