The Problem with Categorical Y

Logistic Regression

Based on Chew C. H. (2020) textbook: AI, Analytics and Data Science. Vol 1., Chap 7.

The most important idea in Linear Regression

- Linear Regression applies only for continuous Y.

  Example: $Y = 5 + 2X_1 - 3X_2$

- Most important idea is the possibility to use other information/predictors (i.e. Xs) to estimate Y:
  - Example: Estimate Price of a Flat (the Y variable).
  - X_1: Location
  - X_2: Level of the Flat
  - X_3: Within 5 mins walking distance to MRT?
  - X_4: Size of the Flat
  - X_5: Years remaining (max 99 years lease).
  - Etc…
- Use other predictors to predict certain disease, stock price, fraud, etc…
- Xs are unrestricted.

What if Y is categorical?

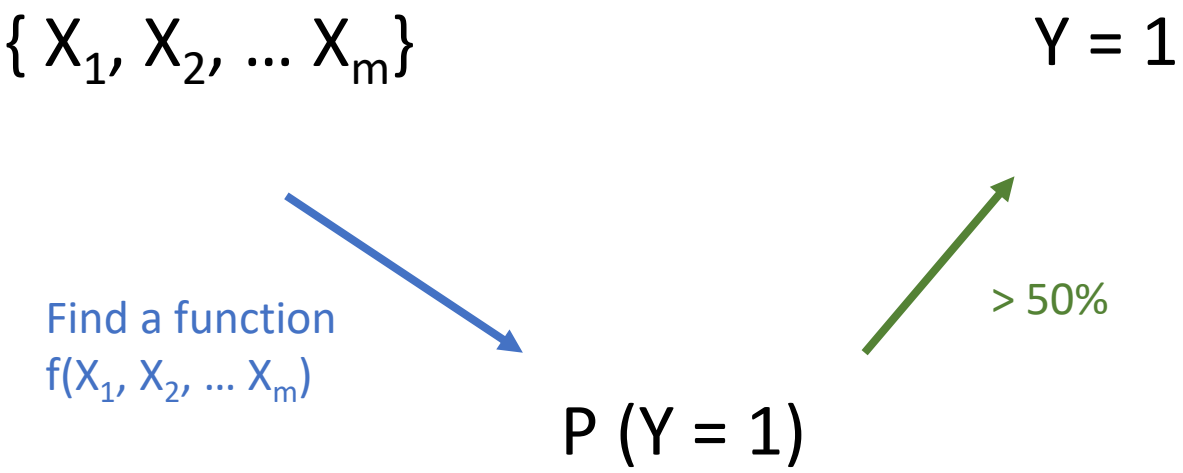
- Simplest Scenario: Y has only 2 Categorical levels (i.e. Binary Y)
  - 0 or 1
  - A or B
  - Yes or No
  - Pass or Fail
- Note: Logistic Regression can handle multi-categorical Y.
- Xs are unrestricted
  - Can be continuous or categorical.

The problem with Categorical Y

- How can we get a model to predict categorical Y, given a list of unrestricted Xs?
- Can $Y = 5 + 2X_1 - 3X_2$?
- In Linear Regression, Y is continuous implies Y can take any value within a certain range.
- Y is categorical means Y can only be A, B, C,….categorical levels.
- It is extremely difficult to find a linear equation involving unrestricted Xs that results in a categorical value for Y.
    - $b_0 + b_1X_1 + b_2X_2$ = category A ?

Reduce to a Simpler Problem: Find P(Y = 1)

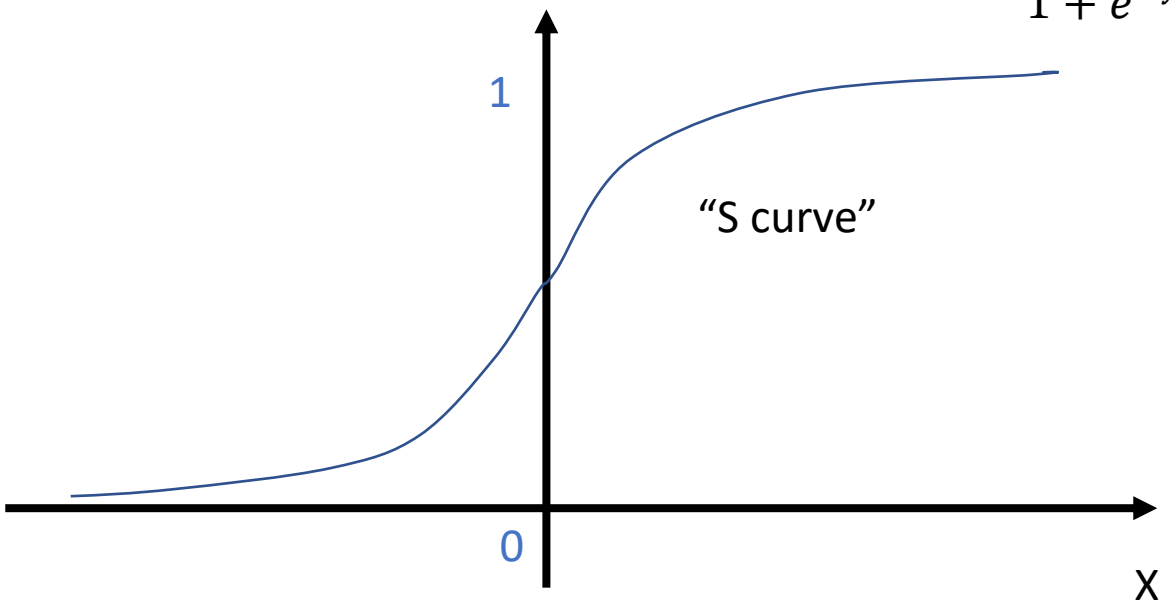
- Y = 0 or 1, and thus, has only two possible categorical values.
- P(Y = 1) has infinitely possible values within 0 to 1. i.e. continuous.
- P(Y = 0) can then be derived from P(Y = 1).



Logistic Function is suitable

- Unrestricted X
- Output is always between 0 to 1.
- Logistic Function f(x) can serve as Probability function P(Y = 1)

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

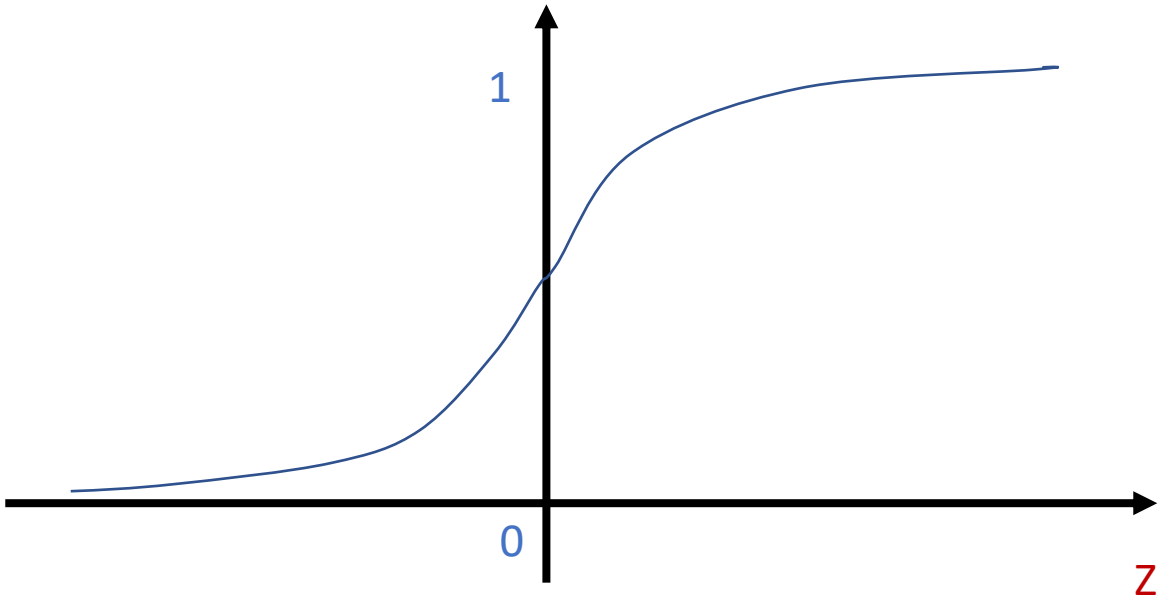


Logistic Function can admit many Xs.

- Unrestricted set of Xs by using Z to combine all Xs into one value.

$$Z = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m$$

$$f(Z) = P(Y = 1) = \frac{1}{1 + e^{-Z}}$$



Logistic Regression Model for Binary Y

Categorical Y = 0 or 1

$$Z = b_0 + b_1X_1 + b_2X_2 + \ldots + b_mX_m$$

$$P(Y = 1) = \frac{1}{1 + e^{-z}}$$

Next: Numerical Example of Logistic Regression

- Strengthen your understanding by
- Applying the Logistic function
- On Simple Data from Wikipedia: passexam.csv, and
- Interpret the results