Introduction to

Binary Logistic Regression - I

Contents

Binary Logistic Regression

DEPENDENT VARIABLE	INDEPENDENT VARIABLE
Categorical (Binary)	Categorical or Continuous
0 or 1	Gender
Death or Survival	Geographical Regions
Absence or Presence	Age
	Income
	Debt Ratio
⋮	⋮

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

Application Areas

Industry	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.

Why Not Use Linear Regression Model?

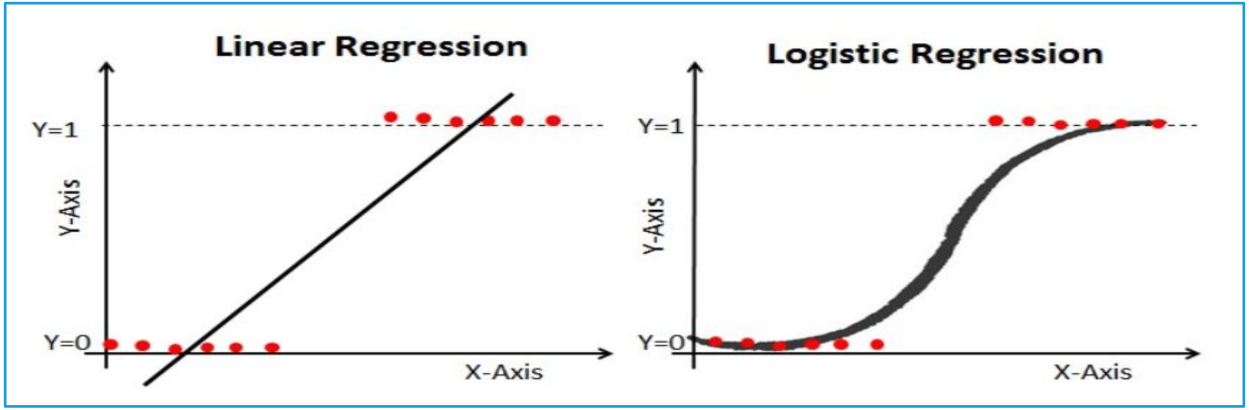
The statistical model for multiple linear regression is,

$$\boxed{Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p + e}$$

- 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.

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**.



Statistical Model

Statistical model for single predictor

$$p = \frac{e^{b_0 + b_1 x_1}}{1 + e^{b_0 + b_1 x_1}}$$

p is the Pr[Y=1/X] and X is the independent variable

$$1 - p = 1 - \frac{e^{b_0 + b_1 x_1}}{1 + e^{b_0 + b_1 x_1}} = \frac{1}{1 + e^{b_0 + b_1 x_1}}$$

$$\frac{p}{1 - p} = e^{b_0 + b_1 x_1} \rightarrow \boxed{\log\left(\frac{p}{1 - p}\right) = b_0 + b_1 X_1}$$

The left hand side uses ‘link function’

Statistical Model – For k Predictors

The model can be extended for k independent variables

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

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

Binary regression model is used to derive predicted probability of outcome.

Error, by definition, is the difference between observed and predicted value.

There is no such thing as comparable "observed probability" and hence **the model does not have any error component.**

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
OTHDEBT	Other Debt	Continuous	-	Positive value
DEFAULTER	Whether customer defaulted on loan	Binary	1(Defaulter), 0(Non-Defaulter)	2

Exploratory Data Analysis

- Before moving to modeling we can undertake some **exploratory data analysis**
- Depending upon the type of variable (Whether continuous or categorical) we can perform bivariate analysis

Cross Table 1: Relationship Between Defaulter and Age Group		
Age Group	**Defaulter**	
	Yes	No
1	86	156
2	61	223
3	36	138

Cross Table 2: Transactional Behaviour of Defaulters v/s Non Defaulters		
Types of Liabilities	**Average Liabilities**	
	Defaulter	Non-Defaulter
Credit Card Debt	2.42	1.25
Debt To Income Ratio	14.73	8.68
Other Debt	3.86	2.77

- Such data insights are also an important aspect of modeling

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

$$\log\left(\frac{p}{1-p}\right) = b_0 + b_1\,(\mathrm{AGE}) + b_2\,(\mathrm{EMPLOY}) + b_3\,(\mathrm{ADDRESS}) + b_4\,(\mathrm{DEBTINC}) + b_5\,(\mathrm{CREDDEBT}) + b_6\,(\mathrm{OTHDEBT})$$

Parameter Estimates

* Note that, this is not the final model. The equation is showed just for the understanding purpose. Only the significant variables will be part of the model.

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)

Parameters, Probability and Odds

Odds = Probability of Success (p) / Probability of Failure (1-p)

- In binary logistic regression, **model LHS is logit(p),** which is the **log of odds.** Hence, estimated parameter gives the **change in log of odds given one unit change in the independent variable.**
  Estimated coefficient of EMPLOY is -0.26172. This means that one unit change in EMPLOY will result in a change of -0.26172 in log of odds. The negative sign implies customer with a relatively steady job, is less likely to default.
- In order to get a more **straightforward and usable association between the independent and dependent variables,** Odds Ratio is calculated.

What is Odds Ratio

- Odds Ratio is a **measure of association between the independent variable and the outcome.**
- It represents the factor by which the **odds (event) change for a one-unit change in the independent variable.**

The odds of outcome being present when $X = x$ is $e^{b_0 + b_1 x}$

The odds of outcome being present when $X = x+1$ is $e^{b_0 + b_1 (x+1)}$

Odds Ratio

$$\frac{e^{b_0 + b_1 (x+1)}}{e^{b_0 + b_1 x}}$$

$$= e^{b_1} = \mathrm{EXP}(b_1)$$

Odds Ratio – Case Study

	Coefficients	Odds Ratio
Intercept	-0.78821	0.4546572
AGE2	0.25202	1.2866254
AGE3	0.62707 `	1.8721087
EMPLOY	**-0.26172**	**0.7697228**
ADDRESS	-0.09964`	0.9051601
DEBTINC	0.08506	1.0887859
CREDDEBT	0.56336 `	1.7565703
OTHDEBT	0.02315 `	1.0234175

- When association between dependent and independent variable is
  - Positive: OR > 1
  - Negative: OR < 1
- OR = 1 indicates no association between variables

- For one unit change in EMPLOY the odds of default will change by 0.7697228 folds.

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): : $bi \neq 0$

$i=1,2\ldots,k$

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**

* Note that, few softwares like SAS provide Wald's chi square since, $z2 \sim \chi2(1)$)

Quick Recap

In this session, we learned about **Binary Logistic Regression** :

Binary logistic regression	• Dependent variable is binary and independent variables are categorical or continuous or mix of both. • Regression line is sigmoid curve. • Parameters are estimated using MLE.
ODDS Ratio	• measure of association between the independent variable and the outcome.