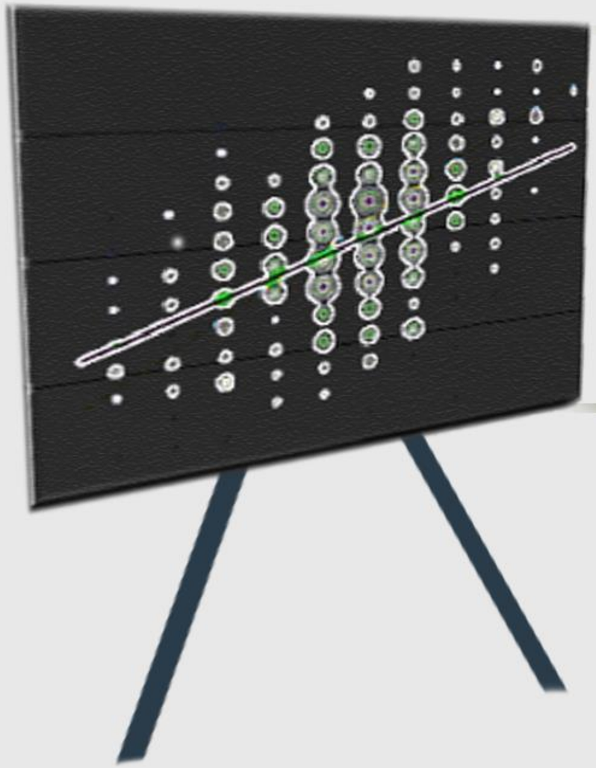


Classification Problems

U. DINESH KUMAR

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Session Outline

- Introduction to classification problems and discrete choice models.
- Introduction to Logistics Regression.
- Logistic function and Logit function.
- Maximum Likelihood Estimator (MLE) for estimation of LR parameters.
- Examples: Challenger Shuttle, German Bank Credit Rating.

Classification Problems

- Classification is an important category of problems in which the decision maker would like to classify the customers into two or more groups.

- Examples of Classification Problems:
  - Customer Churn.
  - Credit Rating (low, high and medium risk)
  - Employee attrition.
  - Fraud (classification of a transaction to fraud/non-fraud)
  - Outcome of any binomial and multinomial experiment.

Always Cheerful

Always Sad

Logistic Regression attempts to classify customers into different categories

Discrete Choice Models

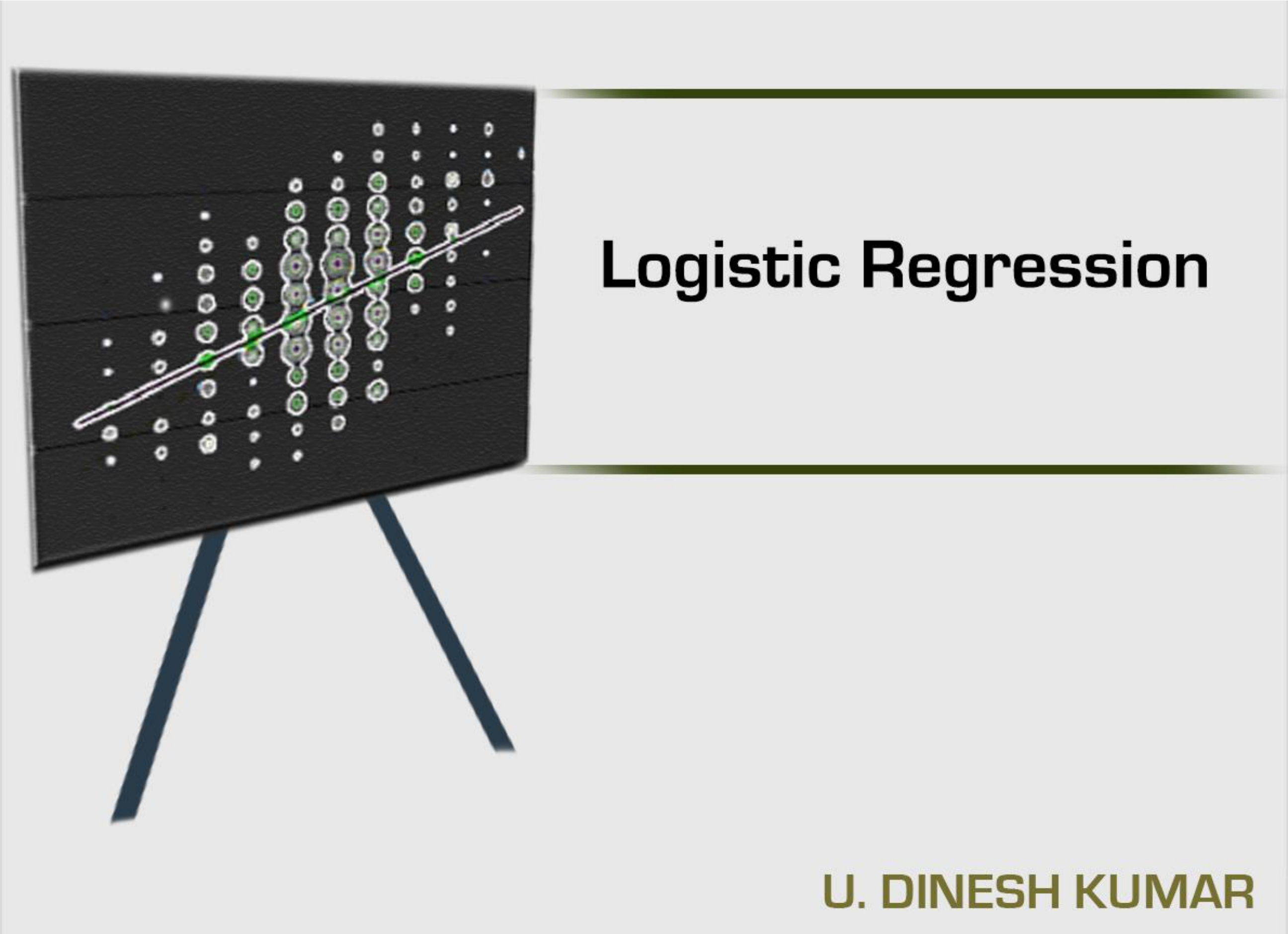
- Problems involving discrete choices available to the decision maker.
- Discrete choice models (in business) examines "which" alternative is chosen by a customer and "why"?
- Most discrete choice models estimate the probability that a customer chooses a particular alternative from several possible alternatives.

Logistic Regression

- Classification Problem
- Discrete Choice Model
- Probability of an event

IIMBX

IIMB
तेजस्वि नावधीतमस्तु



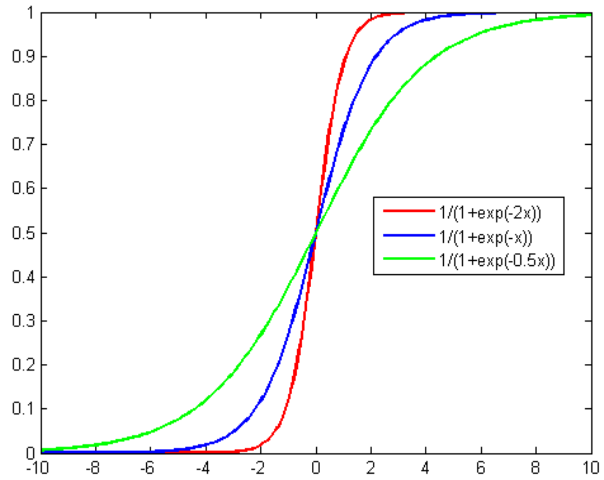
IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Logistic Regression - Introduction

- The name logistic regression emerges from logistic function.

$$P(Y=1)=\pi=\frac{e^{Z}}{1+e^{Z}}$$



- Mathematically, logistic regression attempts to estimate conditional probability of an event.

Logistic Regression

Logistic regression models estimate how probability of an event may be affected by one or more explanatory variables.

Binomial Logistic Regression

- Binomial (or binary) logistic regression is a model in which the dependent variable is dichotomous.

- The independent variables may be of any type.

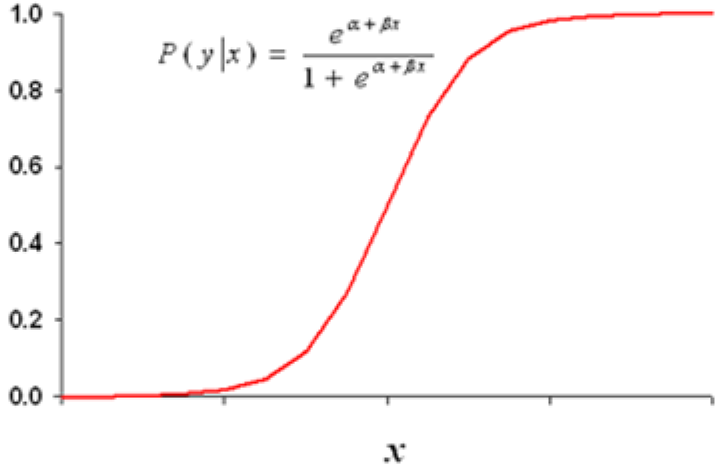
Logistic Function (Sigmoidal function)

$$\pi(z) = \frac{e^z}{1+e^z}$$

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Logistic Regression with one Explanatory Variable

$$P(Y=1 \mid X=x) = \pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$



$\beta = 0$ implies that P(Y | x) is same for each value of x
$\beta > 0$ implies that P(Y | x) is increases as the value of x increases
$\beta < 0$ implies that P(Y | x) is decreases as the value of x increases

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Logit Function

- The Logit function is the logarithmic transformation of the logistic function. It is defined as the natural logarithm of odds.

$$Logit(\pi) = \ln(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1$$

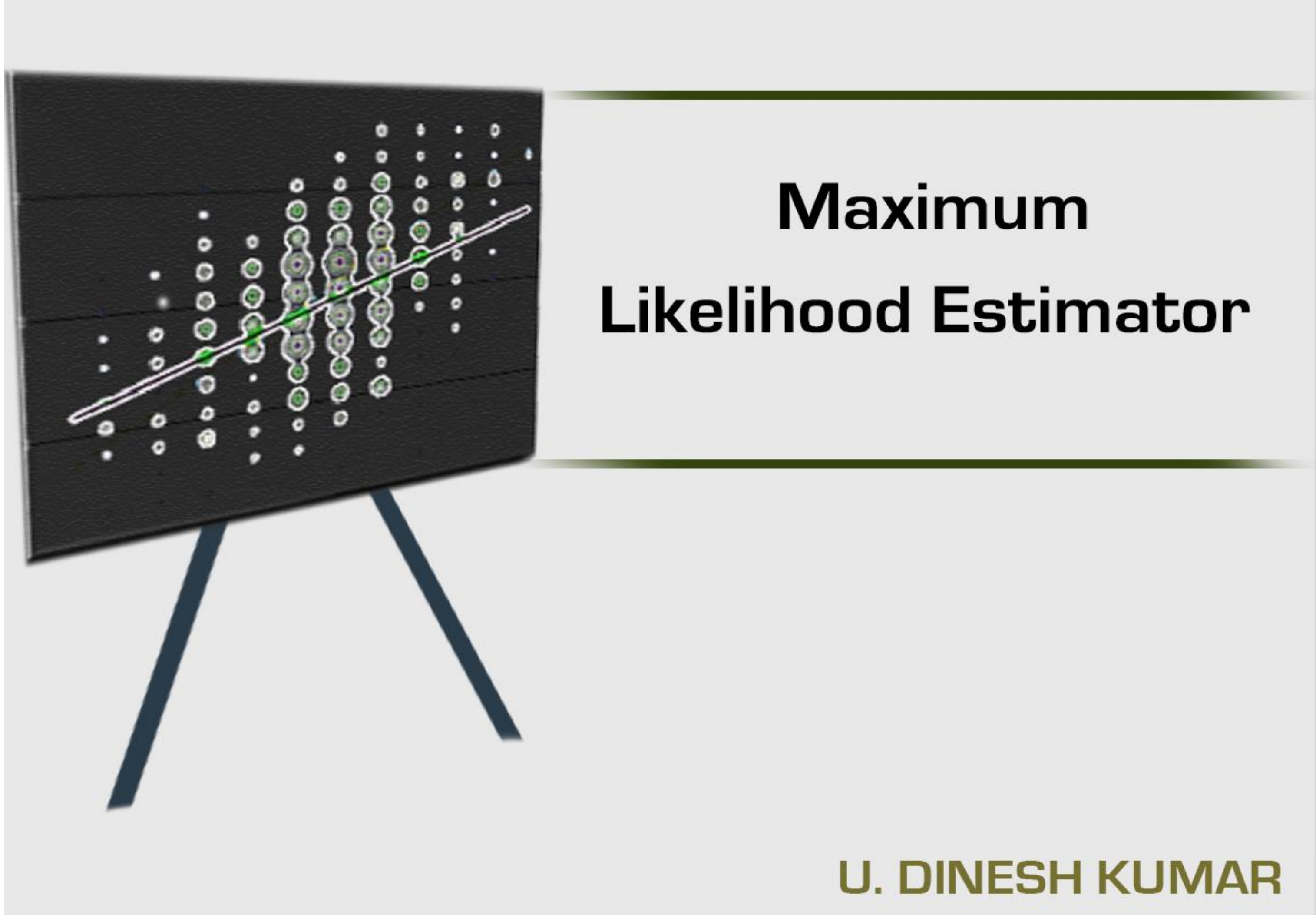
- Logit of a variable π is given by:

$$\frac{\pi}{1-\pi} = \text{odds}$$

Logistic regression

- **More robust**
  - Error terms need not be normal.
  - No requirement for equal variance for error term (homoscedasticity).

- No requirement for linear relationship between dependent and independent variables.


Maximum
Likelihood Estimator
U. DINESH KUMAR

Estimation of parameters in Logistic Regression

- Estimation of parameters in logistic regression is carried out using Maximum Likelihood Estimation (MLE) technique.

- No closed form solution exists for estimation of regression parameters of logistic regression.

Maximum Likelihood Estimator (MLE)

- MLE is a statistical model for estimating model parameters of a function.
- For a given data set, the MLE chooses the values of model parameters that makes the data "more likely", than other parameter values.

Likelihood Function

- Likelihood function L($\boldsymbol{\beta}$) represents the joint probability or likelihood of observing the data that have been collected.

- MLE chooses that estimator of the set of unknown parameters $\boldsymbol{\beta}$ which maximizes the likelihood function L($\boldsymbol{\beta}$).

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Maximum Likelihood Estimator

- Assume that $x_1, x_2, \ldots, x_n$ are some sample observations of a distribution $f(x, \theta)$, where θ is an unknown parameter.

- The likelihood function is $L(\theta) = f(x_1, x_2, \ldots, x_n, \theta)$ which is the joint probability density function of the sample.

- The value of θ, θ^*, which maximizes $L(\theta)$ is called the maximum likelihood estimator of θ.

Example: Exponential Distribution

- Let x1, x2, …, xn be the sample observation that follows exponential distribution with parameter θ. That is:

$$f(x,\theta) = \theta e^{-\theta x}$$

- The likelihood function is given by (assuming independence):

$$L(x,\theta) = f(x_1,\theta) f(x_2,\theta) \bullet\bullet\bullet f(x_n,\theta)$$

$$= \theta e^{-\theta x_1} \times \theta e^{-\theta x_1} \times \bullet\bullet\bullet \theta e^{-\theta x_n} = \theta^n e^{-\theta \sum_{i=1}^{n} x_i}$$

IIMBX

IIMB तेजस्वि नावधीतमस्तु

Log-likelihood function

- Log-likelihood function is given by:

$$Ln(L(x,\theta)) = n\ln\theta - \theta\sum_{i=1}^{n} x_i$$

- The optimal θ, θ^* is given by:

$$\frac{d}{d\theta}(\ln(L(x,\theta)) = \frac{n}{\theta} - \sum_{i=1}^{n} x_i = 0$$

$$\theta^* = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}}$$

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Likelihood function for Binary Logistic Function

- Probability density function for binary logistic regression is given by:

$$f(y_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$L(\beta) = f(y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

$$\ln(L(\beta)) = \sum_{i=1}^{n} y_i \ln[\pi(x_i)] + \sum_{i=1}^{n}(1-y_i)[\ln(1-\pi_i(x_i))]$$

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Likelihood function for Binary Logistic Function

$$\ln\left[L(\beta)\right] = \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) - \sum_{i=1}^{n} \ln(1 + \exp(\beta_0 + \beta_1 x_i))$$

Estimation of LR parameters

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = 0$$

$$\frac{\partial \ln(L(\beta_0, \beta_1))}{\partial \beta_1} = \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \frac{x_i \exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} = 0$$

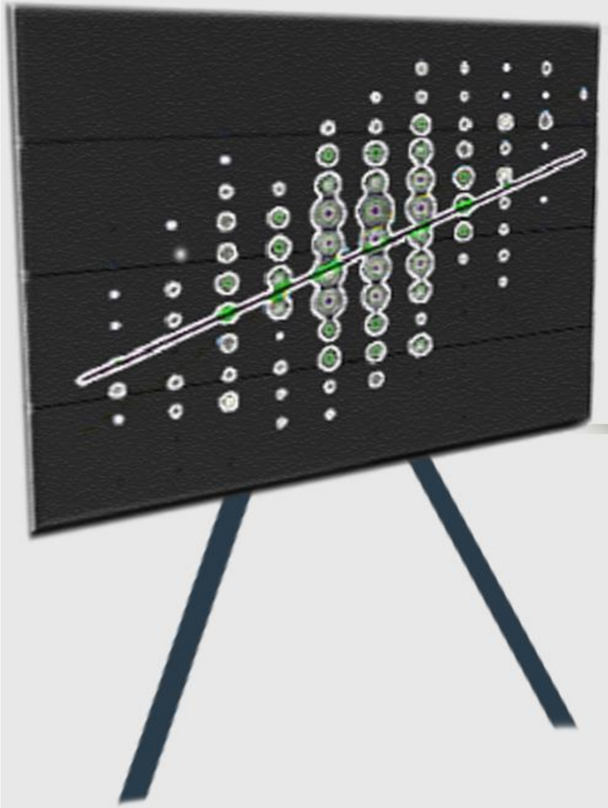
The above system of equations are solved iteratively to estimate β_0 and β_1

Limitations of MLE

- Maximum likelihood estimator may not be unique or may not exist.

- Closed form solution may not exist for many cases, one may have to use iterative procedure to estimate the parameter values.

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB



Challenger Crash Problem

U. DINESH KUMAR

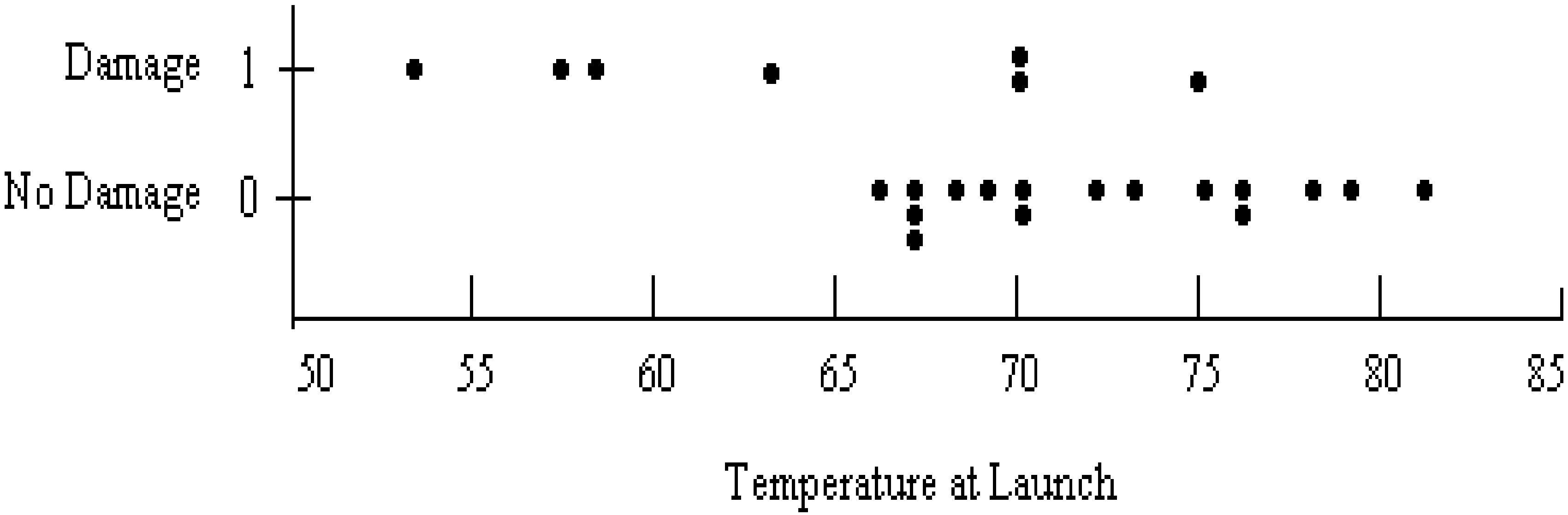
IIMBX

Challenger Data

Flt	Temp	Damage
STS-1	66	No
STS-2	70	Yes
STS-3	69	No
STS-4	80	No
STS-5	68	No
STS-6	67	No
STS-7	72	No
STS-8	73	No
STS-9	70	No
STS-41B	57	Yes
STS-41C	63	Yes
STS-41D	70	Yes

Flt	Temp	Damage
STS-41G	78	No
STS-51-A	67	No
STS-51-C	53	Yes
STS-51-D	67	No
STS-51-B	75	No
STS-51-G	70	No
STS-51-F	81	No
STS-51-I	76	No
STS-51-J	79	No
STS-61-A	75	Yes
STS-61-B	76	No
STS-61-C	58	Yes

Challenger launch temperature vs damage data



Logistic Regression of challenger data

- Let:

  $Y_i = 0$ denote no damage

  $Y_i = 1$ denote damage to the O-ring

$$P(Y_i = 1) = \Pi_i \text{ and } P(Y_i = 0) = 1 - \Pi_i.$$

We predict $P(Y_i = 1 | x_i)$, x_i = launch temperature

Logistic Regression using SPSS

- **Dependent variable:**
  - In Binary logistic regression, the dependent variable can take only two values.
  - In multinomial logistic regression, the dependent variable can take two or more values (but not continuous).

- **Covariate:**
  - All independent (predictor) variables are entered as covariates.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	LaunchTemperature	-.236	.107	4.832	1	.028	.790
	Constant	15.297	7.329	4.357	1	.037	4398676

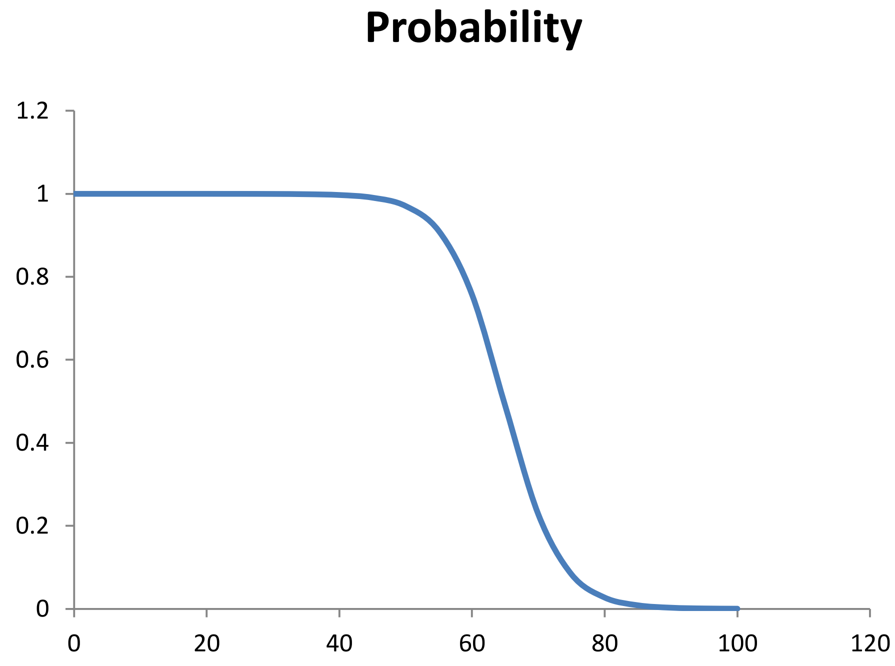
a. Variable(s) entered on step 1: LaunchTemperature.

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = 15.297 - 0.236 X_i$$

$$P(Y_i = 1) = \frac{e^{15.297 - 0.236 X_i}}{1 + e^{15.297 - 0.236 X_i}}$$

Challenger: Probability of failure estimate

$$\pi_i = \frac{e^{15.297 - 0.236 X_i}}{1 + e^{15.297 - 0.236 X_i}}$$



IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Classification table from SPSS

Classification Table[a]

Observed			Predicted: Damage to O-ring 0	Predicted: Damage to O-ring 1	Percentage Correct
Step 1	Damage to O-ring	0	17	0	100.0
		1	3	4	57.1
	Overall Percentage				87.5

a. The cut value is .500

Classification Table[a]

Observed			Predicted: Damage to O-ring 0	Predicted: Damage to O-ring 1	Percentage Correct
Step 1	Damage to O-ring	0	9	8	52.9
		1	1	6	85.7
	Overall Percentage				62.5

a. The cut value is .200

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Accuracy Paradox

- Assume an example of insurance fraud. Past data has revealed that out of 1000 claims in the past, 950 are true claims and 50 are fraudulent claims.

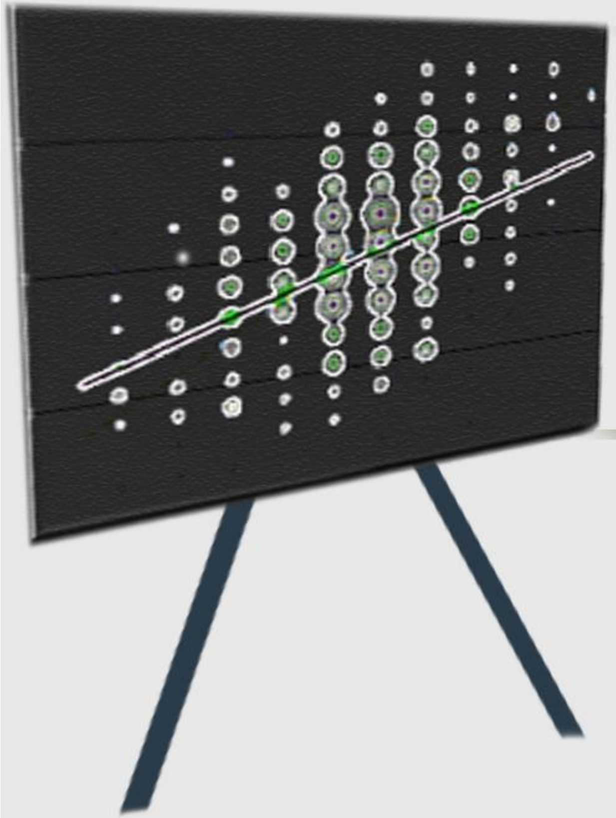
- The classification table using a logistic regression model is given below:

Observed	Predicted		% accuracy
	0	1	
0	900	50	94.73%
1	5	45	90.00%

The overall accuracy is 94.5%. Not predicting fraud will give 95% accuracy!

IIMBX

IIMB
तेजस्वि नावधीतमस्तु



Interpretation of Logistic Regression Parameters

U. DINESH KUMAR

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

ODDS and ODDS RATIO

- **ODDS:** Ratio of two probability values.

$$odds = \frac{\pi}{1-\pi}$$

- **ODDS Ratio:** Ratio of two odds.

ODDS RATIO

Assume that X is an independent variable (covariate). The odds ratio, OR, is defined as the ratio of the odds for X = 1 to the odds for X = 0. The odds ratio is given by:

$$OR = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)}$$

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Interpretation of Beta Coefficient in LR

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 \qquad (1)$$

$$\text{For } x = 0 \quad \ln\left(\frac{\pi(0)}{1-\pi(0)}\right) = \beta_0 \qquad (2)$$

$$\text{For } x = 1 \quad \ln\left(\frac{\pi(1)}{1-\pi(1)}\right) = \beta_0 + \beta_1 \qquad (3)$$

$$\beta_1 = \ln\left(\frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}\right)$$

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Interpretation of LR coefficients

$$\beta_1 = \ln\left(\frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x+1))}\right) = \text{Change in ln odds ratio}$$

$$e^{\beta_1} = \frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x+1))} = \text{Change in odds ratio}$$

IIMBX

IIMB तेजस्वि नावधीतमस्तु

Odds Ratio for Binary Logistic Regression

$$OR = \frac{\pi(1)/1-\pi(1)}{\pi(0)/1-\pi(0)} = e^{\beta_1}$$

If OR = 2, then the event is twice likely to occur when X = 1 compared to X = 0.

Odds ratio approximates the relative risk.

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

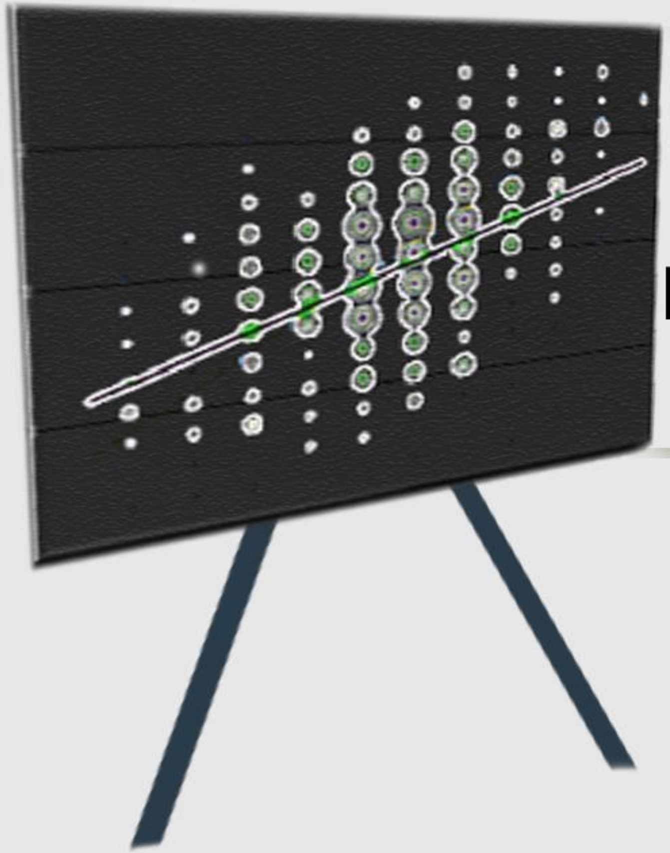
Interpretation of LR coefficients

- β_1 is the change in log-odds ratio for unit change in the explanatory variable.

- β_1 is the change in odds ratio by a factor $\exp(\beta_1)$.

IIMBX

IIMB
तेजस्वि नावधीतमस्तु



Logistic Regression Inference & Diagnostics

U. DINESH KUMAR

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Session Outline

- Measuring the fitness of Logistic Regression Model.
- Testing individual regression parameters (Wald's test).
- Omnibus test for overall model fitness
- Hosmer-Lemeshow Goodness of fit test.
- R^2 in Logistic Regression.
- Confidence Intervals for parameters and probabilities.

Wald Test

- Wald test is used to check the significance of individual explanatory variables (similar to t-statistic in linear regression).

- Wald test statistic is given by:

$$W = \left(\frac{\hat{\beta}_i}{SE(\hat{\beta}_i)} \right)^2$$

W is a chi-square statistic

Wald test hypothesis

- Null Hypothesis H0: $\beta_i = 0$
- Alternative Hypothesis H1: $\beta_i \neq 0$

Wald Test – Challenger Data

The explanatory variable is significant, since the p value is less than 0.05

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	LaunchTemperature	-.236	.107	4.832	1	.028	.790	.640	.975
	Constant	15.297	7.329	4.357	1	.037	4398676.183		

a. Variable(s) entered on step 1: LaunchTemperature.

Wald Test – Challenger Data

For significant variables, the CI for Exp(β) will not contain 1

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	LaunchTemperature	-.236	.107	4.832	1	.028	.790	.640	.975
	Constant	15.297	7.329	4.357	1	.037	4398676.183		

a. Variable(s) entered on step 1: LaunchTemperature.

IIMBX

IIMB तेजस्वि नावधीतमस्तु

Model Chi-Square

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	8.603	1	.003
	Block	8.603	1	.003
	Model	8.603	1	.003

Omnibus test:

H_0: $\beta_1 = \beta_2 = \ldots = \beta_k = 0$

H1: Not all βs are zero

IIMBX

IIMB तेजस्वि नावधीतमस्तु

Hosmer-Lemeshow Goodness of Fit

- Test for overall fitness of the model for a binary logistic regression (similar to chi-square goodness of fit test).

- The observations are grouped into 10 groups based on their predicted probabilities.

Hosmer-Lemeshow Test Statistic

- Hosmer-Lemeshow Test Statistic is given by:

$$C = \sum_{k=1}^{g} \frac{(O_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)}$$

g = Number of groups

n_k = Number of observations in each group

O_k = Sum of the values for k^{th} group.

$\bar{\pi}_k$ = The avearge π in k^{th} group

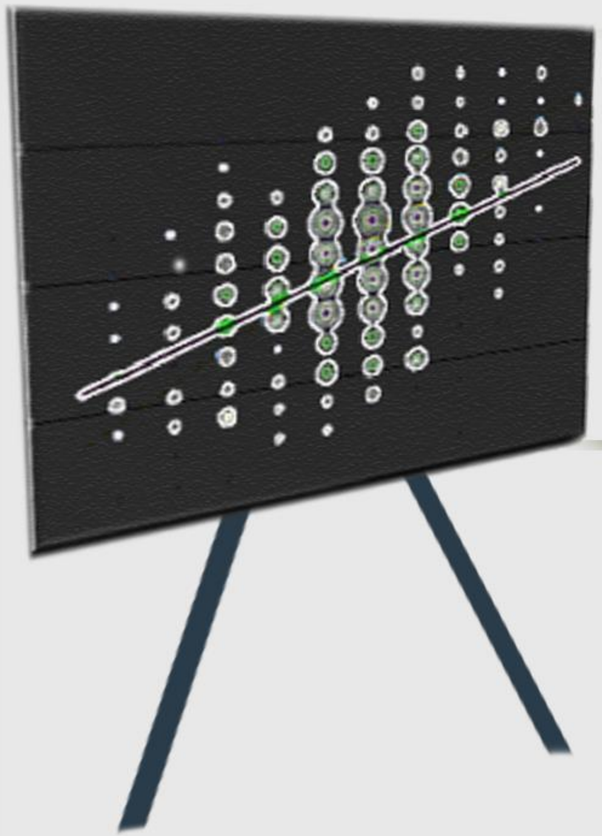
H-L test for Challenger Data

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	9.396	8	.310

Since P (=0.310) is more than 0.05, we accept the null hypothesis that there is no difference between the predicted and observed frequencies (accept the model)

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB



Model
Accuracy Measures

U. DINESH KUMAR

IIMBX

Classification Table[a]

Observed			Predicted: Damage to O-ring 0	Predicted: Damage to O-ring 1	Percentage Correct
Step 1	Damage to O-ring	0	17	0	100.0
		1	3	4	57.1
	Overall Percentage				87.5

a. The cut value is .500

Classification Table[a]

Observed			Predicted: Damage to O-ring 0	Predicted: Damage to O-ring 1	Percentage Correct
Step 1	Damage to O-ring	0	9	8	52.9
		1	1	6	85.7
	Overall Percentage				62.5

a. The cut value is .200

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Classification Table

Prediction (Classification)	Observed	
	1 (Positive) 7	0 (Negative) 17
1 (Positive)	4 [True Positive] TP	0 [False Positive] FP
0 (Negative)	3 [False Negative] FN	17 [True Negative] TN

$$\text{Sensitivity} = \left(\frac{TP}{TP+FN}\right) = \left(\frac{4}{7}\right) = 57.1$$

$$Specificity = \left(\frac{TN}{TN+FP}\right) = \left(\frac{17}{17}\right) = 100$$

IIMBX

Sensitivity & Specificity

$$\text{Sensitivity} = \frac{\text{No of true positives}}{\text{Number of true positives} + \text{Number of false negatives}}$$

Sensitivity is the probability that the predicted value of y = 1 given that the observed value is 1.

$$\text{Specificity} = \frac{\text{No of true negatives}}{\text{Number of true negatives} + \text{Number of false positives}}$$

Specificity is the probability that the predicted value of y = 0 given that the observed value is 0.

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Receiver Operating Characteristics (ROC) Curve

- ROC curve plots the true positive ratio (right positive classification) against the false positive ratio (1- specificity) and compares it with random classification.

- The higher the area under the ROC curve, the better the prediction ability.

IIMBX

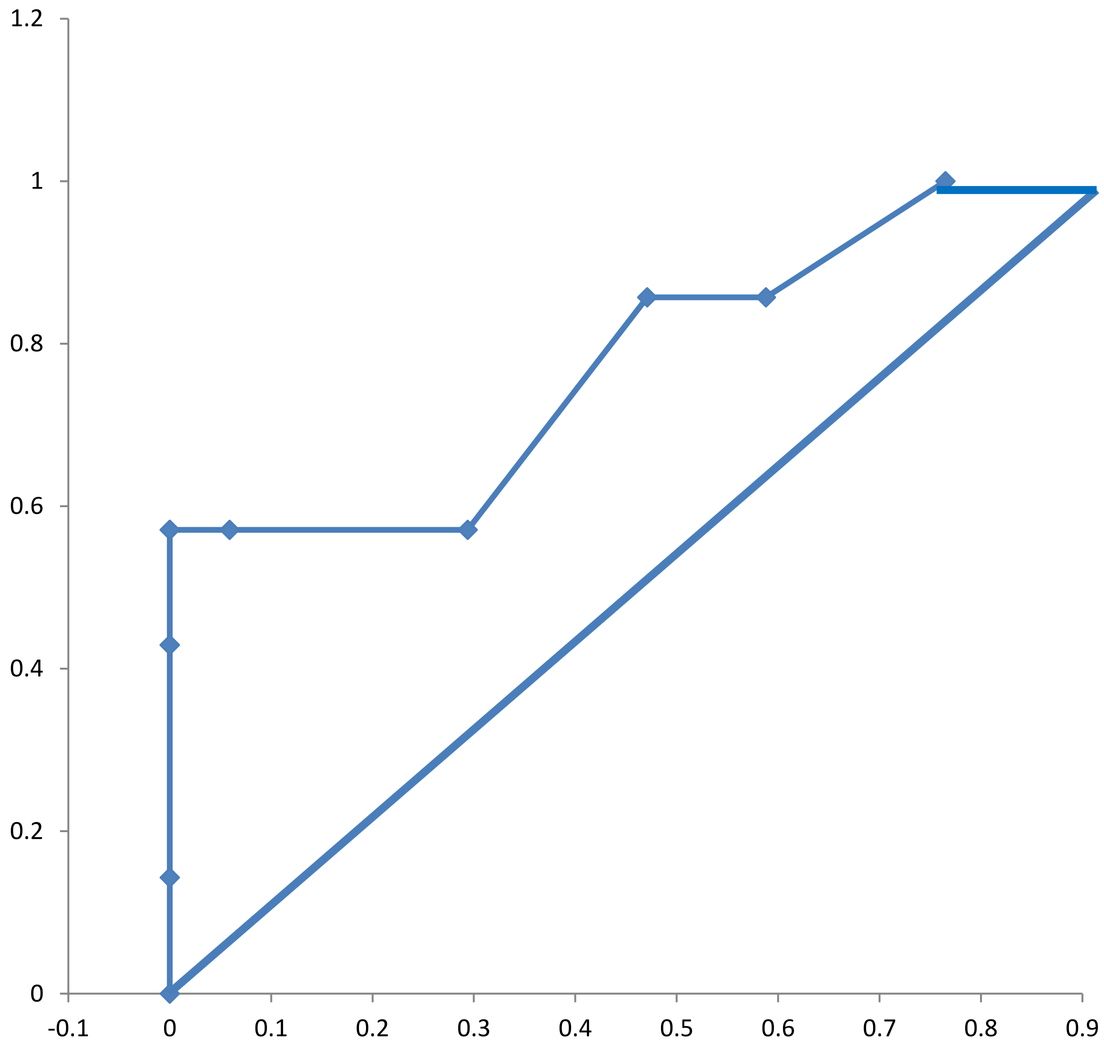
IIMB
तेजस्वि नावधीतमस्तु

Challenger Example: Sensitivity Vs 1-Specificity (True positive Vs False positive)

Cut off Value	Sensitivity	Specificity	1-specificity
0.05	1	0.235	0.765
0.1	0.857	0.412	0.588
0.2	0.857	0.529	0.471
0.3	0.571	0.706	0.294
0.4	0.571	0.941	0.059
0.5	0.571	1	0
0.6	0.571	1	0
0.7	0.429	1	0
0.8	0.429	1	0
0.9	0.143	1	0
0.95	0	1	0

Predictive Analytics : QM901.1x
Prof U Dinesh Kumar, IIMB

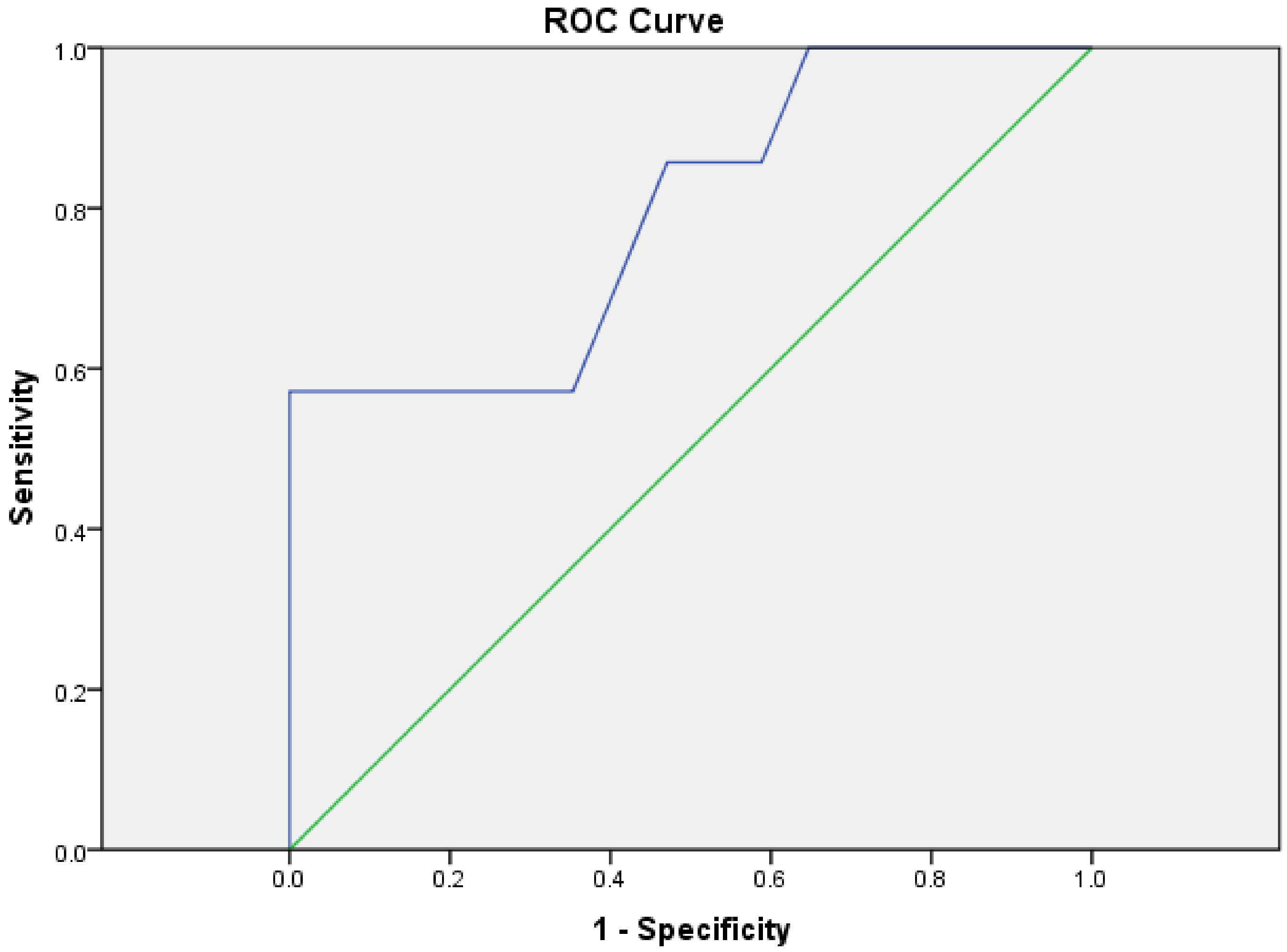
ROC Curve – Challenger Example



IIMBX

IIMB
तेजस्वि नावधीतमस्तु

ROC Curve



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable (s):Predicted probability

Area
.794

The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

IIMBX

IIMB तेजस्वि नावधीतमस्तु

Area Under the ROC Curve

- Area under the ROC curve is interpreted as the probability that the model will rank a randomly chosen positive higher than randomly chosen negative.

- If n1 is the number of positives (1s) and n2 is the number of negatives (0s), then the area under the ROC curve is the proportion of cases in all possible combinations of (n1, n2) such that n1 will have higher probability than n2.

ROC Curve

- General rule for acceptance of the model:

- If the area under ROC is:

0.5 $\Rightarrow$ No discrimination

0.7 $\leq$ ROC area < 0.8 $\Rightarrow$ Acceptable discrimination

0.8 $\leq$ ROC area < 0.9 $\Rightarrow$ Excellent discrimination

ROC area $\geq$ 0.9 $\Rightarrow$ Outstanding discrimination

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Gini Coefficient

- Gini coefficient measures individual impact of the an explanatory variable.
- Gini coefficient = 2 AUC – 1
- AUC = Area under the ROC Curve

Optimal Cut-off probabilities

- Using classification plots.
- Youden’s Index.
- Cost based optimization.

Classification Plots

```
Step number: 1

             Observed Groups and Predicted Probabilities

       4 +                      1                                                                             +
         I                      1                                                                             I
         I                      1                                                                             I
F        I                      1                                                                             I
R      3 +                      1              0                                                              +
E        I                      1              0                                                              I
Q        I                      1              0                                                              I
U        I                      1              0                                                              I
E      2 +  0   0 1             0              0                                                              +
N        I  0   0 1             0              0                                                              I
C        I  0   0 1             0              0                                                              I
Y        I  0   0 1             0              0                                                              I
       1 +  000 0 0   0  0      0    0    0    0     0                1                      1  1       1     +
         I  000 0 0   0  0      0    0    0    0     0                1                      1  1       1     I
         I  000 0 0   0  0      0    0    0    0     0                1                      1  1       1     I
         I  000 0 0   0  0      0    0    0    0     0                1                      1  1       1     I
Predicted ---------+---------+---------+---------+---------+---------+---------+---------+---------+----------
  Prob:   0        .1        .2        .3        .4        .5        .6        .7        .8        .9         1
  Group:  0000000000000000000000000000000000000000000000000011111111111111111111111111111111111111111111111111
```

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Youden's Index

- Youden's index is a measures for diagnostic accuracy.

- Youden's index is calculated by deducting 1 from the sum of test's sensitivity and specificity.

$$\text{Youden's Index } J(p) = \left[\text{Sensitivity}(p) + \text{specificity}(p) - 1\right]$$

IIMBX

IIMB
तेजस्वि नावधीतमस्तु

Cost based Model for Optimal Cut-off

Observed	Predicted	
	0	1
0	P_{00}	P_{01}
1	P_{10}	P_{11}

R_{00} = Cost of classifying 0 as 0
C_{01} = Cost of classifying 0 as 1
C_{10} = Cost of classifying 1 as 0
R_{11} = Cost of classifying 1 as 1

$$\text{Optimal cut-off}$$
$$\underset{p}{\mathrm{Min}}\left[P_{00}R_{00} + P_{01}C_{01} + P_{10}C_{10} + P_{11}R_{11}\right]$$

IIMBX

IIMB
तेजस्वि नावधीतमस्तु