Week 12
Artificial Intelligence (AI) for Investments

Lesson 1: Classification Algorithms: Logit/Probit Regression

Introduction

- Limited dependent variable modeling: background and motivation
- OLS approach: linear probability models (LPMs)
- Issues with LPM models
- Introduction to logit/probit models
- Understanding logit function

Introduction

- Thresholding
- Confusion/classification Matrix
- Receiver operator characteristic (ROC) curve
- Parameter interpretation
- Summary and concluding remarks

Background and Motivation

Limited Dependent Variable/Qualitative Response Regression

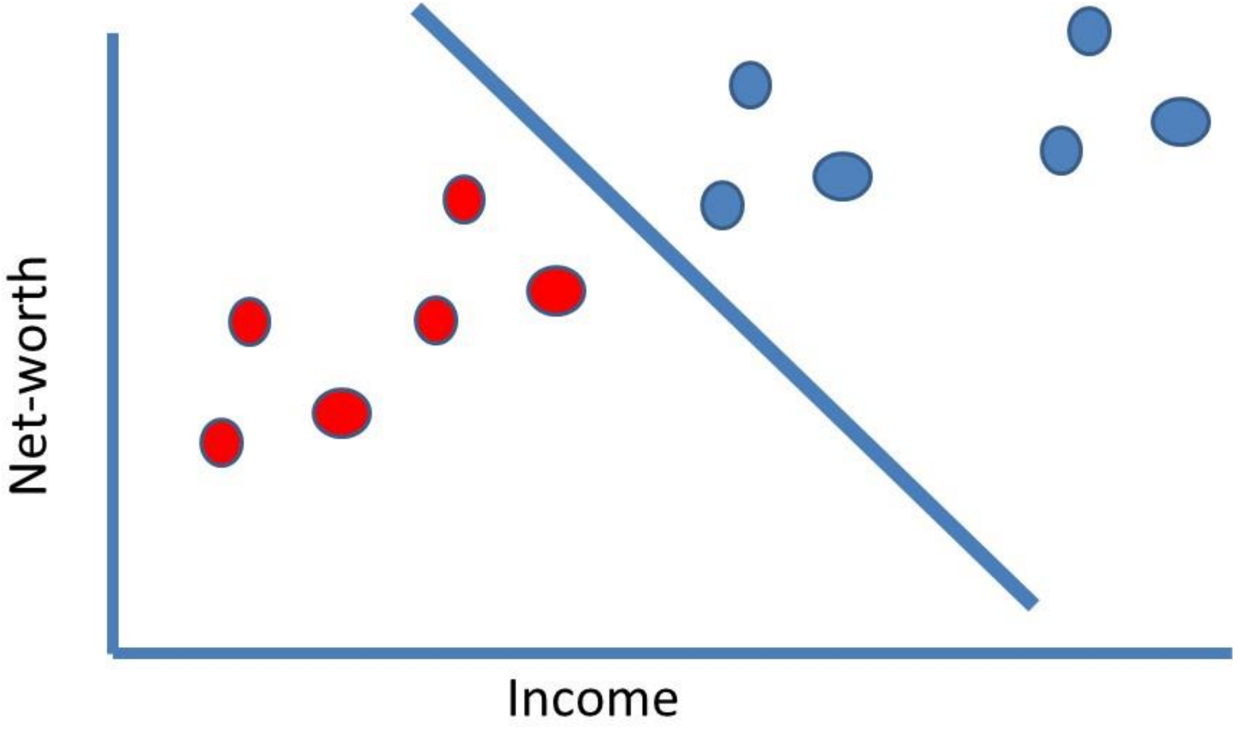
Discrete choice variables, limited dependent variables, or qualitative response variables are not suitable for modeling through linear regression models

Consider the following questions

- Why do firms choose to list their stocks on NSE vs. BSE?
- Why do some stocks pay dividends and others do not?
- What factors affect large corporate borrowers to default?
- What factors affect choices of internal vs. external financing?

Limited Dependent Variable/Qualitative Response Regression

Credit default scoring (classification problem)



Linear Probability Model (LPM)

Linear Probability Model (LPM)

- In such models, the dependent variable is Yes/No or 1/0 kind of variable
- First, we will examine a simple linear regression approach to deal with such models: linear probability model (LPM)
- This is the most simple approach to deal with binary dependent variables
- It is based on the assumption that the probability of an event (P_i) is linearly related to a set of explanatory variables, $x_{1i}, x_{2i}, \dots, x_{ki}$
- $P_i = p(y_i = 1) = \beta_1 + \beta_2 x_2 + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i,\ i = 1, \dots \dots, N$

Linear Probability Model (LPM)

In such models, the actual probabilities cannot be observed, so your estimates (or dependent variables) would be 0s and 1s

- Consider the relationship between the size of a company "*i*" and its ability to pay dividends

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

where X_i= market capitalization of the firm, and Y_i=1 if the dividend is paid and 0 if the dividend is not paid.

Linear Probability Model (LPM)

In such models, the actual probabilities cannot be observed, so your estimates (or dependent variables) would be 0s and 1s

- This is called linear probability model. The conditional expectation of Y_i given X_i, i.e., $E(Y_i|X_i)$, can be interpreted that the event will occur given X_i: that is, $P(Y_i = 1|X_i)$
- $E(Y_i|X_i)= \beta_1 + \beta_2 X_i$ (assuming $E(u_i)$=0)

Summary

Issues with LPM

Issues with LPM

Non-normality and heteroscedasticity of error terms

- Y_i has the following distribution

  $E(Y_i|X_i) = 0\times(1-P_i) + 1\times(P_i) = P_i$

- This kind of model has a number of econometric issues
- What is the nature of errors: $u_i = Y_i - \beta_1 - \beta_2 X_i$?

Y_i	Probability
0	$1 - P_i$
1	P_i
Total	1

	u_i	Probability
When $Y_i = 1$	$1 - \beta_1 - \beta_2 X_i$	P_i
When $Y_i = 0$	$-\beta_1 - \beta_2 X_i$	$(1 - P_i)$

Issues with LPM

Non-normality and heteroscedasticity of error terms

- u_i is not normally distributed; although in large samples, it is not a problem
- u_is are heteroscedastic, i.e., they vary with Y_i

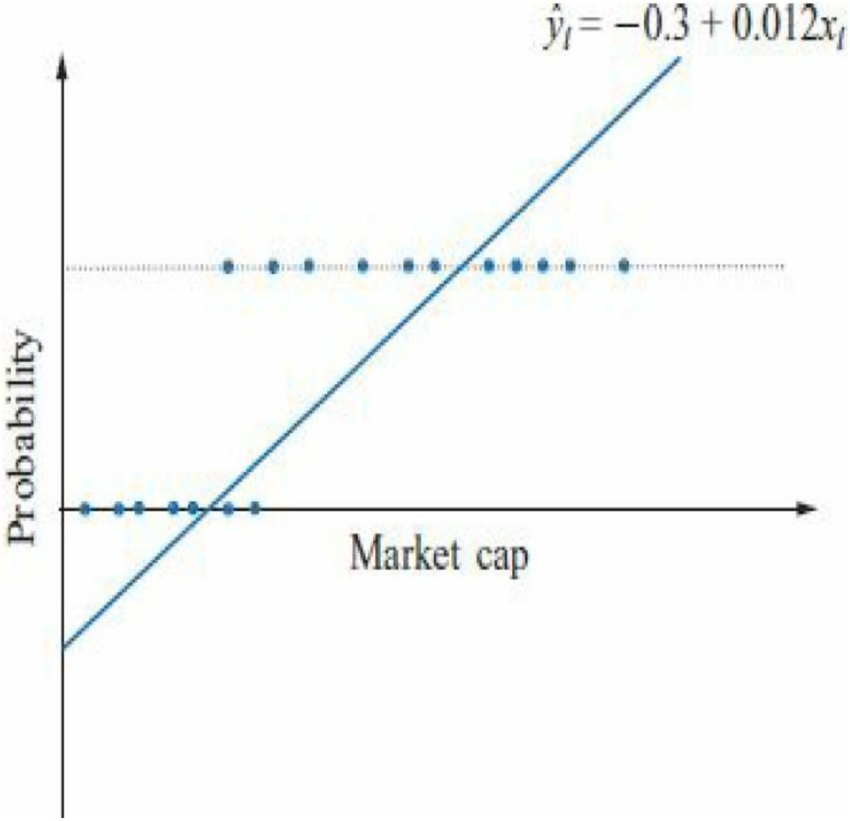
Y_i	Probability
0	$1 - P_i$
1	P_i
Total	1

	u_i	Probability
When $Y_i = 1$	$1 - \beta_1 - \beta_2 X_i$	P_i
When $Y_i = 0$	$-\beta_1 - \beta_2 X_i$	$(1 - P_i)$

Issues with LPM

Nonfulfillment of $0 \leq E(Y_i \mid X) \leq 1$

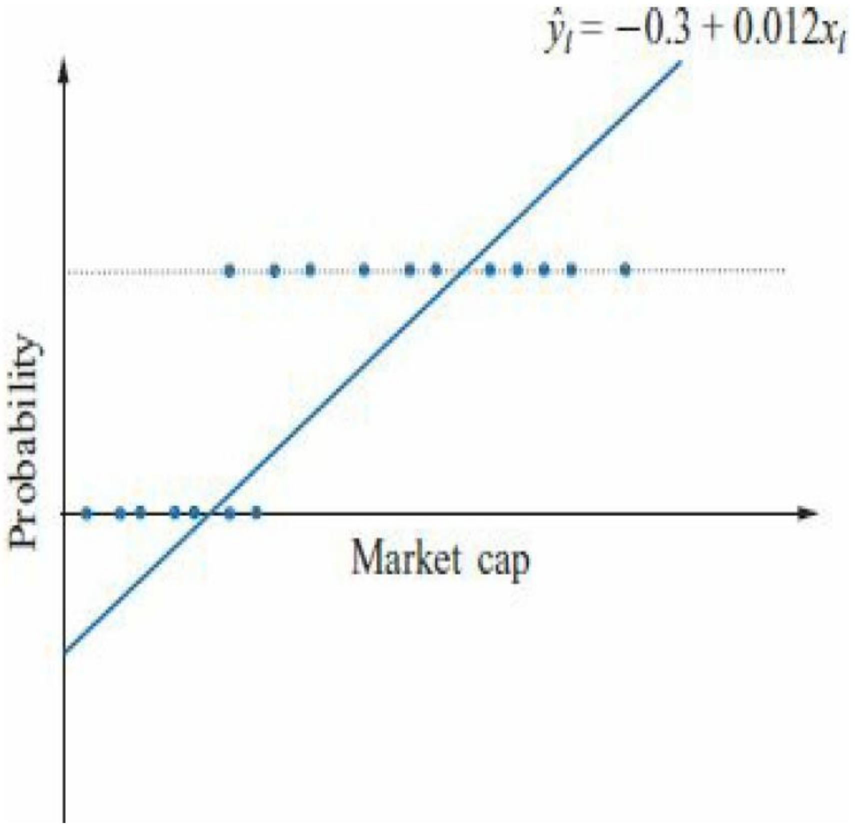
- $Y_i = -0.3 + 0.012X_i$; where X_i is in million dollars
- For every \$1 million increase in size, the probability that the firm will pay dividend increases by 1.2%
- However, for $X <$ \$25 million and $X >$ \$88 million, the probabilities are less than 0 and more than 1



Issues with LPM

Nonfulfillment of $0 \leq E(Y_i \mid X) \leq 1$

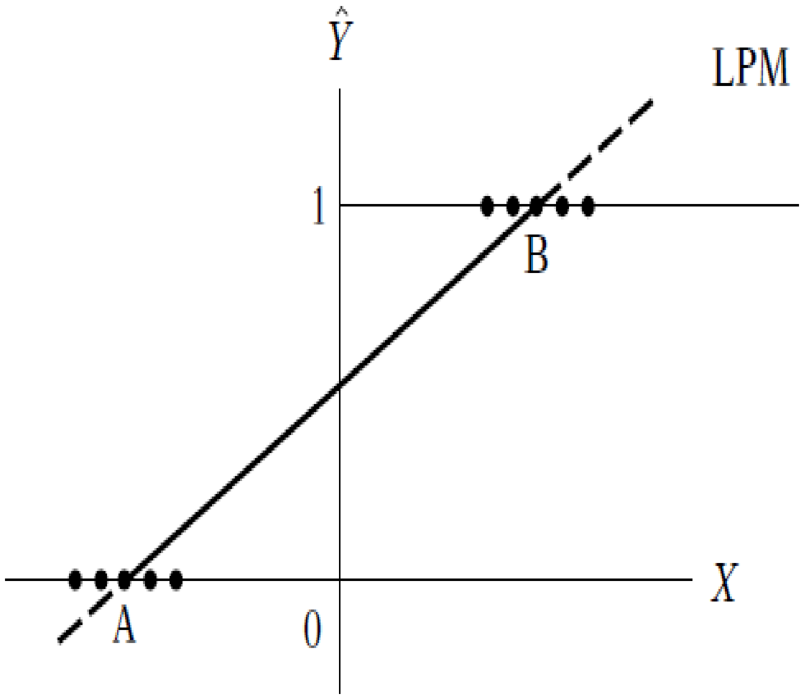
- What to do: set all negative as 0 and a those greater than 1 as 1?
- Implausible to suggest that small firms will never pay dividend and large firms will always pay dividends



Issues with LPM

Diminishing utility of R^2 as a goodness of fit measure

- All the Y values will be on a line $Y = 0$ or $Y = 1$
- The conventional LPM is not expected to fit well with such observations, except those cases where all the observations are scattered closely around points A and B
- Both logit and probit approaches are able to overcome the limitation of LPM that it produces values less than 0 and more than 1



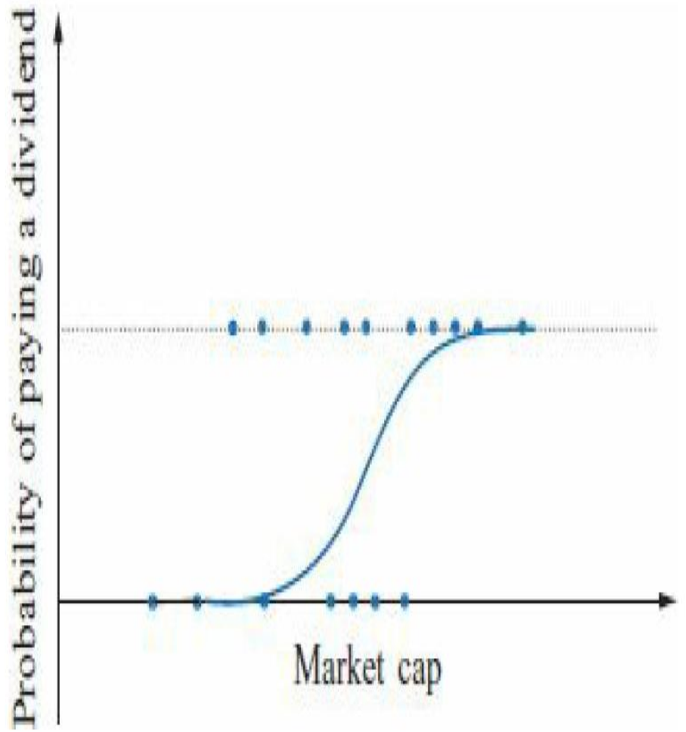
Introduction to Logit Model

Introduction to Logit Model

The logit (and probit) approaches overcome the limitations of the regression model by transforming to a function so that fitted values are bounded within (0,1) interval

- The fitted function looks like an S-shape curve

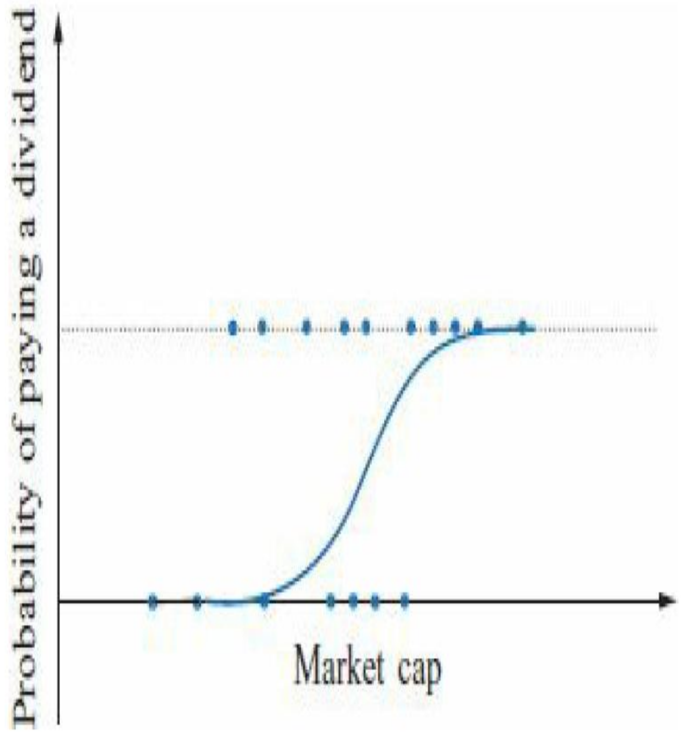
- The logistic function for a random variable z is: $F(z_i) = \frac{(e^{z_i})}{(1+e^{z_i})} = \frac{1}{(1+e^{-z_i})}$



Introduction to Logit Model

The logit (and probit) approaches overcome the limitations of the regression model by transforming to a function so that fitted values are bounded within (0,1) interval

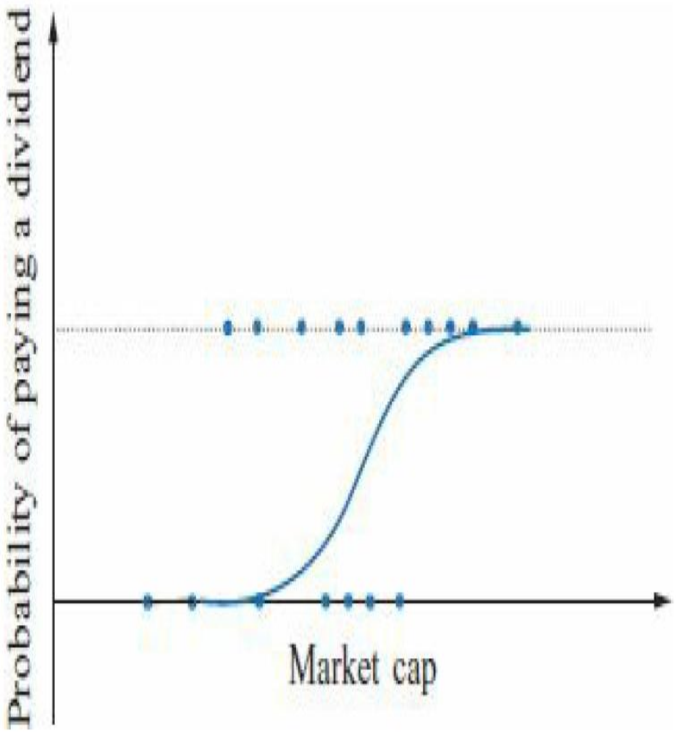
- Here *F* is the cumulative logistic distribution
- The final logit model: $P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$



Introduction to Logit Model

$$P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

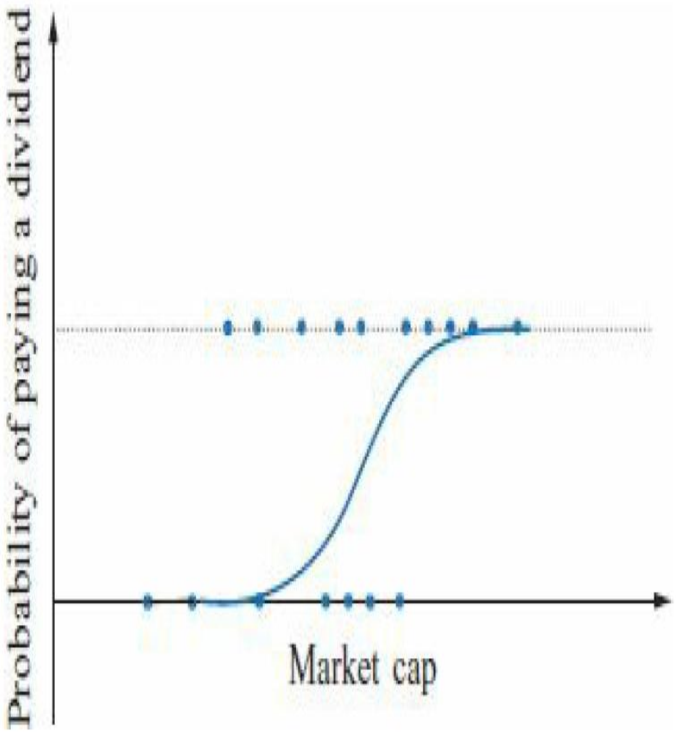
- Model asymptotically touches 0 ($z \to -\infty$) and 1 ($z \to \infty$)
- Is this model linear? Hence, not amenable to OLS estimation
- The model would predict that the probability, e.g., probability of bank loan default (dependent variable = y)



Introduction to Logit Model

$$P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

- $P\,(y = 1)$, then $P(y = 0) = 1 - P(y = 1)$
- Here independent variables are x_{2i}, x_{3i}, x_{4i}, x_{5i}, and so on
- This is essentially a non-linear transformation of the model to produce consistent probability results

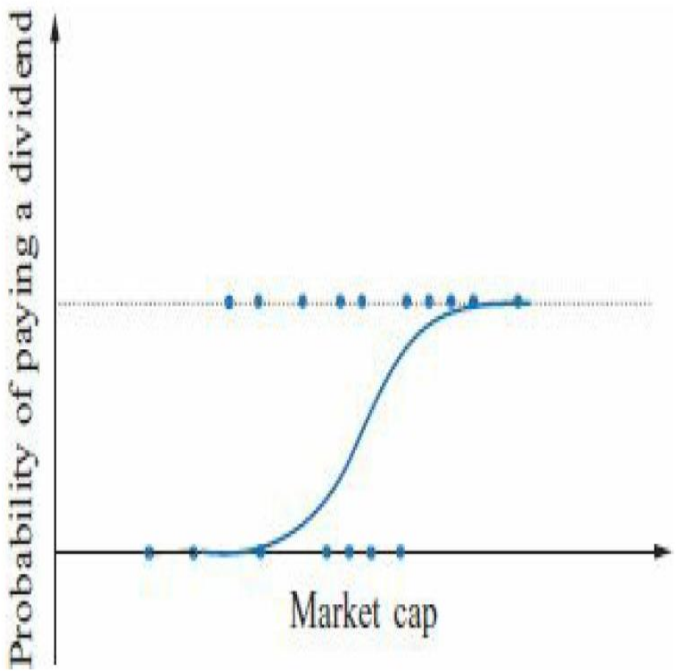


Understanding the Logit Function

Understanding the Logit Function

$$P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

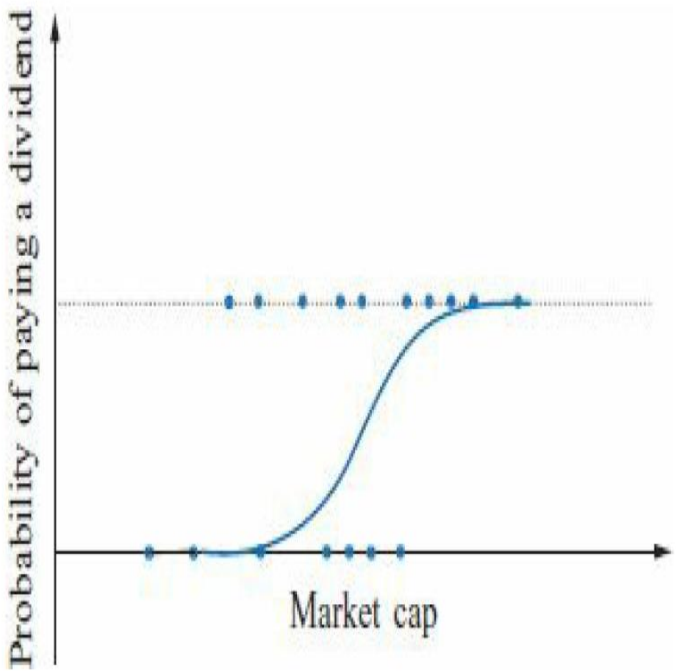
- Here extremely low and negative values of the linear function $\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki}$ would predict No dividend (or non-default cases) with a high probability or $P_i(y_i = 0)$



Understanding the Logit Function

$$P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

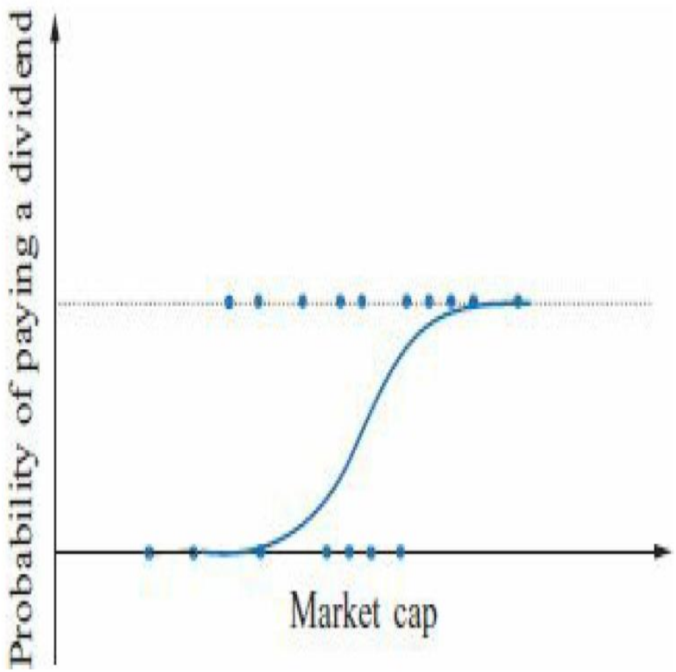
- Extremely high and positive values of the linear function $\beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki}$ would predict dividend payment (or default cases) with high probability or $P_i(y_i = 1)$



Understanding the Logit Function

$$P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

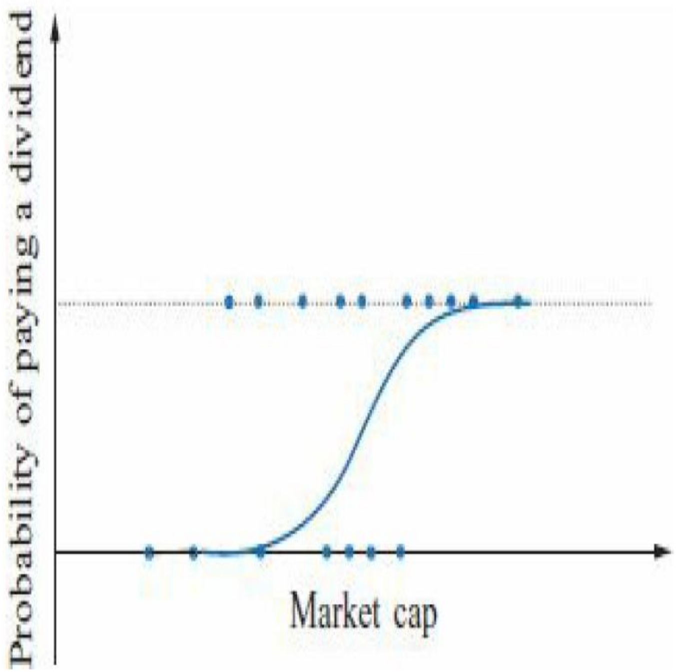
- This can also be expressed in the form of Odds
- $\text{Odds} = \frac{P(y=1)}{P(y=0)};$
- Odds > 1 if $y = 1$ is more likely
- Odds < 1 if $y = 0$ is more likely



Understanding the Logit Function

$$P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$$

- If we substitute the logit function in Odds equation, then
- $\text{Odds} = \exp^{(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)}$ or
- $\ln(\text{Odds}) = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i$
- The higher this logit (or $\ln(\text{Odds})$) form, the higher the probability for $P_i(y_i = 1)$



Thresholding

Thresholding

The outcome of the regression model is a probability

- In real life, you would want to make a binary prediction, e.g., default or no default
- For this, we may consider a threshold value "t"
- If P(Default = 1) >= t, then predict a default case
- If P(Default = 0)< t, then predict a non-default case

Thresholding

What value should we select for "t"? What kind of error do you prefer?

- Given a t value, one can make two types of errors: (1) predict default, but the actual outcome is non-default: false positive; and (2) predict non-default, but the actual outcome is default: false negative
- A large threshold (e.g., $t = 0.8$) will have a very small probability of predicting defaulters and, at the same time, a high probability of predicting cases as non-defaulters

Thresholding

What value should we select for “t”? What kind of error do you prefer?

- A small threshold (e.g., $t = 0.1$) will have a very large probability of predicting defaulters and, at the same time, a small probability of predicting cases as non-defaulters
- An aggressive bank would like to have high t values to increase the possibility of converting a loan

Thresholding

What value should we select for "t"? What kind of error do you prefer?

- A more conservative bank may choose a very low t value to select those loan applications with a very low probability of default
- In the absence of any threshold, t = 0.5 is the correct value to pick

Classification Matrix

Selecting a Threshold: Confusion/Classification Matrix

	Predicted = 0 (Non-Default)	Predicted = 1 (Default)
Actual = 0	True Negatives (TN)	False Positives (FP)
Actual = 1	False Negatives (FN)	True Positives (TP)

Let us compute two outcome measures to determine what kind of errors we are making

- $\text{Sensitivity} = \dfrac{\text{TP}}{\text{TP+FN}} = \text{TP rate}$
- $\text{Specificity} = \dfrac{\text{TN}}{\text{TN+FP}} = \text{TN rate}$

Selecting a Threshold: Confusion/Classification Matrix

Let us compute two outcome measures to determine what kind of errors we are making

- $\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} = \text{TP rate}$
- $\text{Specificity} = \frac{\text{TN}}{\text{TN+FP}} = \text{TN rate}$
- A model with higher *t* will have lower sensitivity and higher specificity
- A model with lower *t* will have higher sensitivity and lower specificity

Selecting a Threshold: Confusion/Classification Matrix

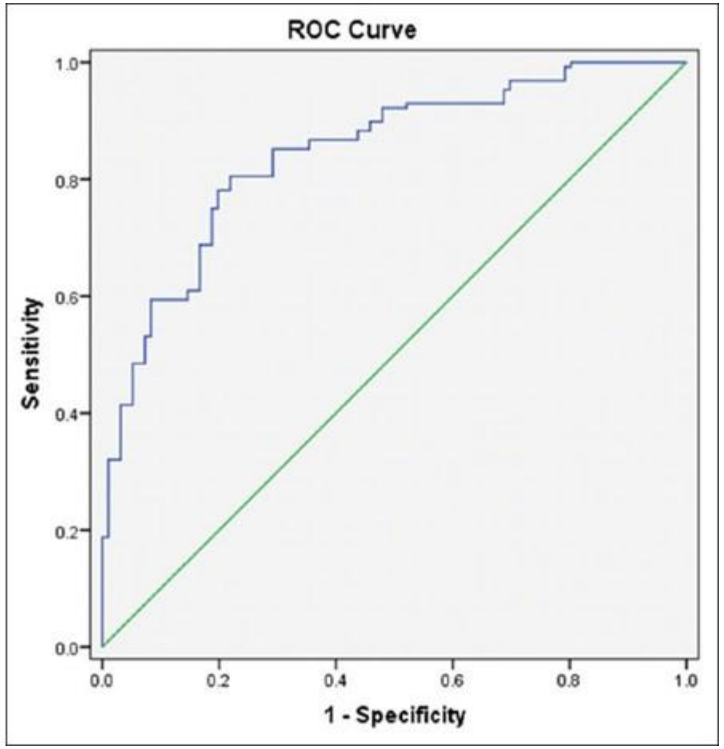
- Overall accuracy $= \frac{(\mathrm{TN}+\mathrm{TP})}{N}$, where $N =$ number of observations
- Overall error rate $= \frac{(\mathrm{FP}+\mathrm{FN})}{N}$
- False negative error rate $= \frac{\mathrm{FN}}{(\mathrm{TP}+\mathrm{FN})}$
- False positive error rate $= \frac{\mathrm{FP}}{(\mathrm{TN}+\mathrm{FP})}$

Receiver Operating Characteristic (ROC) Curve

Receiver Operator Characteristic (ROC) Curve

- True positivity (TP) rate on the *y*-axis, i.e., the proportion of default correctly predicted
- False positive on the *x*-axis, i.e., the proportion non-default incorrectly predicted as default cases
- The curve shows how these two measures vary with different threshold values

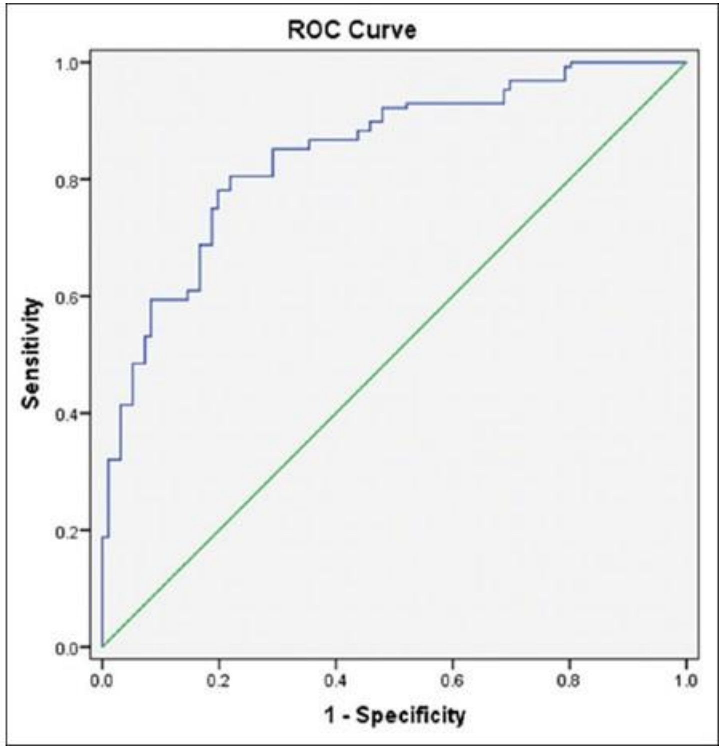
TP Rate



Receiver Operator Characteristic (ROC) Curve

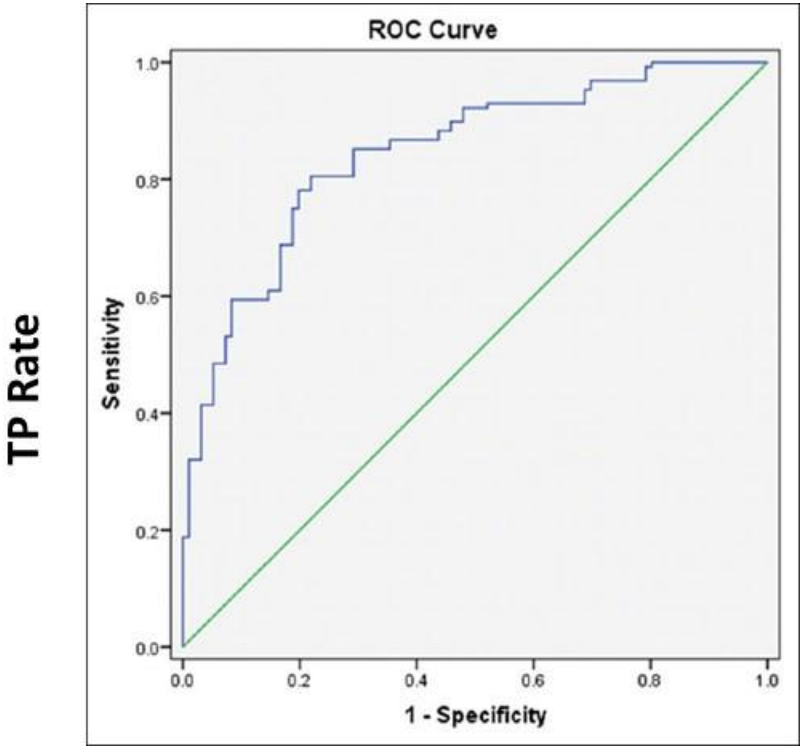
- For t = 1, TP = 0, and FP = 0 → will not be able to predict any default cases but correctly predict all the non-default cases
- For t = 0, TP = 1, and FP = 1 → will be able to correctly predict all the default cases but incorrectly predict all the non-default cases
- As we move from t = 1 to t = 0, different combinations of TP and FP are obtained

TP Rate



Receiver Operator Characteristic (ROC) Curve

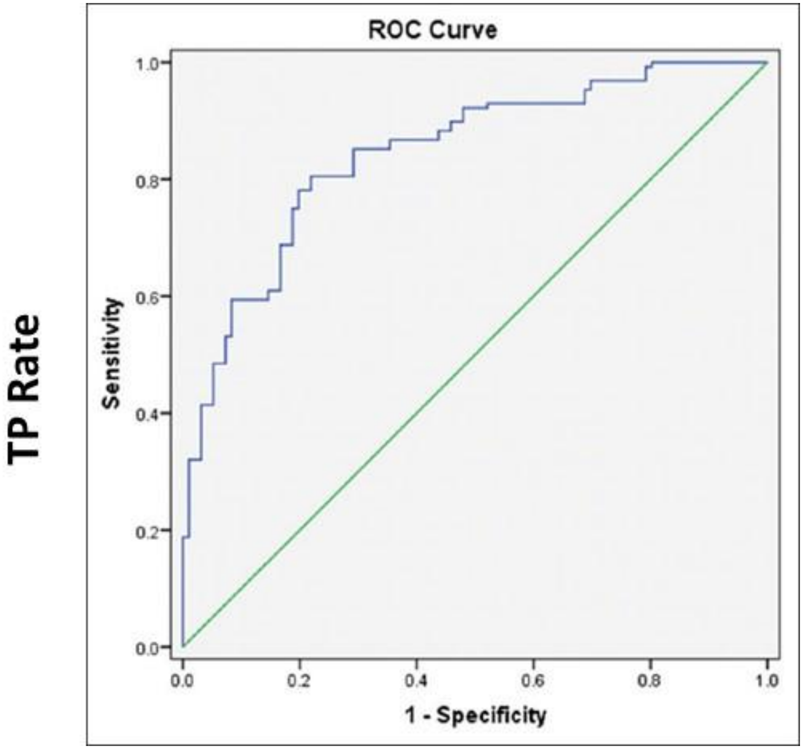
- ROC curve captures all the complete threshold behavior
- High threshold: high specificity and low sensitivity
- Low threshold: low specificity and high sensitivity
- Thus, it is a tradeoff between cost in failing to detect default cases vs. incorrectly considering non-default cases as defaulters



Receiver Operator Characteristic (ROC) Curve

- A 100% score area under the curve will indicate complete accuracy, i.e., all the observations are correctly identified
  TP = 1 and FP = 0

- A 50% score will indicate random guessing, that is, half TP = 0.5 and TN = 0.5 (FP = 0.5)



Parameter Interpretation

Parameter Interpretation

Parameter Interpretation

Unlike LPM, it is incorrect to state that 1 unit increase in x_{2i} will cause 100*β_2% increase in the probability of y_i = 1

- For logit model, we calculate $\frac{dP_i}{dx_{2i}}$; this works out to $\beta_2 F(x_{2i})(1 - F(x_{2i}))$ for the logit model
- So, a 1-unit increase in x_{2i} will increase the probability of y_i= 1 by $\beta_2 F(x_{2i})(1 - F(x_{2i}))$
- Usually, these marginal/incremental impacts are evaluated at mean values

Parameter Interpretation

Example: $P_i(y_i = 1) = \frac{1}{(1+e^{-(\beta_1+\beta_2 x_{2i}+\beta_3 x_{3i}+\cdots+\beta_k x_{ki}+u_i)})}$

- $F(z_i) = \widehat{P_i} = \frac{1}{(1+e^{-(0.1+0.3x_{2i}-0.6x_{3i}+0.9x_{4i})})}$;
- $\beta_1 = 0.1; \beta_2 = 0.3;\ \beta_3 = -0.6;\ \beta_4 = 0.9$
- What is $F(z_i)$? Given $\bar{x}_2 = 1.6$, $\bar{x}_3 = 0.20$, and $\bar{x}_4 = 0.10$?
- Marginal effects of x_{2i} = $\beta_2 F(x_{2i})(1 - F(x_{2i}))$

Parameter Interpretation

Example: $F(z_i) = \widehat{P_i} = \frac{1}{(1+e^{-(0.1+0.3x_{2i}-0.6x_{3i}+0.9x_{4i})})} = \frac{1}{1+e^{-0.55}} = 0.63$

- Thus, a 1-unit increase in x_{2i} will increase the probability of y_i by 0.3*0.63*(1 − 0.63) = 0.07
- Similarly, for x_{3i}, −0.6*0.63*(1 − 0.63), and x_{4i}, 0.9*0.63*(1 − 0.63)
- Sometimes, these are also called marginal effects

Probit Model
Maximum Likelihood Estimation (MLE)
Goodness-of-Fit Measures

Probit Model

- The probit model uses cumulative normal distribution: $F(z_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_i} e^{-(z_i^2)/2}\ dz$
- Model asymptotically touches 0 ($z \rightarrow -\infty$) and 1 ($z \rightarrow \infty$)
- Marginal impact of unit change on an explanatory variable x_{2i} is given as $\beta_2 F(z_i)$, where β_2 is the parameter attached to x_{2i};
$$z_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \cdots + \beta_k x_{ki} + u_i$$
- Both logit and probit models give similar results; differences may occur when data is extremely imbalanced

Maximum Likelihood Estimation (MLE) of Logit/Probit Models

These are non-linear models, hence cannot be estimated with a simple OLS method

- They are estimated with MLE
- In MLE, parameters are chosen to maximize a log-likelihood function
- The log-likelihood function obtains the population estimates that maximize the joint probability of observed sample/sample estimates

Goodness-of-Fit Measures

Conventional R^2 and $\text{adj.} -R^2$ measures do not work well with these models

MLE aims to maximize the log-likelihood function (LLF) and do not minimize RSS

(1) % of y_i values correctly predicted

(2) % of y_i = 1 values correctly predicted + % of y_i = 0 values correctly predicted

Goodness-of-Fit Measures

Conventional R^2 and $\text{adj.}-R^2$ measures do not work well

(3) $\text{Pseudo} - R^2 = 1 - \frac{\text{LLF}}{\text{LLF}_0}$, where LLF is the maximized value of the log-likelihood function for the logit and probit models, and LLF0 is the value of the log-likelihood function for a restricted model

Summary and Concluding Remarks

Summary and Concluding Remarks

- Among supervised learning algorithms, classification algorithm is a very important tool employed in the finance domain for applications such as credit scoring of loan applications
- Classification algorithms are very often implemented through Logit/Probit class of models; these are very simple yet powerful models
- These models account for a number of shortcomings of linear probability models: (a) non-normality and heteroscedasticity of error terms; (b) values of the dependent variable (probability) exceeding the 0–1 range; and (c) diminishing utility of conventional measures of goodness-of-fit (e.g., R^2)

Summary and Concluding Remarks

- Limited dependent variable models (e.g., Logit model) employ cumulative probability functions (e.g., logistic function)
- These models, although non-linear, are very useful for modeling limited dependent variables that are probabilistic in nature
- In the case of the logit model, the logit function is essential the odds ratio
- Since the estimated variable is in the form of probabilities, the thresholding process is needed to convert these probabilities into limited outcomes (e.g., Yes/No)

Summary and Concluding Remarks

- The conventional measures of goodness-of-fit (e.g., R^2) are not very useful for such models
- These measures are evaluated on their ability to accurately classify observations correctly
- For such purposes, a confusion/classification matrix is often employed
- The receiver operator characteristic (ROC) curve provides another useful tool to examine the efficiency of these models, and also facilitates the selection of thresholding values

Summary and Concluding Remarks

- Unlike simple linear models, the parameter estimates are interpreted in a different manner
- Marginal effects are computed to interpret the coefficients and their relationship with the dependent variable
- Other models (e.g., probit model) remain identical in all other aspects, except that a different cumulative probability function is considered (normal distribution in case of probit)
- Since the model is non-linear in nature, OLS cannot be employed for estimation; maximum likelihood method is often employed to estimate these models

Thanks!