# **Multiple Logistic Regression**

### Dr. Wan Nor Arifin

### Unit of Biostatistics and Research Methodology, Universiti Sains Malaysia.

### wnarifin@usm.my / wnarifin.pancakeapps.com

Wan Nor Arifin, 2015. Multiple logistic regression by Wan Nor Arifin is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by-sa/4.0/.

IBM SPSS Statistics Version 22 screenshots are copyrighted to IBM Corp.



### Outlines

- Introduction
- Steps in Multiple Logistic Regression
  - 1. Descriptive Statistics
  - 2. Variable Selection
  - 3. Model Fit Assessment
  - 4. Final Model Interpretation & Presentation



- 1.Understand the reasons behind the use of logistic regression.
- 2.Perform multiple logistic regression in SPSS.
- 3.Identify and interpret the relevant SPSS outputs.
- 4. Summarize important results in a table.

- Logistic regression is used when:
  - Dependent Variable, DV: A binary categorical variable
     [Yes/No], [Disease/No disease] i.e the outcome.
- Simple logistic regression Univariable:
  - Independent Variable, IV: A categorical/numerical variable.
- Multiple logistic regression Multivariable:
  - IVs: Categorical & numerical variables.
- Recall Multiple Linear Regression?

- Multiple Linear Regression
  - $y = a + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$
- Multiple Logistic Regression
  - $-\log(\text{odds}) = a + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$
  - That's why it is called "logistic" regression.

- Binary outcome: Concerned with Odds Ratio.
  - Odds is a measure of chance like probability.
  - Odds = n(Disease)/n(no Disease) among a group.
  - Odds Ratio, OR = Odds(Factor)/Odds(No factor)
  - Applicable to all observational study designs.
- Relative Risk, RR
  - Only cohort study.
- OR ≈ RR for rare disease, useful to determine risk.

Factor vs CAD	CAD	No CAD		
Man	24 [a]	76 [b]		
Woman (i.e. not Man)	13 [c]	87 [d]		

- Odds(man) = a/b = 24/76 = 0.32
- Odds(woman) = c/d = 13/87 = 0.15
- OR(man/woman) = 0.32/0.15 = 2.13
- Shortcut, OR = ad/bc = (24x87)/(76x13) = 2.11

Factor vs CAD	CAD	No CAD		
Man	24 [a]	76 [b]		
Woman (i.e. not Man)	13 [c]	87 [d]		

- Risk(man) = Proportion CAD = a/(a+b) = 0.24
- Risk(woman) = Proportion CAD c/(c+d) = 0.13
- RR(man/woman) = 0.24/0.13 = 1.85 ≈ OR, 2.11

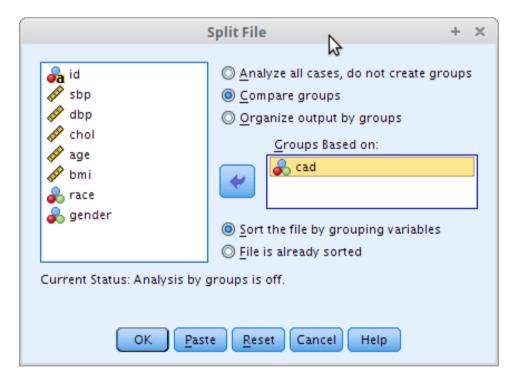
# **Steps in Multiple Logistic Regression**

- Dataset: slog.sav
- Sample size, n=200
- DV: cad (1: Yes, 0: No)
- IVs:
  - Numerical: *sbp* (systolic blood pressure), *dbp* (diastolic blood pressure), *chol* (serum cholesterol in mmol/L), *age* (age in years), *bmi* (Body Mass Index).
  - Categorical: *race* (0: Malay, 1: Chinese, 2: Indian), *gender* (0: Female, 1: Male)

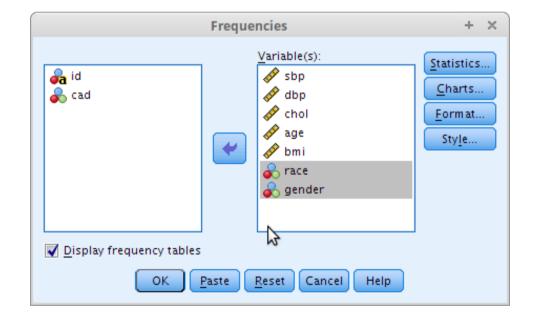
# **Steps in Multiple Logistic Regression**

- **1.Descriptive statistics.**
- 2.Variable selection.
  - a. Univariable analysis.
  - b. Multivariable analysis.
  - c. Multicollinearity.
  - d.Interactions.
- 3.Model fit assessment.
- 4. Final model interpretation & presentation.

- Set outputs by CAD status.
  - Data → Split File →
     Select Compare groups
  - Set Groups Based on:
     cad, OK



- Obtain mean(SD) and n(%) by CAD group.
  - Analyze → Descriptive
     Statistics →
     Frequencies
  - Include relevant variables in Variables

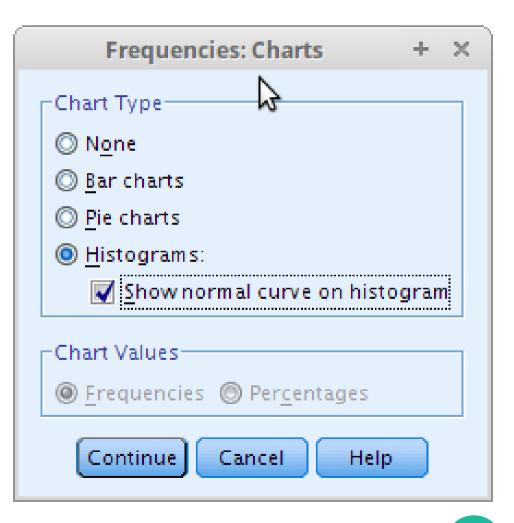


- Cont...
  - Statistics  $\rightarrow \text{tick} \rightarrow$ Continue

Frequencies: Statis	tics + ×
Percentile Values          Quartiles         Cut points for: 10         equal groups         Percentile(s):         Add         Change         Remove	Central Tendency Mean Median Mode Sum
Dispersion Std. deviation Minimum Variance Maximum Range S.E. mean Continue Cancel	<ul> <li>Values are group midpoints</li> <li>Distribution</li> <li>Skewness</li> <li>Kurtosis</li> <li>Help</li> </ul>



- Cont...
  - Charts  $\rightarrow \underline{tick} \rightarrow$ Continue  $\rightarrow$  OK



#### 4 Oct 2016

### • Results

				St	atistics				
cad coron	ary artery dise	ase	sbp Systolic Blood Pressure	dbp Diastolic Blood Pressure	chol serum cholesterol (mmol/l)	age Age in Years	bmi Body Mass Index	race ethnicity	gender gender
0 no cad	N	Valid	163	163	163	163	163	163	163
		Missing	0	0	0	0	0	0	0
	Mean		129.29	80.80	6.0970	45.15	36.9086	.94	.47
	Median		124.00	80.00	6.0500	44.00	37.9000	1.00	.00
	Std. Deviatio	n	22.264	12.607	1.16633	8.412	3.77178	.826	.500
	Minimum		88	56	4.00	31	25.30	0	0
	Maximum		218	120	9.35	62	41.20	2	1
	Percentiles	25	114.00	70.00	5.3350	37.00	36.1000	.00	.00
		50	124.00	80.00	6.0500	44.00	37.9000	1.00	.00
		75	140.00	90.00	6.7650	52.00	39.2000	2.00	1.00
1 cad	N	Valid	37	37	37	37	37	37	37
		Missing	0	0	0	0	0	0	0
	Mean		143.76	88.97	6.6459	47.43	36.4464	.97	.65
	Median		138.00	90.00	6.6550	50.00	37.1248	1.00	1.00
	Std. Deviatio	n	25.611	12.171	1.17041	8.796	3.99414	.833	.484
	Minimum		100	70	4.13	33	25.50	0	0
	Maximum		224	114	9.05	61	45.03	2	1
	Percentiles	25	122.00	78.00	5.9537	38.50	34.0802	.00	.00
		50	138.00	90.00	6.6550	50.00	37.1248	1.00	1.00
		75	159.00	97.00	7.2875	55.00	38.8146	2.00	1.00



### • Results

cad coronary artery disease			Frequency	Percent	Valid Percent	Cumulative Percent
0 no cad	Valid	0 malay	60	36.8	36.8	36.8
		1 chinese	52	31.9	31.9	68.7
		2 indian	51	31.3	31.3	100.0
		Total	163	100.0	100.0	
1 cad	Valid	0 malay	13	35.1	35.1	35.1
		1 chinese	12	32.4	32.4	67.6
		2 indian	12	32.4	32.4	100.0
		Total	37	100.0	100.0	

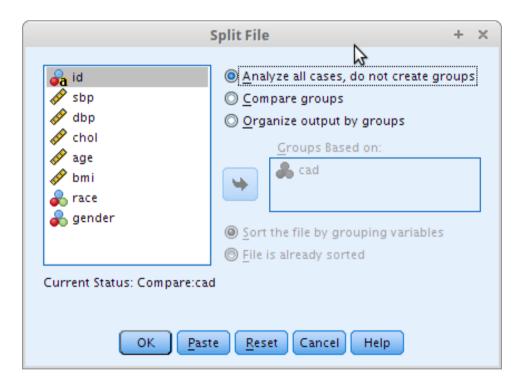
#### race ethnicity

#### gender gender

cad coronary artery disease			Frequency	Percent	Valid Percent	Cumulative Percent
0 no cad	Valid	0 woman	87	53.4	53.4	53.4
		1 man	76	46.6	46.6	100.0
		Total	163	100.0	100.0	
1 cad	Valid	0 woman	13	35.1	35.1	35.1
		1 man	24	64.9	64.9	100.0
		Total	37	100.0	100.0	

### • Results

- Look at histograms to decide data normality for numerical variables.
   Remember your Basic Stats!
- Caution! Reset back the data.
  - Data → Split File → Select
     Analyze all cases



- **OK** 

### • Present the results in a table.

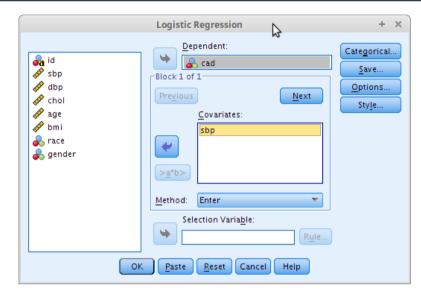
Fac	tors	CAD, n=37 mean(SD)	No CAD, n=163 mean(SD)	
Systolic Blo	od Pressure	143.8(25.61)	129.3(22.26)	
Diastolic Blo	ood Pressure	89.0(12.17)	80.8(12.61)	
Cholesterol		6.6(1.17)	6.1(1.17)	
A	ge	47.4(8.80)	45.2(8.41)	
В	MI	36.4(3.99)	36.9(3.77)	
Race*	Malay Chinese Indian	13(35.1%) 12(32.4%) 12(32.4%)	60(36.8%) 52(31.9%) 51(31.3%)	
Gender* Male Female		24(64.9%) 13(35.1%)	76(46.6%) 87(53.4%)	

## 2. Variable selection

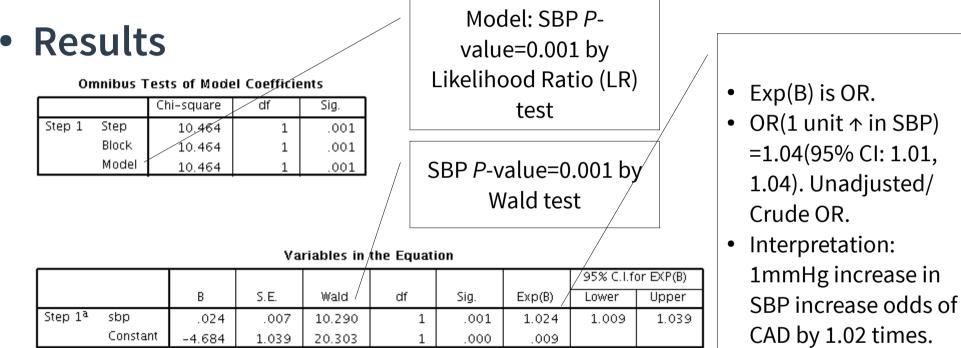
- To select best variables to predict the outcome.
- Sub-steps:
  - a. Univariable analysis.
  - b. Multivariable analysis.
  - c. Checking multicollinearity & interactions.

- Perform Simple Logistic Regression on each IV.
- Select IVs which fullfill:
  - *P*-value <  $0.25 \rightarrow$  Statistical significance.
  - Clinically significant IVs  $\rightarrow$  You decide.

- Analyze <u>numerical</u> variables:
  - Analyze → Regression →
     Binary Logistic
  - Dependent: cad,Covariates: sbp
  - Click Options → Tick
     Iteration history, CI for
     exp(B) → Continue → OK
  - Repeat for *dbp*, *chol*, *age*, *bmi*



Logistic Regres	Logistic Regression: Options + ×								
Statistics and Plots									
Classification plots	Correlations of estimates								
Hosmer-Lemeshow goodness-of-fit	Iteration history								
Casewise listing of residuals	✓ CI for exp(B): 95 %								
Outliers outside     Z     std. dev.     All cases     Display	5								
At <u>each step</u> At <u>last step</u>									
Probability for Stepwise Entry: 0.05 Remo <u>v</u> al: 0.10	Classification c <u>u</u> toff: 0.5 <u>M</u> aximum Iterations: 20								
Conserve memory for complex analyses or large <u>d</u> atasets									
✓ Include constant in model									
Continue Cancel Help									

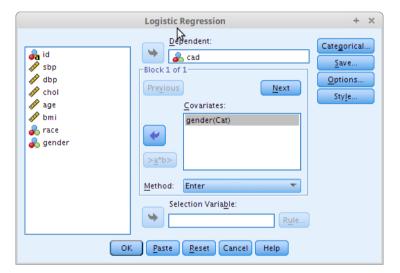


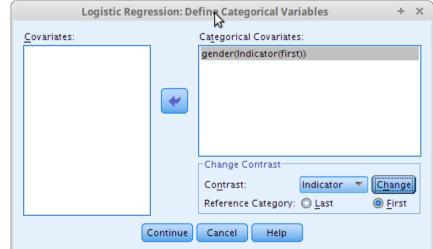
a. Variable(s) entered on step 1: sbp.

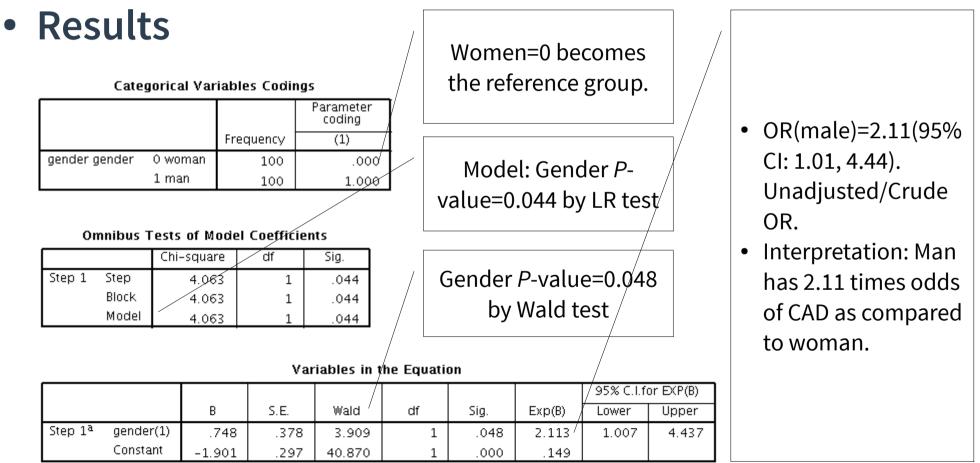
 In variable selection context, less concern about OR & interpretation.

#### 4 Oct 2016

- Analyze <u>categorical</u> variables:
  - Dependent: cad,
     Covariates: gender
  - Click Categorical →
     Categorical Covariates:
     gender → Change Contrast
     → Reference Category:
     First → Change → Continue.
  - Repeat for race







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

#### 4 Oct 2016

### P-values of IVs – select P-value < 0.25</li>

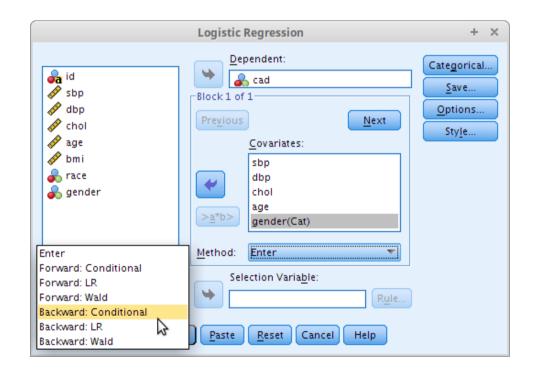
	Factors	<i>P</i> -value (Wald test)	<i>P</i> -value (LR test)	
Systo	lic Blood Pressure	0.001	0.001	
Diasto	lic Blood Pressure	0.001	0.001	
	Cholesterol	0.012	0.011	
	Age	0.143	0.141	
	BMI	0.505	0.511	
Race Chinese-vs-Malay Indian-vs-Malay		0.887 0.852	0.981*	
Gender Man-Woman		0.048	0.044	

\*For both variables

4 Oct 2016

- Selected variables:
  - sbp, dbp, chol, age, gender
- Perform Multiple logistic regression of the selected variables (multivariable) in on go.
- Variable selection is now proceed at multivariable level.
- Some may remain significant, some become insignificant.

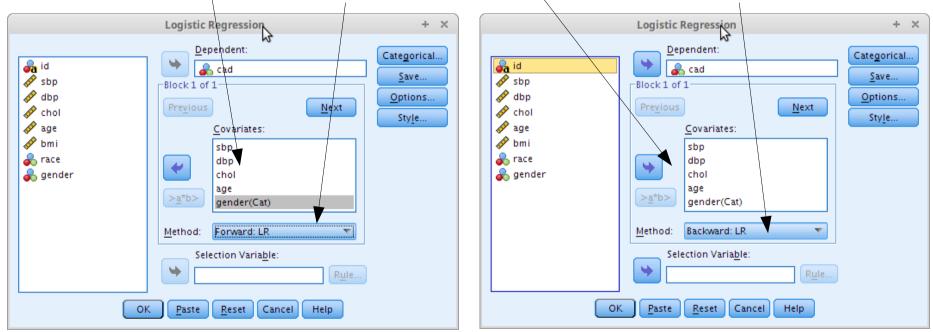
- Variable Selection Methods:
  - Automatic.
    - Forward: Conditional, LR, Wald. Enters variables.
    - Backward: Conditional, LR, Wald. Removes variables.
  - Manual.
    - Enter. Entry & removal of variables done manually. (Recommended, but leave to experts/statisticians).



#### 4 Oct 2016

- Variable Selection in this workshop:
  - Automatic by Forward & Backward LR.
  - Selection of variables by *P*-values based on LR test.

- Enter all <u>selected variables</u>.
- Perform 2x + 1x Forward LR, 1x Backward LR.



**Options**: Just leave at the default values.

#### 4 Oct 2016

Probability for Stepwise

Entry: 0.05 Removal: 0.10

				Va	riables in	the Equati	ion			
									95% C.I.f	or EXP(B)
			В	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
	Step 1ª	dbp	.049	.015	11.298	1	.001	1.050	1.021	1.080
		Constant	-5.620	1.277	19.358	1	.000	.004		
Results	Step 2 <sup>b</sup>	dbp	.050	.015	11.444	1	.001	1.051	1.021	1.081
Results		gender	.806	.391	4.250	1	.039	2.238	1.040	4.815
		Constant	-6.120	1.317	21.606	1	.000	.002		
Forward LR	/ a. Varial b. Varia	ble(s) entere ble(s) entere	d on step 1: d on step 2:	dbp. : gender.						
TOTWARE EN			·		riables in 1	the Equati	on			
									95% C.I.fo	or EXP(B)
			В	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
<ul> <li>Both methods</li> </ul>	Step 1ª	sbp	.009	.014	.371	1	.542	1.009	.981	1.037
• Dotti methous		dbp	.034	.025	1.799	1	.180	1.034	.985	1.086
keep same		chol	.187	.188	.987	1	.321	1.205	.834	1.742
•		age	016	.028	.335	1	.563	.984	.931	1.040
IVs: dbp &		gender -	.755	.401	3.544	1	.060	2.127	.969	4.667
gender.		Constant	-6.334	1.524	17.272	1	.000	.002		
•	Step 2ª	sbp	.006	.013	.183	1	.668	1.006	.980	1.031
<ul> <li><i>P</i>-values by</li> </ul>		dbp chol	.035	.025	1.965	1	.161	1.036	.986	1.087
<b>,</b>		gender	.162 .728	.182 .398	.796 3.351	1	.372 .067	1.176 2.070	.823 .950	1.681 4.512
Wald test.		Constant	-6.623	.398 1.449	20.882	1	.007	.001	.950	4.512
	Step 3ª	dbp	-0.023	.016	7.290	1	.000	1.044	1.012	1.077
	2.26.2	chol	.175	.180	.948	1	.330	1.191	.838	1.694
		gender	.741	.396	3.495	1	.062	2.098	.965	4.564
		- Constant	-6.657	1.452	21.017	1	.000	.001		
	Step 4ª	dbp	.050	.015	11.444	1	.001	1.051	1.021	1.081
Backward LR		gender	.806	.391	4.250	1	.039	2.238	1.040	4.815
		Constant	-6.120	1.317	21.606	1	.000	.002		

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

### Results

Forward LR
------------

- Both methods keep same IVs: *dbp* & *gender*.
- *P*-values by LR test.

#### Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	dbp	-95.778	11.933	1	.001
 Step 2	dbp	-93.747	12.289	1	.000
	gender	-89.812	4.419	1	.036

#### Model if Term Removed

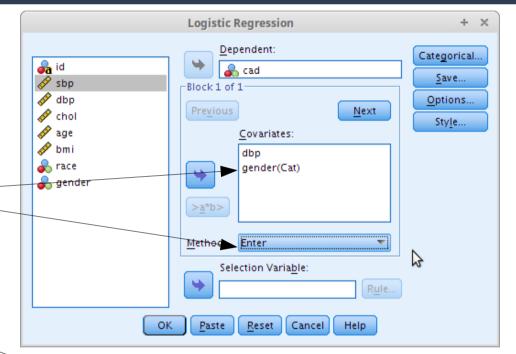
Variable	1	Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	sbp	-87.054	.370	1	.543
	dbp	-87.799	1.858	1	.173
	chol	-87.363	.988	1	.320
	age	-87.039	.339	1	.560
	gender	-88.698	3.657	1	.056
Step 2	sbp	-87.130	. 182	1	.670
	dbp	-88.056	2.034	1	. 154
	chol	-87.436	.793	1	.373
	gender	-88.765	3.451	1	.063
Step 3	dbp	-90.934	7.609	1	.006
	chol	-87.602	.945	1	.331
	gender	-88.932	3.604	1	.058
Step 4	dbp	-93.747	12.289	1	.000
	gender	-89.812	4.419	1	.036

Backward LR

4 Oct 2016

# **2c.** Multicollinearity

- Indicates redundant variables highly correlated IVs.
- Perform <u>Enter</u> method with <u>dbp &</u> <u>gender</u>.
- Look at <u>coefficients (B)</u> & <u>std</u> <u>errors (SE) / ORs (95% CIs)</u> if they are suspiciously large.



### Results

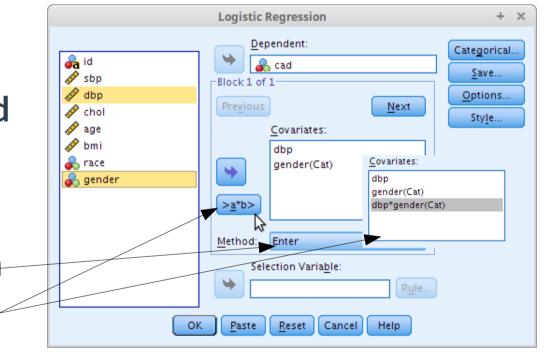
	Variables in the Equation											
								95% C.I.f	or EXP(B)			
		В	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper			
Step 1ª	dbp	.050	.015	11.444	1	.001	1.051	1.021	1.081			
	gender	.806	.391	4.250	1	.039	2.238	1.040	4.815			
	Constant	-6.120	1.317	21.606	1	.000	.002					

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

- SEs are quite small relative to Bs.
- 95% CIs are not too wide.
- No multicollinearity.

### 2d. Interactions

- IVs combination that requires interpretation of regression separately based on levels of IV → making things complicated.
- Perform <u>Enter method with</u> <u>dbp, gender & dbp x gender</u>.
   Select both dbp & gender (hold Ctrl on keyboard) → Click ><u>a</u>\*b>



### 2d. Interactions

### Results

Variables in the Equation

								95% C.I.	for EXP(B)
		В	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper
Step 1ª	dbp	.060	.028	4.615	1	.032	1.062	1.005	1.122
	gender(1)	2.117	2.911	.529	1	.467	8.308	.028	2495.947
	dbp by gender(1)	015	.033	.208	1	.648	.985	.924	1.051
	Constant	-7.070	2.502	7.987	1	005	.001		

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

Wald test for dbp by gender (dbp\*gender) not sig. Can remove the interaction term from model.

### 2. Variable selection

### • At the end of Variable Selection Step $\rightarrow$ *Preliminary* Final Model.

			Va	riables in t	he Equati	ion				• <i>P</i> -values by Wald test per variable
				wata	-16		Euro (D)	95% C.I.f		by Enter method.
		В	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper	
Step 1ª	dbp	.050	.015	11.444	1	.001	1.051	1.021	1.081	• Take this adjusted
	gender(1)	.806	.391	4.250	1	.039	2.238	1.040	4.815	OR.
	Constant	-6.120	1.317	21.606	1	.000	.002			
a. Variak	ole(s) entered	d on step 1:	dbp, gende	er.						 _
Om	nibus Test	s of Model	Coefficier	its		P-valı	ies bv	LR tes	st for	
	Ch	i-square	df	Sig.				gende		
Step 1	Step	16.352	2	.000 /		DOUL	upp a	genue	i Dy	

							_						-
	<b>b.</b> 4	 del	÷.	-	· ~ ·	-	 р.	~ •	-	~ •	 	-1	
	IN I	пър			- 14	r <b>F</b> F	 ĸ	ы	гн		 		

16.352

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change	
Step 1	dbp	-95.778	11.933	1	.001	//
Step 2	dbp	-93.747	12.289	1	.000	
	gender	-89.812	4.419	1	.036	

2

.000

Enter method.

P-values by LR per variable. Obtained with Forward LR method.

#### 4 Oct 2016

Block

Model

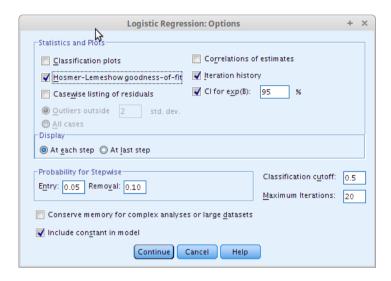
### 3. Model fit assessment

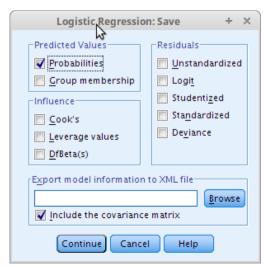
### By these 3 goodness-of-fit assessment methods:

- a. Hosmer-Lemeshow test
- b. Classification table.
- c. Area under Receiver Operating Characteristics (ROC) curve.
- At the end → *Final Model*.

### 3. Model fit assessment

- Perform Enter method with *dbp* & *gender*.
- Additionally
  - Click Options... → Tick
     Hosmer-Lemeshow
     goodness-of-fit
  - Click Save... → Tick
     Probabilities under
     Predicted Values
  - A new variable *PRE\_1* will be created.





#### 4 Oct 2016

### **3a. Hosmer-Lemeshow test**

### • Indicates fit of Preliminary Final Model to data.

### Results

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.	$\bigvee$
1	13.626	8	.092 /	

*P*-value 0.09 > 0.05 → Good model fit to the data.

Observed counts in data.

#### Contingency Table for Hosmer and Lemeshow Test

		cad coronary a = 0 n		cad coronary a		
		Observed	Expected	Observed	Expected	Total
Step 1	1	20	20.712	2	1.288	22
	2	18	18.369	2	1.631	20
	3	22	19.644	0	2.356	22
	4	24	20.787	0	3.213	24
	5	15	16.019	4	2.981	19
	б	16	18.009	б	3.991	22
	7	12	14.276	6	3.724	18
	8	17	15.260	3	4.740	20
	9	11	13.648	9	6.352	20
	10	8	6.277	5	6.723	13

- Expected/predicted counts by model.
- The smaller the differences
   between Observed vs
  - Expected  $\rightarrow$  Better model fit to data.

#### 4 Oct 2016

### **3b. Classification table**

- CAD & No CAD subjects observed vs predicted/classified by *Preliminary Final Model*.
- % correctly classified > 70% is expected for good model fit.

### Results

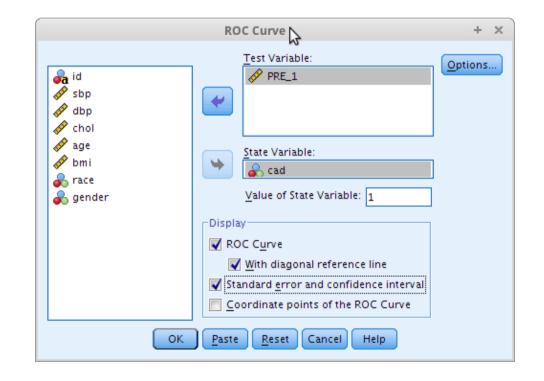
Classification Table <sup>a</sup>										
Predicted										
			cad coronary a	artery disease						
	Observed		0 no cad	1 cad	Percentage Correct	2				
Step 1	cad coronary artery	0 no cad	157	6	96.	.3				
	disease	1 cad	34	3	8.	.1				
	Overall Percentage				80.	.0				

- 80% of subjects are correctly classified by the model.
- Good model fit to the data.

a. The cut value is .500

## **3c.** Area under ROC curve (AUC)

- A measure of ability of the model to discriminate CAD vs Non CAD subjects.
- AUC > 0.7 is acceptable fit.
- AUC ≤ 0.5 no discrimination at all, not acceptable.
- Steps
  - Analyze → ROC curve... → Assign Test
     Variable: Predicted probability (PRE\_1),
     State Variable: cad, Value of State
     Variable: 1.
  - Under Display tick ROC Curve, With diagonal reference line and Standard Error and confidence interval.



#### 4 Oct 2016

### **3c.** Area under ROC curve (AUC)

### Results

#### Area Under the Curve

Test Result Variable(s):PRE\_1 Predicted probability

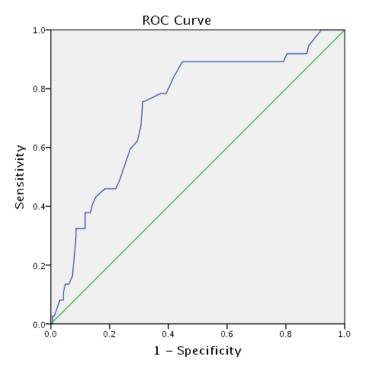
			Asymptotic 95 Inte	
Area	Std. Error <sup>a</sup>	Asymptotic Sig. <sup>b</sup>	Lower Bound	Upper Bound
<sub>I</sub> .732	.045	.000	.643	.821

The test result variable(s): PRE\_1 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

- AUC=0.73 > 0.7.
- 95% CI: 0.64, 0.82.
- Lower limit slightly < 0.7, still acceptable > 0.5.
- Good model fit to the data.



Diagonal segments are produced by ties.

### 3. Model fit assessment

- All 3 methods indicate good model fit of *Preliminary Final Model*.
- Can conclude the model with dbp & gender → Final Model.



# 2. Final Model interpretation & presentation

### • The Final Model.

	Variables in the Equation											
						_		95% C.I.f	or EXP(B)			
		В	S.E.	Wald	df	Sig.	Exp(B)	Lower	Upper			
Step 1ª	dbp	.050	.015	11.444	1	.001	1.051	1.021	1.081			
	gender(1)	.806	.391	4.250	1	.039	2.238	1.040	4.815			
	Constant	-6.120	1.317	21.606	1	.000	.002					

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

#### Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	16.352	2	.000
	Block	16.352	2	.000
	Model	16.352	2	.000

#### Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change	
Step 1	dbp	-95.778	11.933	1	.001	} ∕
Step 2	dbp	-93.747	12.289	1	.000	Y
	gender	-89.812	4.419	1	.036	

*P*-values by LR test for both *dbp* & *gender* by Enter method. • *P*-values by Wald test per variable by Enter method.

• Take this adjusted OR.

*P*-values by LR per variable. Obtained with Forward LR method.

#### 4 Oct 2016

## 4. Final Model interpretation & presentation

• Associated factors of coronary artery disease.

Fac	ctors	b	Adjusted OR (95% CI)	P-value <sup>a</sup>
Diastolic Bl	ood Pressure	0.05	1.05 (1.02, 1.08)	< 0.001
Gender	Man vs Woman	0.81	2.24 (1.04, 4.82)	0.036
<sup>a</sup> LR test 1mmHg increase increase odds o by 1.05 times, v controlling for g	of CAD while		CAD as com	24 times odds o pared to woma trolling for DBP.
	)mmHg increase i ) = exp(10 x 0.05)		= 1.65 times.	



