Logistic Regression

Logistic Regression

Logistic regression is used in cases where the response is binary (e.g., yes or no, 1 or 0). Consider an example where we are trying to predict whether someone will pass a test based on the number of hours studied.

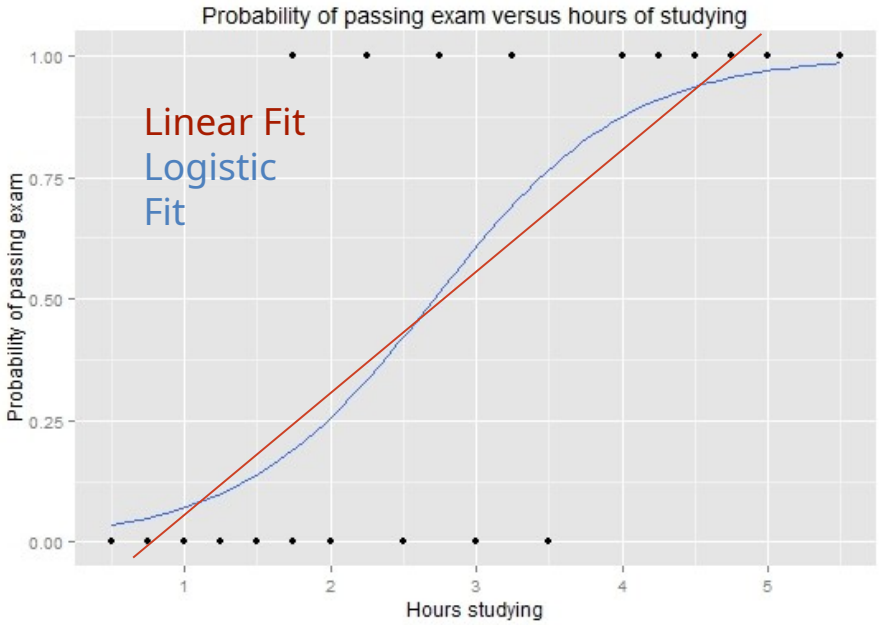
Logistic regression is used to model the probability the response falls into those categories.

$$Pr(pass = Yes|hours)$$

The prediction is made using a threshold of probability. For example, we may predict that the student passes the test if:

$$Pr(pass = Yes|hours) > 0.5$$

Logistic Model



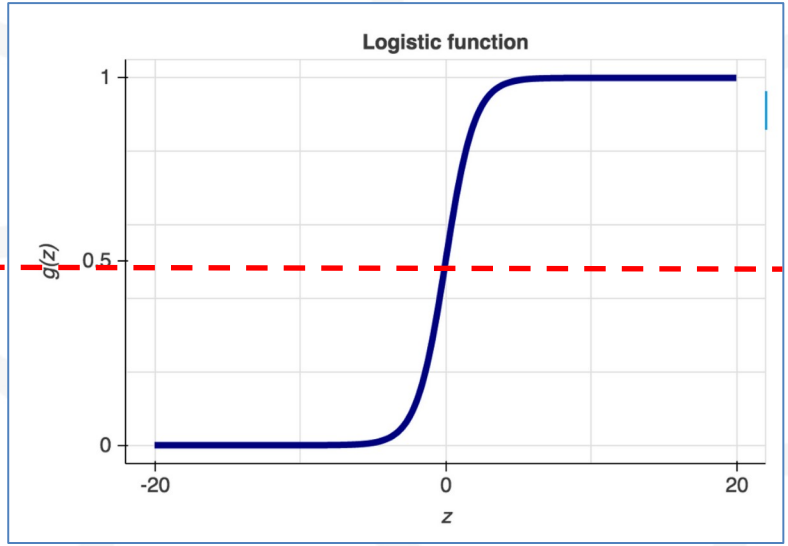
If we used linear regression to model this, we would get predictions below 0 and above 1, which is not reasonable since we are modeling a probability.

Therefore, we have to use a function that only gives predictions between 0 and 1.

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Logistic Regression Predictions

1. Set a decision threshold
2. Fit logistic model. Model fits regression parameters to optimize separation.
3. Compare the model output to the decision threshold.
    a. Above threshold is class ‘1’
    b. Below threshold is class ‘0’



Threshold can be varied if needed for specific requirements

Logistic Regression Assumptions

- Logistic regression does not have assumptions on normality, linearity or constant variance of the predictors.
- It requires that the response be binary.
- Assumes independent errors and uncorrelated predictors.
- Needs larger sample size than linear regression.

A Note on Binary Data

Binary data contains less information than continuous data. If you have influence on the data collection process, encourage the collection of continuous data. You can convert continuous data to binary data but you cannot convert binary data to continuous.

Ex. Pass vs Fail - Binary Data
Ex. Test Score - Continuous Data