ISSN No:-2456-2165

LOGISTIC REGRESSION MODEL –A REVIEW

Mrinalini Smita
Assistant Professor,
Intermediate Section, St. Xavier's College, Ranchi, Jharkhand

Abstract:- The method of determining future values of a company's stocks and other financial values is called stock price prediction. The movements of stock prices and stock indices are influenced by many macro-economic variables such as political events, policies of the corporate enterprises, general economic conditions, commodity price index, bank rate, loan rates, foreign exchange rates, investors' expectations, investors' choices and the human psychology of stock market investors. [Miao et al,2007**] Hence to develop predictive models for stock market prediction is a difficult task due to the uncertainty involved in the movement of stock market. That is why it requires continuous improvement in forecasting models. Forecasting accuracy is the most important factor in selecting any forecasting methods. Financial ratios influence investment decision-making. This is the reason that stock market prediction with the help of binary logistic regression using relation between financial ratios and stock performance can enhance an investor's stock price forecasting ability. This paper is presenting a review on Logistic regression Model (LRM).**

Keywords:- *Stock Price Prediction, Financial Ratios, Logistic Regression Model.*

I. INTRODUCTION

LOGISTIC REGRESSION

Regression analysis is one of the most useful and the most frequently used statistical methods. The aim of the regression methods is to describe the relationship between a response variable and one or more explanatory variables. Among the different regression models, logistic regression plays a particular role.[**Ngunyi et. al, 2014**] Logistic regression extends the ideas of linear regression to the situations where the dependent variable Y is categorical. Logistic Regression is applied to categorize a bunch of independent variables into either two or more mutually exclusive classes. [**Ali et. al, 2018**]Logistic regression seeks to

- **MODEL** the probability of an event occurring depending on the values of the independent variables , which can be categorical or numerical .
- **ESTIMATE** the probability that an event occurs for a randomly selected observations verses that the probability that the event does not occur.
- **PREDICT** the effect of a series of variables on binary response variables.
- **CLASSIFY** observations by estimating the probability that an observation is in a particular category (such as GOOD or POOR performance of Company of S&P BSE 30 in our case).

OBJECTIVE OF THE STUDY:

- To help investors to get idea where to invest their valuable money.
- To enable stock brokers and the investing public to make better informed decisions.

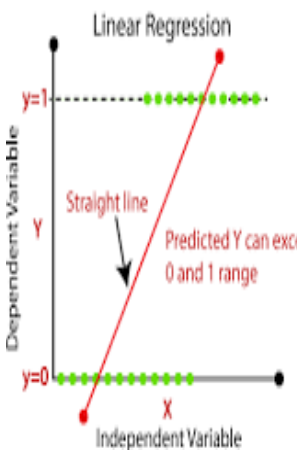
II. LOGISTIC REGRESSION MODEL

Binary logistic regression model is also known ***predictive model***. It is used where data in dichotomous or binary (0 or 1) dependent variables like good/poor, success/failure etc. [**Ali et. al, 2018**] In logistic regression, the goal is to predict classification the observations into one of the classes.

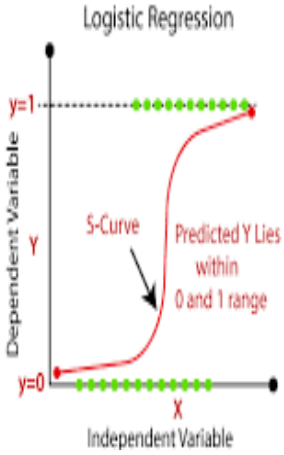
ASSUMPTIONS OF LOGISTIC REGRESSION MODEL:

- Dependent variables should be measured on a dichotomous scale. For example, performance of stock-GOOD or POOR
- We must have one or more continuous or categorical independent variables .
- The dependent variable should have mutually exclusive and exhaustive categories.
- A linear relationship should be between any continuous independent variables along with the logit transformation of the dependent variable.
- There should be high correlation with two or more independent variables i.e. data must show multicollinearity .

Logistic regression is a predictive analysis which extends the idea of linear regression to the situation where dependent variable is ***categorical variable with two levels, including Y/N, High/ low, Good/ Poor*** while predictor can be continuous or dichotomous, just as in regression analysis. Since the probability of an event must be between 0 and 1, it is impractical to model probabilities with linear regression techniques because the ***linear regression model allows the dependent variable to take values greater than 1 or less than 0*** .[**Dutta et. al 2012**] . Moreover, calculations using linear regression are very complex. In linear regression , accuracy is low.[**Navale et. al, 2016**]



Graph1. Linear Regression model



Graph2. Logistic Regression model

Graph depicting Difference between LINEAR Regression and LOGISTIC Regression :

In Logistic regression, instead of predicting ***the value of the variable Y*** from a predictor variable X or several predictor variable Xs , we predict the ***Probability of Y occurring*** , given known values of Xs. Hence the logistic regression model is a type of generalised linear model(GLM) that extends the linear regression model by linking the range of real numbers to 0-1 range.

The ***logistic regression model or logit probit model*** is given below.

$$p(x_i) = P(y_i = 1 : x_i)$$

$$= [1 + \exp(X^T \beta)]^{-1}$$

and $$X^T\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots .. \beta_n x_n$$

Where x_1, x_2 ..... are independent variables and β is the coefficient.

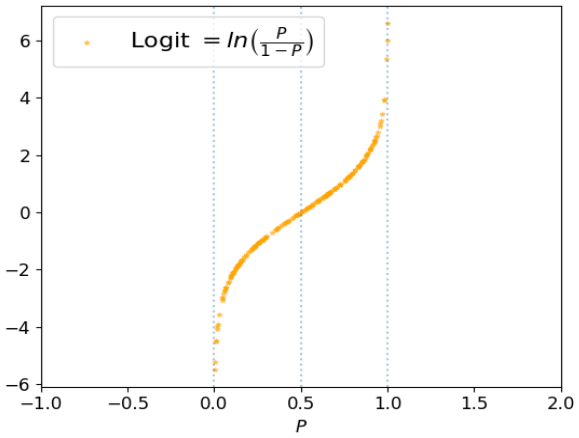
Logistic regression model the relationship between the dichotomous dependent variables depending upon odds ratios. ODDS RATIO for a variable in logistic regression represents ***how the odds change with one unit increase in that variable holding all others variables constant.*** In logistic regression , the dependent variable is a *log odds or logit*, which is the natural log of odds. [Dutta et. al., 2012]

$$\frac{p(x)}{p(1-x)} = [\exp(-X^T\beta)]^{-1} = \textbf{ODDS}$$

Taking natural log both sides

$$\log \frac{p(x)}{p(1-x)} = [-(X^T\beta)]^{-1}$$

Above transformation in log is known as ***logistic transformation (logit).***
[Ali et. al, 2018]

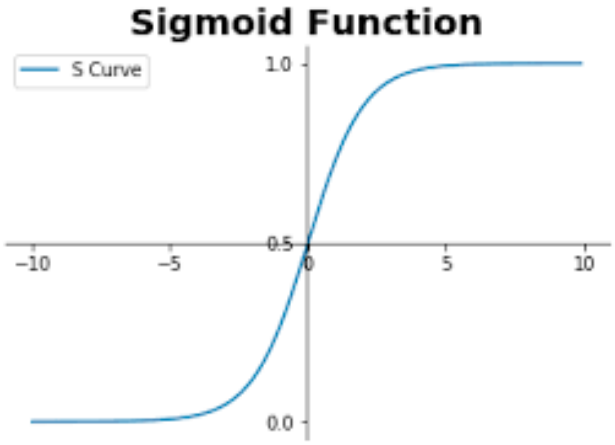


Graph 3. Graph of Logit

Classification in case of linear regression is not possible as $h(x)$ given by

$h(x) = \sum_{i=0}^{N} \beta_i x_i$, where N is the number of predictors.

Always gives real values . So we can apply another function on the linear function so that we can use the result for classification. That another function is ***Logistic Function (Sigmoid Function) ,*** which is a S-shaped curve .



Graph 4: Sigmoid Function

In the graph , as $z \to \infty, g(z) \to 1$

and $z \to -\infty, g(z) \to 0$

g(z) or $\sigma(z) = \frac{1}{1+e^{-z}}$

Just like in regression h(x), in logistic regression for classification, we have

$h(x) = g(\sum \beta_i x_i)$

$= g(\beta^T X)$ [Matrix notation]

Or $\sigma(z) = \frac{1}{1+e^{-\beta^T X}}$

We can use a linear function of β, pass it through the ***Sigmoid function*** (S-shaped curve) and use it for ***classification.***

This ***derivative*** of Sigmoid function, given by
$g'(z) = g(z)(1 - g(z))$,

is the most attractive feature of Sigmoid function which is extremely simple to compute and making use of it for classification problem.

III. PARAMETER ESTIMATION

Maximum Likelihood Estimation (MLE) is a method of estimating the parameter of probability distribution by maximizing a likelihood function, in order to increase the probability of occurring the observed data. The goal of maximum likelihood is the optimal way to fit a distribution to the data.

In ***Maximum Likelihood Estimation***

- Consider N samples with labels either 0 or 1
- ***For samples labelled "1"***: estimate $\hat{\beta}$(β-hat) such that $p(X)$ is as closed to 1 as possible
- ***For samples labelled "0":*** estimate $\hat{\beta}$(β-hat) such that

1- $p(X)$ is as closed to 1 as possible i.e. the maximum value.

On combining these requirements ,we want to find β parameters such that both these product is maximum over all elements of dataset. This function we need to optimize is called ***likelihood function.***

Thus , on combining the products

$$L(\beta) = \prod_{x\,for\,y_i=1} p(x_i) \cdot \prod_{x\,for\,y_i=0} (1 - p(x_i))$$

$$L(\beta) = \prod p(x_i)^{y_i} \cdot (1 - p(x_i))^{1-y_i}$$

$$l(\beta) = \sum_{i=1}^{n} y_i \log(p(x_i)) + (1 - y_i)\log(1 - p(x_i))$$

[Here l represents log-likelihood]

$$l(\beta) = \sum_{i=1}^{n} y_i \log\left(\frac{1}{1 + exp^{-\beta x_i}}\right) + (1 - y_i)\log\left(\frac{exp^{-\beta x_i}}{1 + exp^{-\beta x_i}}\right)$$

$$where\ p(x_i) = \frac{1}{1+exp^{-\beta x_i}}$$

$$l(\beta) = \sum_{i=1}^{n} y_i \left[\log\left(\frac{1}{1 + exp^{-\beta x_i}}\right) - \log\left(\frac{exp^{-\beta x_i}}{1 + exp^{-\beta x_i}}\right)\right] + \log\left(\frac{exp^{-\beta x_i}}{1 + exp^{-\beta x_i}}\right)$$

$$l(\beta) = \sum_{i=1}^{n} y_i [\log(exp^{\beta x_i})] + \log\left(\frac{exp^{-\beta x_i}}{1 + exp^{-\beta x_i}} \cdot \frac{exp^{\beta x_i}}{exp^{\beta x_i}}\right)$$

$$l(\beta) = \sum_{i=1}^{n} y_i\ \beta x_i + \log\left(\frac{1}{1 + exp^{\beta x_i}}\right)$$

$l(\beta) = \sum_{i=1}^{n} y_i\ \beta x_i - \log(1 + exp^{\beta x_i})$ [transcendental equation]

This is the final form of ***likelihood function which is to be optimized.*** The goal is to find the value of β that maximizes this function. This is called transcendental equation (involving logarithm and exponents term) which is computationally expensive. However , we can use numerical methods for approximation. To update β parameters to maximise this function we use Newton-Raphson Method to converge to the maximum of this function for estimation.

IV. NEWTON RAPHSON METHOD FOR PARAMETER ESTIMATION

One of the most common method to determine the beta β values for the logistic regression equation and then to make predictions on new data is done by Newton Raphson Method. Newton Raphson is a deterministic numerical optimization technique. The primary advantage of using Newton Raphson compared to probabilistic alternatives is that in most iterations this method is fast.

We consider the ***Newton Raphson Method***.

$$\nabla_\beta l(\beta) = \nabla_\beta l(\beta^*) + (\beta - \beta^*)\nabla_{\beta\beta} l(\beta^*)$$

$$\nabla_\beta l(\beta^*) + (\beta - \beta^*)\nabla_{\beta\beta} l(\beta^*) = 0$$

$$\beta = \beta^* - \frac{\nabla_\beta l(\beta^*)}{\nabla_{\beta\beta} l(\beta^*)}$$

$$\beta^{t+1} = \beta^t - \frac{\nabla_\beta l(\beta^t)}{\nabla_{\beta\beta} l(\beta^t)} \qquad [\textbf{\textit{Newton raphson Equation}}]$$

We need to compute this for t iterations then data will eventually converge to the approximate coefficient vector.

Now we compute the ***Gradient*** with respect toβ:

$$\nabla_\beta l = \nabla_\beta \sum_{i=1}^{n} y_i \beta x_i - \log(1 + \exp^{\beta x_i})$$

$$\nabla_\beta l = \sum_{i=1}^{n} \nabla_\beta [y_i \beta x_i - \log(1 + \exp^{\beta x_i})]$$

$$\nabla_\beta l = \sum_{i=1}^{n} \nabla_\beta [y_i \beta x_i] - \nabla_\beta [\log(1 + \exp^{\beta x_i})]$$

$$\nabla_\beta l = \sum_{i=1}^{n} y_i x_i - \left[\frac{1}{1+\exp^{\beta x_i}} . \exp^{\beta x_i} . x_i\right]$$

$$\nabla_\beta l = \sum_{i=1}^{n} y_i x_i - \left[\frac{1}{1+\exp^{-\beta x_i}} .. x_i\right]$$

$$\nabla_\beta = \sum_{i=0}^{n} y_i x_i - [p(x_i)] x_i]$$

$$\nabla_\beta l = \sum_{i=1}^{n} [y_i - [p(x_i)] . x_i \quad [\textbf{\textit{Gradient Vector}}]$$

This is the ***numerator term*** of Newton- Raphson Equation.

Next we compute the ***denominator term*** ($\nabla_{\beta\beta}\, l(\beta^t)$) called the ***Hessian matrix*** which is a matrix of second order derivatives with respect to data coefficients.

$$\nabla_{\beta\beta} l = \nabla_\beta \sum_{i=1}^{n} [y_i - p(x_i)] x_i$$

$$\nabla_{\beta\beta} l = \sum_{i=1}^{n} \nabla_\beta [y_i - p(x_i)] x_i$$

$$\nabla_{\beta\beta} l = \sum_{i=1}^{n} \nabla_\beta - p(x_i) x_i$$ [remove y, as it is independent of β]

$$\nabla_{\beta\beta} l = \sum_{i=1}^{n} \nabla_\beta - \left[\frac{1}{1+exp^{-\beta x_i}}\right] x_i$$

$$where\ p(x_i) = \frac{1}{1+exp^{-\beta x_i}}$$

$$\nabla_{\beta\beta} l = \sum_{i=1}^{n} \nabla_\beta \left[\frac{1}{1+exp^{-\beta x_i}}\right]^2 . exp^{-\beta x_i} . (-x_i) . x_i$$

$$\nabla_{\beta\beta} l = -\sum_{i=1}^{n} [\frac{exp^{-\beta x_i}}{1+exp^{-\beta x_i}}] . [\frac{1}{1+exp^{-\beta x_i}}] . x_i{}^T x_i$$

$$\nabla_{\beta\beta} = -\sum_{i=1}^{n} p(x_i)(1-p(x_i) x_i{}^T x_i \ [\textbf{\textit{Hessian Matrix}}]$$

Now converting these into Matrix Notation:

$$\nabla_\beta l = X^T (Y - Y^\wedge)$$

$$\nabla_{\beta\beta} l = -X^T P(1-P) X$$

$\nabla_{\beta\beta} l = -X^T W X$ [replacing P(1-P) by W as diagonal matrix]

Substituting these values in Newton Raphson Equation, we have

$$\beta^{t+1} = \beta^t + (X^T W^{(t)} X)^{-1} X (Y - Y^{\wedge(t)})$$

Then we have to execute it for number of iterations 't' until the value of converges. Once the coefficients have been estimated, we can substitute the values of some feature vectors X to estimate the probabilities of it belonging to a specific class (by choosing a parameter above which it is ***class 1 (GOOD***) and below which it is ***class 0 (POOR***).

The final logistic regression equation estimated by using the maximum likelihood estimation for classifying the dependent variable, z for given independent variables , $x_1, x_2, \ldots\ldots\ldots x_n$ is:

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots . \beta_n x_n + \in$$

where: $z = \log(\frac{p}{1-p})$ and '$p = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots.\beta_n x_n)}}$' is the probability that the outcome is GOOD(1) and $\in$ is the error term.[**Ali et al, 2018**]

Hence the three important steps for classification prediction through logistic regression model is given by

1. $logit = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots . \beta_n x_n + \in$

Where β_0 is the intercept and $\beta_1, \beta_2, \beta_3, \ldots\ldots\ldots \beta_n$ are co-efficients to be determined .

2. ODDS = e^{logit}

3. $P(Y) = \frac{odds}{1+0dds}$

MERIT/ RELEVANCE OF THE LOGISTIC REGRESSION MODEL :

- The variables may be either continuous or discrete or any combination of both types and they do not necessarily have normal distributions.
- The logit function is particularly popular because its results are relatively easy to interpret.
- The contribution of Logistic Regression (co-efficient and intercept) can be interpreted.
- It is possible to test the statistical significance of the coefficients in the model. These statistical tests are Kolmogorov-Smirnov Test, Hosmer and Lemeshow Test, Omnibus test of model coefficients which can be used to build models incrementally .
- This model helps investors to form an opinion about the right time to invest with better decisions.

DEMERIT/LIMITATION OF THE LOGISTIC REGRESSION MODEL:

- These models can predict linear patterns only.

V. CONCLUSION

The logistic regression model is very interesting and important in the field of stock market forecasting. The ultimate goal is to increase the yield from the investment by providing useful information to shareholders and potential investors to enable them to make better decisions regarding investments. This model has overcome the tedious, expensive and time-consuming process of traditional techniques of prediction. This model can be a stepping stone for future prediction technologies. Moreover, the logistic regression model can be used for comparative study with other models like Artificial Neural Network , Time –Series Model etc. as well, in future.

REFERENCES

[1]. Miao, K., Chen, F. & Zhao, Z.G. (2007). "*Stock price forecast based on bacterial colony RBF neural network*". *Journal of Qingdao University (Natural Science Edition), 2*, 11.
[2]. Ngunyi, A., Mwita, P.N., Odhiambo,R.O., "*On the estimation of properties of logistic regression Parameters", IOSR Journals of Mathematics.,e-ISSN:2278-5728,Volume10,Issue 4,(2014) :57-68*
[3]. Ali, S.S. , Mubeen, M., Lal I., Hussain A. , "*Prediction of stock performance by using logistic regression model: evidence from Pakistan Stock Exchange (PSX)*" , Asian Journal of Empirical Research ,Volume 8, Issue 7 (2018): 247-258 ISSN (P): 2306-983X, ISSN (E): 2224-4425
[4]. Navale, G., Dudhwala, N., Jadhav, K. , Gabda, P. , Vihangam , B.K., "*Prediction of stockmarket using Data mining and Artificial Intelligence*" , IJESC, Vol. 6,ISSN: 2321-3361,2016, 1-6,6539-6544.
[5]. Upadhyay, A., Bandyopadhyay, G., & Dutta, A. (2012). "*Forecasting stock performance in indian market using multinomial logistic regression.", Journal of Business Studies Quarterly, 3*(3), 16-39.