DATA ANALYTICS

Unit 2:Logistic Regression

Mamatha.H.R
Department of Computer Science and Engineering

DATA ANALYTICS

Unit 2:Logistic Regression

Mamatha H R
Department of Computer Science and Engineering

CLASSIFICATION PROBLEMS

- Classification problems are an important category of problems in analytics in which the response variable (*Y*) takes a discrete value.

- The primary objective is to predict the class of a customer (or class probability) based on the values of explanatory variables or predictors.

Classification Problems

- Classification is an important category of problems in which the decision maker would like to classify the case/entity/customers into two or more groups.

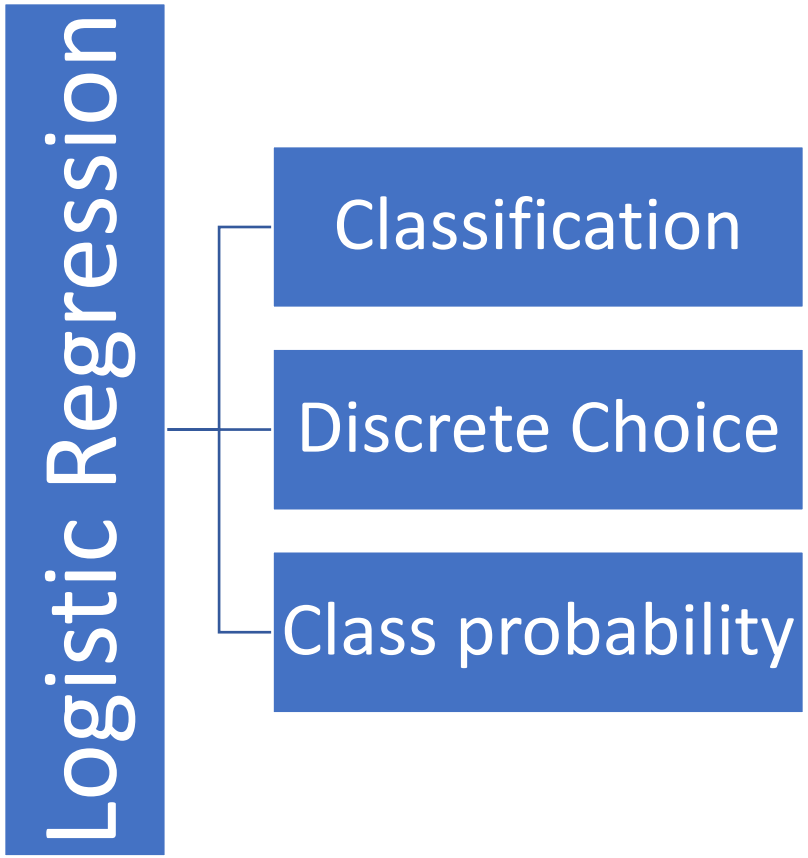
- Examples of Classification Problems:

  - ✓Customer profiling (customer segmentation)
  - ✓Customer Churn.
  - ✓Credit Classification (low, high and medium risk)
  - ✓Employee attrition.
  - ✓Fraud (classification of transaction to fraud/no-fraud)
  - ✓Stress levels
  - ✓Text Classification (Sentiment Analysis)
  - ✓Outcome of any binomial and multinomial experiment.

Challenging Classification Problems

- Ransomware
- Anomaly Detection
- Image Classification (Medical Devices, Satellite images)
- Text Classification

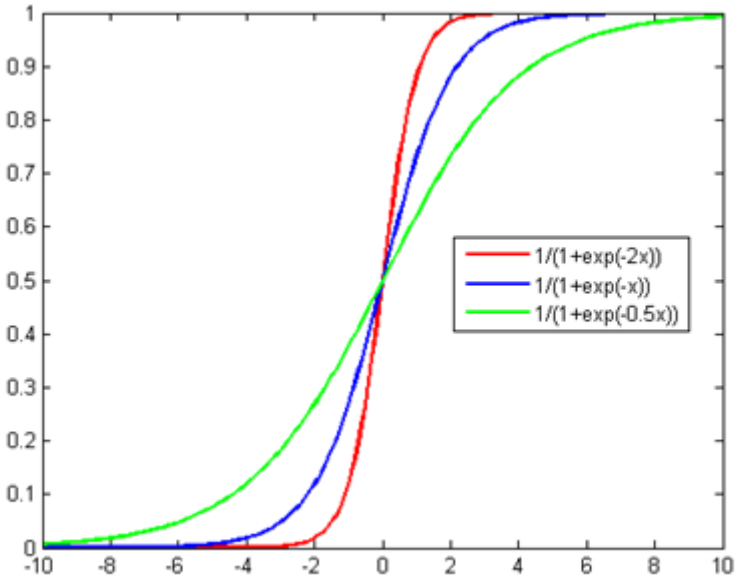
Logistic Regression - Supervised Learning Algorithm



Logistic Regression - Introduction

- The name logistic regression emerges from logistic distribution function.

$$\frac{e^Z}{1+e^Z}$$

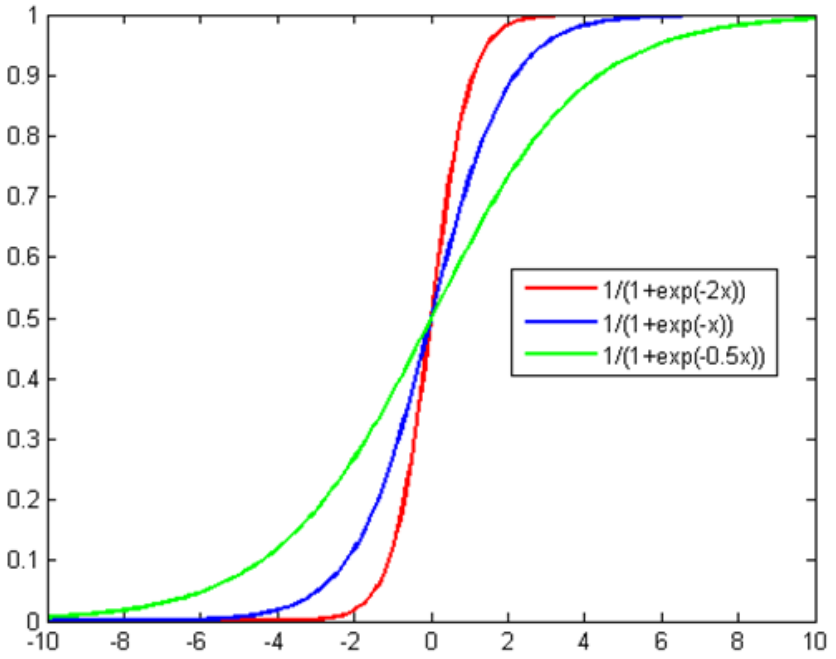


- Mathematically, logistic regression attempts to estimate conditional probability of an event (or class probability).

Logistic Distribution

$$f(t) = \frac{e^Z}{\sigma\left(1+e^Z\right)^2}$$

$$F(Z) = \frac{e^Z}{1+e^Z}$$



It is a symmetrical distribution (density function)

F(Z) is S-shaped curve

Logistic Regression

- Logistic regression models estimate how probability of an event may be affected by one or more explanatory variables.

- Logistic regression is a technique used for predicting "**class probability**", that is the probability that the case belongs to a particular class.

Types of Logistic Regression

- **Binomial (or binary) logistic regression** is a model in which the dependent variable is dichotomous.
- The categorical response has only two 2 possible outcomes. Example: Spam or Not.
- **In multinomial logistic regression model**, the dependent variable can take more than two values.
- The independent variables may be of any type.
- Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)
- **Ordinal Logistic Regression**
- Three or more categories with ordering. Example: Movie rating from 1 to 5

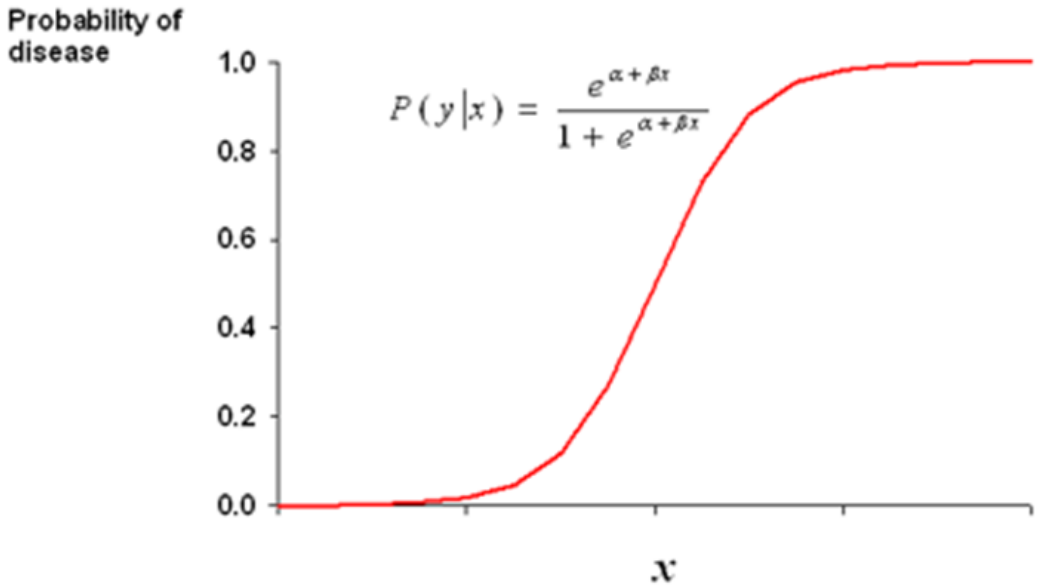
Logistic Function (Sigmoidal function)

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

$$z=\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n$$

Logistic Regression with one Explanatory Variable

b= $$P(Y=1|\ X=x)=\pi(x)=\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$



b = 0 implies that P(Y|x) is same for each value of x

b > 0 implies that P(Y|x) is increases as the value of x increases

b < 0 implies that P(Y|x) is decreases as the value of x increases

Generalized Linear Model (GLM)

The term general linear model (GLM) usually refers to conventional linear regression models for a continuous response variable given continuous and/or categorical predictors.

It includes multiple linear regression, as well as ANOVA and ANCOVA

The form is $$\mathrm{yi} \sim \mathrm{N}(X_i^T, \beta, \sigma^2)$$

where xi contains known covariates and β contains the coefficients to be estimated. These models are fit by least squares and weighted least squares

Generalized Linear Model (GLM)

The term generalized linear model (GLIM or GLM) refers to a larger class of models popularized by McCullagh and Nelder (1982, 2nd edition 1989).

In these models, the response variable yi is assumed to follow an exponential family distribution with mean μi, which is assumed to be some (often nonlinear) function of xTiβ.

Some would call these “nonlinear” because μi is often a nonlinear function of the covariates, but McCullagh and Nelder consider them to be linear, because the covariates affect the distribution of yi only through the linear combination xTiβ.

Generalized Linear Model (GLM)

Random Component – refers to the probability distribution of the response variable (Y); e.g. normal distribution for Y in the linear regression, or binomial distribution for Y in the binary logistic regression. Also called a noise model or error model.

Systematic Component - specifies the explanatory variables (X1, X2, ... Xk) in the model, more specifically their linear combination in creating the so called linear predictor

Link Function, η or $g(\mu)$ - specifies the link between random and systematic components. It says how the expected value of the response relates to the linear predictor of explanatory variables; e.g., $\eta = g(E(Yi)) = E(Yi)$ for linear regression, or $\eta = \text{logit}(\pi)$ for logistic regression.

DATA ANALYTICS

Generalized Linear Model (GLM)

The generalized linear models (GLMs) are a broad class of models that include linear regression, ANOVA, Poisson regression, log-linear models etc. The table below provides a good summary of GLMs

Model	Random	Link	Systematic
Linear Regression	Normal	Identity	Continuous
ANOVA	Normal	Identity	Categorical
ANCOVA	Normal	Identity	Mixed
Logistic Regression	Binomial	Logit	Mixed
Loglinear	Poisson	Log	Categorical
Poisson Regression	Poisson	Log	Mixed
Multinomial response	Multinomial	Generalized Logit	Mixed

DATA ANALYTICS

ODDS and ODDS RATIO

We will consider a data-set that tells us about depending on the gender, whether a customer will purchase a product or not

Table of frequency of ‘yes’ and ‘no’ depending on the gender

Gender	Purchase	
	Yes	No
Female	159	106
Male	121	125

Odds, which describes the ratio of success to ratio of failure

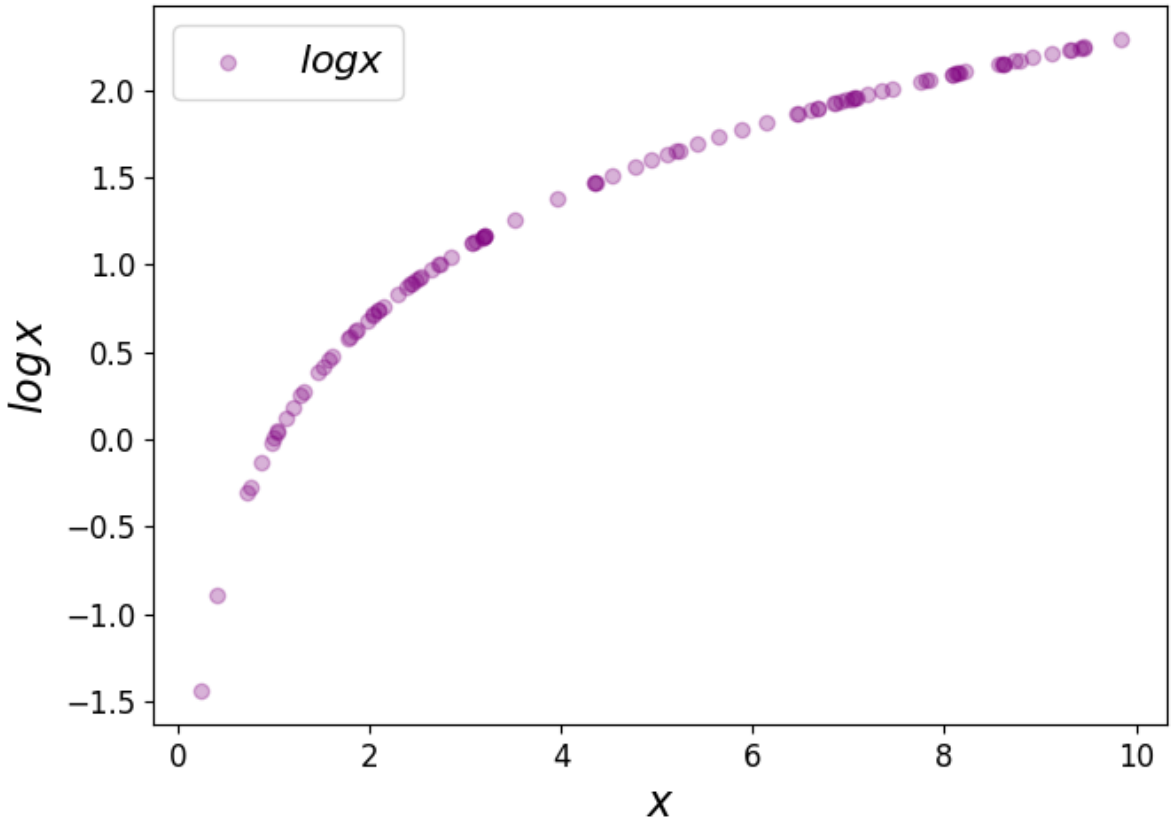
ODDS and ODDS RATIO

Considering females group,

- we see that probability that a female will purchase (success) the product is = 159/265 (yes/total number of females).
- Probability of failure (no purchase) for female is 106/265.
- In this case the odds is defined as (159/265)/(106/265) = 1.5.
- Higher the odds, better is the chance for success. Range of odds can be any number between [0 , ∞].

ODDS and ODDS RATIO

log(x) is defined for x≥0 but the range varies from [-∞, ∞]



log x vs x; for all +'ve' values of x, log x can vary between -∞ to + ∞.

ODDS and ODDS RATIO

Odds ratio, is the ratio of odds.

Considering the example, Odds ratio, represents which group (male/female) has better odds of success, and it's given by calculating the ratio of odds for each group.

So odds ratio for females = odds of successful purchase by female / odds of successful purchase by male = (159/106)/(121/125).

Odds ratio for males will be the reciprocal of the above number.

Odds ratio can vary between 0 to positive infinity, log (odds ratio) will vary between $[-\infty, \infty]$

ODDS and ODDS RATIO

- ODDS: Ratio of two probability values.

$$odds = \frac{\pi}{1-\pi}$$

- ODDS RATIO: Ratio of two odds.

Assume that X is an independent variable (covariate). The odds ratio, OR, is defined as the ratio of the odds for X = 1 to the odds for X = 0. The odds ratio is given by:

$$OR = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)}$$

Logistic Transformation

- The logistic regression model is given by:

$$\pi_i = \frac{e^{(\beta_0 + \beta_1 X_i)}}{1 + e^{(\beta_0 + \beta_1 X_i)}}$$

$$\frac{\pi_i}{1 - \pi_i} = e^{(\beta_0 + \beta_1 X_i)}$$

$$\ln\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \beta_1 X_i$$

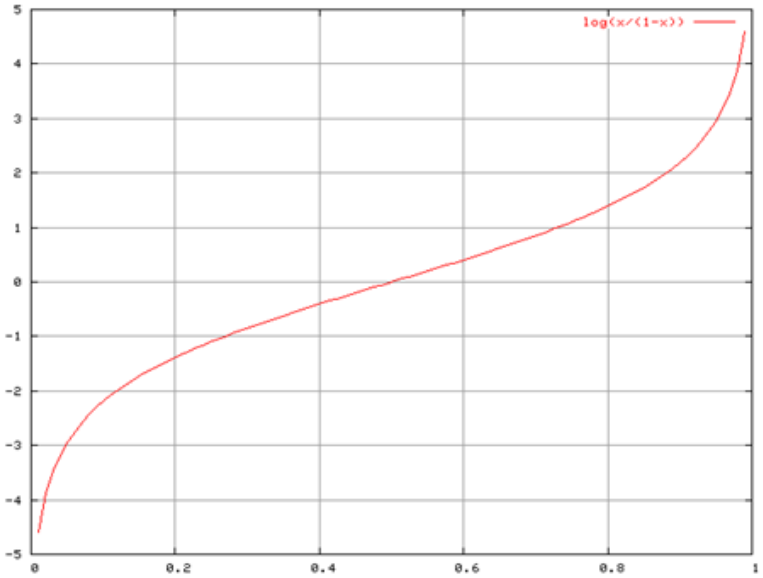
Function with linear properties (Link Function)

Logit Function

- The logit function is the logarithmic transformation of the logistic function. It is defined as the natural logarithm of odds.

- Logit of a variable π (with value between 0 and 1) is given by:

$$Logit(\pi) = \ln(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x$$



Odds Ratio for Binary Logistic Regression

$$OR = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)} = e^{\beta_1}$$

If OR = 2, then the event is twice likely to occur when X = 1 compared to X = 0.

Odds ratio approximates the relative risk.

Interpretation of LR coefficients

- β_1 is the change in log-odds ratio for unit change in the explanatory variable.

- β_1 is the change in odds ratio by a factor exp(β_1).

$$\beta_1 = \ln\left(\frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x+1))}\right) = \text{Change in ln odds ratio}$$

$$e^{\beta_1} = \frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x+1))} = \text{Change in odds ratio}$$

Parameter Estimation in Logistic Regression (Maximum Likelihood Estimate)

Likelihood function for Binary Logistic Function

- Probability density function for binary logistic regression is given by:

$$f(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$L(\beta) = f(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$\ln(L(\beta)) = \sum_{i=1}^{n} y_i \ln[\pi(x_i)] + \sum_{i=1}^{n} (1-y_i)[\ln(1-\pi_i(x_i))]$$

Likelihood function for Binary Logistic Function

$$\ln\left[L(\beta)\right] = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) -$$

$$\sum_{i=1}^{n} \ln(1 + \exp(\beta_0 + \beta_1 x_i))$$

Estimation of LR parameters

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = 0$$

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_1} = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \frac{x_i \exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = 0$$

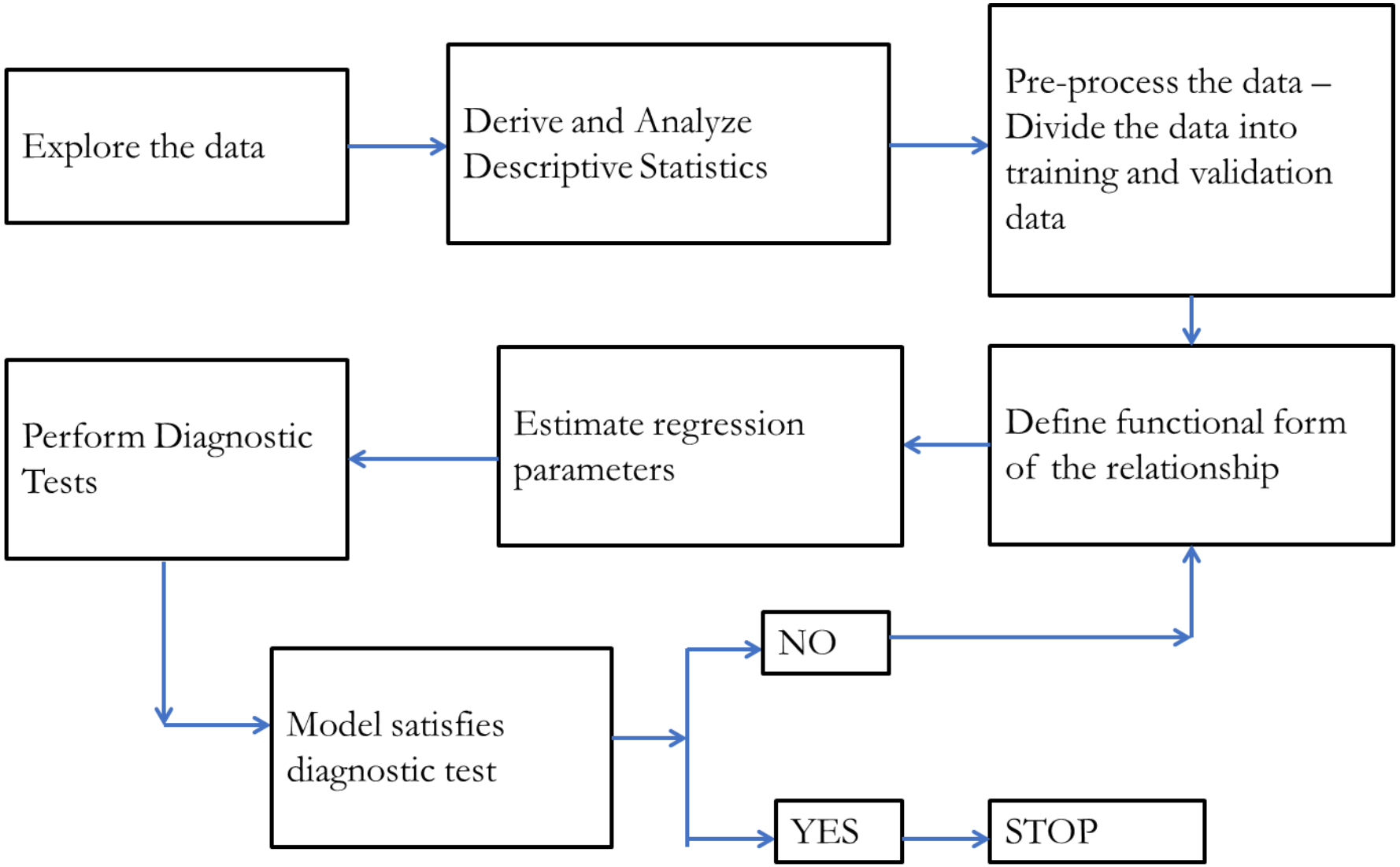
The above system of equations are solved iteratively to estimate β_0 and β_1

Limitations of MLE

- Maximum likelihood estimator may not be unique or may not exist.

- Closed form solution may not exist for many cases, one may have to use iterative procedure to estimate the parameter values.

Logistic Regression Model Development



LOGISTIC REGRESSION MODEL DIAGNOSTICS

Omnibus tests are generic statistical tests used for checking whether the variance explained by the model is more than the unexplained variance.

The log likelihood function for binary logistic regression model is given by

$$LL = \sum_{i=1}^{n} Y_i \ln[\pi(Z)] + \sum_{i=1}^{n} (1 - Y_i)[\ln(1 - \pi(Z))]$$

Wald's test

Wald's test is used for checking statistical significance of individual predictor variables (equivalent to *t*-test in MLR model). The null and alternative hypotheses for Wald's test are:

H_0: $\beta_i = 0$

H_1: $\beta_i \neq 0$

Wald's test statistic is given by

$$W = \left[\frac{\hat{\beta}_i}{S_e(\hat{\beta}_i)} \right]^2$$

Hosmer-Lemeshow Test

Hosmer–Lemeshow (H–L) is a chi-square goodness of fit test used for checking the goodness of logistic regression model (Hosmer and Lemeshow, 2000).

The null and alternative hypotheses in H–L test are

H_0: The logistic regression model fits the data

H_1: The logistic regression model does not fit the data

The H–L test statistic is given by

$$H = \sum_{k=1}^{G} \left[\frac{(O_k - E_k)^2}{N_k \pi_k (1 - \pi_k)} \right]$$

R^2 in Logistic Regression

- In linear regression R^2 is the proportion of variation explained by the regression model.

- It is not possible to develop a R^2 type measure for Logistic Regression since the variance of the error term is not constant.

- Many Pseudo R^2 values are used in Logistic Regression. Pseudo R2 is an indicator of strength of relationship.

R^2 in Logistic Regression

- **R-squared is a measure of improvement from null model to fitted model** - The denominator of the ratio can be thought of as the sum of squared errors from the null model--a model predicting the dependent variable without any independent variables.

- In the null model, each *y* value is predicted to be the mean of the *y* values.

Pseudo R^2

It is not possible to calculate R^2 as in the case of continuous dependent variable in a logistic regression model.

However, many pseudo R^2 values are used which compare the intercept-only model to the model with independent variables.

Cox and Snell R^2

Cox and Snell R^2 is given by

$$R^2 = 1 - \left\{ \frac{L(\text{Intercept only model})}{L(\text{Full Model})} \right\}^{2/N}$$

Based on Log-likelihood ratio.

$$R^2 = 1 - \left(\frac{LL(\text{Null Model})}{LL(\text{ Model})} \right)^{2/n}$$

Null Model : Model without predictors

Full Model: Model with predictors

n is the number of observations

Nagelkerke's R^2

Nagelkerke R^2 is an adjustment over Cox and Snell R^2, so that the maximum value Pseudo R^2 is 1.

The maximum value of Cox and Snell R2 may not be 1. Nagelkerke modified Cox and Snell R2 to the maximum value = 1.

$$R^2 = \frac{1-\left\{\frac{\mathrm{L(Intercept\text{-}only\ model)}}{\mathrm{L(Full\ model)}}\right\}^{2/N}}{1-\{\mathrm{L(Intercept\text{-}only\ model)}\}^{2/N}}$$

$$\text{Nagelkerke } R^2 = \left[\frac{1-\left(\frac{LL(\text{Null Model})}{LL(\text{Full Model})}\right)^{2/n}}{1-LL(\text{null model})^{2/n}}\right]$$

Exercise

- Bring out differences between Linear and Logistic Regression

- Find an application where Logistic regression is suitable and build a regression model for the same.

References

Text Book:

"Business Analytics, The Science of Data-Driven Decision Making", U. Dinesh Kumar, Wiley 2017

https://online.stat.psu.edu/stat504/node/216/

https://towardsdatascience.com/logit-of-logistic-regression-understanding-the-fundamentals-f384152a33d1

THANK YOU

Dr.Mamatha H R

Professor,Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834