Multiple Logistic Regression

Dr. Wan Nor Arifin

Unit of Biostatistics and Research Methodology,
Universiti Sains Malaysia.

wnarifin@usm.my / wnarifin.github.io

Wan Nor Arifin. 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

Objectives

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.

Introduction

- **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?**

Introduction

- **Multiple Linear Regression**
  - $y = a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$
- **Multiple Logistic Regression**
  - log(odds) $= a + b_1x_1 + b_2x_2 + \ldots + b_nx_n$
  - That's why it is called “logistic” regression.

Introduction

- **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.**

Introduction

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

Introduction

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

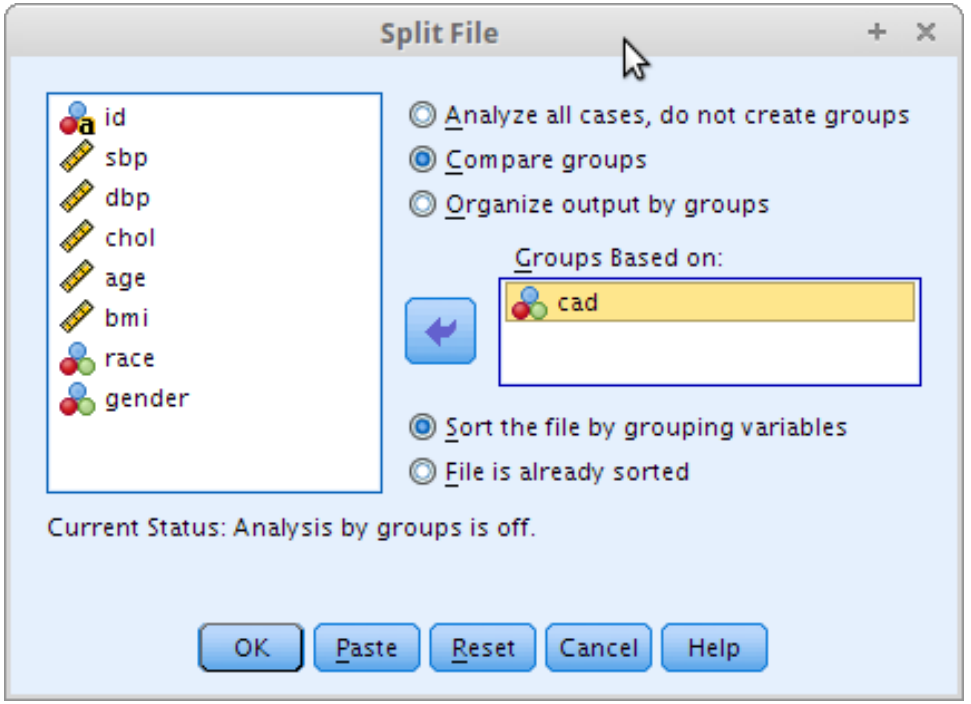
- **Dataset: *slog.sav***
- **Sample size, n=200**
- **DV: *cad* (1: Yes, 0: No)**
- **IVs:**
  - Numerical: *sbp* (systolic blood pressure), *dbp* (diastolic blood pressure), *chol* (serum cholesterol in mmol/L), *age* (age in years), *bmi* (Body Mass Index).
  - Categorical: *race* (0: Malay, 1: Chinese,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

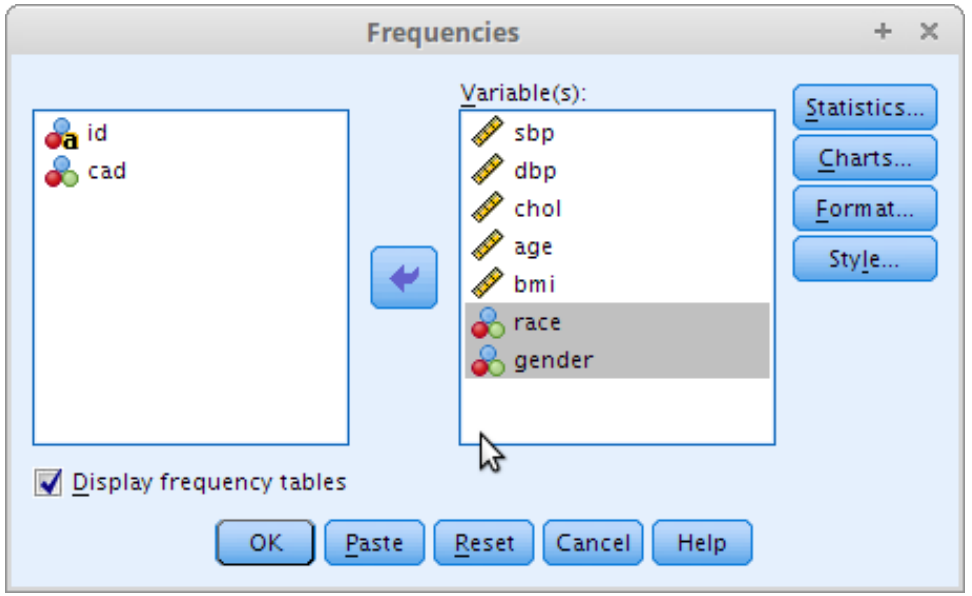
1. Descriptive statistics

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



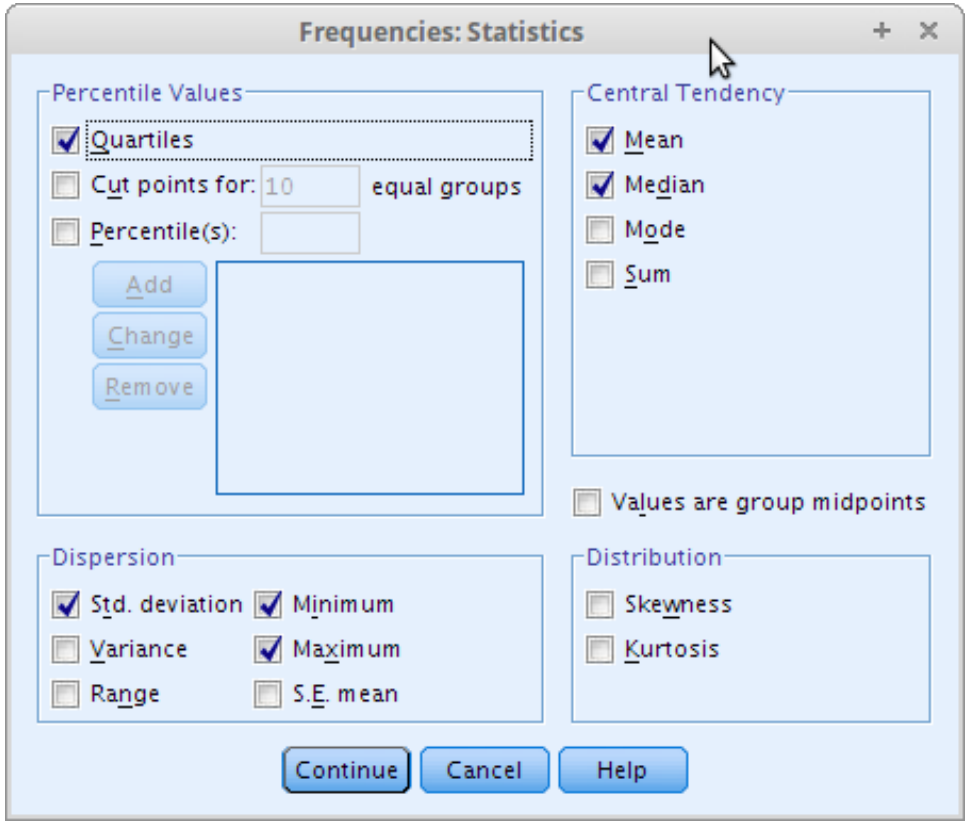
1. Descriptive statistics

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



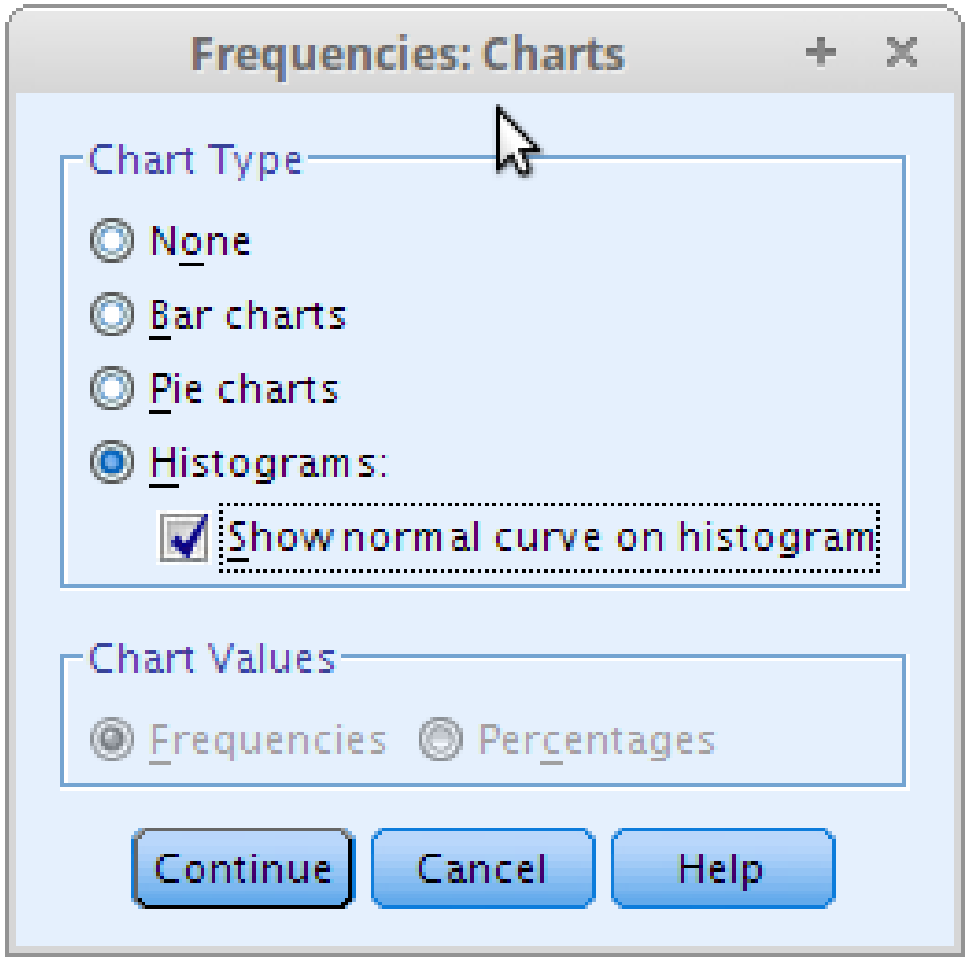
1. Descriptive statistics

- Cont...
  - **Statistics** → tick → **Continue**



1. Descriptive statistics

- Cont...
  - **Charts** → tick → **Continue** → **OK**



1. Descriptive statistics

- Results

Statistics

cad coronary artery disease			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. Deviation		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. Deviation		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

1. Descriptive statistics

- Results

race ethnicity

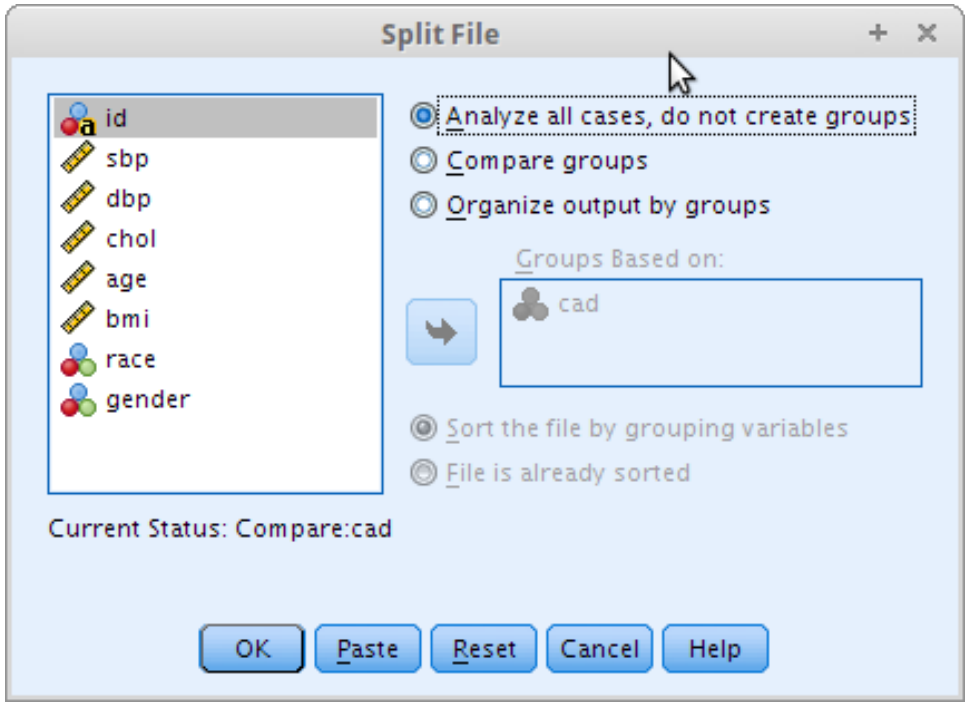
cad coronary artery disease			Frequency	Percent	Valid Percent	Cumulative Percent
0 no cad	Valid	0 malay	60	36.8	36.8	36.8
		1 chinese	52	31.9	31.9	68.7
		2 indian	51	31.3	31.3	100.0
		Total	163	100.0	100.0	
1 cad	Valid	0 malay	13	35.1	35.1	35.1
		1 chinese	12	32.4	32.4	67.6
		2 indian	12	32.4	32.4	100.0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

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	

1. Descriptive statistics

- **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**



1. Descriptive statistics

- Present the results in a table.

Factors		CAD, n=37 mean(SD)	No CAD, n=163 mean(SD)
Systolic Blood Pressure		143.8(25.61)	129.3(22.26)
Diastolic Blood Pressure		89.0(12.17)	80.8(12.61)
Cholesterol		6.6(1.17)	6.1(1.17)
Age		47.4(8.80)	45.2(8.41)
BMI		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

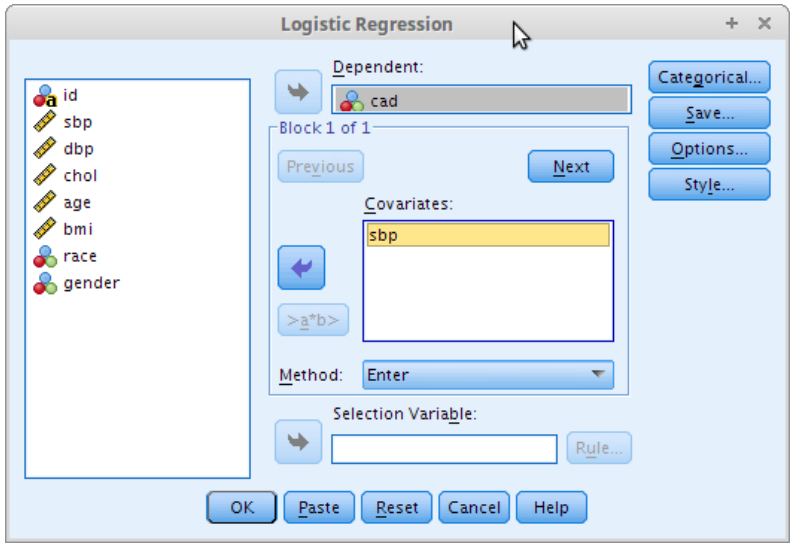
- **To select best variables to predict the outcome.**
- **Sub-steps:**
  a. Univariable analysis.
  b. Multivariable analysis.
  c. Checking multicollinearity & interactions.

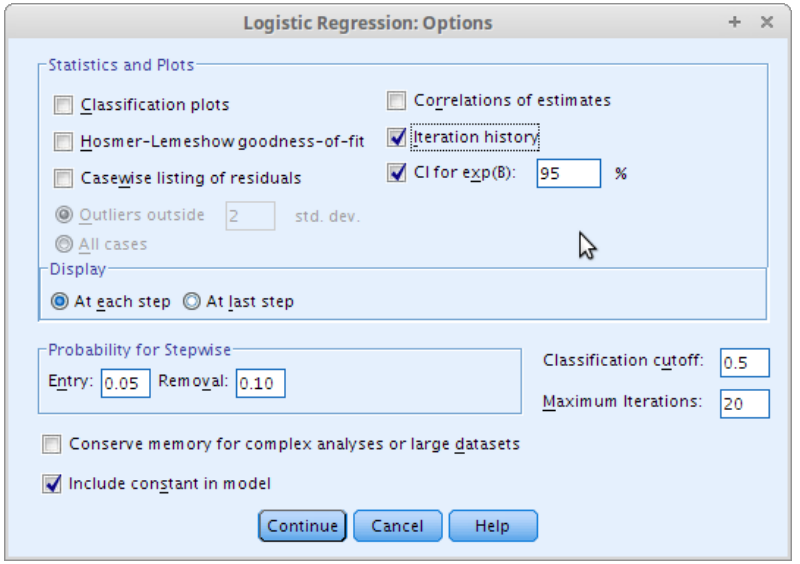
2a. Univariable analysis

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

2a. Univariable analysis

- 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*





2a. Univariable analysis

- **Results**

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	10.464	1	.001
	Block	10.464	1	.001
	Model	10.464	1	.001

Model: SBP *P*-value=0.001 by Likelihood Ratio (LR) test

SBP *P*-value=0.001 by Wald test

Variables in the Equation

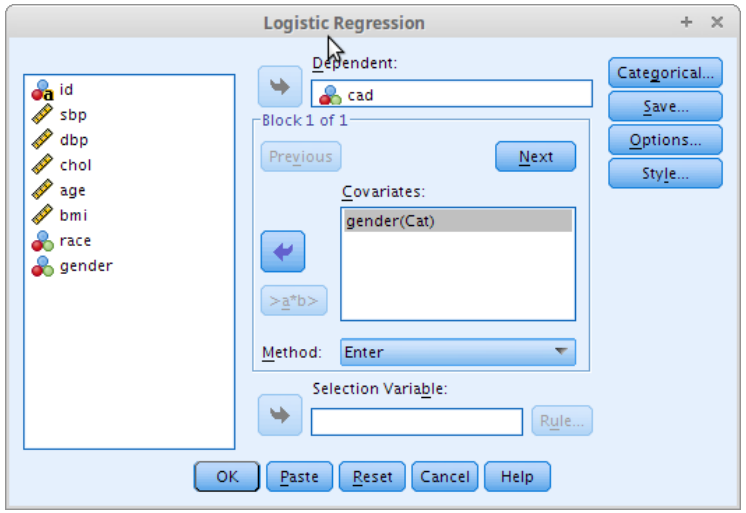
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	sbp	.024	.007	10.290	1	.001	1.024	1.009	1.039
	Constant	-4.684	1.039	20.303	1	.000	.009		

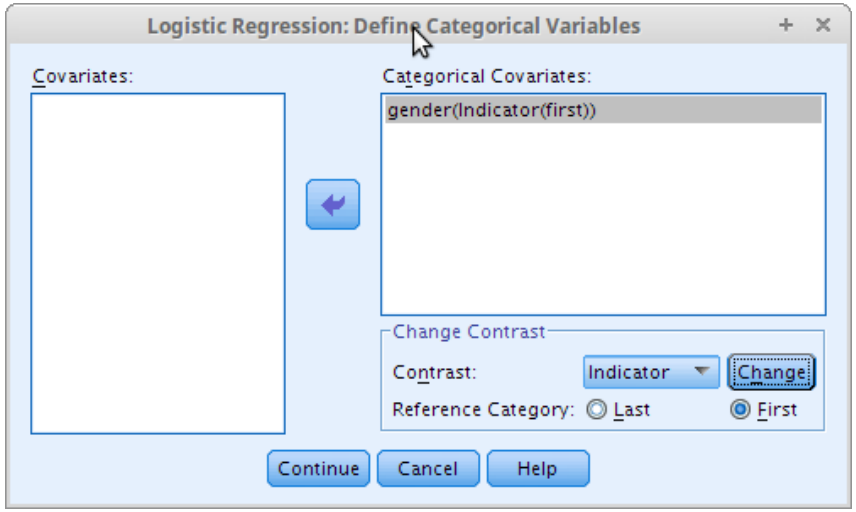
a. Variable(s) entered on step 1: sbp.

- Exp(B) is OR.
- OR(1 unit ↑ in SBP) =1.04(95% CI: 1.01, 1.04). Unadjusted/ Crude OR.
- Interpretation: 1mmHg increase in SBP increase odds of CAD by 1.02 times.
- In variable selection context, less concern about OR & interpretation.

2a. Univariable analysis

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





2a. Univariable analysis

- **Results**

Categorical Variables Codings

		Frequency	Parameter coding (1)
gender gender	0 woman	100	.000
	1 man	100	1.000

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	4.063	1	.044
	Block	4.063	1	.044
	Model	4.063	1	.044

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)	.748	.378	3.909	1	.048	2.113	1.007	4.437
	Constant	-1.901	.297	40.870	1	.000	.149		

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

Women=0 becomes the reference group.

Model: Gender *P*-value=0.044 by LR test

Gender *P*-value=0.048 by Wald test

- OR(male)=2.11(95% CI: 1.01, 4.44). Unadjusted/Crude OR.
- Interpretation: Man has 2.11 times odds of CAD as compared to woman.

2a. Univariable analysis

- *P*-values of IVs – select *P*-value < 0.25

Factors		*P*-value (Wald test)	*P*-value (LR test)
Systolic Blood Pressure		***0.001***	***0.001***
Diastolic 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***

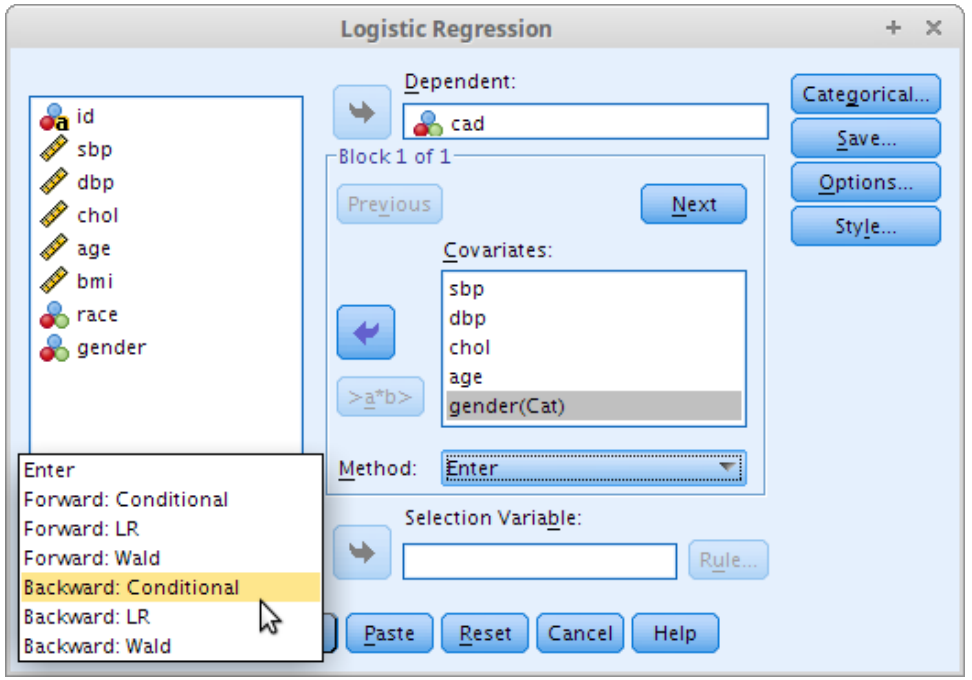
*For both variables

2b. Multivariable analysis

- **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.**

2b. Multivariable analysis

- **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).

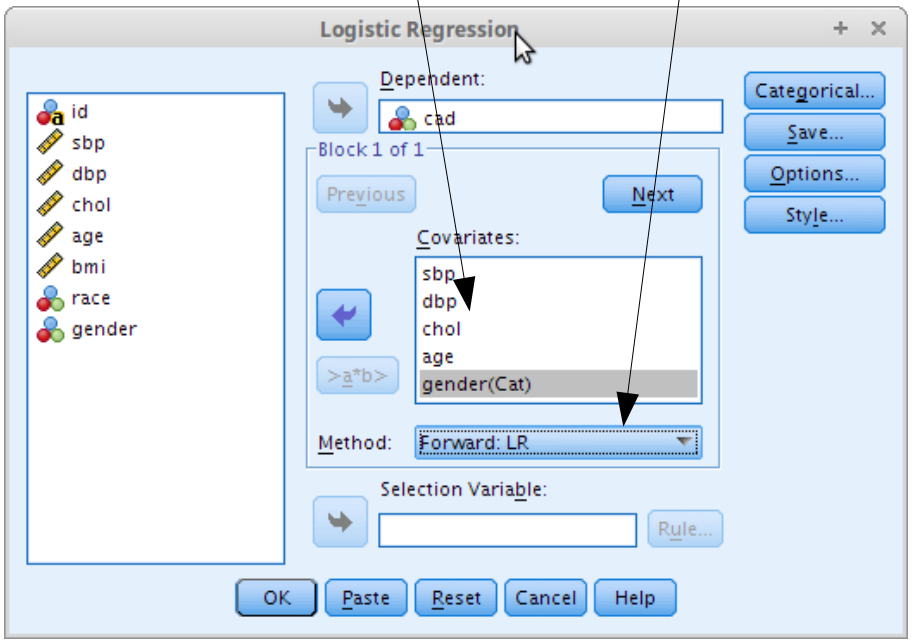


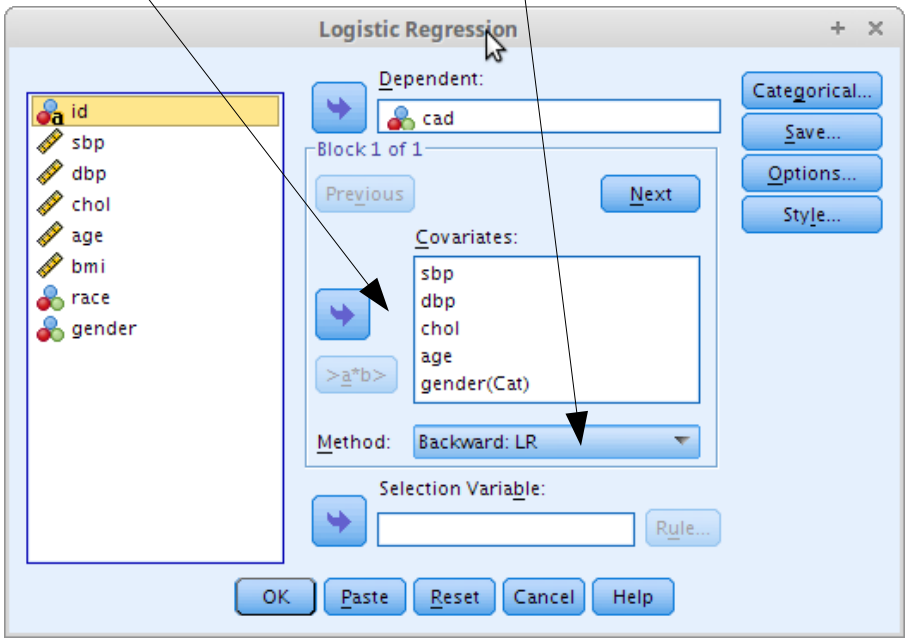
2b. Multivariable analysis

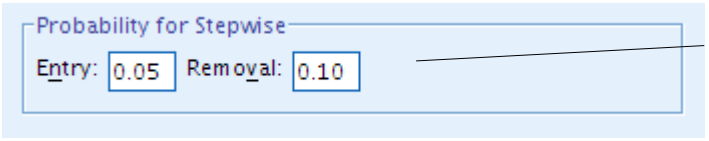
- **Variable Selection in this workshop:**
  - Automatic by Forward & Backward LR.
  - Selection of variables by *P*-values based on LR test.

2b. Multivariable analysis

- Enter all selected variables.
- Perform 2x – 1x Forward LR, 1x Backward LR.







Options: Just leave at the default values.

2b. Multivariable analysis

- **Results**

Forward LR

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	dbp	.049	.015	11.298	1	.001	1.050	1.021	1.080
	Constant	-5.620	1.277	19.358	1	.000	.004		
Step 2[b]	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.
b. Variable(s) entered on step 2: gender.

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

Backward 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	.009	.014	.371	1	.542	1.009	.981	1.037
	dbp	.034	.025	1.799	1	.180	1.034	.985	1.086
	chol	.187	.188	.987	1	.321	1.205	.834	1.742
	age	-.016	.028	.335	1	.563	.984	.931	1.040
	gender	.755	.401	3.544	1	.060	2.127	.969	4.667
	Constant	-6.334	1.524	17.272	1	.000	.002		
Step 2[a]	sbp	.006	.013	.183	1	.668	1.006	.980	1.031
	dbp	.035	.025	1.965	1	.161	1.036	.986	1.087
	chol	.162	.182	.796	1	.372	1.176	.823	1.681
	gender	.728	.398	3.351	1	.067	2.070	.950	4.512
	Constant	-6.623	1.449	20.882	1	.000	.001		
Step 3[a]	dbp	.043	.016	7.290	1	.007	1.044	1.012	1.077
	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 4[a]	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: sbp, dbp, chol, age, gender.

2b. Multivariable analysis

- **Results**

Forward LR

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

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

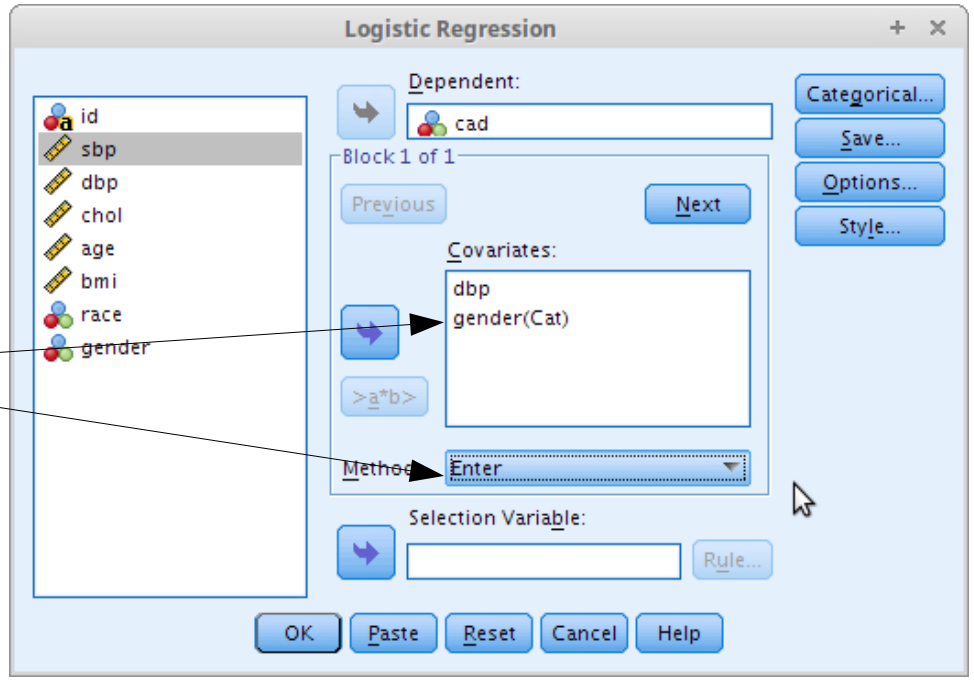
Model if Term Removed

Variable		Model Log Likelihood	Change in -2 Log Likelihood	df	Sig. of the Change
Step 1	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

2c. Multicollinearity

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



Variables in the Equation

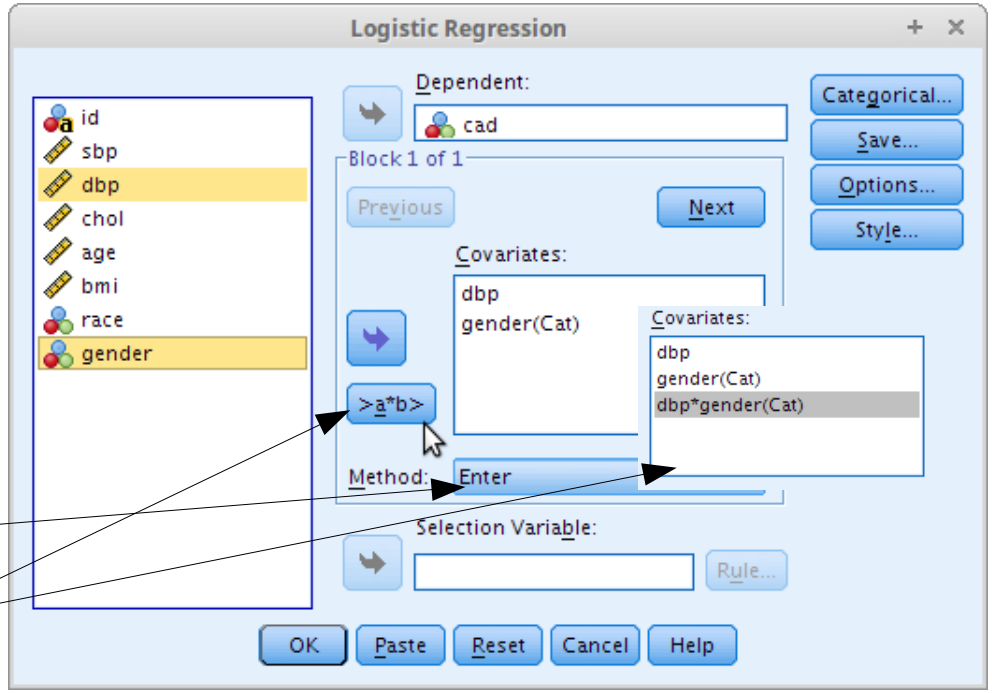
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	dbp	.050	.015	11.444	1	.001	1.051	1.021	1.081
	gender	.806	.391	4.250	1	.039	2.238	1.040	4.815
	Constant	-6.120	1.317	21.606	1	.000	.002		

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 Enter method with *dbp, gender & dbp x gender*. Select both *dbp* & *gender* (hold Ctrl on keyboard) → Click >a*b>



2d. Interactions

- 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	.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 → *Preliminary 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	.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.

- *P*-values by Wald test per variable by Enter method.
- Take this adjusted OR.

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

P-values by LR test for both *dbp* & *gender* by Enter method.

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

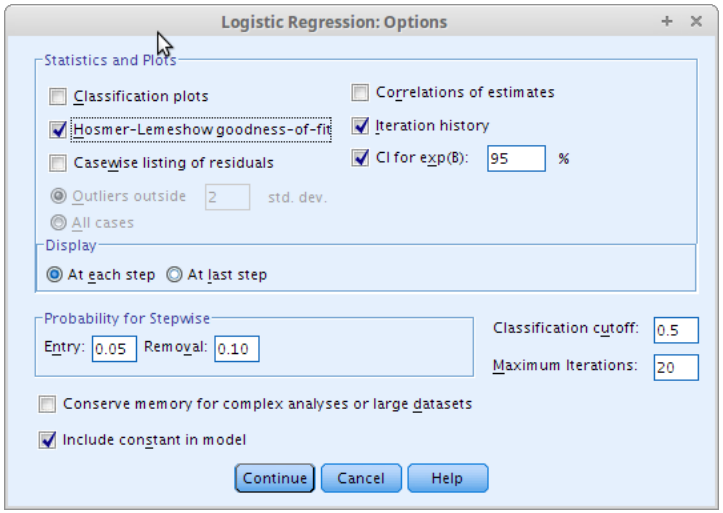
P-values by LR per variable. Obtained with Forward LR method.

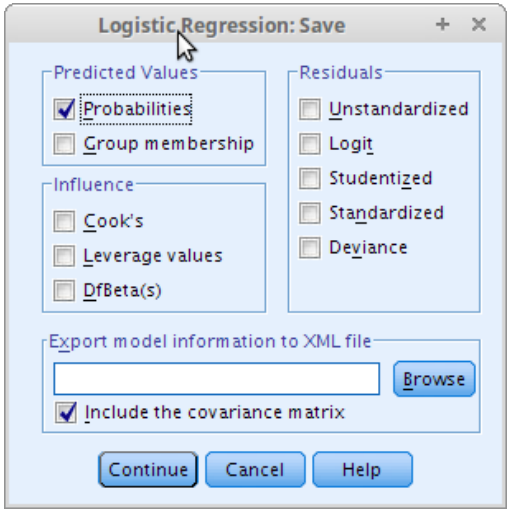
3. Model fit assessment

- **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.





3a. Hosmer-Lemeshow test

- Indicates fit of *Preliminary Final Model* to data.
- Results

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	13.626	8	.092

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

Contingency Table for Hosmer and Lemeshow Test

		cad coronary artery disease = 0 no cad		cad coronary artery disease = 1 cad		
		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
	6	16	18.009	6	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

Observed counts in data.

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

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[a]

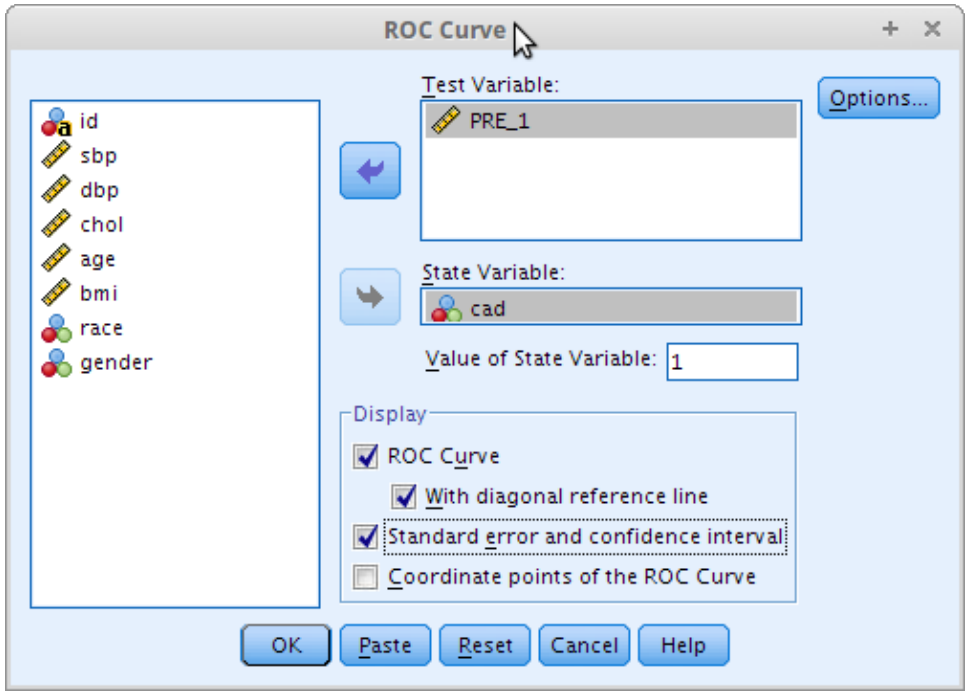
Observed			Predicted		
			cad coronary artery disease		Percentage Correct
			0 no cad	1 cad	
Step 1	cad coronary artery disease	0 no cad	157	6	96.3
		1 cad	34	3	8.1
	Overall Percentage				80.0

a. The cut value is .500

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

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** → **Classify** → **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**.



3c. Area under ROC curve (AUC)

- Results

Area Under the Curve

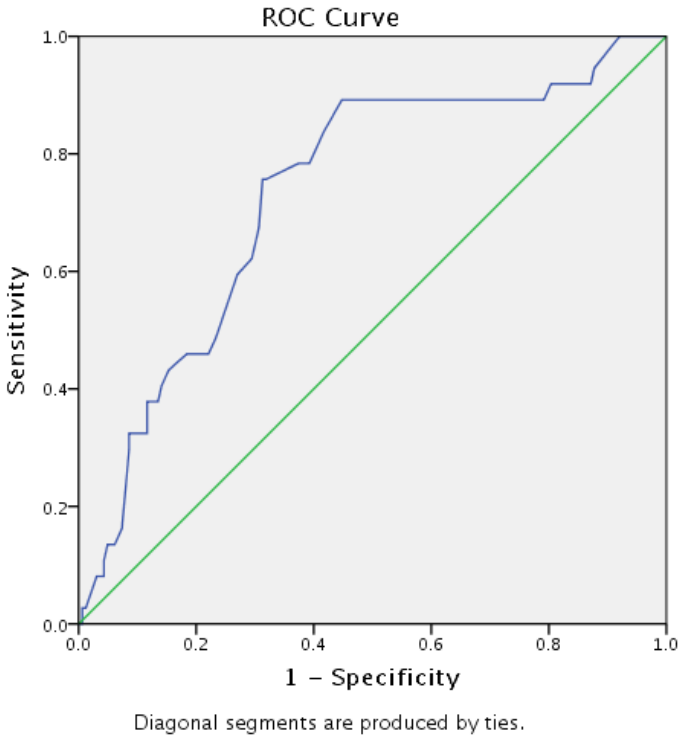
Test Result Variable(s):PRE_1 Predicted probability

Area	Std. Error[a]	Asymptotic Sig.[b]	Asymptotic 95% Confidence Interval	
			Lower Bound	Upper Bound
.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.



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

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	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.

- *P*-values by Wald test per variable by Enter method.
- Take this adjusted OR.

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

P-values by LR test for both *dbp* & *gender* by Enter method.

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

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

4. Final Model interpretation & presentation

- Associated factors of coronary artery disease.

Factors		*b*	Adjusted OR (95% CI)	*P*-value[a]
Diastolic Blood 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

[a]LR test

1mmHg increase in DBP increase odds of CAD by 1.05 times, while controlling for gender.

Man has 2.24 times odds of CAD as compared to woman, while controlling for DBP.

To obtain for 10mmHg increase in DBP
OR = exp(c x *b*) = exp(10 x 0.05) = exp(0.5) = 1.65 times.

Q&A