Logistic Regression

TAG-J

AIMS

When and why do we use Logistic Regression?

- Binary
- Multinomial

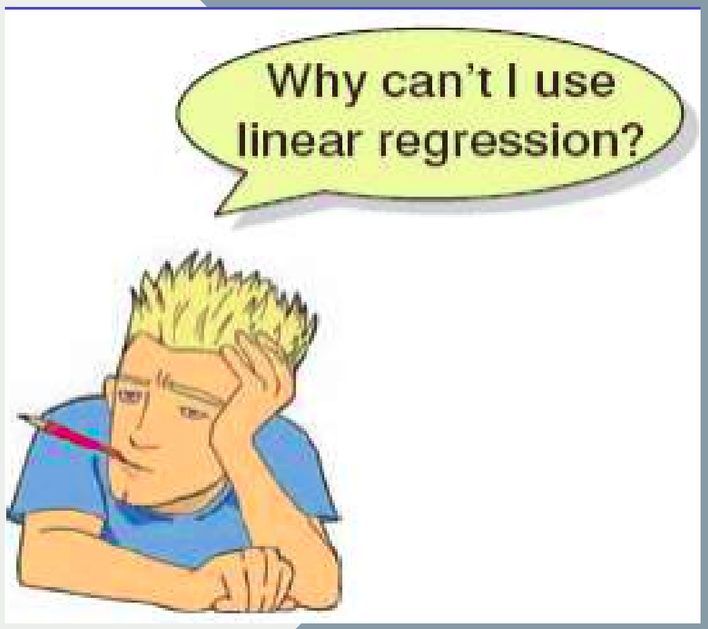
Theory Behind Logistic Regression

- Assesing the Model
- Assesing predictors
- Things that can go Wrong

Interpreting Logistic Regression

When And Why?

- To predict an outcome variable that is categorical from one or more categorical or continuous predictor variables.
- Used because having a categorical outcome variable violates the assumption of linearity in normal regression.
- Does not assume a linear relationship between DV and IV



When And Why

- No assumptions about the distributions of the predictor variables.
- Predictors do not have to be normally distributed
- Logistic regression does not make any assumptions of normality, linearity,

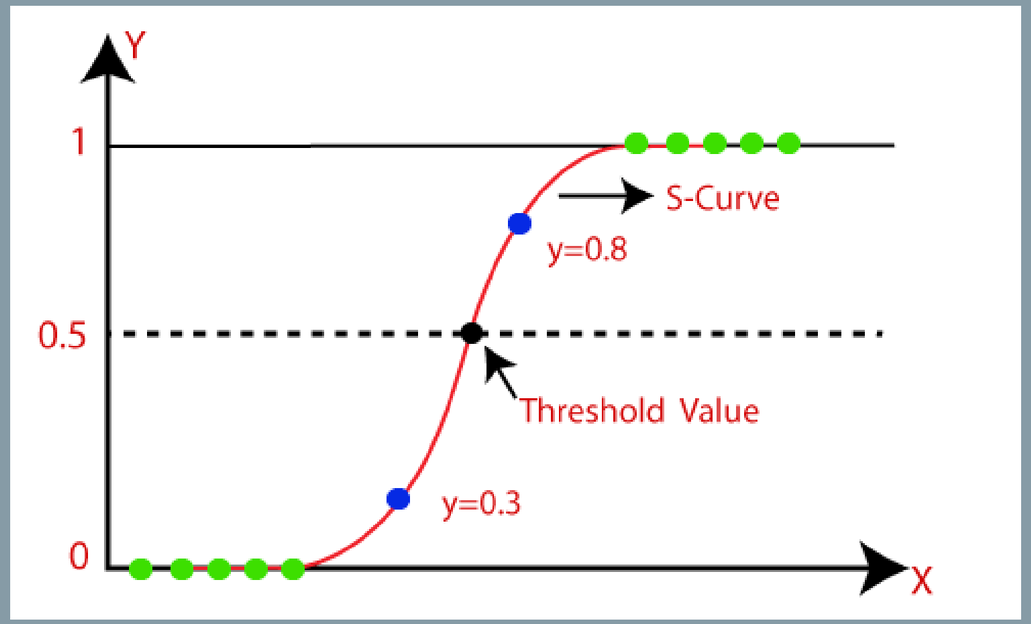
  and homogeneity of variance for the independent variables.
- Because it does not impose these requirements, it is preferred to discriminant analysis when the data does not satisfy these assumptions.

Logistic Regression

- Logistic regression is a supervised learning classification algorithm.
- It is used for predicting the categorical dependent variable using a given set of independent variables.
- A logistic regression model predicts P(Y=1) as a function of X.
- It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

- The below image is showing the logistic function:



Types of Logistic Regression

- Binomial: In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
- Multinomial: In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep“
- Ordinal: In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

With One Predictor

$$P(Y) = \frac{1}{1+e^{-(b_0+b_1X_1+\varepsilon_i)}}$$

- Outcome
  - We predict the probability of the outcome occurring
- b0 and b0
  - Can be thought of in much the same way as multiple regression
  - Note the normal regression equation forms part of the logistic regression equation

With Several Predictor

$$P(Y) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_nX_n+\varepsilon_i)}}$$

- Outcome
  - We still predict the probability of the outcome occurring
- Differences
  - Note the multiple regression equation forms part of the logistic regression equation
  - This part of the equation expands to accommodate additional predictors

Measuring the Probability of Outcome

- The probability of the outcome is measured by the odds of occurrence of an event.
- If P is the probability of an event, then (1-P) is the probability of it not occurring.
- Odds of success = P / 1-P

Measuring the Probability of Outcome

$$P(Y) = \frac{1}{1+e^{-(b_0+b_1X_1+b_2X_2+\ldots+b_nX_n+\varepsilon_i)}}$$

- Outcome
  - We still predict the probability of the outcome occurring
- Differences
  - Note the multiple regression equation forms part of the logistic regression equation
  - This part of the equation expands to accommodate additional predictors

Methods of Regression

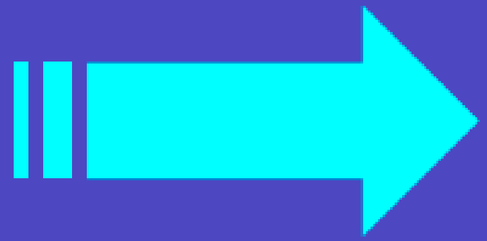
- Forced Entry: All variables entered simultaneously.
- Hierarchical: Variables entered in blocks.
  - Blocks should be based on past research, or theory being tested. Good Method.
- Stepwise: Variables entered on the basis of statistical criteria (i.e.relative contribution to predicting outcome).
  - Should be used only for exploratory analysis

DECISION PROCESS Stage 1:

Objectives Of logistic regression

- Identify the independent variable that impact in the dependent variable
- Establishing classification system based on the logistic model for determining the group membership

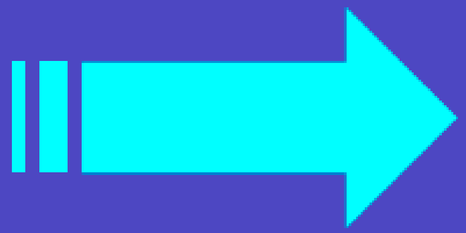
Linear Regression

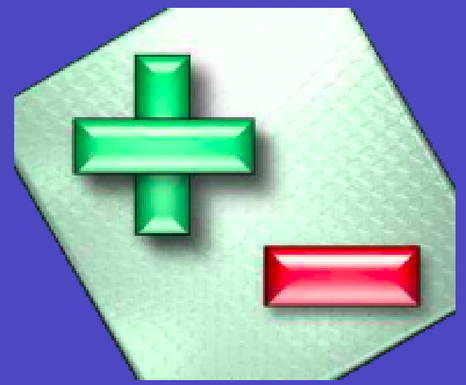


Independent Variable

Dependent Variable

Logistic Regression





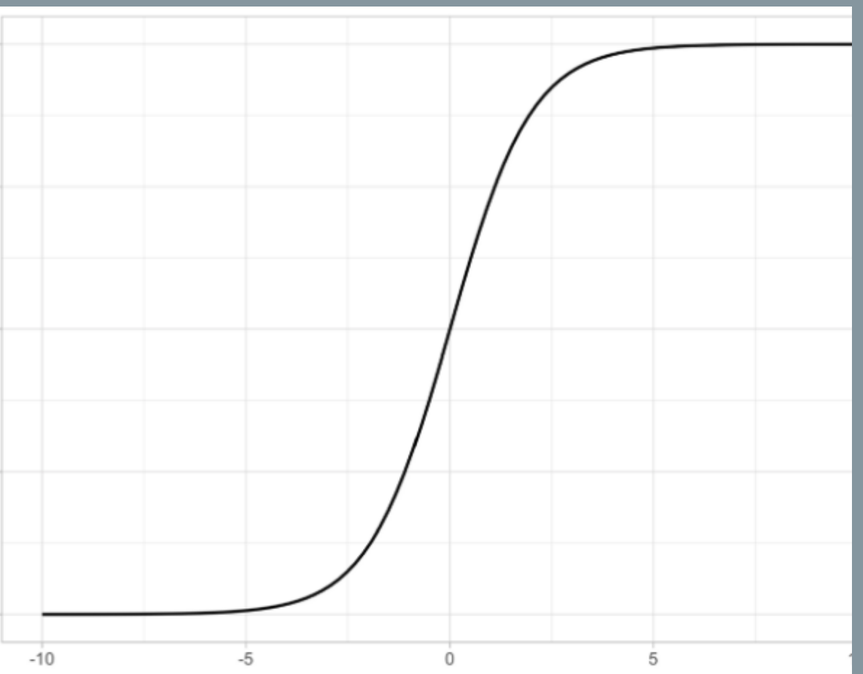
Independent Variable

Dependent Variable

The standard logistic function (or sigmoid function)

- Let z be any continuous value whose domain is $(-\infty, \infty)$. If you plug z into the sigmoid function like

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



A nice property of the output is that it is always within 0 and 1.

Here are some properties of θ(z):

1. When z = 0, θ = .5

$$\theta(z) = \frac{1}{1+e^0} = \frac{1}{1+1} = .5$$

2. When z is very large, θ is approximately 1

$$\theta(z) \approx \frac{1}{1+0} = 1$$

3.When z is very small/negative, θ is approximately 0

$$\theta(z) \approx \frac{1}{1+\infty} = 0$$

We can use the sigmoid function to convert a continuous, unbounded output z to a decimal number $\theta \in (0,1)$, which is advantageous for representing probabilities.

Converting linear regression outputs into logistic regression outputs with the sigmoid function

To go from a linear regression to a logistic regression, we can substitute the OLS output:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

For z like so:

$$\theta(z = \hat{y}) = \frac{1}{1 + e^{-\hat{y}}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)}}$$

- This function reinterprets the OLS output as a probability.
- The formula above represents the output of a logistic regression model.

Mathematical properties of logistic regression

By isolating y term ,we have:

$$\hat{y} = \log(\frac{\theta}{1-\theta})$$

By substituing y:

$$\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p = \log(\frac{\theta}{1-\theta})$$

Terms to know:

$$\text{Odds} = \frac{\theta}{1-\theta}$$

- The odds ratio specifies is defined as the probability of success as compared to the probability of failure.
- It is another way to represent probability, and is key to the interpretation of logistic regression coefficients.

Two key observations on these terms

1. In logistic regression, the logit must be linearly related to the independent variables.
2. This follows from equation A, where the left-hand side is a linear combination of x.
3. This is analogous to the OLS assumption that y be linearly related to x.

If you increase an independent variable x_i by 1, your odds grow by a factor of exp(β_i). This follows from equation B.

Point (2) follows from the algebra below:

$$e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p} = e^{\beta_1 x_1} \cdot e^{\beta_0+\beta_2 x_2+\ldots+\beta_p x_p}$$

Say we want to increase x_i by 1:

$$\begin{aligned} e^{\beta_1(x_1+1)} \cdot e^{\beta_0+\beta_2 x_2+\ldots+\beta_p x_p} &= e^{\beta_1 x_1} e^{\beta_1} \cdot e^{\beta_0+\beta_2 x_2+\ldots+\beta_p x_p} \\ &= e^{\beta_1} \cdot e^{\beta_0+e^{\beta_1 x_1}+\ldots+\beta_p x_p} \\ &= e^{\beta_1} \cdot \frac{\theta}{1-\theta} \end{aligned}$$

Then the odds of success increase by a factor of exp(β_i).

Assumptions of logistic regression

Dependent variable is binary:

- If this is not true, then logistic regression outputs do not apply.

Linearity between logit and independent variables:

- This follows from equation A — if this condition is not met, logistic regression is invalid.

No multicollinearity:

- Multicollinearity distorts tests of statistical significance on regression coefficients.

Large sample size:

- This is more of a rule of thumb.