BINARAY LOGISTIC REGRESSION INTRODUCTION

Multiple Linear Regression-Quick Recap

- Multiple linear regression is used to explain the relationship between one continuous dependent variable and two or more independent variables.
- The independent variables can be continuous or categorical.
- Multiple Linear Regression is used when we want to predict the value of a variable based on the values of two or more other variables.
- The variable we want to predict is called the dependent variable
- The variables used to predict the value of dependent variable are called independent variables (or explanatory variables/predictors).

- Example: The price house in USD can be dependent variable and area of house, location of house , air quality index in the area, distance from airport etc. can be independent variables.

$$\boxed{Model: Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + e}$$

DATA SCIENCE INSTITUTE

Binary Logistic Regression

DEPENDENT VARIABLE

↓

Categorical (Binary)

0 or 1

Death or Survival

Absence or Presence

⋮

INDEPENDENT VARIABLES

↓

Categorical or Continuous

Gender

Geographical Regions

Age

Income

Debt Ratio

⋮

Binary logistic regression **models the dependent variable as a logit of p, where p is the conditional probability that dependent variable takes value 1**

DATA SCIENCE INSTITUTE

Application Areas

Industry/Function	Objective	Based on Information such as:
Marketing Analytics	Identify the potential customers who will buy the product	• Age, Gender, Payment Mode, Purchase Frequency, Historical purchase details, etc.
Churn Management	Identify the employees who are likely to leave the company	• Gender, Qualification, Source of Hiring, Department, Compensation, etc.
Risk Management (Credit Scoring/ Fraud Detection)	Predict defaulters	• Age, occupation, annual income, other loan details etc.
Insurance	Predict policy Lapse	• Age, Gender, occupation, premium amount etc.

DATA SCIENCE INSTITUTE

Why Not Use Linear Regression Model?

The statistical model for multiple linear regression is,

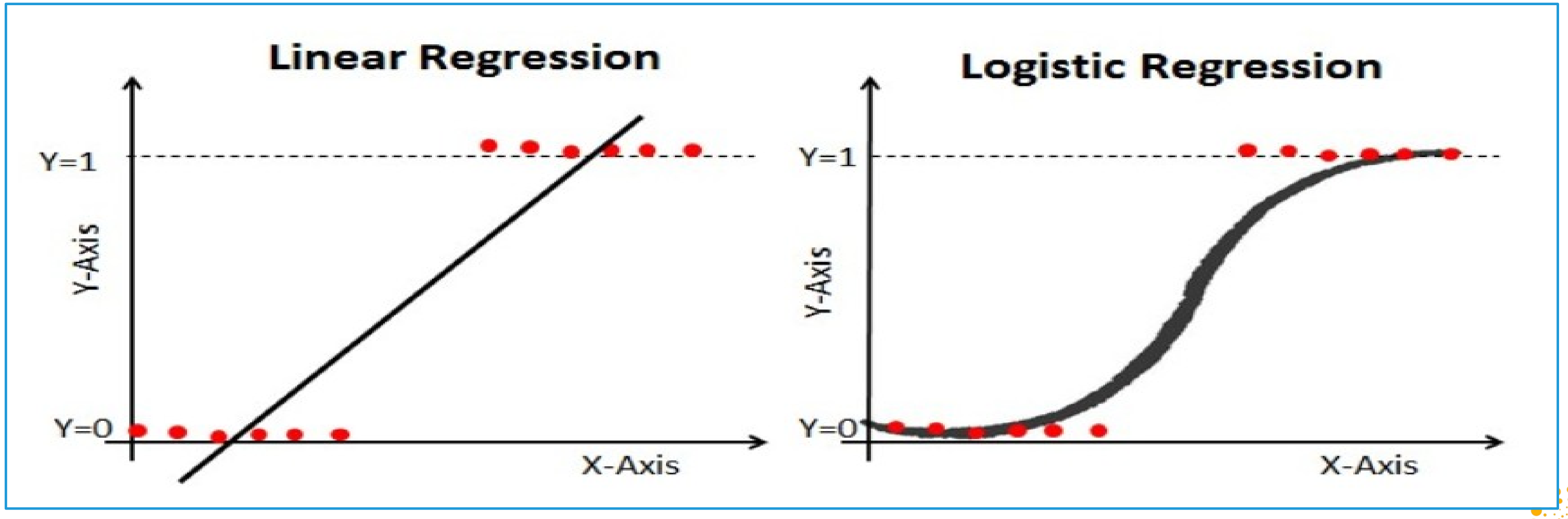
$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p + e$$

- If binary variable Y is used on left hand side of the model, then the two sides are not comparable. Right hand side is a continuous term.
- If probability 'P' is used instead of Y then linearity may not hold true. The relationship assumed in logistic regression is a 'S' shaped curve.

DATA SCIENCE INSTITUTE

Why Not Use Linear Regression Model?

- Linear regression is suitable for predicting outcome which is continuous value.
  For example, predicting the price of a property based on area in Sq. Feet.
- The regression line is a **straight line**.
- Whereas logistic regression is for classification problems, which predicts a probability range between 0 to 1 (or predicts categories Yes or no).
  For example, predict whether a customer will make a purchase or not.
- The regression curve is a **sigmoid curve**.



DATA SCIENCE INSTITUTE

Statistical Model – For k Predictors

$$\log\left(\frac{p}{1-p}\right)=b_0+b_1X_1+\ldots+b_kX_k$$

where,

p : Probability that Y=1 given X

Y : Dependent Variable

$X_1, X_2, \ldots, X_k$: Independent Variables

$b_0, b_1, \ldots, b_k$: Parameters of Model

Note that **LHS of the model can lie between - ∞ to ∞**

Parameters of the model are estimated by Maximum Likelihood Method

Case Study – Modeling Loan Defaults

Background

- A bank possesses demographic and transactional data of its loan customers. If the bank has a model to predict defaulters it can help in loan disbursal decision making.

Objective

- To predict whether the customer applying for the loan will be a defaulter or not.

Available Information

- Sample size is 700
- **Independent Variables**: Age group, Years at current address, Years at current employer, Debt to Income Ratio, Credit Card Debts, Other Debts. The information on predictors was collected at the time of loan application process.
- **Dependent Variable**: Defaulter (=1 if defaulter ,0 otherwise). The status is observed after loan is disbursed.

Data Snapshot

Bank Loan Data

Independent Variables

Dependent Variable

SN	AGE	EMPLOY	ADDRESS	DEBTINC	CREDDEBT	OTHDEBT	DEFAULTE
1	3	17	12	9.3	11.36	5.01	1
2	1	10	6	17.3	1.36	4	0
3	2	15	14	5.5	0.86	2.17	0
4	3	15	14	2.9	2.66	0.82	0

Column	Description	Type	Measurement	Possible Values
SN	Serial Number		-	-
AGE	Age Groups	Categorical	1(<28 years), 2(28-40 years), 3(>40 years)	3
EMPLOY	Number of years customer working at current employer	Continuous	-	Positive value
ADDRESS	Number of years customer staying at current address	Continuous	-	Positive value
DEBTINC	Debt to Income Ratio	Continuous	-	Positive value
CREDDEBT	Credit Card Debt	Continuous	-	Positive value

DATA SCIENCE INSTITUTE

Binary Logistic Regression Model for the bank loan data

- Model of default on the predictors will look like this:

Where p is the Probability of Dependent Variable = 1 (here, Default=1)

Intercept

Independent Variables

$$\log\,() = b_0 + b_1\,(AGE) + b_2\,(EMPLOY) + b_3\,(ADDRESS) + b_4\,(DEBTINC) + b_5\,(CREDDEBT) + b_6\,(OTHDEBT)$$

Parameter Estimates

DATA SCIENCE INSTITUTE

Likelihood Function

- The parameters of the logistic model are estimated using **maximum likelihood estimation (MLE)**.
- The Likelihood function is as below:

$$L = \prod_{i=1}^{n} p^{y_i}(1-p)^{1-y_i}$$

n is the number of observations

- The likelihood function is a **joint probability** of Yi's.
- It is expressed as a function of regression parameters after substituting known X and Y value.
- Parameters are estimated by maximizing L.
- Two commonly used iterative maximum likelihood algorithms are **Fisher scoring method** and **Newton-Raphson method**. Both algorithms give the same parameter estimates; however, the estimated covariance matrix of the parameter estimators can differ slightly.

Maximum Likelihood Estimates of Parameters

	Coefficients
Intercept	-0.78821
AGE2	0.25202
AGE3	0.62707
EMPLOY	-0.26172
ADDRESS	-0.09964
DEBTINC	0.08506
CREDDEBT	0.56336
OTHDEBT	0.02315

log ((p/(1-p)) = **-0.78821**+ **0.25202** (AGE2)+ **0.62707** (AGE3) **-0.26172** (EMPLOY) - **0.09964** (ADDRESS) + **0.08506**(DEBTINC) + **0.56336**(CREDDEBT) + **0.02315** (OTHDEBT)

DATA SCIENCE INSTITUTE

Individual testing using Wald's test

Individual testing is used for checking significance of each independent variable separately.

Objective	To test the **null hypothesis** that **each variable is insignificant**

Null Hypothesis (H_0): $b_i = 0$

Alternate Hypothesis (H_1): : bi ≠ 0

i=1,2…,k

Test Statistic	Z =(Estimate of bi)/(Standard Error of estimated bi) Under H0, Z is assumed to follow standard normal distribution.
Decision Criteria	Reject the null hypothesis **if p-value < 0.05**

DATA SCIENCE INSTITUTE

Binary Logistic Regression in R

```
# Import data and check data structure before running model
```

```r
data<-read.csv("BANK LOAN.csv",header=TRUE)
str(data)
```

```
# Output:
```

```
$ SN       : int  1 2 3 4 5 6 7 8 9 10 ...
$ AGE      : int  3 1 2 3 1 3 2 3 1 2 ...
$ EMPLOY   : int  17 10 15 15 2 5 20 12 3 0 ...
$ ADDRESS  : int  12 6 14 14 0 5 9 11 4 13 ...
$ DEBTINC  : num  9.3 17.3 5.5 2.9 17.3 10.2 30.6 3.6 24.4 19.7 ...
$ CREDDEBT : num  11.36 1.36 0.86 2.66 1.79 ...
$ OTHDEBT  : num  5.01 4 2.17 0.82 3.06 ...
$ DEFAULTER: int  1 0 0 0 1 0 0 0 1 0 ...
```

```r
data$AGE<-factor(data$AGE)
str(data)
```

Age is an integer and need to convert into factor. Since, it is a categorical variable.

```
# Output:
```

```
'data.frame':	700 obs. of  8 variables:
 $ SN       : int  1 2 3 4 5 6 7 8 9 10 ...
 $ AGE      : Factor w/ 3 levels "1","2","3": 3 1 2 3 1 3 2 3 1 2 ...
 $ EMPLOY   : int  17 10 15 15 2 5 20 12 3 0 ...
 $ ADDRESS  : int  12 6 14 14 0 5 9 11 4 13 ...
 $ DEBTINC  : num  9.3 17.3 5.5 2.9 17.3 10.2 30.6 3.6 24.4 19.7 ...
 $ CREDDEBT : num  11.36 1.36 0.86 2.66 1.79 ...
 $ OTHDEBT  : num  5.01 4 2.17 0.82 3.06 ...
 $ DEFAULTER: int  1 0 0 0 1 0 0 0 1 0 ...
```

DATA SCIENCE INSTITUTE

Logistic Regression in R

```
# Using glm function to develop binary logistic regression model

riskmodel<-glm(DEFAULTER~AGE+EMPLOY+ADDRESS+DEBTINC+CREDDEBT+OTHDEBT,
               family=binomial,data=data)
```

- **glm** is Generalized Linear Model. Logistic regression is type of GLM.
- LHS of ~ is dependent variable and independent variables on RHS are separated by '+'.
- **riskmodel** is the model object
- By setting the **family =binomial**, **glm()** fits a logistic regression model

Individual Hypothesis Testing in R

```
# Individual Testing
summary(riskmodel)
```

- **summary()** function gives the output of glm.

```
# Output:
```

```
Call:
glm(formula = DEFAULTER ~ AGE + EMPLOY + ADDRESS + DEBTINC +
    CREDDEBT + OTHDEBT, family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.3495  -0.6601  -0.2974   0.2509   2.8583

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.78821    0.26407  -2.985  0.00284 **
AGE2         0.25202    0.26651   0.946  0.34433
AGE3         0.62707    0.36056   1.739  0.08201 .
EMPLOY      -0.26172    0.03188  -8.211  < 2e-16 ***
ADDRESS     -0.09964    0.02234  -4.459 8.22e-06 ***
DEBTINC      0.08506    0.02212   3.845  0.00012 ***
CREDDEBT     0.56336    0.08877   6.347 2.20e-10 ***
OTHDEBT      0.02315    0.05709   0.405  0.68517
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 553.41  on 692  degrees of freedom
AIC: 569.41

Number of Fisher Scoring iterations: 6
```

Interpretation :

- Since p-value is <0.05 for Employ, Address, Debtinc, Creddebt, these independent variables are statistically

Re-run Model in R

- Once variables to be retained are finalized ,re-run the model with these final variables and obtain revised coefficients for the model.
- Re-run the model with employ, address, debtinc, creddebt.

```
riskmodel<-glm(DEFAULTER~EMPLOY+ADDRESS+DEBTINC+CREDDEBT,
               family=binomial,data=data)

summary(riskmodel)
```

Re-run Model in R

```
# Output:
```

```
Call:
glm(formula = DEFAULTER ~ EMPLOY + ADDRESS + DEBTINC + CREDDEBT,
    family = binomial, data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4483  -0.6396  -0.3108   0.2583   2.8496

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.79107    0.25154  -3.145  0.00166 **
EMPLOY      -0.24258    0.02806  -8.646  < 2e-16 ***
ADDRESS     -0.08122    0.01960  -4.144 3.41e-05 ***
DEBTINC      0.08827    0.01854   4.760 1.93e-06 ***
CREDDEBT     0.57290    0.08725   6.566 5.17e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 804.36  on 699  degrees of freedom
Residual deviance: 556.74  on 695  degrees of freedom
AIC: 566.74

Number of Fisher Scoring iterations: 6
```

Interpretation :

- Since p-value is <0.05 for Employ, Address, Debtinc, Creddebt, these independent variables are significant and sign of the coefficients are also logical.

Final Model

Final Model is :

$$\log() = -0.79107 - 0.24258 * (\text{EMPLOY}) - 0.08122 * (\text{ADDRESS}) + 0.08827 * (\text{DEBTINC}) + 0.57290\ (\text{CREDDEBT})$$

This model is used for predicting the probabilities.

DATA SCIENCE INSTITUTE

Predicting Probabilities in R

```
# Predicting Probabilities
data$predprob<-round(fitted(riskmodel),2)
head(data,n=10)
```

- **fitted** function generates the predicted probabilities based on the final riskmodel.
- **round** function helps rounding the probabilities to 2 decimal
- **data$predprob:** Predicted probabilities are saved in the same dataset ‘data’ in new variable ‘predprob’.

Predicting Probabilities in R

```
# Output:
```

	SN	AGE	EMPLOY	ADDRESS	DEBTINC	CREDDEBT	OTHDEBT	DEFAULTER	predprob
1	1	3	17	12	9.3	11.36	5.01	1	0.81
2	2	1	10	6	17.3	1.36	4.00	0	0.20
3	3	2	15	14	5.5	0.86	2.17	0	0.01
4	4	3	15	14	2.9	2.66	0.82	0	0.02
5	5	1	2	0	17.3	1.79	3.06	1	0.78
6	6	3	5	5	10.2	0.39	2.16	0	0.22
7	7	2	20	9	30.6	3.83	16.67	0	0.19
8	8	3	12	11	3.6	0.13	1.24	0	0.01
9	9	1	3	4	24.4	1.36	3.28	1	0.75
10	10	2	0	13	19.7	2.78	2.15	0	0.82

Interpretation :

- Last column in the data ‘predprob;’ is the probabilities generated using final model.

Classification Table

- Based on **cut-off value** of p, Y is estimated to be either 1 or 0

  Ex. $p>0.5$; $Y=1$

  $p\leq 0.5$; $Y=0$
- **Cross tabulation** of observed values of Y and predicted values of Y is called as **Classification Table.**
- The predictive success of the logistic regression can be assessed by looking at the classification table, but classification table is not always a good measure of goodness fit since it **varies with the cut off value set.**
- Accuracy Rate measures **how accurate a model is in predicting outcomes.**
- In the adjoining table, 479 times Y=0 was observed as well as predicted. Similarly, Y=1 was observed and predicted 92 times.

Accuracy Rate = (479+92)/700 = 81.57

Observed \ Expected	0	1
0	479	38
1	91	92

DATA SCIENCE INSTITUTE

Misclassification

- Misclassification Rate ✉ Percentage of wrongly predicted observations
- Note that misclassification rate depends on cut off used for predictions

Suppose our classification table looks as follows:

		Expected	
		0	1
Observed	0	479	38
	1	91	92

- Here **misclassification rate is : (38 +91) / 700=18.43%**