Introduction to Logistic Regression

Content

- Simple and multiple linear regression
- Simple and multiple Discriminant Analysis
- Simple logistic regression
  - The logistic function
  - Estimation of parameters
  - Interpretation of coefficients
- Multiple logistic regression
  - Interpretation of coefficients
  - Coding of variables

What are Discriminant Analysis (DA) and Logistic Regression (LR)

We sometimes encounter a problem that involves a categorycal dependent variable and several matric independent variables. Example: Credit Risk (god or bad), Consumer Decision (Buying or Not, Like or dislike). HRD (Succes or Fail), General Managemen (Succes or Fail).

DA and LR are the appropriate statistical techniques when the dependent variable is categorial (nominal or non metric) and the independent variables are metric.

DA, capable to handling either two groups or multiple (more than two groups). When involved two group is refered two-group discriminant analysis (simple DA), when more than two indetified group is refered to multiple discriminant analysis.

What is Logistic Regression (LR)……

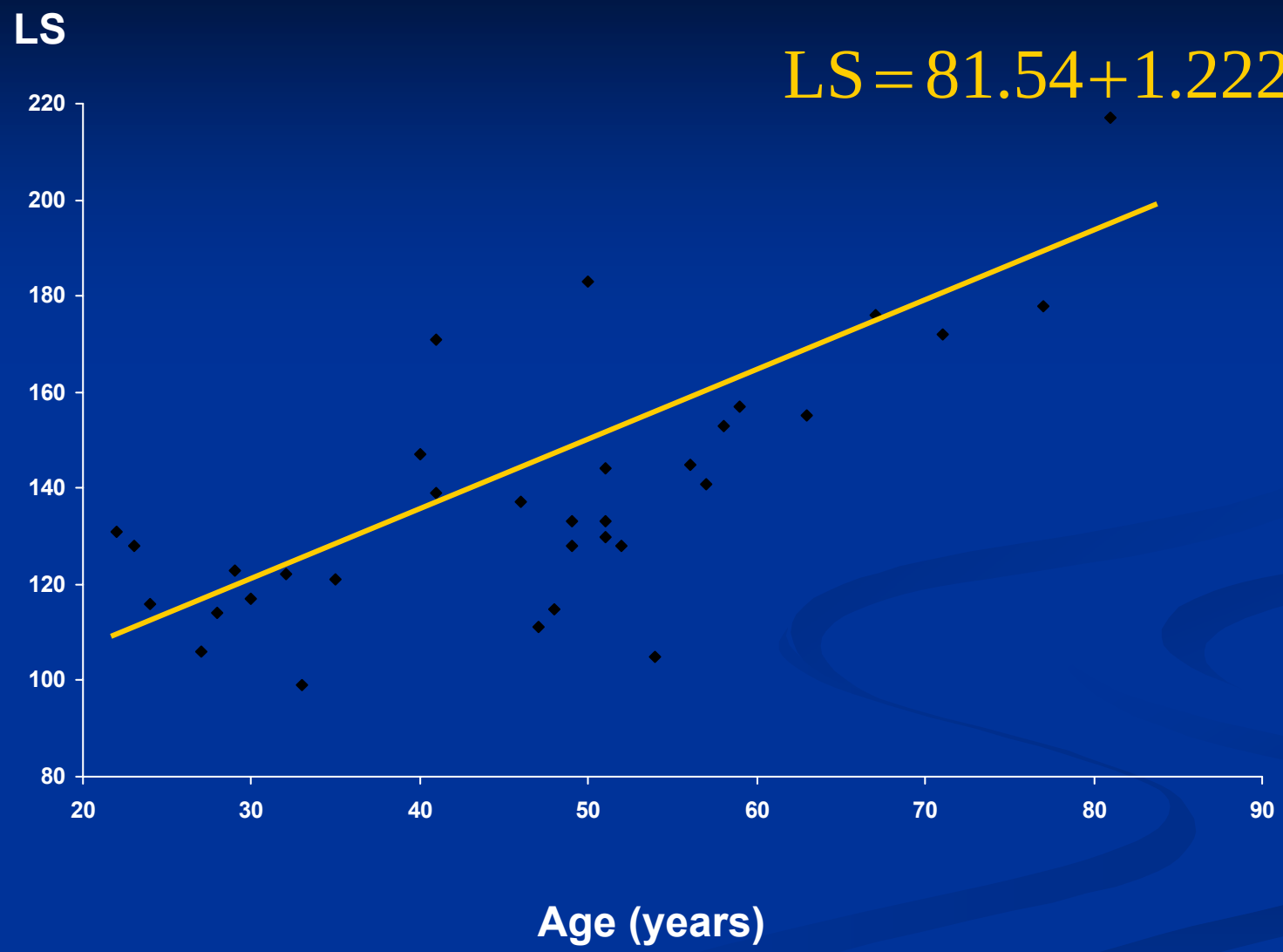
However, when the dependent variables has only two groups, logistic regression may be prefered for several reason:

1. DA,relies on stricly meeting the assumptions of multivariate normality and equal variance-covariance matrices across group. LR does not face these strict assumptions.
2. Beacouse similar to linear regression, so researcher more prefer.
3. In DA, the nonmetric character of dichotomous dependent variables is accommodated by making predictions of group membership based on discriminant Z scores. Calculating of cutting scores and the assigment of observation to group.
4. LR, similar to linear regression, but it can direcly predicts the probability of an event accuring. ALthought probability is emetric measure is fundamental differences between Linear regression. (See Picture slide 15)

Simple linear regression

Table 1 Age and Leadership (LD) among 33

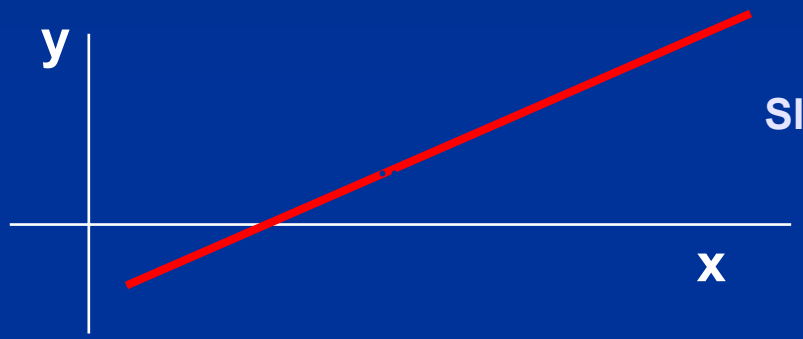
Age	LS	Age	LS	Age	LS
22	131	41	139	52	128
23	128	41	171	54	105
24	116	46	137	56	145
27	106	47	111	57	141
28	114	48	115	58	153
29	123	49	133	59	157
30	117	49	128	63	155
32	122	50	183	67	176
33	99	51	130	71	172
35	121	51	133	77	178
40	147	51	144	81	217



$$LS = 81.54 + 1.222 \cdot Age$$

Simple linear regression

- Relation between 2 continuous variables (LD and age)



$$y = \alpha + \beta_1 x_1$$

- Regression coefficient β_1
  - Measures association between y and x
  - Amount by which y changes on average when x changes by one unit
  - Least squares method

Multiple linear regression

- Relation between a continuous variable and a set of i continuous variables

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

- Partial regression coefficients β_i
  - Amount by which y changes on average when x_i changes by one unit and all the other x_is remain constant
  - Measures association between x_i and y adjusted for all other x_i

Multiple linear regression

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

Predicted	Predictor variables
Response variable	Explanatory variables
Outcome variable	Covariables
Dependent	Independent variables

General linear models

- Family of regression models
- Outcome variable determines choice of model

Outcome	Model
Continuous	Linear regression
Binomial	Logistic regression

- Uses
  - Model building, risk prediction

Logistic regression

- Models relationship between set of variables x_i
  - dichotomous (yes/no)
  - categorical (social class, ...)
  - continuous (age, ...)

and

  - dichotomous (binary) variable Y

- Dichotomous outcome most common situation in business (Marketing, HRD, Finance)

Logistic regression (1)

Table 2 Age and signs of Stress (SS)

Age	SS	Age	SS	Age	SS
22	0	40	0	54	0
23	0	41	1	55	1
24	0	46	0	58	1
27	0	47	0	60	1
28	0	48	0	60	0
30	0	49	1	62	1
30	0	49	0	65	1
32	0	50	1	67	1
33	0	51	0	71	1
35	1	51	1	77	1
38	0	52	0	81	1

How can we analyse these data?

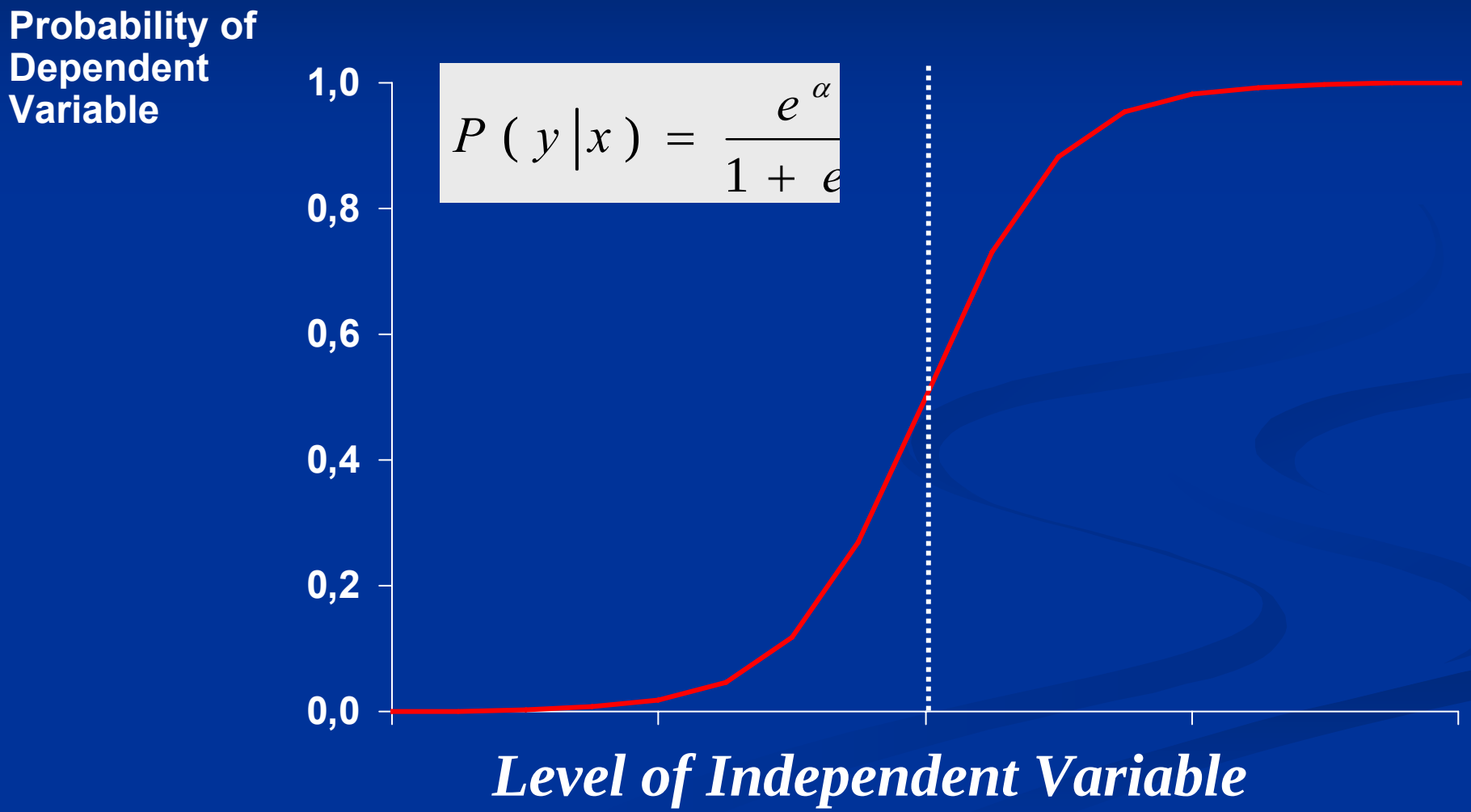
- Compare mean age of Yes and No
  - NO: 38.6 years
  - Yes: 58.7 years
- Linear regression?

Logistic regression (2)

Table 3 Prevalence (%) of signs of SS according to age group

		SS	
Age group	# in group	#	%
20 - 29	5	0	0
30 - 39	6	1	17
40 - 49	7	2	29
50 - 59	7	4	57
60 - 69	5	4	80
70 - 79	2	2	100
80 - 89	1	1	100

Logistic function (1)



Logistic transformation

$$P(y|x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

$$\ln\left[\frac{P(y|x)}{1-P(y|x)}\right] = \alpha + \beta x$$

logit of $P(y|x)$

Advantages of Logit

- Properties of a linear regression model
- Logit between - ∞ and + ∞
- Probability (P) constrained between 0 and 1

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta x \qquad \frac{P}{1-P} = e^{\alpha+\beta x}$$

- Directly related to nation of odds

Interpretation of coefficient β

	Exposure x	
SS y	yes	no
yes	$P(y\|x=1)$	$P(y\|x=0)$
no	$1-P(y\|x=1)$	$1-P(y\|x=0)$

$$\frac{P}{1-P} = e^{\alpha+\beta x}$$

$$Odds_{d|e} = e^{\alpha+\beta}$$

$$Odds_{d|\bar{e}} = e^{\alpha}$$

$$OR = \frac{e^{\alpha+\beta}}{e^{\alpha}} = e^{\beta}$$

$$\ln(OR) = \beta$$

Interpretation of coefficient β

- β = increase in logarithm of odds ratio for a one unit increase in x
- Test of the hypothesis that β=0 (Wald test)

$$\chi 2 = \frac{\beta^2}{\text{Variance}(\beta)} \quad \text{(1df)}$$

- Interval testing $95\%\,\text{CI} = e^{(\beta \pm 1.96 SE_\beta)}$

Example

- Risk of developing Stress (ss) by age (<55 and 55+ years)

SS	55+ (1)	< 55 (0)
Present (1)	21	22
Absent (0)	6	51

Odds of …… among exposed = 21/6
Odds of …….. among unexposed = 22/51

Odds ratio = 8.1

Logistic Regression Model

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 \times \text{Age} = -0.841 + 2.094 \times \text{Age}$$

	Coefficient	SE	Coeff/SE
Age	2.094	0.529	3.96
Constant	-0.841	0.255	-3.30

$$\text{OR} = e^{2.094} = 8.1$$

$$\text{Wald Test} = 3.96^2 \text{ with } 1\text{df} \quad (p < 0.05)$$

$$95\%\ \text{CI} = e^{(2.094 \pm 1.96 \times 0.529)} = 2.9,\ 22.9$$

Fitting equation to the data

- Linear regression: Least squares
- Logistic regression: Maximum likelihood
- Likelihood function
  - Estimates parameters α and β with property that likelihood (probability) of observed data is higher than for any other values
  - Practically easier to work with log-likelihood

$$L(\mathrm{B}) = \ln\left[l(\mathrm{B})\right] = \sum_{i=1}^{n}\left\{y_i \ln\left[\pi(x_i)\right] + (1-y_i)\ln\left[1-\pi(x_i)\right]\right\}$$

Maximum likelihood

- Iterative computing
  - Choice of an arbitrary value for the coefficients (usually 0)
  - Computing of log-likelihood
  - Variation of coefficients' values
  - Reiteration until maximisation (plateau)
- Results
  - Maximum Likelihood Estimates (MLE) for α and β
  - Estimates of P(y) for a given value of x

Multiple logistic regression

- More than one independent variable
  - Dichotomous, ordinal, nominal, continuous …

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots \beta_i x_i$$

- Interpretation of β_i
  - Increase in log-odds for a one unit increase in x_i with all the other x_is constant
  - Measures association between x_i and log-odds adjusted for all other x_i

Effect modification

- Effect modification
  - Can be modelled by including interaction terms

$$\ln\left(\frac{P}{1-P}\right)=\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 \times x_2$$

Statistical testing

- Question
  - Does model including given independent variable provide more information about dependent variable than model without this variable?
- Three tests
  - Likelihood ratio statistic (LRS)
  - Wald test
  - Score test

Likelihood ratio statistic

- Compares two nested models

  $\text{Log(odds)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ (model 1)

  $\text{Log(odds)} = \alpha + \beta_1 x_1 + \beta_2 x_2$ (model 2)

- LR statistic

  -2 log (likelihood model 2 / likelihood model 1) =

  -2 log (likelihood model 2) *minus* -2log (likelihood model 1)

  LR statistic is a χ^2 with DF = number of extra parameters in model

Example

P	Probability for Succes
Gender	1= Male, 0 = Female
Smk	1= smk, 0= non-smu

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1\,\text{Gender} + \beta_2\,\text{Smk}$$

$$= 0.7102 + 1.0047\,\text{Gender} + 0.7005\,\text{Smk}$$

$$(\text{SE}\,0.2614) \quad (\text{SE}\,0.2664)$$

$$\text{OR for lack of Male} = e^{1.0047} = 2.73 \ \textit{(adjusted for smk)}$$

$$95\%\,\text{CI} = e^{(1.0047 \pm 1.96 \times 0.2614)} = 1.64 - 4.56$$

- Interaction between smoking and exercise?

$$\ln\left(\frac{P}{1-P}\right) = \alpha + \beta_1\,\text{Gend} + \beta_2\,\text{Smk} + \beta_3\,\text{Smk} \times \text{Gend}$$

- Product term $\beta_3 = -0.4604$ (SE 0.5332)

  Wald test = 0.75 (1df)

  -2log(L) = 342.092 with interaction term

  = 342.836 without interaction term

  $\Rightarrow$ LR statistic = 0.74 (1df), p = 0.39

  $\Rightarrow$ No evidence of any interaction

Stages in Logistic Regression

- Stage1. 2,3 Research Objective, Research Design and Statistical Asumptions
- Stage 4. Estimation of the Logistic Regression Model and Assessing Overall Fit
- Stage 5. Interpretaion of the results
- Stage 6. Validation of the result

Stages in Logistic Regression

- Stage1. 2,3 Research Objective, Research Design and Statistical Asumptions

Dependent Variable (non metric, single or multiple)

Independent Variable (metric or non metric)

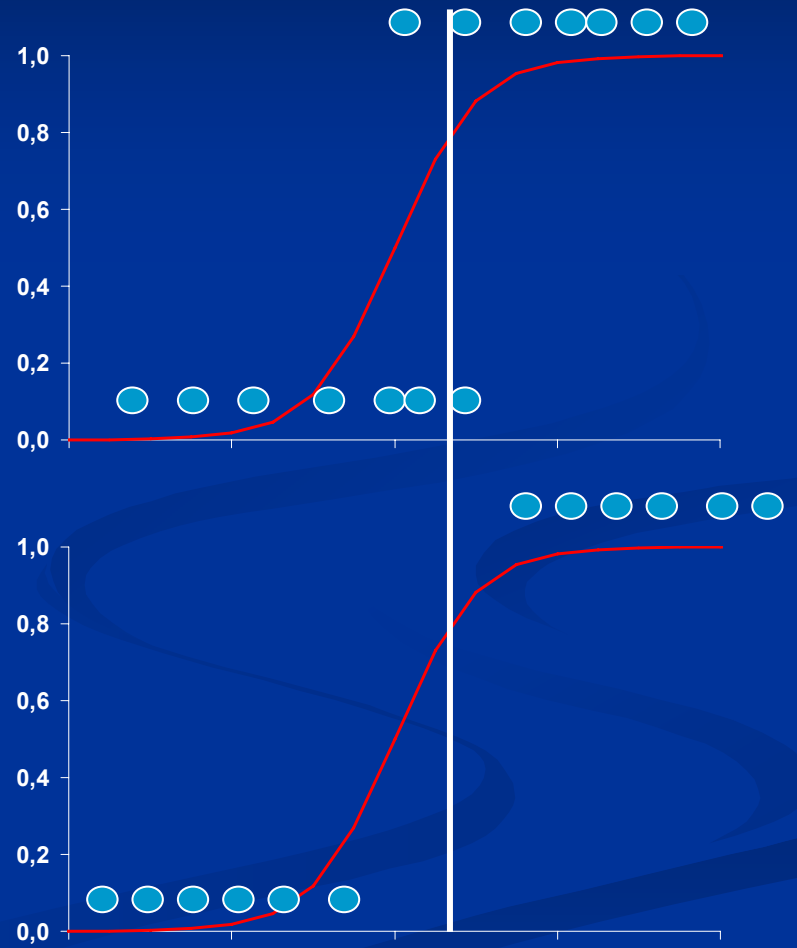
Sample Size (n=…….)

Stage 4. Estimation of the Logistic Regression Model and Assessing Overall Fit

- **Predict probability of an event accuring**
- **Assumed relationship between independent and dependent variables that resebles an S-shaped curve(see slide 15)**
- **The error term of as discrete variables follows the binomial distribution, invalidting normality**
- **The variance of dichotomous variables is not constant, creating instance of heterodasticity as well**

Stage 4. Estimation of the Logistic Regression Model and Assessing Overall Fit

- **Estimated coefficients for each independent variables by using logistic tranformation , the maximum likelihood procedure ("Most likely")**
- **The result in the use of likelihood value when calculating measure of overal fit model.**



Stages in Logistic Regression

- Stage1. 2,3 Research Objective, Research Design and Statistical Asumptions
- Stage 4. Estimation of the Logistic Regression Model and Assessing Overall Fit
- Stage 5. Interpretaion of the results
- Stage 6. Validation of the result

Stages in Logistic Regression

- Stage1. 2,3 Research Objective, Research Design and Statistical Asumptions
- Stage 4. Estimation of the Logistic Regression Model and Assessing Overall Fit
- Stage 5. Interpretaion of the results
- Stage 6. Validation of the result

Stages in Logistic Regression

- Stage1. 2,3 Research Objective, Research Design and Statistical Asumptions
- Stage 4. Estimation of the Logistic Regression Model and Assessing Overall Fit
- Stage 5. Interpretaion of the results
- Stage 6. Validation of the result