Logistic Regression

Start Date	@17 November 2025
Weeks	

- Used for classification → Either 0 or 1
- Can work with continuous data + discrete data
- **Wald's Test** → Statistical hypothesis test → check if set of variables in model is significant → tested with difference between estimated parameter vs. hypothesised value reactive to parameter variability
- Logistic Regression has no "residual" → can't calculate R^2

Odds + Log(Odds)

⚠️ **Odds $\neq$ Probability**

- Example:
  - Team has $\frac{5}{3}$ odds of winning → 1.7
  - Team has $\frac{5}{8}$ probability of winning → 0.625
- Odds from probability:

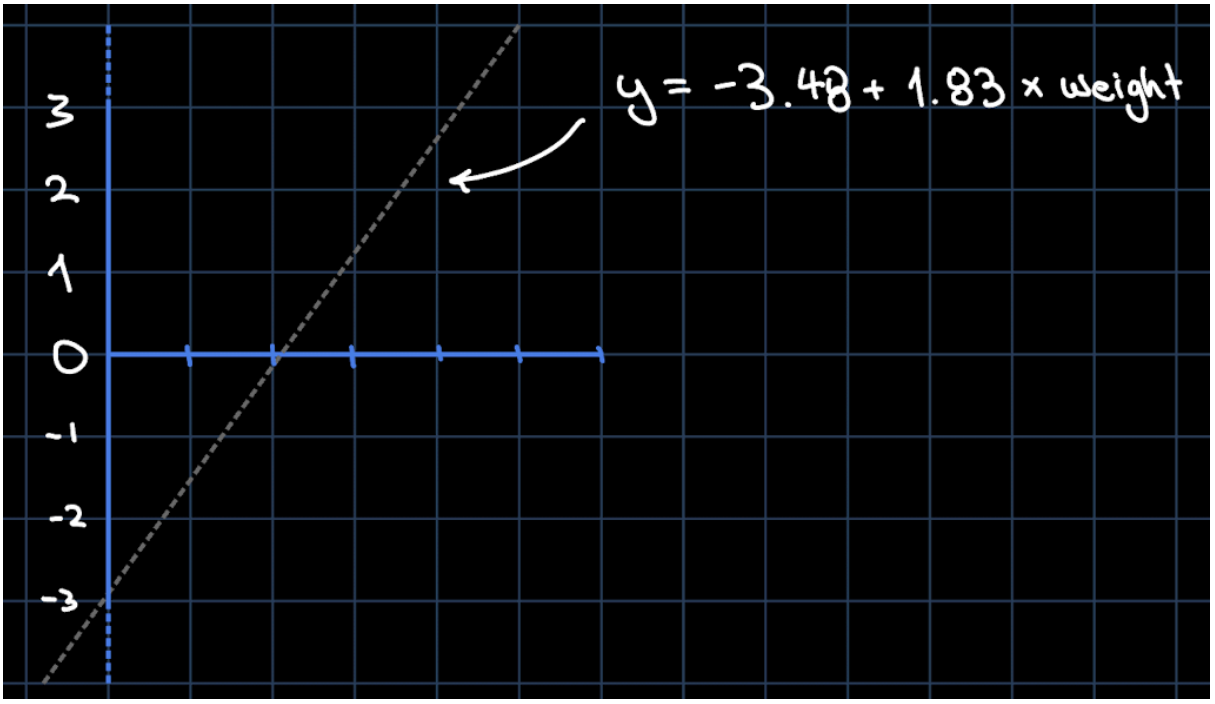
$$\frac{\textbf{Ratio of probability of winning}}{(1-\textbf{Ratio of probability of winning})} = \frac{5/8}{3/8} = \frac{5}{3} = 1.7$$

- Log(Odds) Makes everything symmetrical
  - Odds of winning → 6/1 ⇒ log(6) = 1.79
  - Odds of winning → 1/6 ⇒ log(1/6) = -1.79

$$\log(Odds) = \log\left(\frac{5}{3}\right) = \log\left(\frac{P}{(1-P)}\right) = \log(1.7)$$

$\log\left(\frac{P}{(1-P)}\right)$ = Logit Function → **Basis of Logistic Regression**

- Log(Odds) → Normal Distribution Shape
  - Good for statistical problems → Good for yes/no situations
- Odds are ratio $\neq$ odds ratio
- Odds ratio + log(odds ratio)
  - Example
    - $\frac{2/4}{3/1} = \log \frac{(2/4)}{(3/1)} = -1.79$
    - $\frac{3/1}{2/4} = \log \frac{(3/1)}{(2/4)} = 1.79$
- Odds ratio + log(odds ratio) are like R-Squared → Indicate relationship between two things
- 3 Ways to determine if odds ratio + log(odds ratio) is statistically significant:
  - Fisher's Exact Test
  - Chi-Square Test
  - The Wald Test
- Logistic Regression → Generalised Linear Model (GLM) → Generalisation of concepts + abilities of linear models.
- Y-Axis of Logistic Regression → Becomes Log(odds ratio) → Logit Function
- Axis from → New Y-Axis (look at the figures)
  - p = 0.5 to 1 ⇒ 0 → $+\infty$
  - p = 0.5 to 0 ⇒ 0 → $-\infty$



Coefficients

- Coefficient presented on log(odds) graph
- **Continuous variable:**
  - Closely reacted to “Linear Regression”
  - Logistic Regression assumes a linear relationship between variable and the log-ggs of outcome
    - Changes ins the continuous variable translate into proportional changes in log-odds
- **Discrete variable:**
  - With multiple categories → represented by dummy or indicator variables for each except one to avoid redundancy
  - Discrete inputs → model learns separate weights for each category level → influence probability of positive outcome

Types of Logistic Regression

1. **Binomial Logistic Regression:**
   - Variable only has two options. Yes/No, Pass/Fail, etc...

- Most common Logistic Regression → Used in Binary Classification problems

2. **Multinomial Logistic Regression:**
   - Dependent variables = Three or more
   - Extends Binary Logistic Regression to handle multiple classes
3. **Ordinal Logistic Regression:**
   - Dependent variables = three or more + natural order or ranking
   - E.g. “Low”, “Medium”, “High” → Takes order of categories into account when modelling

Logistic Regression Assumptions

1. **Independent Observation** ⇒ No correlation or dependence of data points
2. **Binary Dependent Variables** ⇒ Binary variable → takes only two values → **Softmax Function** for more than two categories
3. **Linearity Relationship Between Independent Variables + Log Odds** ⇒ Predictor affects the log odds in a linear way
4. **No Outlier** ⇒ Dataset should not contain extreme outliers → Distort estimation of Logistic Regression coefficient
5. **Large Sample Size** ⇒ Requires large sample size to produce reliable and stables results

Sigmoid Function

1. Important in Logistic Regression → Convert output of model into a probability between 0-1
2. Takes any real number → maps to range(0,1) → forms “s” shape → sigmoid curve
3. Logistic Regression → 0.5 → Threshold value to decide class label
   - This approach → transform continuous input values → meaningful class predictions

Does Logistic Regression Work?

- Transforms Linear Regression function → to categorical value output

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{bmatrix} \qquad Y = \begin{cases} 0 & \text{class 1} \\ 1 & \text{class 2} \end{cases}$$

- Apply multi-linear function

$$z = \left(\sum_{i=1}^{n} w_i x_i \right) + b$$

- x_i = ith observation of X
- w_i = weights or coefficient
- b_i = Bias term (intercept)

$$z = w \cdot x + b$$

$$\sigma(x) = \frac{1}{1 + e^{-z}}$$

- $\sigma(z)$ - Tends towards 1 as $z \to \infty$
- $\sigma(z)$ - Tends towards 0 as $z \to -\infty$
- $\sigma(z)$ - Always bound between 0 and 1

Logistic Regression Equation + Odds

- Odds of dependent event occurring (ratio of probability of the event) to probability of it not occurring:

$$\frac{p(x)}{1 - p(x)} = e^{x}$$

Natural logarithm of odds → Log-odds (Logit):

$$\log\left[\frac{p(x)}{1-p(x)}\right] = z$$

$$\log\left[\frac{p(x)}{1-p(x)}\right] = w \cdot X + b$$

$$\frac{p(x)}{1-p(x)} = e^{w \cdot X + b} \qquad \cdots \text{Exponentiate both sides}$$

$$p(x) = e^{w \cdot X + b} \cdot (1 - p(x))$$

$$p(x) = e^{w \cdot X + b} - e^{w \cdot X + b} \cdot p(x)$$

$$p(x) + e^{w \cdot X + b} \cdot p(x) = e^{w \cdot X + b}$$

$$p(x)\left(1 + e^{w \cdot X + b}\right) = e^{w \cdot X + b}$$

$$p(x) = \frac{e^{w \cdot X + b}}{1 + e^{w \cdot X + b}}$$

Final Logistic Regression equation:

$$P(X; b, w) = \frac{e^{w \cdot X + b}}{1 + e^{w \cdot X + b}} = \frac{1}{1 + e^{-w \cdot X + b}}$$

Likehood Function

- Goal → Find weights w + bias b → maximises the likehood of observing the data
- For each data points i:
  - $y = 1$ - Probability $= P(X; b, w) = p(x)$
  - $y = 0$ - Probability $= 1 - P(X; b, w) = 1 - p(x)$

$$L(w,b) = \prod_{i=1}^{n} P(x_i)^{y^i}(1-p(x_i))^{1-y_i}$$

Natural Log on both sides

$$\begin{aligned}
\log(L(b,w)) &= \sum_{i=1}^{n} \left[y_i \log p(x_i) + (1-y_i) \log(1-p(x_i)) \right] \\
&= \sum_{i=1}^{n} y_i \log p(x_i) + \log(1-p(x_i)) - y_i \log(1-p(x_i)) \\
&= \sum_{i=1}^{n} \log(1-p(x_i)) + \sum_{i=1}^{n} y_i \log \left(\frac{p(x_i)}{1-p(x_i)} \right) \\
&= \sum_{i=1}^{n} -\log \left(1 - e^{-(w \cdot x_i + b)} \right) + \sum_{i=1}^{n} y_i (w \cdot x_i + b) \\
&= \sum_{i=1}^{n} -\log \left(1 + e^{w \cdot x_i + b} \right) + \sum_{i=1}^{n} y_i (w \cdot x_i + b)
\end{aligned}$$