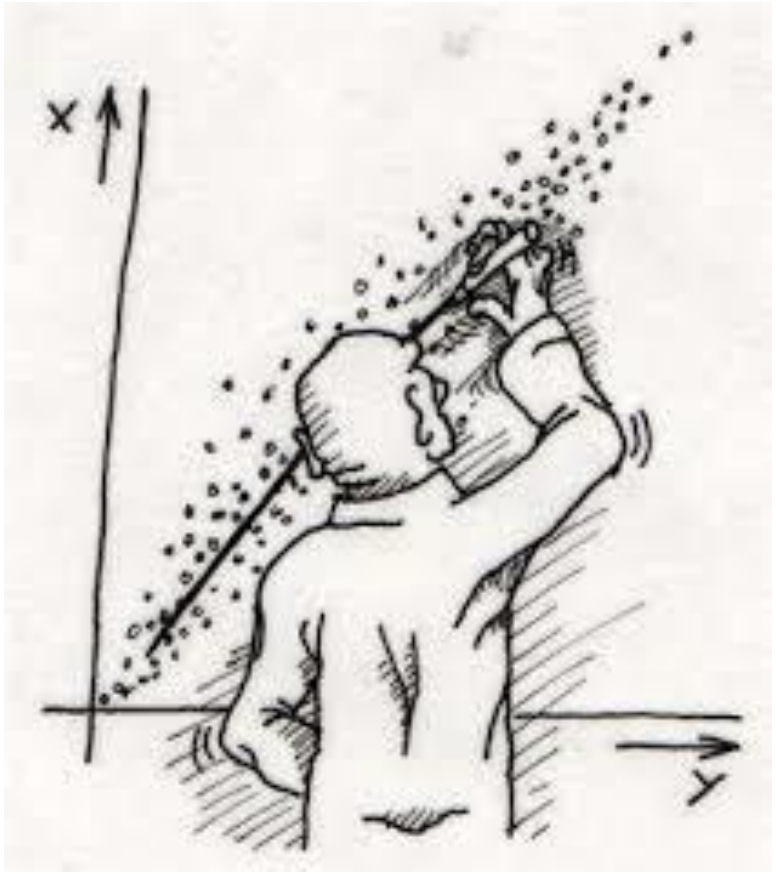


Logistic Regression Modelling

Why do we need Logistic Regression?

We deal with data which:

- Violates the assumption of Linear Regression!
- Assumption says that the residuals should be normally distributed.
- The error term can only take on two values, hence it's impossible for it to have a normal distribution.
- Violates the assumption of Homoscedasticity!
- Homoscedasticity describes a situation in which the error term is the same across all values of the independent variables

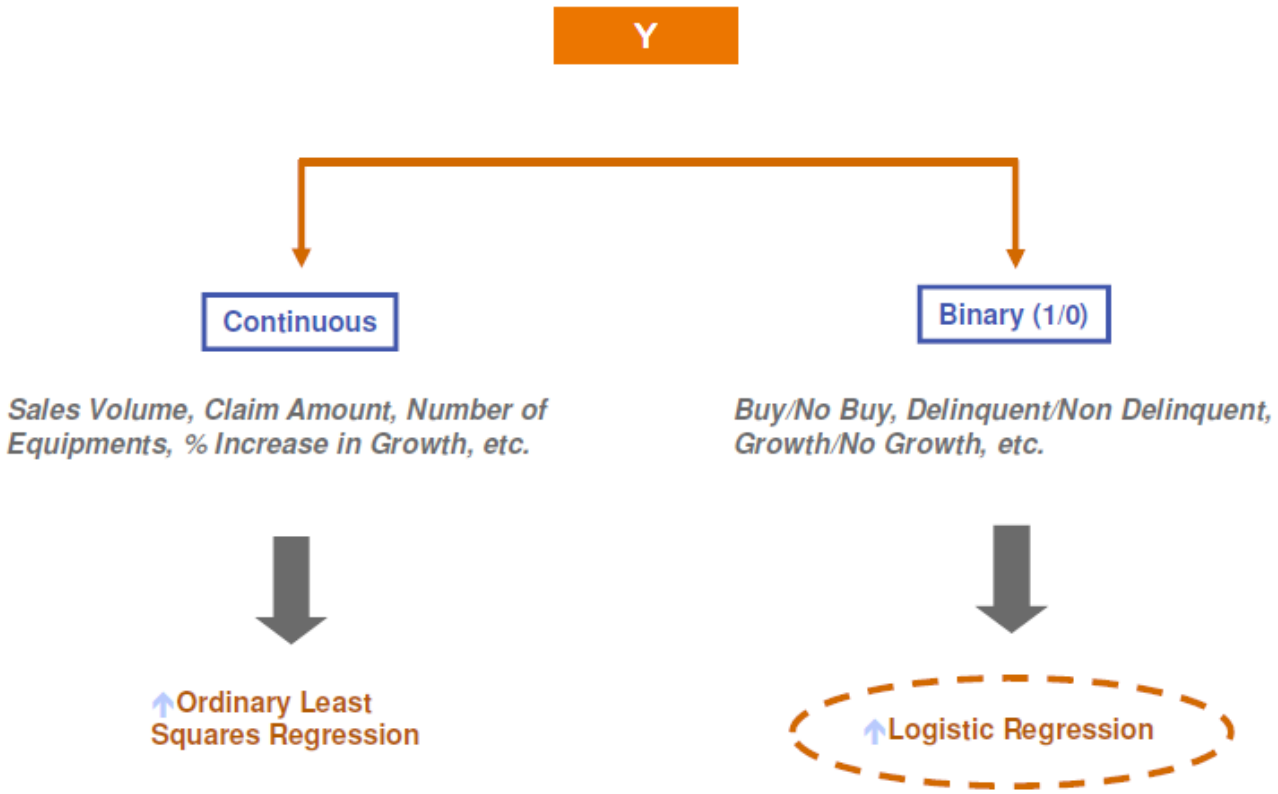
Why do we need Logistic Regression?

Always Cheerful

Always Sad

Logistic Regression attempts to classify customers into different categories

Types of Regression



Logistic Regression

Model equation	$P_i = Prob(Y_i=0) = \frac{e^{Li}}{(1+e^{Li})}$ Where, $L_i = a+b_1X_{1i}+b_2X_{2i}+\ldots+b_pX_{pi}$
Assumption	Y_i and Y_j independent for all $i \neq j$
Parameters to be Estimated	a, b_1, b_2, ..., b_p
Method of Estimation	Maximum Likelihood

Logistic Regression: Introduction

- The name logistic regression emerges from **logistic function**.
- Mathematically, logistic regression attempts to estimate conditional probability of an event.

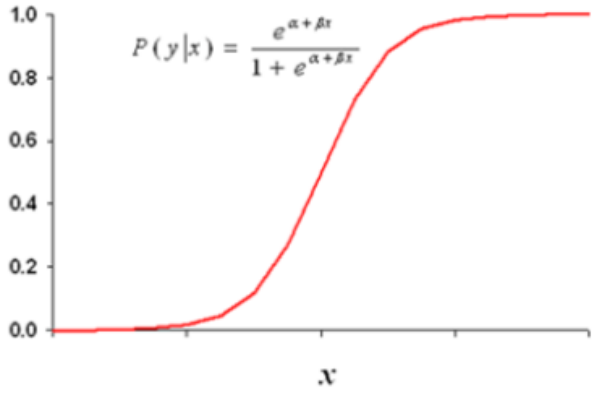
$$P(Y=1)=\pi=\frac{e^{Z}}{1+e^{Z}}$$

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

Logistic Regression: Introduction

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

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



$\beta = 0$ 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

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:

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

The odds ratio, OR, is defined as the ratio of the odds for X = 1 to the odds for X = 0.

Odds Ratio= P(Y=1)/(1-P(Y=1))

Odds and Odds Ratio (Example)

Professional School
ISO 9001 : 2000 Organisation

Suppose that seven out of 10 males are admitted to an Business school while three of 10 females are admitted.

The probabilities for admitting a male are, $p = 7/10 = .7$ $q = 1 - .7 = .3$

If you are male, the probability of being admitted is 0.7 and the probability of not being admitted is 0.3.

Here are the same probabilities for females, $p = 3/10 = .3$ $q = 1 - .3 = .7$

If you are female it is just the opposite, the probability of being admitted is 0.3 and the probability of not being admitted is 0.7.

Now we can use the probabilities to compute the odds of admission for both males and females, odds(male) = .7/.3 = 2.33333
odds(female) = .3/.7 = .42857

Next, we compute the odds ratio for admission, OR = 2.3333/.42857 = 5.44
Thus, for a male, the odds of being admitted are 5.44 times larger than the odds for a female being admitted.

Logistic Regression: Methodology

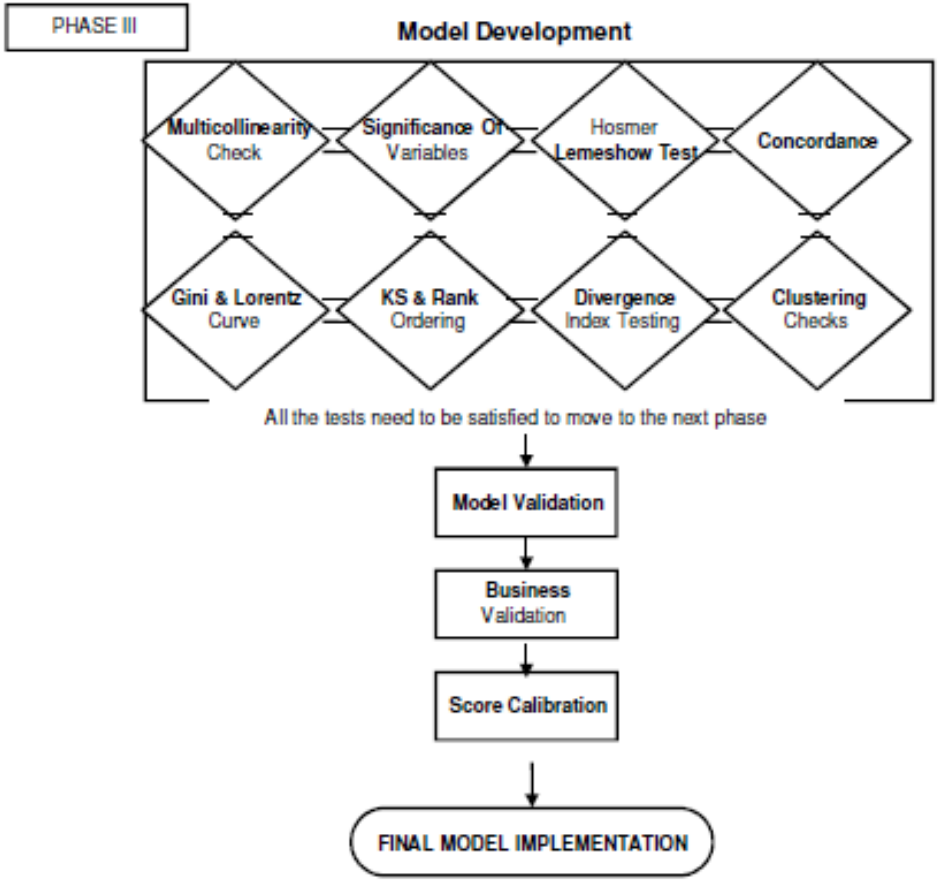
Maximum Likelihood Estimation Technique

- Estimation of parameters in logistic regression is carried out using **Maximum Likelihood Estimation (MLE) technique.**
- MLE is a statistical model for estimating model parameters of a function.
- For a given dataset, the MLE chooses the values of model parameters that makes the data "**more likely**", than other parameter values.

Maximum Likelihood Estimator

- Assume that x1,x2,…,xn are some sample observations of a distribution f(x,α),where α is an unknown parameter.
- The likelihood function is L(α)=f(x1,x2,…,xn, α) which is the joint probability density function of the sample.
- The value of α, α*,which maximizes L(α) is called the maximum likelihood estimator of α.

Logistic Procedure



Logistic Regression: Interpretation of Model Statistics

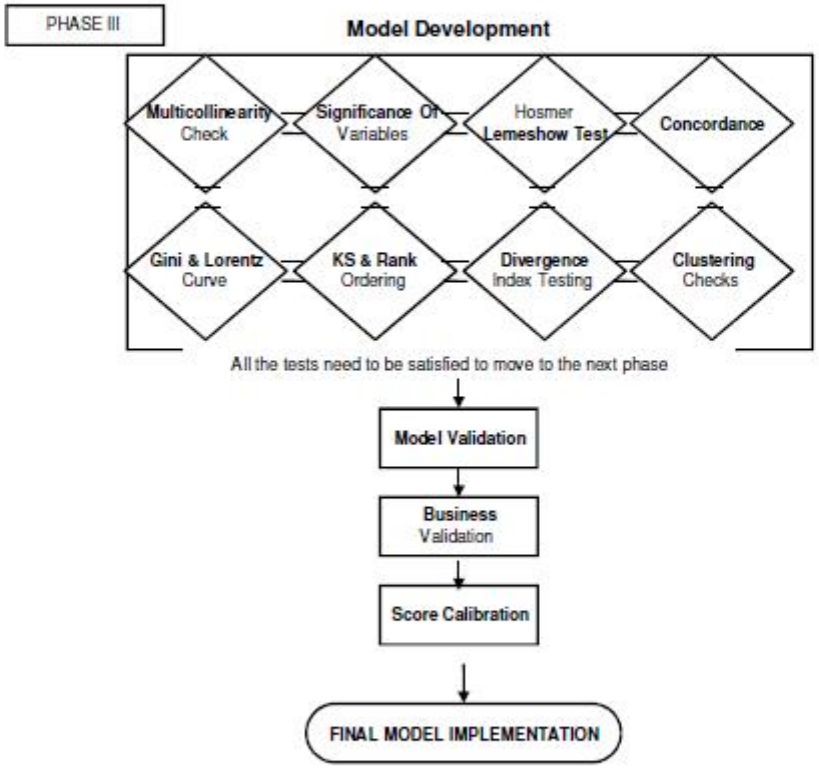
Interpretation of LR coefficients

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

exp (β_1).= Change in odds ratio

Logistic Regression:
Procedure, Inference and Diagnostics

Logistic Procedure



Logistic Procedure - Multocollinearity

What is Multicollinearity ?

Multicollinearity is a phenomenon when there is a linear relationship between a set of variable

Why is Multicollinearity a problem ?

Multicollinearity affects the parameter estimates making them unreliable.

How to detect Multicollinearity ?

Variance Inflation Factor (VIF) = $1/(1- R^2)$

How to remove Multicollinearity ?

- Look into Variance proportions table for the row with highest CI
- Identify variables with highest factor loadings in the row
- Drop the variable which is least significant

VIF>1.75 => Multicollinearity

Logistic Procedure – Variables Significance

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	0.6010	0.1423	17.8279	<.0001
d1_cons_cd_grt_1	1	1.0016	0.1326	57.0378	<.0001
d3_max_cdlevel	1	-1.0768	0.2338	21.2164	<.0001
d1_Payment_method	1	1.6529	0.1449	130.1012	<.0001
d3_OTB_jun04	1	0.6993	0.1176	35.3416	<.0001
d2_crlimit_may04	1	0.3627	0.1156	9.8523	0.0017
d2_avg_pay_bal	1	0.4720	0.1084	18.9700	<.0001
d2_max_payment	1	0.2424	0.1110	4.7691	0.0290
d4_age	1	0.4141	0.1094	14.3331	0.0002

Chi – Square value for each explanatory variable – the chi-square value indicates the level of significance, i.e – the impact of independent (explanatory) variable on the dependent variable.

The p-value cut-off should be decided in discussion with the business. Ideally the p-value<0.0001. However in case of smaller population size p-value could be <0.05 or p-value<0.1.

Logistic Procedure – Concordance

Association of Predicted Probabilities and Observed Responses

Percent Concordant	79.0
Percent Discordant	19.1
Percent Tied	1.9
Pairs	3627468

- ✓ Concordance is used to assess how well scorecards are separating the good and bad accounts in the development sample.
- ✓ The higher is the concordance, the larger is the separation of scores between good and bad accounts.
- ✓ The concordance ratio is a non-negative number, which theoretically may lie between 0 and 1.

Concordance Determination:

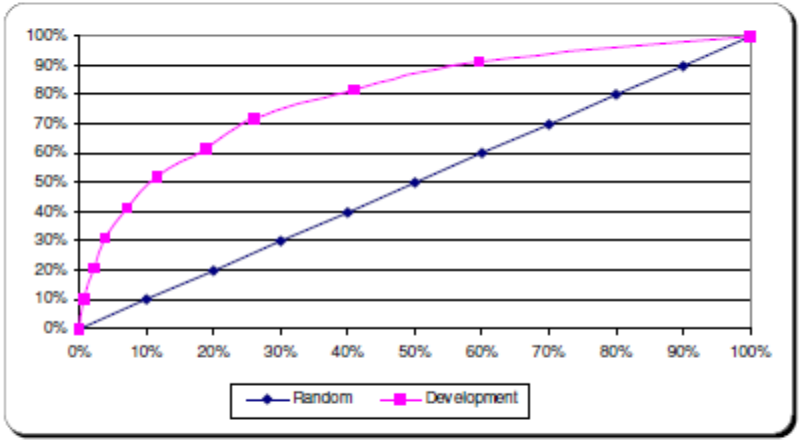
Among all pairs formed from 0 & 1 observations from the dependent variable, the % of pairs where the probability assigned to an observation with value 1 for the dependent variable is greater than that assigned to an observation with value 0.

Percentage of concordant pairs should be *at least greater than 60*.

Logistic Procedure – Lorenz Curve, Gini, KS

Lorenz curve indicates the lift provided by the model over random selection.

Gini coefficient represents the area covered under the Lorenz curve. A good model would have a Gini coefficient between 0.2 - 0.35



Lorenz Curve

Kolmogorov-Smirnoff (KS) statistic is defined as the absolute difference of cumulative % of Goods and cumulative % of Bads.

KS statistic value *should not be less than 20*. Higher the KS – better is the model.

Logistic Procedure – Divergence Index Test

Good	_FREQ_	ave	variance		Ho: Bad Score => Good Score	p- value
	41338	752.67	4070.44		Null Hypothesis is Rejected	
0	856	654.55	10578.1225	DI	T - Statistic	
1	40482	754.75	3725.8816	**1.4038**	**-28.398**	**<0.0001**

Divergence Index is an indicator of how well the means of the goods and bads are differentiated.

Null Hypothesis: The means of Good accounts / population = The means of Bad accounts / population

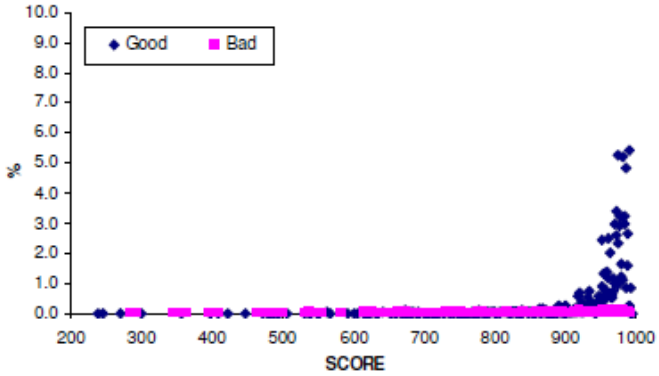
Alternative Hypothesis: The means of Good accounts / population is not equal to the means of Bad accounts / population

For a robust model – we need to *reject* the null hypothesis. Hence, *lower the p-value better the model*.

Logistic Procedure – Clustering Check

The concept behind Clustering check is that a good model should be sensitive enough to differentiate between 2 Good/Bad accounts.

i.e the model should be able to identify differences between seemingly same type of accounts/sample observations and assign them different scores.



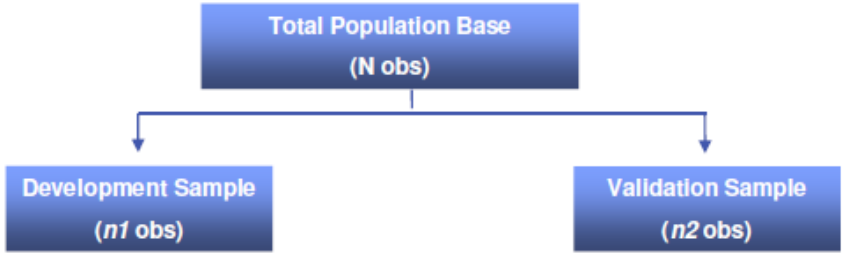
A good model should not have significant clustering of the population at any particular score and the population must be well scattered across.

Ideally the clustering should be as low as possible. A thumb-rule would be to contain the clustering so that it is *within 5-6%*.

Logistic Procedure – Validation

Validation could be done in 2 ways:

- ✓ Validation Re-run
- ✓ Scoring the Validation sample



Validation Re-run

- Rerun the model on the validation sample.
- Check the chi-sq values and level of significances and p-values for each explanatory variable.
- The p-values should not change significantly from the development sample to the validation sample.
- Check the signs of the parameter estimates. They should not change from development sample to the validation sample.

Validation sample scoring

- Score the validation sample using the parameter estimates obtained from the scorecard developed on the development sample.
- Check rank ordering. Both development and validation samples should rank order.

Thank You