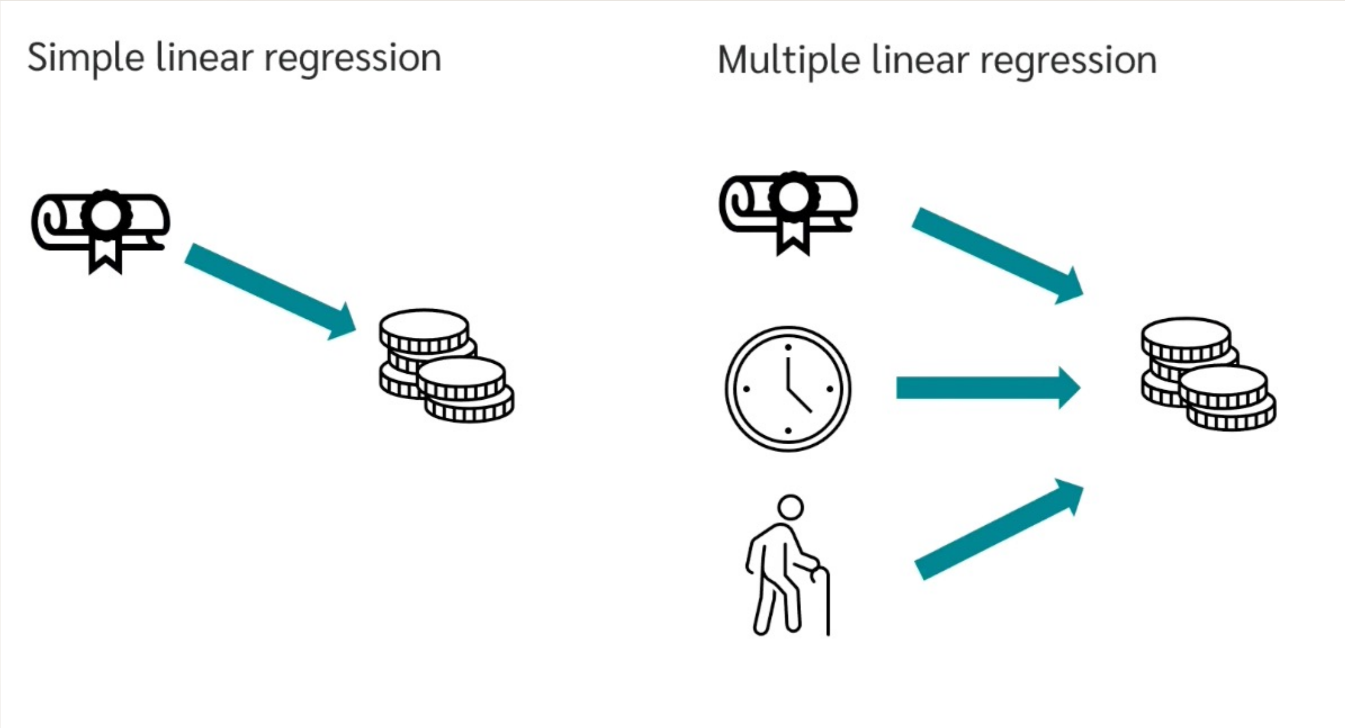
# Logistic Regression Analysis

Dr. Khaing S Htun

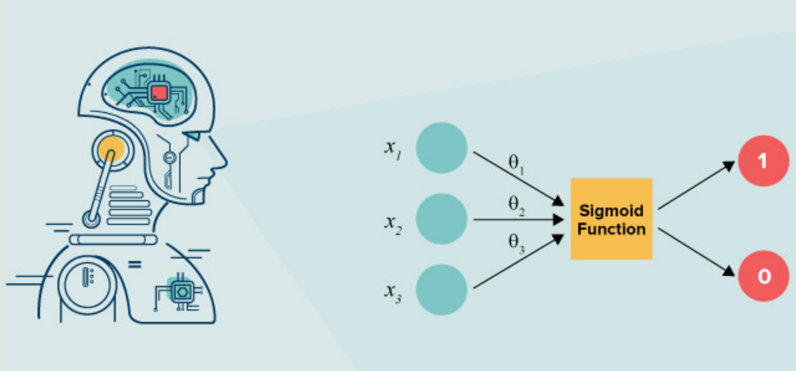


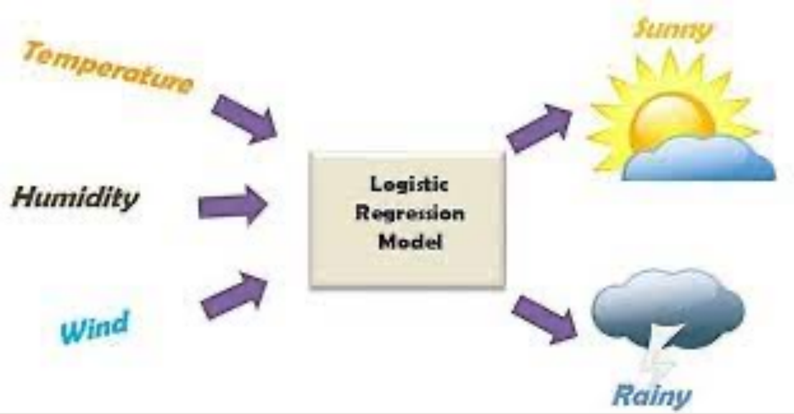

# Linear Regression & Multiple Linear Regression

# Logistic Regression

- classification algorithm.
- predict a discrete/binary outcome (0 or 1)
  - output/dependent variable, y, is **categorical**





# Logistic Regression

# Key assumptions

- The dependent variable is binary or dichotomous
- There should be **no**, or very little, multicollinearity between the predictor variables
- The independent variables should be linearly related to the log odds *(an alternate way of expressing probabilities)*
- Logistic regression requires fairly large sample sizes

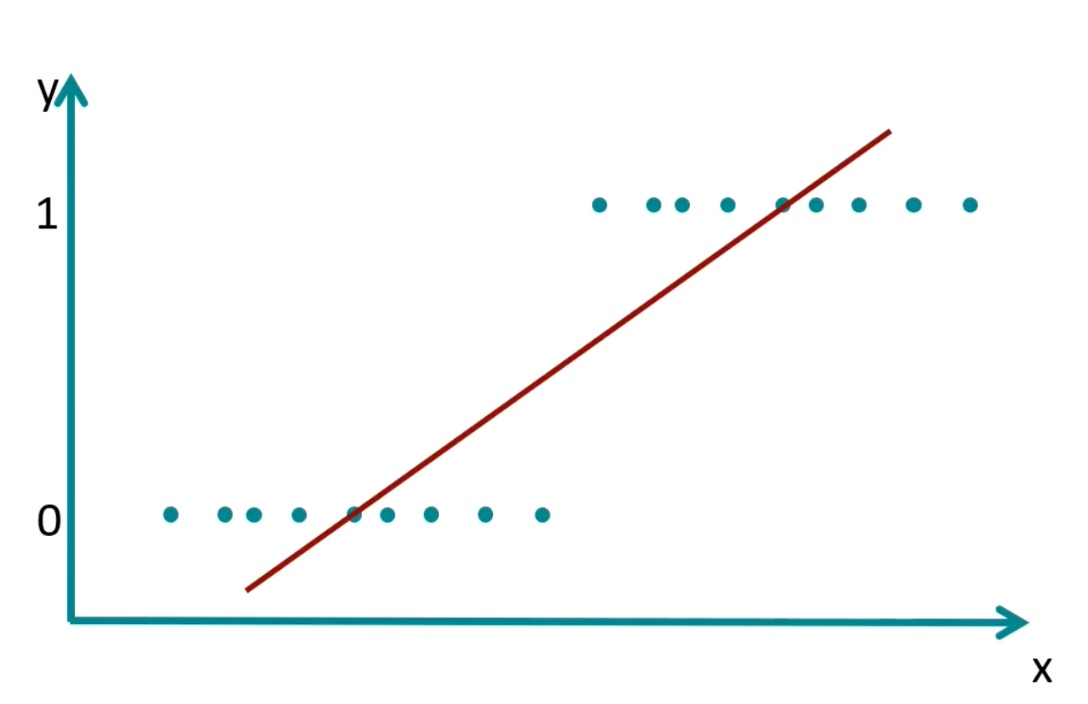
# Type of Logistic Regression Models

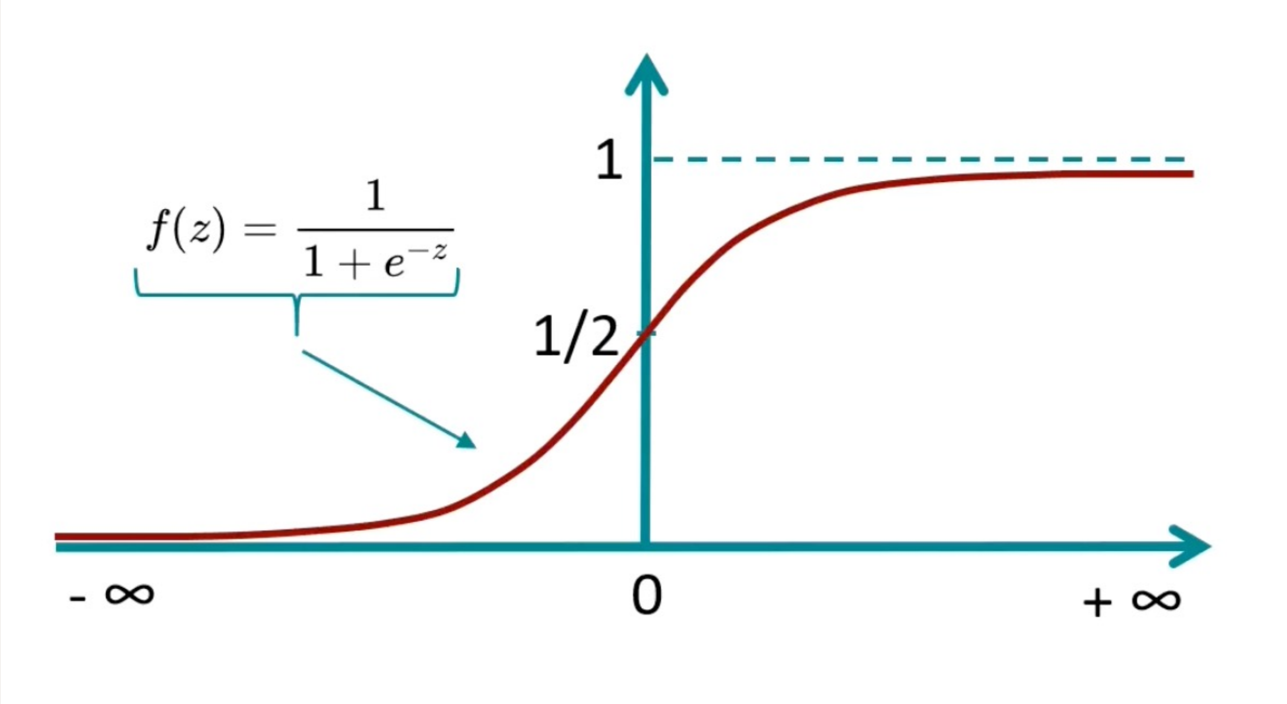
Three types of logistic regression models:

1. **Binary logistic regression**: The response variable can only belong to one of two categories.
2. **Multinomial logistic regression**: The response variable can belong to one of three or more categories and there is no natural ordering among the categories.
3. **Ordinal logistic regression**: The response variable can belong to one of three or more categories and there *is* a natural ordering among the categories.

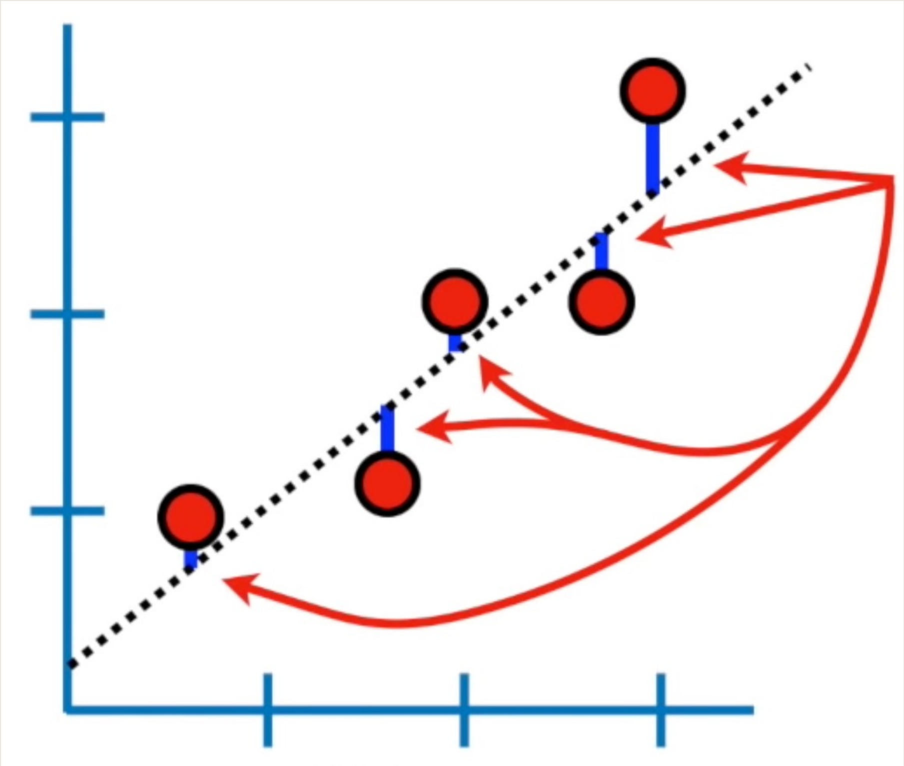
| | Binomial Logistic Regression | Multinomial Logistic Regression | Ordinal Logistic Regression |
|---|---|---|---|
| Number of categories for y | 2 | 2 or more | 3 or more |
| Does order of categories matter? | No | No | Yes |

# Why not Linear but Logistics Regression

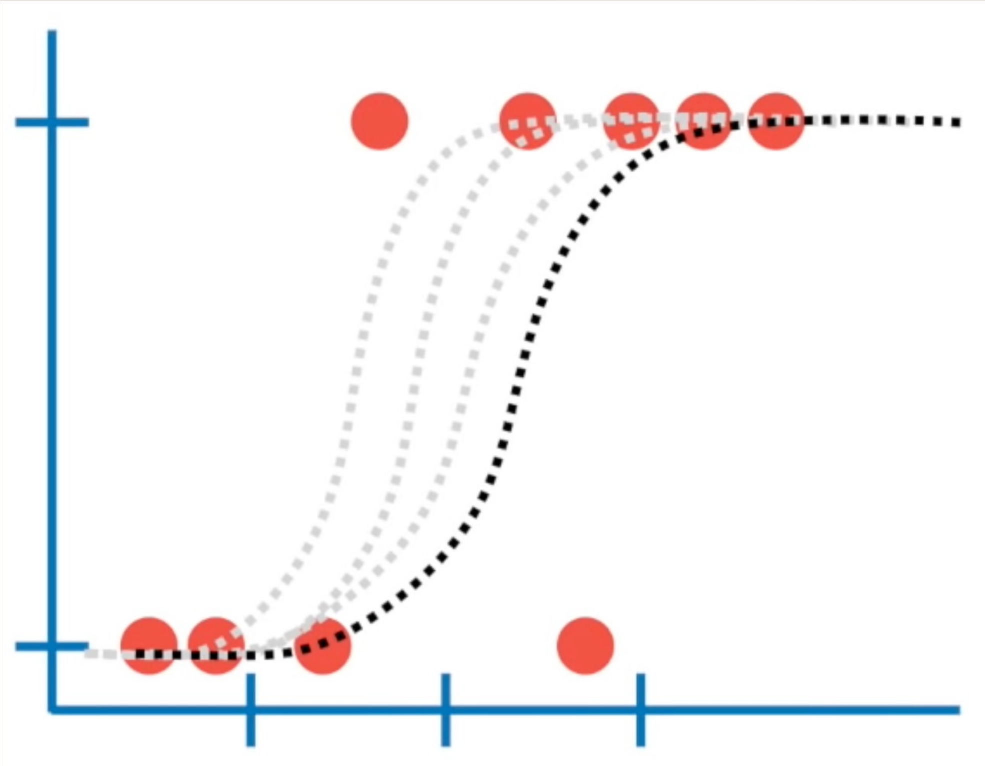


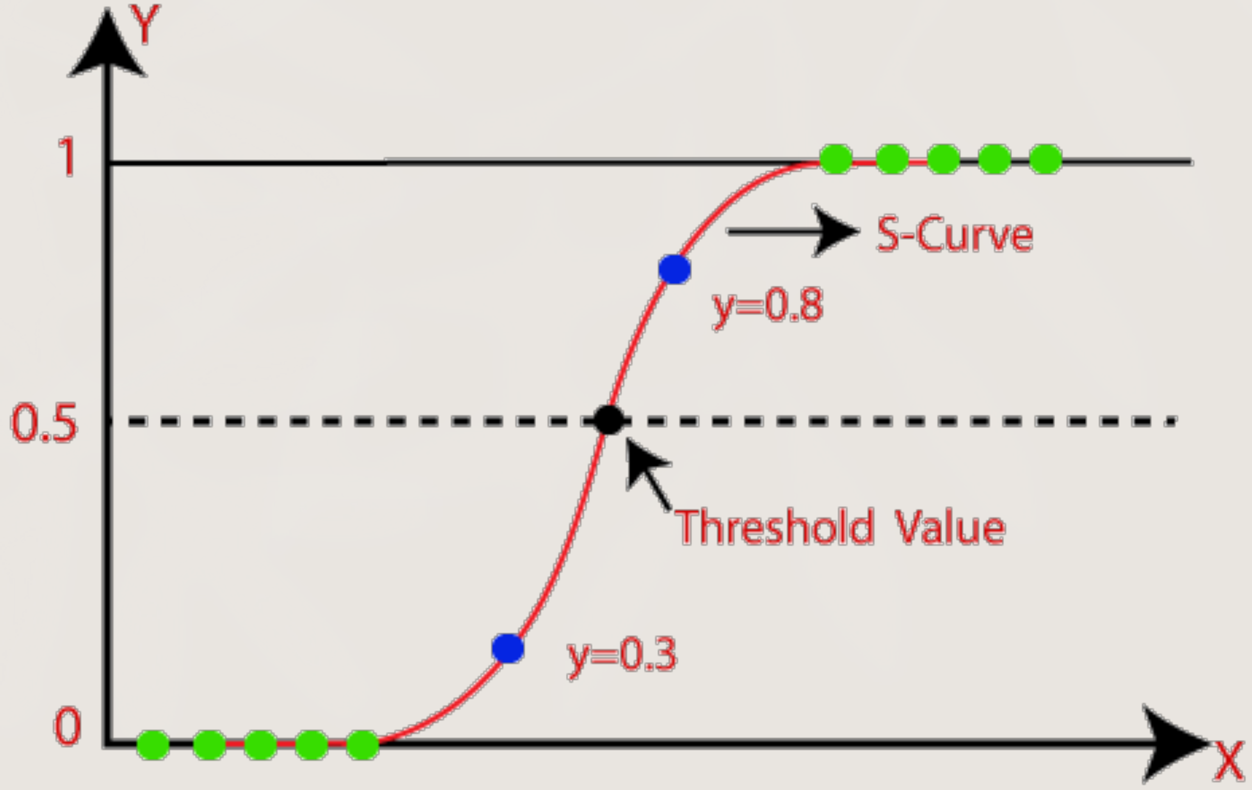
# Maximum Likelihood



"Least Sum of the Squares Errors"



**"Maximum Likelihood"**




# Maximum Likelihood Estimation (MLE)

| Linear Regression | Logistic Regression |
|---|---|
| Dependent variable’s are continuous in nature. | Dependent variable is binary in nature. |
| Linear Relationship | 1(True,Success) , 0(False,Failure) |
| y=m*x+c | Independent variable can be continuous or binary. |
| Output will be value of the variable. | Output will be probability of occurance of events |
| Accuracy & goodness of fit is measured by loss,R squared etc. | Accuracy & goodness of fit is measured by Precision, Recall, F1 score , Confusion matrix etc. |

# Logistics Regression Model

**Logistic regression model:** $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$

**Logistic regression equation:** $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$

$$\boxed{E(y) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}}$$

$$E(y) = \frac{e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}{1 + e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$$

Interpretation of E(y) as a **probability** in logistic regression

$$E(y) = P(y = 1 | x_1, x_2, \dots, x_k) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}}$$

**Estimated Logistic regression equation:**

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

$$\boxed{\hat{y} = P(y = 1 | x_1, x_2, \dots, x_k) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k)}}}$$

# Interpreting the Logistic Regression Equation

$$odd = \frac{P(y = 1|x_1, x_2, \dots, x_k)}{P(y = 0|x_1, x_2, \dots, x_k)} = \frac{P(y = 1|x_1, x_2, \dots, x_k)}{1 - P(y = 1|x_1, x_2, \dots, x_k)}$$

$$\boxed{Odds\, ratio = \frac{odds_1}{odds_0}}$$