Multiple Logistic Regression

Professor Dr. Syed Hatim Noor
Unit Biostatistics and Research Methodology
School of Medical Sciences
Universiti Sains Malaysia

Introduction

- Multiple logistic regression is the estimation of the relationship between a dichotomous dependent variable and more than one independent variables or covariates
- Applied in exploratory studies, explanatory studies
- **Independent** variables are the combination of **numerical** and **categorical** variables
- **Outcome** is **binary categorical** variable
- If the outcome is dichotomous (called Multiple Logistic Regression)
- If the outcome is polytomous (called Multinomial Logistic Regression)

Introduction

- The goals of the regression analysis is to establish a model that is
  - Best fit
  - Parsimonious
  - Biologically sound (Biological plausibility)
  - Statistically significant
- To answer the research question: What are the factors associated to the dependent variable (event-yes/no)??
  - Eg.: What are the factors that associate with coronary artery disease (CAD)

Odds

- The odds = the chance
- The odds of an event is the ratio of the number of ways the event can occur to the number of ways the event can not occur
- Eg.: On average 54 girls are born in every 100 births. What is the odds of any randomly chosen delivery to be is a girl?
- number of girls/number of boys=54/46
- So, about 1.17 the odds to get a baby girl

Odds Ratio

- Odds ratio is calculated by dividing the 2 odds
- Eg.: What is the odds ratio of men to have CAD compared to women?
- The odds of men having CAD/The odds of women having CAD

	Have CAD = 1	No CAD = 0
Men = 1	a	b
Women = 0	c	d

- The odds for men having CAD = $\frac{n \text{ of men having CAD (a)}}{n \text{ of men not having CAD (b)}}$

	Have CAD = 1	No CAD = 0
Men = 1	a	b
Women = 0	c	d

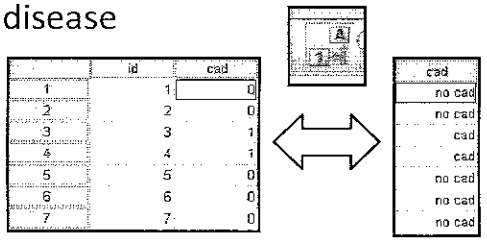
- The odds for women having CAD = $\frac{n \text{ of women having CAD (c)}}{n \text{ of women not having CAD (d)}}$

	Have CAD = 1	No CAD = 0
Men = 1	a	b
Women = 0	c	d

- The odds ratio (OR) of men to have CAD compared to women =
  (a/b) / (c/d)
- Thus, OR = ad/bc

The Dataset

- Using dataset: coronary.sav
- Objective of the study: to determine the factors that associated with CAD
- The dependent variable:
  - CAD: coronary artery disease
  - Label: 0 – no CAD
    1 – has CAD



- Independent variable
  - systolic blood pressure (sbp): mmHg
  - diastolic blood pressure (dbp): mmHg
  - serum cholesterol (chol): mmol/l
  - body mass index (bmi): unit
  - age of the patient (age): years
  - race of the patient (race): 0 – Malays
    1 – Chinese
    2 – India
  - gender of the patient (gender): 0 – women
    1 – men

How to Code Categorical Variables

- Always start with 0
- 0: reference group (low risk, non-diseased, normal)
- For 2 leveled-categorical variable, eg.:
  - smoking status: 0 (non-smoker), 1 (smoker)
  - cancer status: 0 (no cancer), 1 (has cancer)
- For more than 3 leveled-categorical variable
  - race: 0 (Malay), 1 (Chinese), 2 (India)
  - income level: 0 (low), 1 (medium), 2 (high)

Steps in Multiple Logistic Regression

(1) Data exploration & cleaning

(2) Univariable analysis (**Simple Logistic Regression)**

(3) Variable selection (**Multiple Logistic Regression**)
(preliminary main effect model)

(4) Checking multicollinearity & interaction
(preliminary final model)

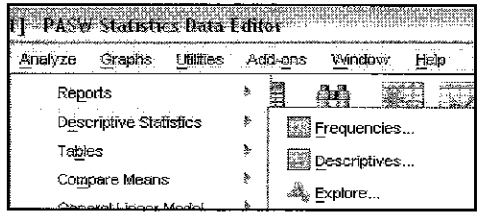
(5) Checking assumptions (final model)

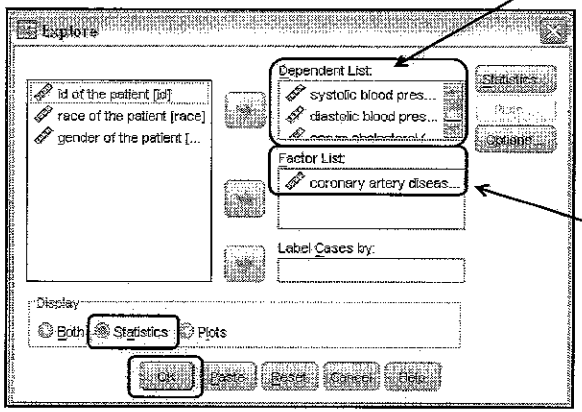
(6) Interpretation, conclusion & presentation

Step 1: Data exploration and cleaning

- Descriptive statistics
  - Numerical independent variable: mean(SD)

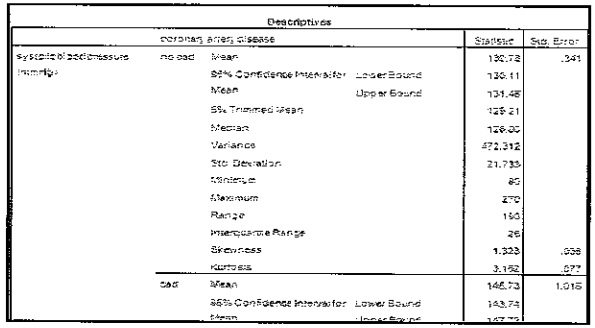
Analyze > Descriptive Statistics > Explore





Put all the **numerical** variable in the *Dependent List* box:
- systolic blood pressure
- diastolic blood pressure
- serum cholesterol
- age of the patient
- body mass index

Put **dependent** variable in the *Factor List* box:
- coronary artery disease



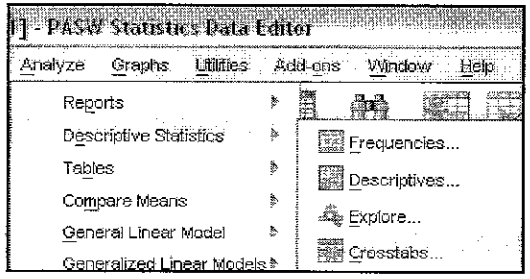
Summarize as below:-

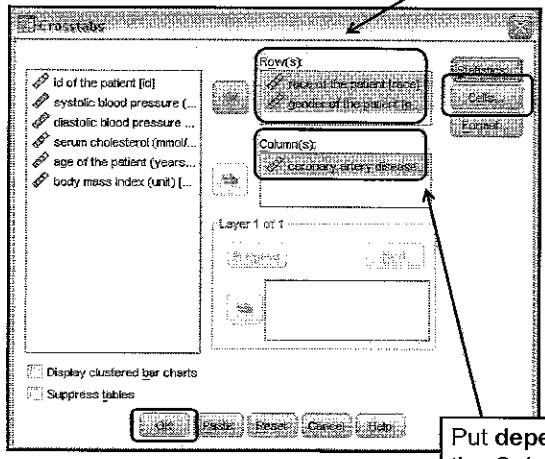
Table I: Descriptive statistics for numerical variables

Variables	No CAD Mean (SD)	Has CAD Mean (SD)
systolic blood pressure (mmHg)	130.78 (21.73)	145.73 (25.31)
diastolic blood pressure (mmHg)	81.32 (12.18)	90.45 (13.43)
serum cholesterol (mmol/l)	6.12 (1.32)	6.57 (1.26)
age (years)	45.70 (8.41)	48.19 (8.76)
bmi (unit)	36.91 (3.75)	36.73 (3.83)

– Categorical independent variable: n(%)

Analyze > Descriptive Statistics > Crosstabs





Put all the categorical variable in *Row(s)* box:
- race of the patient
- gender of the patient

Put **dependent** variable in the *Column(s)* box:
- coronary artery disease

race of the patient * coronary artery disease Crosstabulation

			coronary artery disease		
			no cad	cad	Total
race of the patient	malay	Count	1315	209	1524
		% within race of the patient	86.3%	13.7%	100.0%
	chinese	Count	1375	212	1587
		% within race of the patient	86.6%	13.4%	100.0%
	india	Count	1379	200	1579
		% within race of the patient	87.3%	12.7%	100.0%
Total		Count	4069	621	4690
		% within race of the patient	86.8%	13.2%	100.0%

gender of the patient * coronary artery disease Crosstabulation

			coronary artery disease		
			no cad	cad	Total
gender of the patient	woman	Count	1819	228	2047
		% within gender of the patient	88.9%	11.1%	100.0%
	man	Count	2250	393	2643
		% within gender of the patient	85.1%	14.9%	100.0%
Total		Count	4069	621	4690
		% within gender of the patient	86.8%	13.2%	100.0%

Summarize as below:-

Table II: Descriptive statistics for variable gender

	No CAD	Has CAD
Gender	**n (%)**	**n (%)**
Woman	1819 (88.9)	228 (11.1)
Man	2250 (85.1)	393 (14.9)

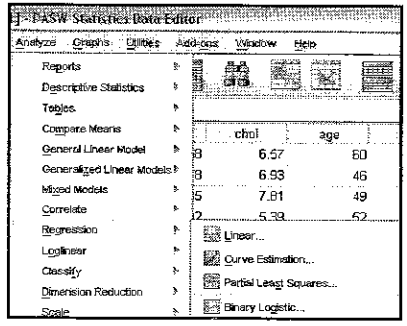
Table III: Descriptive statistics for variable race

	No CAD	Has CAD
Race	**n (%)**	**n (%)**
Malay	1315 (86.3)	209 (13.7)
Chinese	1375 (86.6)	212 (13.4)
Indian	1379 (87.3)	200 (12.7)

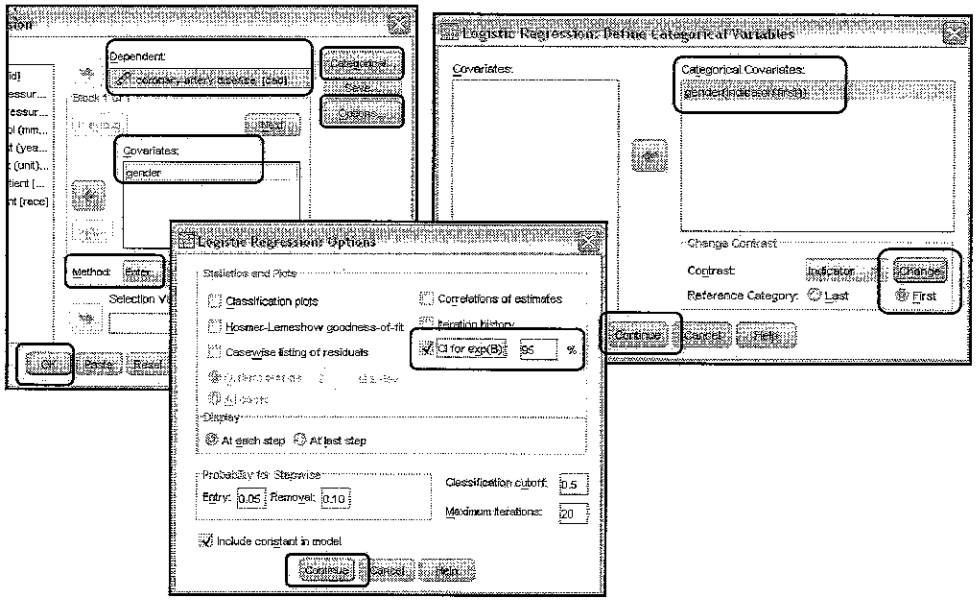
Step 2: Univariable analysis (Simple Logistic Regression)

- To screen for important independent variables
  - Look for variables with p-value < 0.25 and/or clinically important

    Analyze > Regression > Binary Logistic



- For categorical independent variable



Case Processing Summary

		N	Percent
Unweighted Cases[a]			
Selected Cases	Included in Analysis	4690	100.0
	Missing Cases	0	.0
	Total	4690	100.0
Unselected Cases		0	.0
Total		4690	100.0

a. If weight is in effect, see classification table for the total number of cases.

- **N** – the number of cases in the dataset
- **Included in Analysis** – This row gives the number of cases that were included in the analysis
- **Missing Cases** – This row gives the number of missing cases (not include in the analysis)

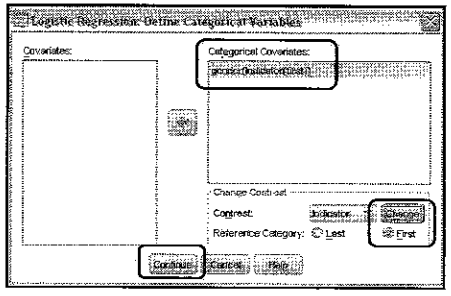
Dependent Variable Encoding

Original Value	Internal Value
no cad	0
cad	1

- Showing the coding of dependent variable
  - 0: no cad
  - 1: has cad
- Our interest is the chance of someone who has cad (comparing ***has cad*** to ***no cad***)
- Thus, ***no cad*** is the reference group

Categorical Variables Codings

		Frequency	Parameter coding (1)
gender of the patient	woman	2047	.000
	man	2643	1.000



This table showing the reference group defined in the analysis

By choosing the option ***first*** and click ***Change***, SPSS will analyze the data based on coding 0 as your reference group

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	14.166	1	.000
	Block	14.166	1	.000
	Model	14.166	1	.000

- Omnibus test tests the significance of the independent variable in the model

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	gender(1)	.332	.089	13.894	1	.000	1.394	1.170	1.659
	Constant	-2.077	.070	873.767	1	.000	.125		

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

- B – regression coefficient
- S.E. – standard error
- Wald and Sig. – a test to test the null hypothesis that the regression coefficient equals 0. The hypothesis is rejected when p-value (Sig.) is <0.05.

$$Wald = \left(\frac{B}{S.E.}\right)^2$$

- Thus, the independent variable is significant to the model

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	gender(1)	.332	.089	13.894	1	.000	1.394	1.170	1.659
	Constant	-2.077	.070	873.767	1	.000	.125		

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

- df – degrees of freedom
- Exp (B) – exponentiation of the B coefficient
  - ODDS RATIO
  - Exp(0.332) = 1.394
- 95% C.I. for EXP (B) – 95% confidence interval for odds ratio

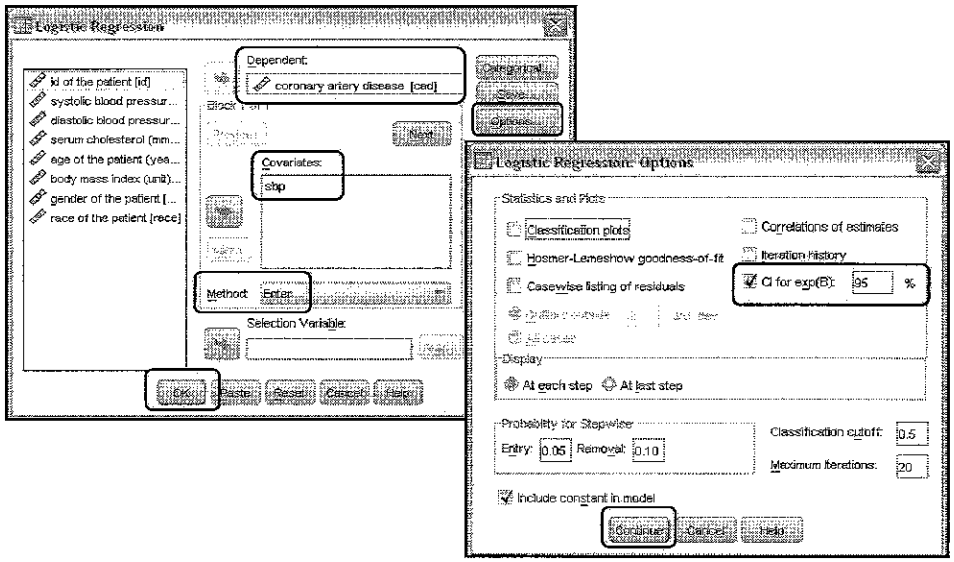
Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	gender(1)	.332	.089	13.894	1	.000	1.394	1.170	1.659
	Constant	-2.077	.070	873.767	1	.000	.125		

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

- Interpretation:-
  - 95% confidence interval does not include 1 (1.17,1.66), coefficient is positive (0.332), variable gender is significant to the model (p<0.001). Men has 39.4% higher odds (chance) to have coronary artery disease compared to women when other confounders were not adjusted.
  - If the odds ratio less than 1, it is protective (less risk)

- For numerical independent variable



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	sbp	.024	.002	203.487	1	.000	1.025	1.021	1.028
	Constant	-5.212	.245	453.437	1	.000	.005		

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

- Interpretation:-
  - 95% confidence interval does not include 1 (1.02,1.03), coefficient is positive (0.024), variable systolic blood pressure is significant to the model ($p<0.001$). A person with 1mmHg increase in systolic blood pressure has 2.5% higher odds (chance) to have coronary artery disease when other confounders were not adjusted.

Hands-on

- Try to do Simple Logistic Regression to other variables in the dataset:-
  - dbp
  - chol
  - age
  - bmi
  - race

Results from Simple Logistic Regression

Table IV: Associated factors of coronary artery disease by Simple Logistic Regression model

Variable	Regression coefficient (b)	Crude Odds Ratio (95%CI)	Wald statistic	p-value
systolic blood pressure (mmHg)	0.02	1.025 (1.02,1.03)	203.49	**<0.001**
diastolic blood pressure (mmHg)	0.05	1.053 (1.05,1.06)	245.58	**<0.001**
serum cholesterol (mmol/l)	0.25	1.28 (1.21,1.37)	63.07	**<0.001**
age of the patient (years)	0.03	1.04 (1.02,1.05)	45.50	**<0.001**
body mass index (unit)	-0.01	0.99 (0.97,1.01)	1.15	0.285
gender of the patient				
women	0	1		
men	0.33	1.39 (1.17,1.66)	13.89	**<0.001**
race of the patient				
Malay	0	1		
Chinese	-0.03	0.97 (0.79,1.19)	0.08	0.772
India	-0.09	0.91 (0.74,1.12)	0.74	0.389

Step 3: Variable selection (Multiple Logistic Regression)

- Review all the p-values from univariable analysis (simple logistic regression)
- Select the candidate variables with p-value <0.25
- May select variable with p-value >0.25 **BUT** clinically important
- In this dataset, we select
  - sbp - age
  - dbp - bmi
  - chol - gender

Methods of Variable Selection

- Enter – manual
  - Enter or remove manually the independent variable
- Forward selection
  - Automatically enters the IMPORTANT independent variable into the model
- Backward elimination
  - Automatically removes the UNIMPORTANT independent variable out of the model

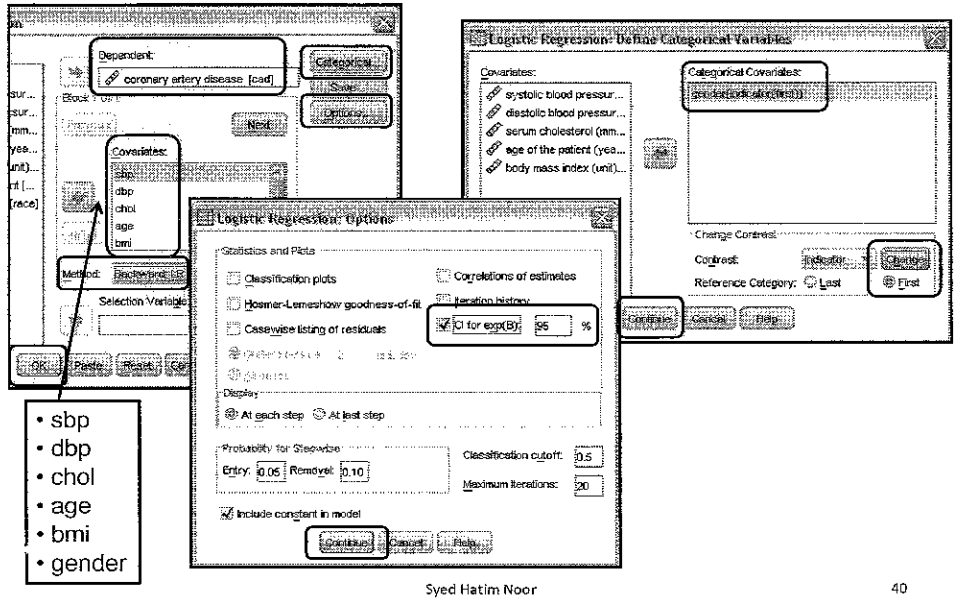
- Forward selection
  - Conditional
    - Stepwise selection method with entry testing based on the significance of the score statistic
    - Removal testing based on the probability of a likelihood-ratio statistic based on conditional parameter estimates
  - Likelihood Ratio (LR)
    - Stepwise selection method with entry testing based on the significance of the score statistic
    - Removal testing based on the probability of a likelihood-ratio statistic based on the maximum partial likelihood estimates

- Forward selection
  - Wald
    - Stepwise selection method with entry testing based on the significance of the score statistic
    - Removal testing based on the probability of the Wald statistic

- Backward elimination
  - Conditional
    - Backward stepwise selection
    - Removal testing based on the probability of a likelihood-ratio statistic based on conditional parameter estimates
  - Likelihood Ratio (LR)
    - Backward stepwise selection
    - Removal testing based on the probability of a likelihood-ratio statistic based on the maximum partial likelihood estimates

- Backward elimination
  - Wald
    - Backward stepwise selection
    - Removal testing based on the probability of the Wald statistic

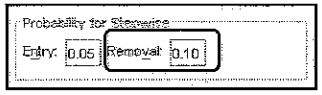
Backward Elimination (LR)



Variables in the Equation

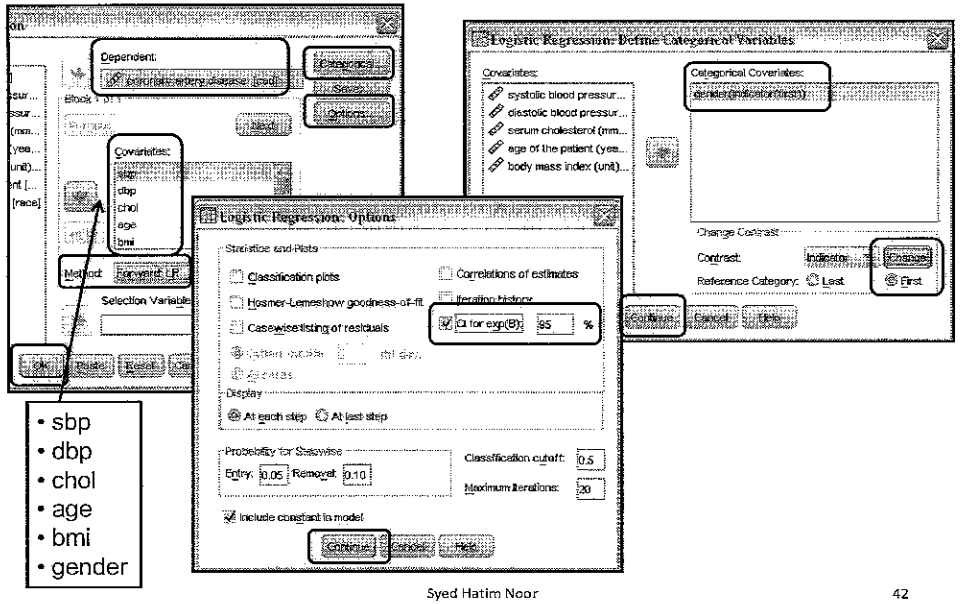
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	sbp	.003	.003	.694	1	.405	1.003	.996	1.009
	dbp	.045	.006	62.369	1	.000	1.046	1.034	1.057
	chol	.124	.035	12.289	1	.000	1.132	1.056	1.213
	age	.008	.006	1.715	1	.190	1.008	.996	1.020
	bmi	-.007	.012	.335	1	.563	.993	.971	1.016
	gender(1)	.371	.094	15.703	1	.000	1.450	1.207	1.742
	Constant	-7.192	.579	154.254	1	.000	.001		
Step 2[a]	sbp	.003	.003	.690	1	.406	1.003	.996	1.009
	dbp	.045	.006	62.520	1	.000	1.046	1.034	1.057
	chol	.125	.035	12.422	1	.000	1.133	1.057	1.214
	age	.008	.006	1.709	1	.191	1.008	.996	1.020
	gender(1)	.371	.094	15.644	1	.000	1.449	1.206	1.741
	Constant	-7.444	.384	375.788	1	.000	.001		
Step 3[a]	dbp	.048	.003	192.543	1	.000	1.050	1.042	1.057
	chol	.124	.035	12.392	1	.000	1.132	1.057	1.214
	age	.010	.006	2.896	1	.089	1.010	.999	1.021
	gender(1)	.382	.093	16.959	1	.000	1.465	1.222	1.757
	Constant	-7.492	.381	387.625	1	.000	.001		

a. Variable(s) entered on step 1: sbp, dbp, chol, age, bmi, gender.



- Start with all variables
- Based on removal probability 0.10 available in the ***Option***, eliminate one variable at each step
- At the final step, variable dbp, chol, age and gender retain in the model

Forward Selection (LR)

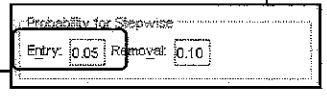


- sbp
- dbp
- chol
- age
- bmi
- gender

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	dbp	.052	.003	245.578	1	.000	1.053	1.046	1.060
	Constant	-6.317	.295	457.462	1	.000	.002		
Step 2[b]	dbp	.053	.003	250.436	1	.000	1.054	1.047	1.061
	gender(1)	.417	.092	20.530	1	.000	1.518	1.267	1.818
	Constant	-6.622	.305	471.442	1	.000	.001		
Step 3[c]	dbp	.050	.003	212.621	1	.000	1.051	1.044	1.058
	chol	.137	.035	15.663	1	.000	1.146	1.071	1.227
	gender(1)	.398	.092	18.552	1	.000	1.488	1.242	1.783
	Constant	-7.242	.349	429.940	1	.000	.001		

a. Variable(s) entered on step 1: dbp.
b. Variable(s) entered on step 2: gender.
c. Variable(s) entered on step 3: chol.



- dbp has the smallest p-value from change in -2 log likelihood (LR test). It is included first (step 1)
- Followed by gender in step 2 and chol in step 3
- At the final step, the variable being included in the model are dbp, chol and gender

Comparing Forward & Backward Results

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	dbp	.052	.003	245.578	1	.000	1.053	1.046	1.060
	Constant	-6.317	.295	457.462	1	.000	.002		
Step 2[b]	dbp	.053	.003	250.436	1	.000	1.054	1.047	1.061
	gender(1)	.417	.092	20.530	1	.000	1.518	1.267	1.818
	Constant	-6.622	.305	471.442	1	.000	.001		
Step 3[c]	dbp	.050	.003	212.621	1	.000	1.051	1.044	1.058
	chol	.137	.035	15.663	1	.000	1.146	1.071	1.227
	gender(1)	.398	.092	18.552	1	.000	1.488	1.242	1.783
	Constant	-7.242	.349	429.940	1	.000	.001		

a. Variable(s) entered on step 1: dbp.
b. Variable(s) entered on step 2: gender
c. Variable(s) entered on step 3: chol.

Forward Selection (LR)

Backward Elimination(LR)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	sbp	.003	.003	.694	1	.405	1.003	.996	1.009
	dbp	.045	.006	62.369	1	.000	1.046	1.034	1.057
	chol	.124	.035	12.289	1	.000	1.132	1.056	1.213
	age	.008	.006	1.715	1	.190	1.008	.996	1.020
	bmi	-.007	.012	.335	1	.563	.993	.971	1.016
	gender(1)	.371	.094	15.703	1	.000	1.450	1.207	1.742
	Constant	-7.192	.579	154.254	1	.000	.001		
Step 2[a]	sbp	.003	.003	.690	1	.406	1.003	.996	1.009
	dbp	.045	.006	62.520	1	.000	1.046	1.034	1.057
	chol	.125	.035	12.422	1	.000	1.133	1.057	1.214
	age	.008	.006	1.709	1	.191	1.008	.996	1.020
	gender(1)	.371	.094	15.644	1	.000	1.449	1.206	1.741
	Constant	-7.444	.384	375.788	1	.000	.001		
Step 3[a]	dbp	.048	.003	192.543	1	.000	1.050	1.042	1.057
	chol	.124	.035	12.392	1	.000	1.132	1.057	1.214
	age	.010	.006	2.896	1	.089	1.010	.999	1.021
	gender(1)	.382	.093	16.959	1	.000	1.465	1.222	1.757
	Constant	-7.492	.381	387.625	1	.000	.001		

a. Variable(s) entered on step 1: sbp, dbp, chol, age, bmi, gender

- Using backward elimination, independent variable age is retained
- However, the p-value of age is 0.089 (>0.05)
- If researcher uses p-value of 0.05 as a selection criteria (the cut off point), then age may need to be excluded from the model
- The researcher's decision on removal based on
  - p-value from Wald statistic
  - clinical importance
- At the variable selection step, **preliminary main effect model** is obtained

Variable Selection Methods

- Researcher should use various methods
- Each model may differ from the other
- **Advisable to do both forward** selection and **backward** elimination method to compare which model is the best model in consideration of
  - The model which is the most biologically parsimonious
  - The model which is the most fit (check assumptions)

Interpretation

Variables in the Equation

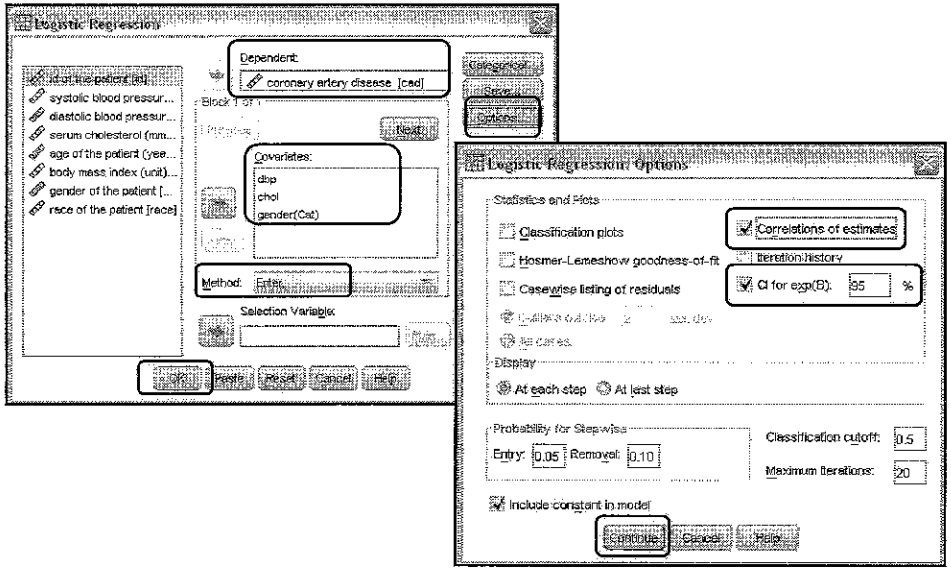
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	dbp	.050	.003	212.621	1	.000	1.051	1.044	1.058
	chol	.137	.035	15.663	1	.000	1.146	1.071	1.227
	gender(1)	.398	.092	18.552	1	.000	1.488	1.242	1.783
	Constant	-7.242	.349	429.940	1	.000	.001		

a. Variable(s) entered on step 1: dbp, chol, gender.

- A person with 1 mmHg increase in dbp has 1.05 times the odds to have cad
  (b=0.05, OR=1.05, 95%CI 1.04,1.06, p<0.001)
- A person with 1 mmol/l increase in chol has 1.15 times the odds to have cad
  (b=0.14, OR=1.15, 95%CI 1.07,1.23, p<0.001)
- Men has 1.49 times the odds to have cad
  (b=0.40, OR=1.49, 95%CI 1.24,1.78, p<0.001)

Step 4: Checking multicollinearity & interaction

- Check multicollinearity
  - Checked to assess which variable (2 or more) correlate highly
  - Check correlation estimates
  - Check standard errors
- May omit the variable if standard error is big
- Decision is subjective (depend on researcher)



Correlation Matrix

		Constant	dbp	chol	gender(1)
Step 1	Constant	1.000	-.753	-.473	-.187
	dbp	-.753	1.000	-.186	.068
	chol	-.473	-.186	1.000	-.053
	gender(1)	-.187	.068	-.053	1.000

- Based on the SPSS output, the correlation between variables are relatively small
  - dbp & chol: -0.19
  - dbp & gender: 0.07
  - chol & gender: -0.05

Variables in the Equation

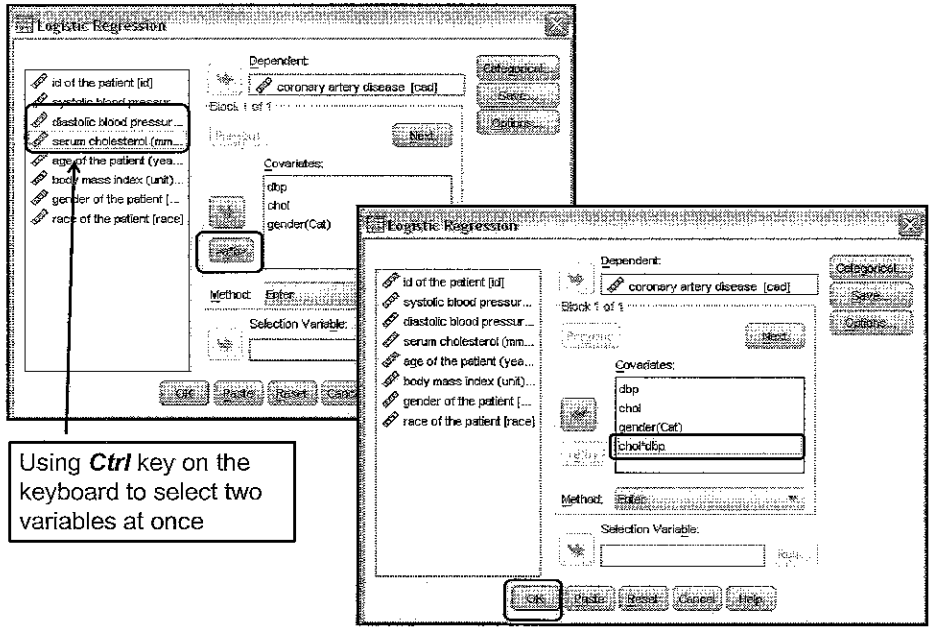
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	dbp	.050	.003	212.621	1	.000	1.051	1.044	1.058
	chol	.137	.035	15.663	1	.000	1.146	1.071	1.227
	gender(1)	.398	.092	18.552	1	.000	1.488	1.242	1.783
	Constant	-7.242	.349	429.940	1	.000	.001		

a. Variable(s) entered on step 1: dbp, chol, gender.

- Based on the SPSS output, the standard error of variables are relatively small
  - dbp: 0.003
  - chol: 0.035
  - gender: 0.092

- Check interaction
  - Test 2-way biologically / clinically meaningful interaction term one at a time
  - Choose 2 independent variables in the model based on practical consideration
  - Create an interaction term
  - Add it into model one at a time and check p-value
    - if <0.05, include it in the model

- Possible 2-way interaction in this model:-
  1. dbp & chol
  2. dbp & gender
  3. chol & gender



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	dbp	.080	.016	24.466	1	.000	1.084	1.050	1.119
	chol	.555	.219	6.426	1	.011	1.742	1.134	2.675
	gender(1)	.407	.092	19.395	1	.000	1.502	1.253	1.800
	chol by dbp	-.005	.002	3.747	1	.053	.995	.990	1.000
	Constant	-9.814	1.434	47.814	1	.000	.000		

a. Variable(s) entered on step 1: dbp, chol, gender, chol * dbp .

- The interaction term (cholesterol and diastolic blood pressure) is not significant (p=0.053)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	dbp	.044	.008	59.945	1	.000	1.045	1.033	1.057
	chol	.135	.035	15.195	1	.000	1.144	1.069	1.225
	gender(1)	-.396	.829	.397	1	.528	.673	.198	2.308
	dbp by gender(1)	.009	.007	1.623	1	.203	1.009	.995	1.023
	Constant	-6.714	.538	155.936	1	.000	.001		

a. Variable(s) entered on step 1: dbp, chol, gender, dbp * gender .

- The interaction term (diastolic blood pressure and gender) is not significant (p=0.203)

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	dbp	.050	.003	212.491	1	.000	1.051	1.044	1.058
	chol	.151	.057	7.135	1	.008	1.163	1.041	1.300
	gender(1)	.546	.467	1.367	1	.242	1.727	.691	4.317
	chol by gender(1)	-.023	.071	.106	1	.745	.977	.851	1.123
	Constant	-7.341	.464	250.205	1	.000	.001		

a. Variable(s) entered on step 1: dbp, chol, gender, chol * gender .

- The interaction term (cholesterol and gender) is not significant (p=0.745)

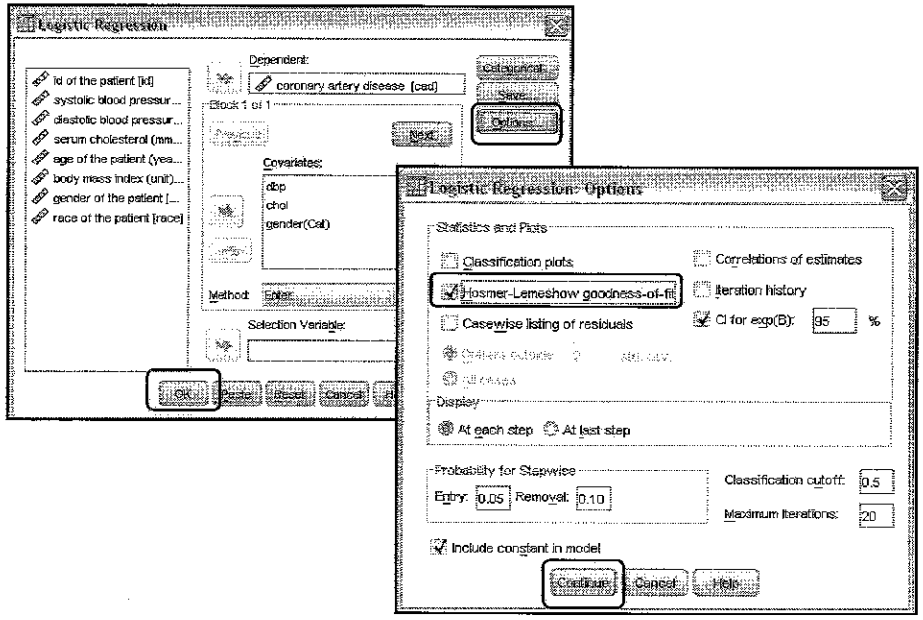
- As a conclusion, the standard error and correlation are relatively small for three of the independent variables in the model
- There is no significant interaction effect in the model
- **Preliminary final model** is obtained

Step 5: Checking Assumptions

- Assessing the goodness of fit
  1. The Hosmer-Lemeshow test
  2. Classification table
  3. Area under the Receiver Operating Characteristic (ROC) curve

Hosmer-Lemeshow test

- It is based on grouping cases into deciles of risk
- It compares the observed probability with the expected probability within each deciles
- Check the p-value. If it is >0.05, there is no significant difference between the observed probability and the expected probability
- Thus, assumption is met



Contingency Table for Hosmer and Lemeshow Test

		coronary artery disease = no cad		coronary artery disease = cad		
		Observed	Expected	Observed	Expected	Total
Step 1	1	455	451.305	14	17.695	469
	2	451	441.348	18	27.652	469
	3	439	435.523	31	34.477	470
	4	427	428.513	42	40.487	469
	5	417	421.639	52	47.361	469
	6	416	414.107	53	54.893	469
	7	398	403.998	70	64.002	468
	8	392	393.645	78	76.355	470
	9	356	372.313	113	96.687	469
	10	318	306.609	150	161.391	468

deciles

- Compare the discrepancy between the observed and expected probability
- Better (fitter) if there is small discrepancy

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	10.790	8	.214

- The p-value is >0.05, which is 0.214, assumption is met
- The model is fit

Classification table

- Default in SPSS logistic regression
- Overall correctly classified percentage is good if above 70%
- You can manually calculate
  - Sensitivity
  - Specificity
  - PPV
  - NPV

Classification Table[a]

Observed			Predicted		
			coronary artery disease		Percentage Correct
			no cad	cad	
Step 1	coronary artery disease	no cad	4036	33	99.2
		cad	604	17	2.7
	Overall Percentage				86.4

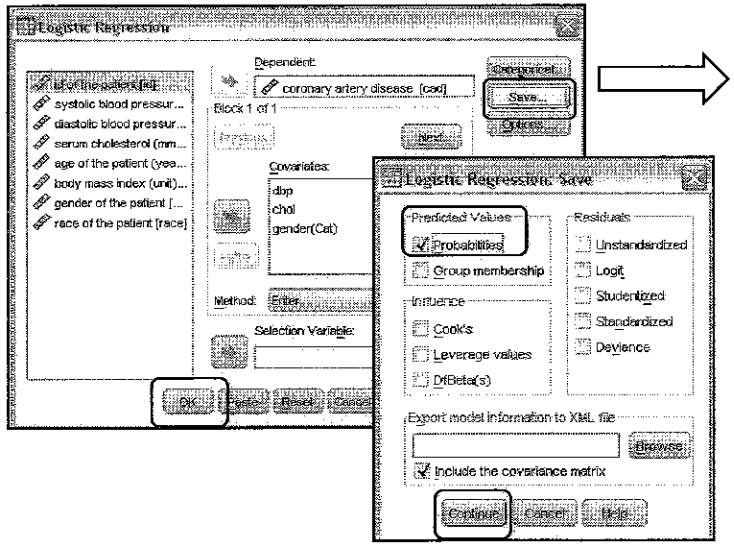
a. The cut value is .500

- In this context, the overall correctly classified percentage is 86.4%
- Assumption is met
- Model is fit

Area under the ROC curve

- Ranges from 0 to 1
- Able to assess the model discrimination
- A value of 0.5 means the model is useless for discrimination
- The recommended area under the ROC curve is at least 0.70
- Values near to 1 is better

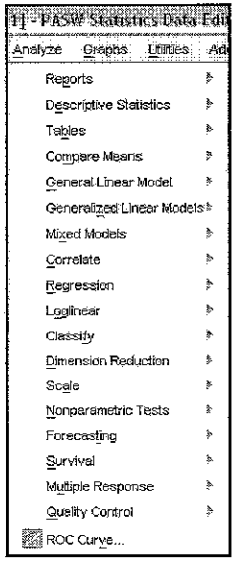
- Create predicted value

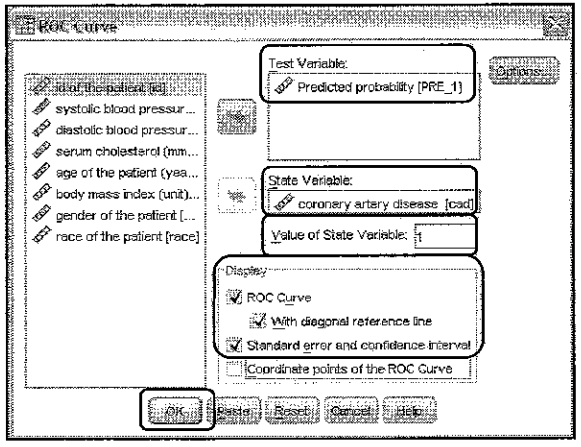


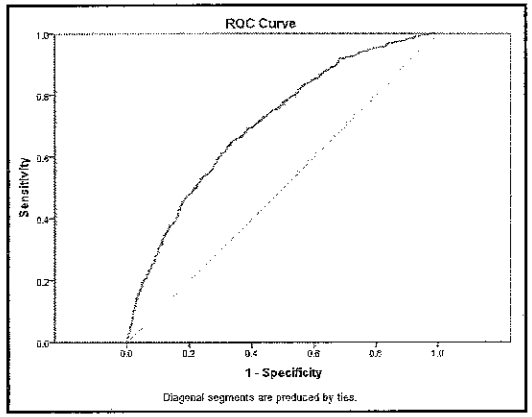
PRE_1
.04933
.08227
.17571
.17830
.24403
.08451
.16911
.04446
.20566
.11499
.15548
.18635
.23431
.07624

- Create ROC curve

  Analyze > ROC Curve







Area Under the Curve

Test Result Variable(s):Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.709	.011	.000	.688	.730

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.

a. Under the nonparametric assumption

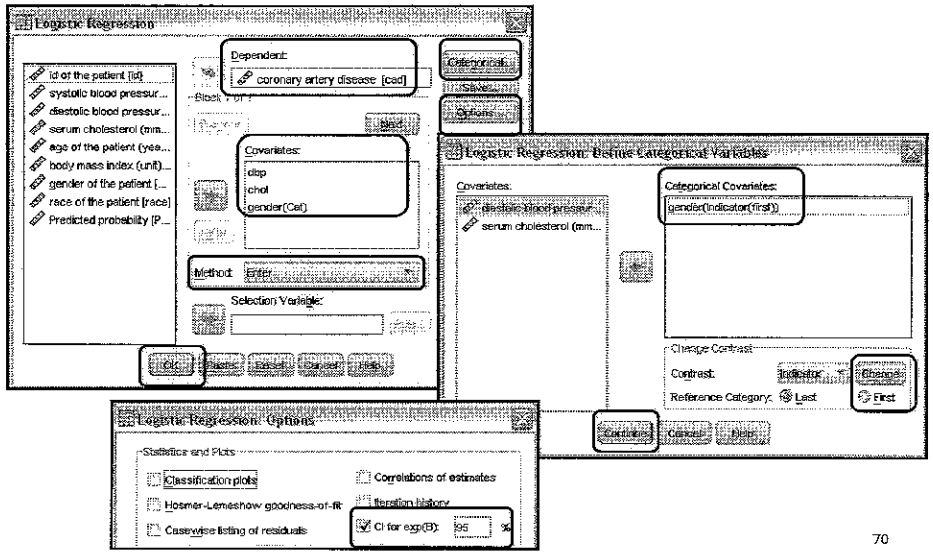
b. Null hypothesis: true area = 0.5

- Area under the ROC curve is 0.709 (95% CI 0.69,0.73)
- It is significantly different from 0.5 (p-value<0.05)
- The model can accurately discriminate 70.9% of the cases

- As a conclusion,
  - Hosmer-Lemeshow test: p-value=0.214, which is >0.05
  - Classification table: overall correctly classified percentage is 86.4%, which is >70%
  - ROC curve: Area under the curve is 70.9%, which is >70%
- Assumptions are met
- **Final model** is achieved

Step 6: Interpretation, Conclusion & Presentation

- Establish final model



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	dbp	.050	.003	212.821	1	.000	1.051	1.044	1.058
	chol	.137	.035	15.663	1	.000	1.146	1.071	1.227
	gender(1)	.398	.092	18.552	1	.000	1.488	1.242	1.783
	Constant	-7.242	.349	429.940	1	.000	.001		

a. Variable(s) entered on step 1: dbp, chol, gender.

- diastolic blood pressure, serum cholesterol and gender have significant association with presence of coronary artery disease

Table V: Associated factors of coronary artery disease by Multiple Logistic Regression model

Variable	Regression coefficient (b)	Adjusted Odds Ratio [a] (95%CI)	Wald statistic	p-value
diastolic blood pressure (mmHg)	0.05	1.05 (1.04,1.06)	212.62	<0.001
serum cholesterol (mmol/l)	0.14	1.15 (1.07,1.23)	15.66	<0.001
gender of the patient				
women	0	1		
men	0.40	1.49 (1.24,1.78)	18.55	<0.001

[a] Forward LR Multiple Logistic Regression model was applied
Multicollinearity and interaction term were checked and not found
Hosmer-Lemeshow test, (p=0.214), classification table (overall correctly classified percentage=86.4%) and area under the ROC curve (70.9%) were applied to check the model fit

Table VI: Associated factors of coronary artery disease by simple and multiple logistic regression model

Variable	Simple Logistic Regression			Multiple Logistic Regression [a]		
	b	Crude OR (95% CI)	p	b	Adjusted OR (95% CI)	P
diastolic blood pressure (mmHg)	0.05	1.053 (1.05,1.06)	<0.001	0.05	1.05 (1.04,1.06)	<0.001
serum cholesterol (mmol/l)	0.25	1.28 (1.21,1.37)	<0.001	0.14	1.15 (1.07,1.23)	<0.001
gender						
women	0	1		0	1	
men	0.33	1.39 (1.17,1.66)	<0.001	0.40	1.49 (1.24,1.78)	<0.001

[a] Forward LR Multiple Logistic Regression model was applied
Multicollinearity and interaction term were checked and not found
Hosmer-Lemeshow test, (p=0.214), classification table (overall correctly classified percentage=86.4%) and area under the ROC curve (70.9%) were applied to check the model fit

Interpretation

- A person with an increase in 1 mmHg of diastolic blood pressure has 5% higher odds to have coronary artery disease (95% CI 1.04,1.06, p<0.001) when adjusted for serum cholesterol and gender
- A person with an increase in 1 mmol/l of serum cholesterol has a 15% higher odds to have coronary artery disease (95% CI 1.07,1.23, p<0.001) when adjusted for diastolic blood pressure and gender
- Men has 49% higher odds compared to women to have coronary artery disease (95% CI 1.24,1.78, p<0.001) when adjusted for diastolic blood pressure and serum cholesterol

Prediction

- B (regression coefficient)
  - This is the value for the logistic regression equation for predicting the dependent variable from the independent variable
  - It is in log-odds unit
- The prediction equation is

$$\log (p/1-p) = b_0 + b_1*x_1 + b_2*x_2 + b_3*x_3 + b_4*x_4$$

- where p is the probability of being the honors composition
- Expressed in terms of the variables used in this example, the logistic regression equation is

$$\log (p/1-p) = -7.242 + 0.050*dbp + 0.137*chol + 0.398*gender(1)$$

Logit Equation

Conclusion

- Binary logistic regression deals with a dichotomous outcome variable
- Odds ratio help the interpretation of association between independent and dependent variables
- Follow proper steps to ensure the best model can be obtained
- Proper coding must be practiced
- Understand the statistical importance and clinical important