Data Mining for Business

Logistic Regression Model

Dr. Shipra Maurya
Department of Management Studies
IIT (ISM) Dhanbad
Email: shipra@iitism.ac.in

When limited dependent variables may be used?

- Why firms choose to list their shares on the NSE rather than the BSE?
- Why some stocks pay dividends while others do not?
- What factors help in identifying whether a product is defective or not?
- Why some firms choose to issue new stock to finance an expansion while others issue bonds
- Why some firms choose to engage in stock splits while others do not.
- **It is fairly easy to see in all these cases that the appropriate form for the dependent variable would be a 0-1 dummy variable since there are only two possible outcomes. There are cases when dependent variable may take on other values too.**

The Linear Probability Model (1/3)

- Works like a normal linear regression model, but the interpretations change because now Y (target variable) is binary
- It is based on an assumption that the probability of an event occurring, Pi, is linearly related to a set of explanatory variables

$$P_i = p(y_i = 1) = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \ldots + \beta_k x_{ki} + u_i$$

- The actual probabilities cannot be observed, so we would estimate a model where the outcomes, yi (the series of zeros and ones), would be the dependent variable.
- The set of explanatory variables could include either quantitative variables or dummies or both.
- A predicted value (Yˆ) is the predicted probability that the dependent variable equals one, given X

The Linear Probability Model (2/3)

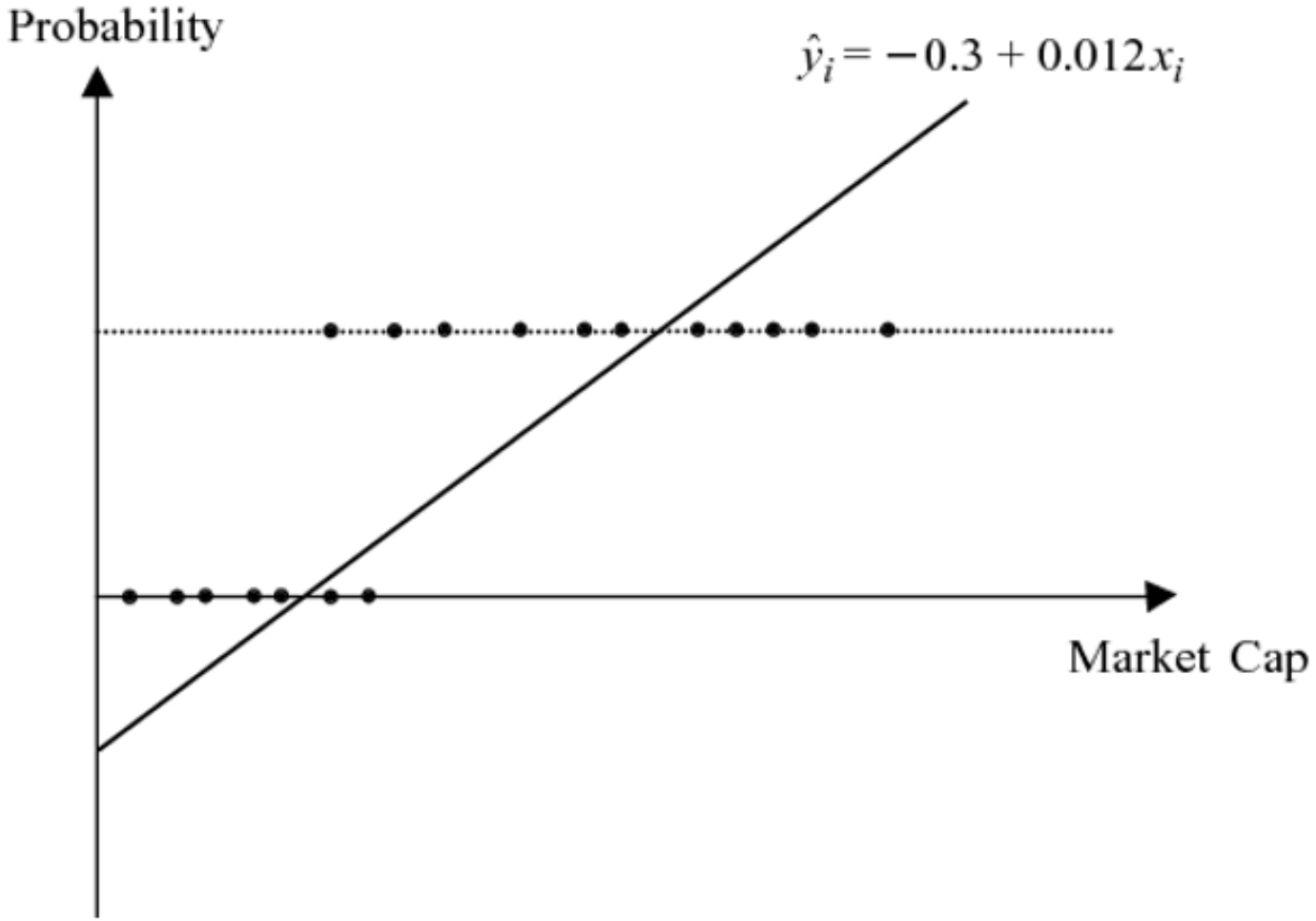
- **How to Interpret Slope coefficient:** the change in the probability that the dependent variable will equal 1 for a one-unit change in a given explanatory variable, holding the effect of all other explanatory variables fixed.
- Suppose, for example, that we wanted to model the probability that a firm *i* will pay a dividend p(y_i = 1) as a function of its market capitalization (x_{2i} , measured in millions of US dollars), and we fit the following line:

$$\hat{P}_i = -0.3 + 0.012x_{2i}$$

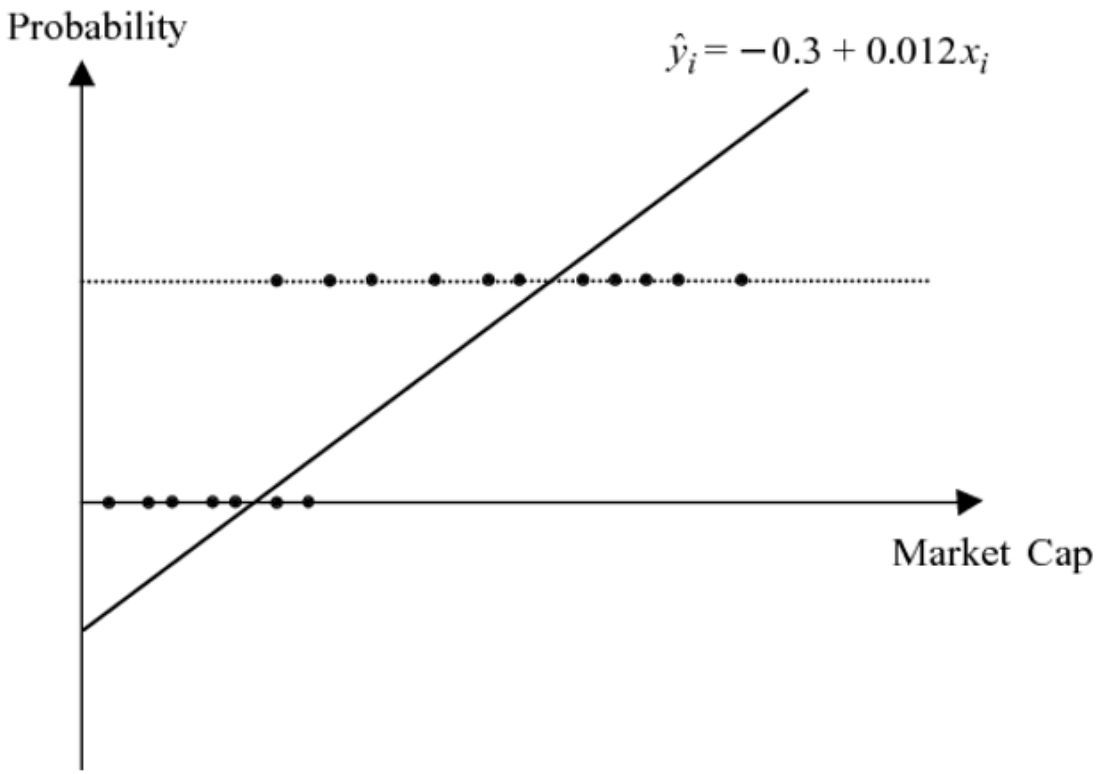
where $\hat{P}_i$ denotes the fitted or estimated probability for firm *i*.

- This model suggests that for every $1m increase in size, the probability that the firm will pay a dividend increases by 0.012 (or 1.2%).
- A firm whose stock is valued at $50m will have a -0.3+0.012x50=0.3 (or 30%) probability of making a dividend payment.

The Linear Probability Model (3/3)



Major Drawback of Linear Probability Model (1/2)



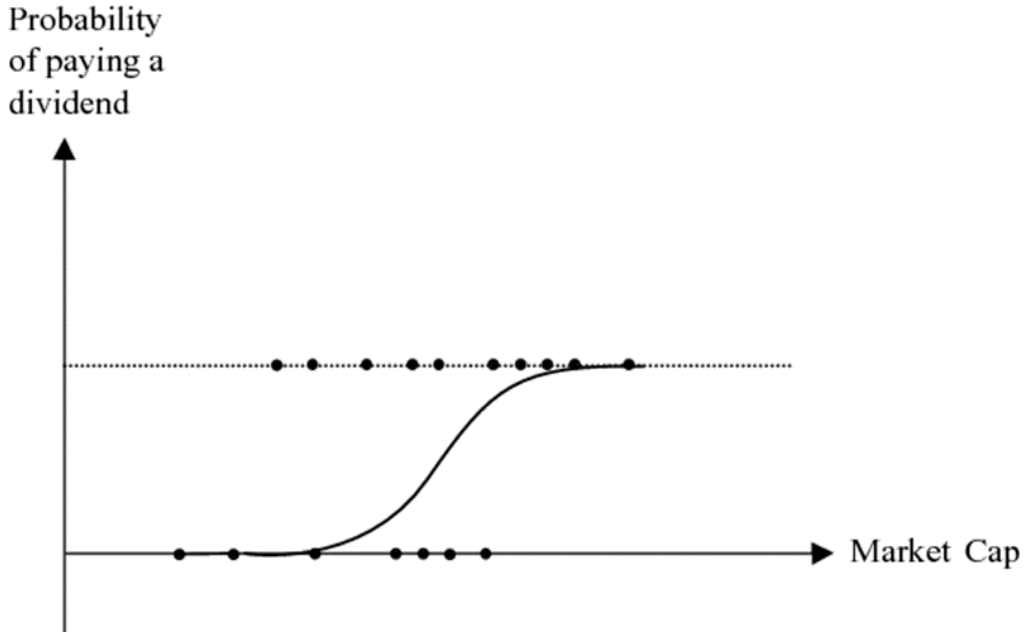
- For any firm whose value is less than $25m, the model-predicted probability of dividend payment is negative, while for any firm worth more than $88m, the probability is greater than one.
- Clearly, such predictions cannot be allowed to stand, since the probabilities should lie within the range (0,1).
- An obvious solution is to truncate the probabilities at 0 or 1, so that a probability of -0.3, say, would be set to zero, and a probability of, say, 1.2, would be set to 1.

Major Drawback of Linear Probability Model (2/2)

- However, there are at least two reasons why this is still not adequate.
- The process of truncation will result in too many observations for which the estimated probabilities are exactly zero or one.
- More importantly, it is simply not plausible to suggest that the firm's probability of paying a dividend is either exactly zero or exactly one. Are we really certain that very small firms will definitely never pay a dividend and that large firms will always make a payout?
- Probably not, and so a different kind of model is usually used for binary dependent variables i.e. Logistic Regression model.

Advantages of Logit over LPM

- The logit model can produce estimated probabilities that are negative or greater than one.
- It does this by using a function that effectively transforms the regression model so that the fitted values are bounded within the (0,1) interval.
- Visually, the fitted regression model will appear as an S-shape rather than a straight line, as was the case for the LPM.



Logit Model (1/2)

- Also known as Logistic Regression
- It is implemented when Dependent variable is categorical
- It uses the cumulative logistic distribution to transform the model so that the probabilities follow the S-shape curve (sigmoid function)
- It is non-linear model and hence OLS can not be used for estimation
- Instead, maximum likelihood is usually used to estimate the parameters of the model
- Probability can be converted into log(odds ratio) using following formula: log(odds ratio)= log(p/1-p)

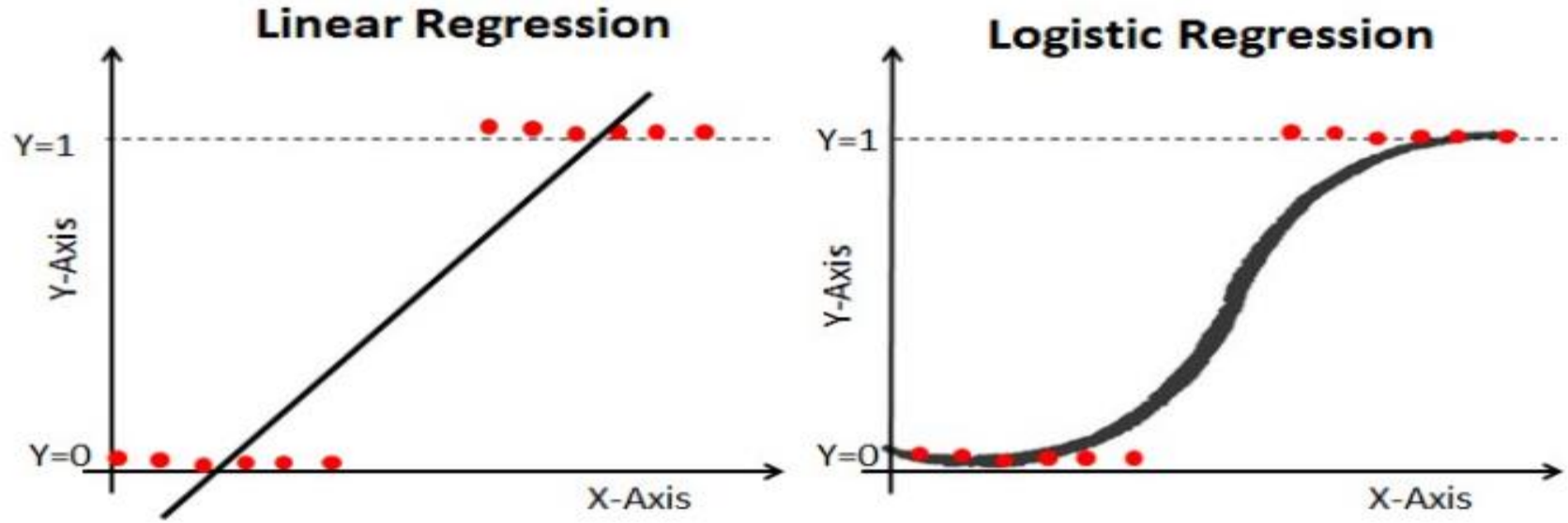
Logit Model (2/2)

- The Logistic function F (**cumulative logistic distribution**), which is a function of any random variable, *z (z=* $\beta_1+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}+u_i$) would be

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

- Where e is the exponential under the logit approach. Following logistic model would be estimated. We can use following formula to calculate the overall probability (predicted yhat).

$$P_i=\frac{1}{1+e^{-(\beta_1+\beta_2 x_{2i}+\cdots+\beta_k x_{ki}+u_i)}}$$



Parameter Interpretation for Logistic Regression Model

- **Interpretation of Coefficients:**

  - **For Logit coefficients:** Change in log odds ratio (z) by beta coefficient for one unit increase/decrease in x, holding constant all other k-1 independent variables.

Parameter Interpretation for Logit Models

- **Marginal Effects:**
  - Marginal effects are reported after reporting the coefficients in the logit model.
  - The marginal effects reflect the change in the probability of y=1 given one unit change in the x variable.
  - In OLS regression model, coefficients are the marginal effects. But in Logit model, coefficients are not the marginal effects, it has to be calculated from the coefficients
- **Interpretation of Marginal Effects:**
  - An increase in x increases (decreases) the probability that y=1 by the marginal effect expressed as a percentage holding constant all other k-1 independent variables.
    - For dummy independent variables, the marginal effect is expressed in comparison to the base category (x=0)
    - For continuous independent variables, the marginal effect is expressed for one-unit change in x.
  - We interpret both the sign and magnitude of the marginal effects

Types of Logistic Regression Model

- **Binary Regression:** Only two possible outcomes of the target variable (dichotomous). For example, Supplier is risky and not risky; Satisfied customer and not satisfied customer; customer may default and not default etc. Binary logit would be used.

- **Multinomial Regression:** three or more nominal categories in the target variable (without ordering). For example, which type of financing is preferred by firms (Pecking order theory) etc. Multinomial logit would be used.

- **Ordinal Regression:** three or more ordinal categories in the target variable. For example, customer satisfaction rating from 1 to 5 etc. Ordinal logit would be used.

Pseudo-R-square

- In logistic regression, an equivalent statistic to R-squared (covered in OLS) does not exist.
- "The model estimates from a logistic regression are maximum likelihood estimates arrived at through an iterative process. They are not calculated to minimize variance, so the OLS approach to goodness-of-fit does not apply."
- To evaluate the goodness-of-fit of logistic models, several pseudo R-squared have been developed.
- Pseudo R-squared value should range between 0 and 1 but there might be cases when it does not fall within 0 and 1. Pseudo R-squared can not be interpreted like R-squared.

McFadden's Pseudo-R-square

- $McFadden's\ R - Square = 1 - \frac{Log\ Likelihood\ of\ Full\ model}{Log\ Likelihood\ of\ Null\ model}$

Null model only has intercept

Full model has intercept plus independent variables (features)

- Higher value indicates a better fit.

- A pseudo R-squared only has meaning when compared to another pseudo R-squared of the same type, on the same data, predicting the same outcome.

Logit Models

- Convert the log(odds ratio) into probability by following: Probability=1/(1+e^-log(odds ratio))

- Convert the log(odds ratio) into Odds ratio by following: Odds ratio=e^log(odds ratio)

Thank you!

You can reach me on :

Email : shipra@iitism.ac.in

LinkedIn : https://www.linkedin.com/in/shipra-maurya1205/?originalSubdomain=in