Simple Logistic Regression

Dr. Wan Nor Arifin

Biostatistics and Research Methodology Unit,
Universiti Sains Malaysia

wnarifin@usm.my / wnarifin.github.io

Outlines

- Introduction
- Odds ratio vs relative risk
- Simple logistic regression
- Analysis in SPSS
    1. Descriptive Statistics
    2. Univariable Analysis
    3. Interpretation

Objectives

1. Understand the concepts of odds and risk, and their relations with logistic regression
2. Perform simple logistic regression in SPSS
3. Identify and interpret the results

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.
- **Linear Regression?**
  - Dependent Variable, DV: ???

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.
- **Linear Regression?**
  - Dependent Variable, DV: Numerical

Introduction

- **Simple Linear Regression**
  - $y = a + bx$
- **Simple Logistic Regression**
  - log(odds) = $a + bx$
  - That's why it is called “logistic” regression
  - Allows us to obtain ***odds ratio***

Odds ratio vs relative risk

- Association analysis for cross-tabulation of a binary factor with a binary outcome can be expressed as odds ratio.
- Odds is a measure of chance of disease occurrence in a specified group,

$$Odds = \frac{n_{disease}}{n_{no\ disease}}$$

Odds ratio vs relative risk

- Odds ratio, OR is the ratio between the odds of two groups; the group with the risk factor and the group without the risk factor,

$$Odds\,ratio\,,\mathrm{OR} = \frac{Odds_{factor}}{Odds_{no\,factor}}$$

- Odds ratio is applicable to all observational study designs (cohort, cross-sectional and case-control) -- does not imply a cause-effect association, but only plain association.

Odds ratio vs relative risk

- In epidemiology, the association between a risk factor and a disease is expressed in terms of risk and relative risk.
- Risk is a measure of chance of disease occurrence in a specific group,

$$Risk = \frac{n_{disease}}{n_{group}}$$

Odds ratio vs relative risk

- Relative risk is the ration between the risk in the group with the factor and the risk in the group without the risk factor,

$$Relative\ risk\ , \mathrm{RR} = \frac{Risk_{factor}}{Risk_{no\ factor}}$$

- Relative risk is only appropriate to calculate risk and relative risk for cohort studies, because the cause-effect relationship is well defined.

Odds ratio vs relative risk

- **Odds Ratio, OR**
  - Applicable to all observational studies.
- **Relative Risk, RR**
  - Only cohort study.
- **OR ≈ RR for rare disease, useful to determine risk from a case-control study.**

Odds ratio vs relative risk

Factor vs Disease	Lung CA	No Lung CA
Smoker	24 [a]	76 [b]
Non-smoker	13 [c]	87 [d]

- Odds(smoker) = a/b = 24/76 = 0.32
- Odds(non-smoker) = c/d = 13/87 = 0.15
- OR($Odds_{smoker}/Odds_{non\text{-}smoker}$) = 0.32/0.15 = 2.13
- Shortcut, OR = ad/bc = (24x87)/(76x13) = 2.11

Odds ratio vs relative risk

Factor vs Disease	Lung CA	No Lung CA
Smoker	24 [a]	76 [b]
Non-smoker	13 [c]	87 [d]

- Risk(smoker) = Proportion CAD = a/(a+b) = 0.24
- Risk(non-smoker) = Proportion CAD c/(c+d) = 0.13
- RR($Risk_{smoker}$/$Risk_{non\text{-}smoker}$) = 0.24/0.13 = 1.85 ≈ OR, 2.11

Simple Logistic Regression

- **Simple Logistic Regression**
  - log(odds) = $a + bx$
  - That's why it is called “logistic” regression
  - Allows us to obtain ***odds ratio***
- **Odds ratio,**

$$\mathrm{OR} = \exp(b)$$

Analysis in SPSS

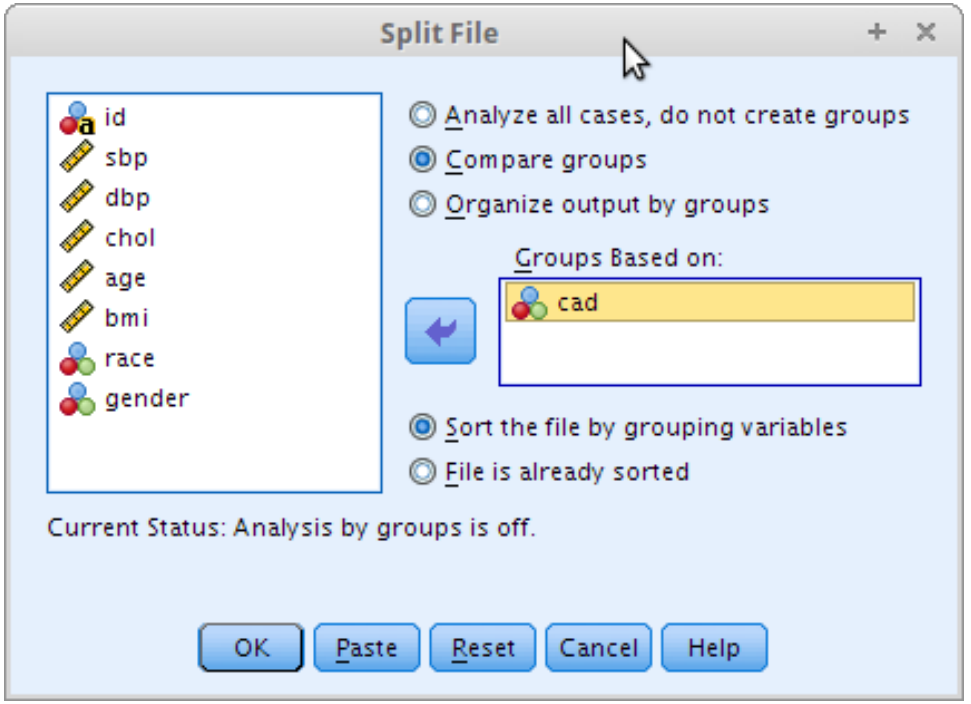
- **Dataset: *slog.sav***
- **Sample size, n=200**
- **DV: *cad* (1: Yes, 0: No)**
- **IVs:**
  - Numerical: *sbp* (systolic blood pressure), *dbp* (diastolic blood pressure), *chol* (serum cholesterol in mmol/L), *age* (age in years), *bmi* (Body Mass Index).
  - Categorical: *race* (0: Malay, 1: Chinese,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

Steps in Multiple Logistic Regression

1. Descriptive statistics
2. Univariable analysis
3. Interpretation

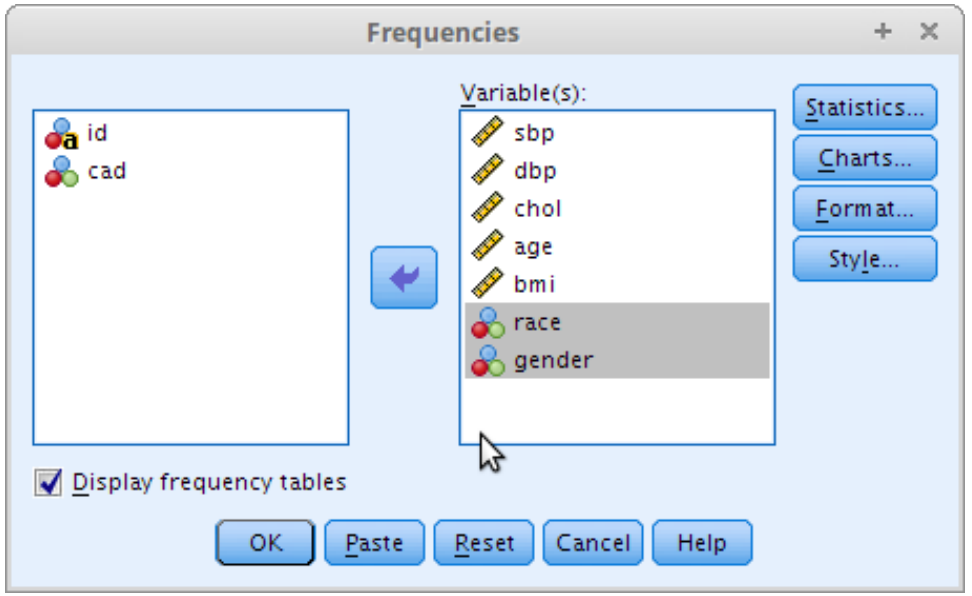
1. Descriptive statistics

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



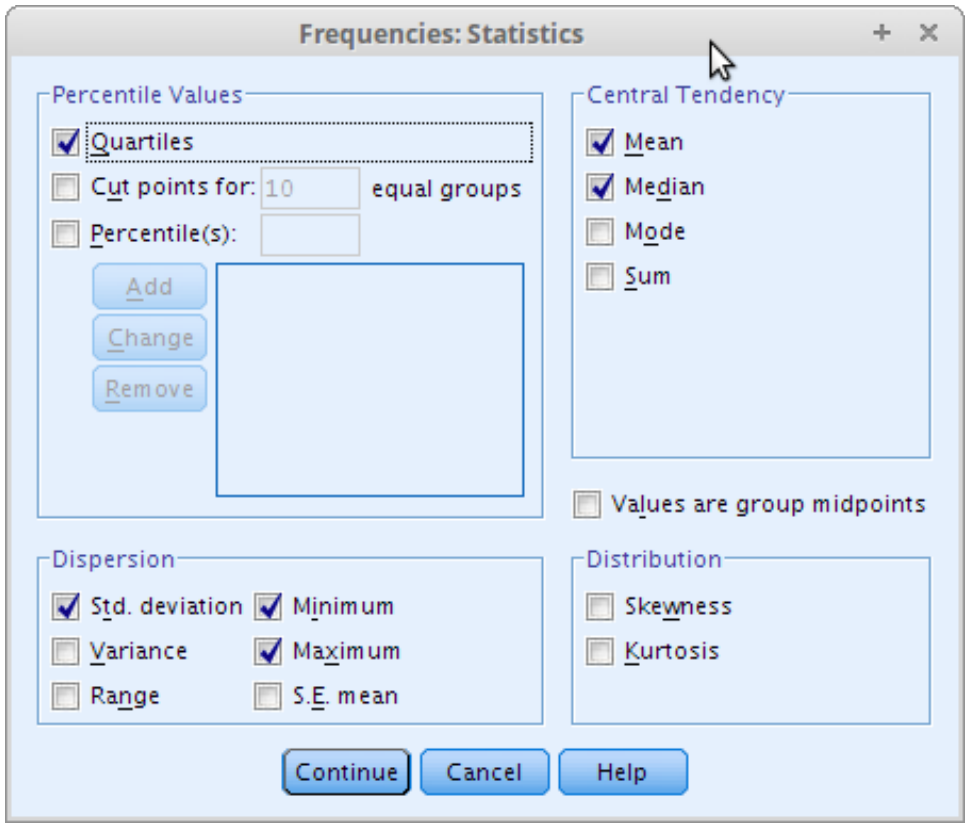
1. Descriptive statistics

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



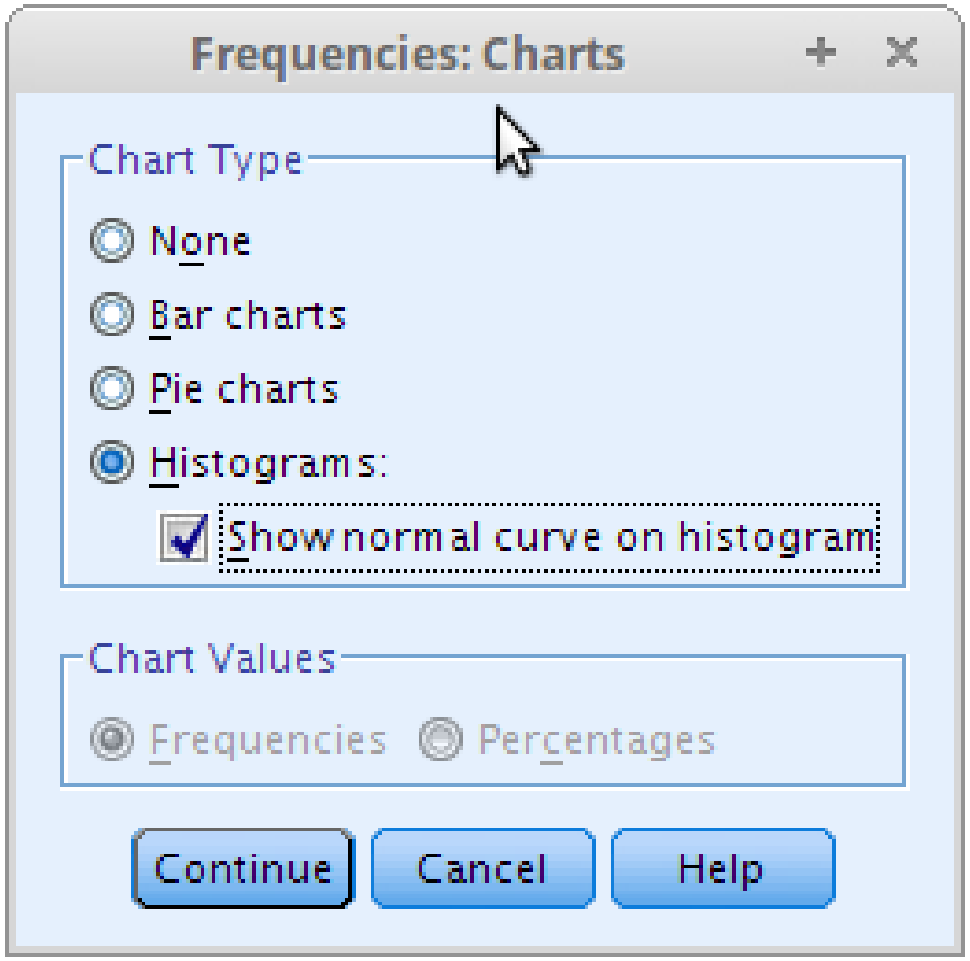
1. Descriptive statistics

- 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

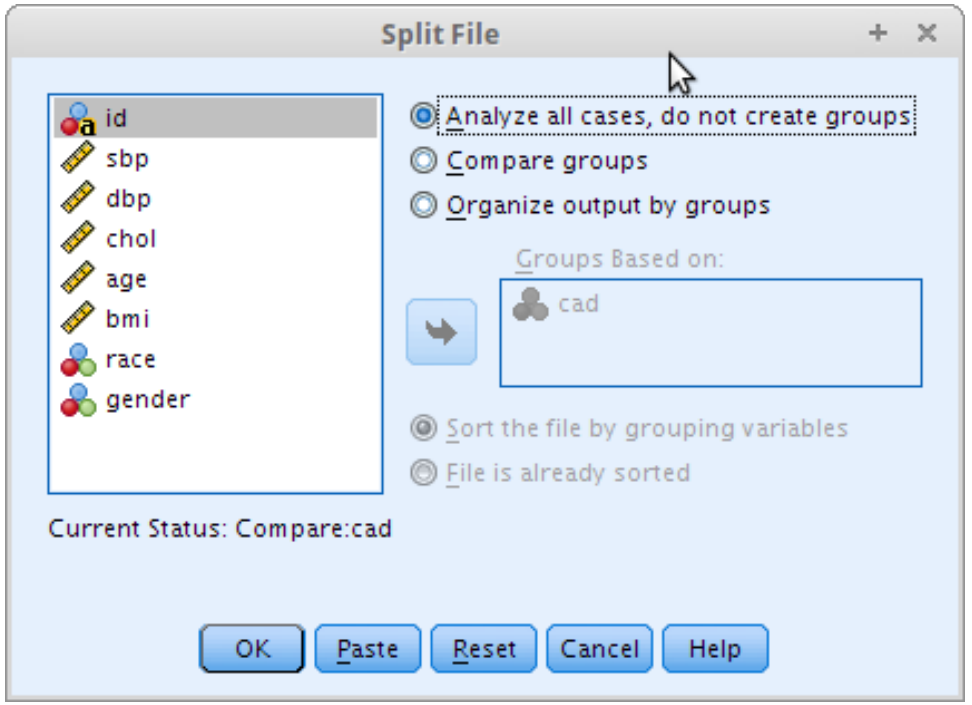
cad coronary artery disease			Frequency	Percent	Valid Percent	Cumulative Percent
0 no cad	Valid	0 malay	60	36.8	36.8	36.8
		1 chinese	52	31.9	31.9	68.7
		2 indian	51	31.3	31.3	100.0
		Total	163	100.0	100.0	
1 cad	Valid	0 malay	13	35.1	35.1	35.1
		1 chinese	12	32.4	32.4	67.6
		2 indian	12	32.4	32.4	100.0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

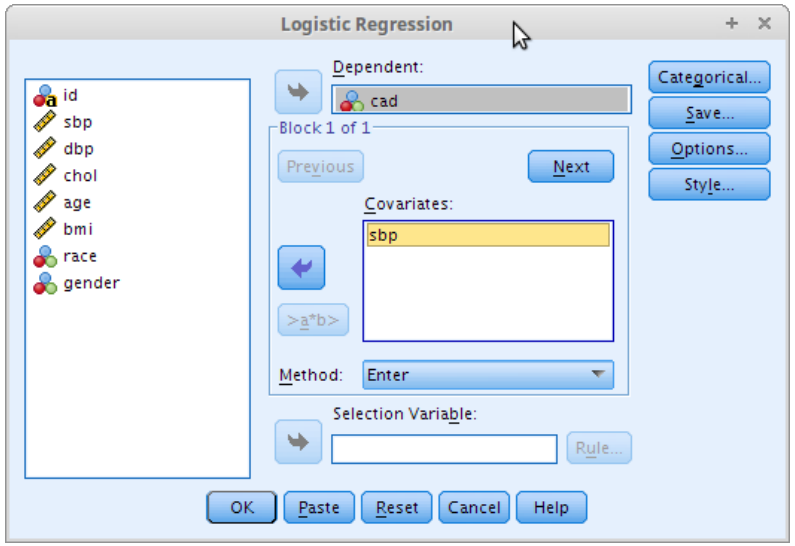
**n* (%)

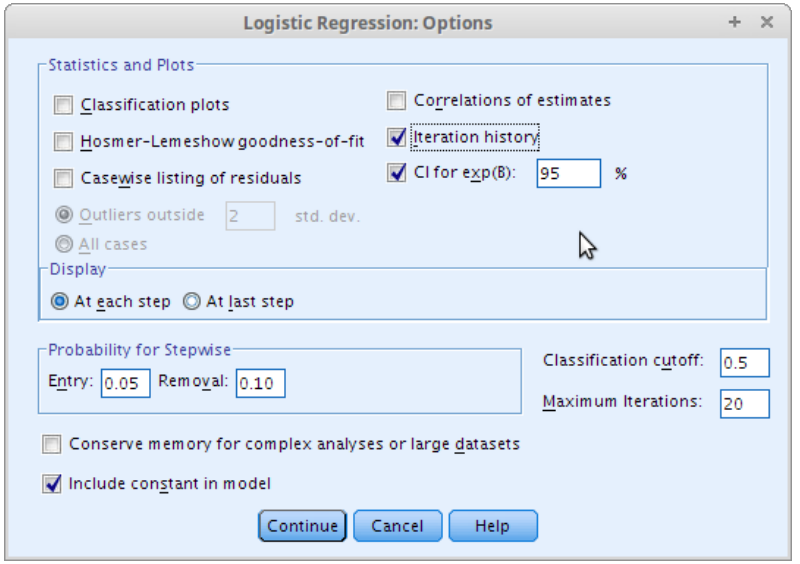
2. Univariable analysis

- Perform Simple Logistic Regression on each IV
- Pay attention to whether IV is numerical or categorical

2. 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*





2. Univariable analysis

- **Results**

SBP *P*-value=0.001 by Wald test

Variables in the Equation

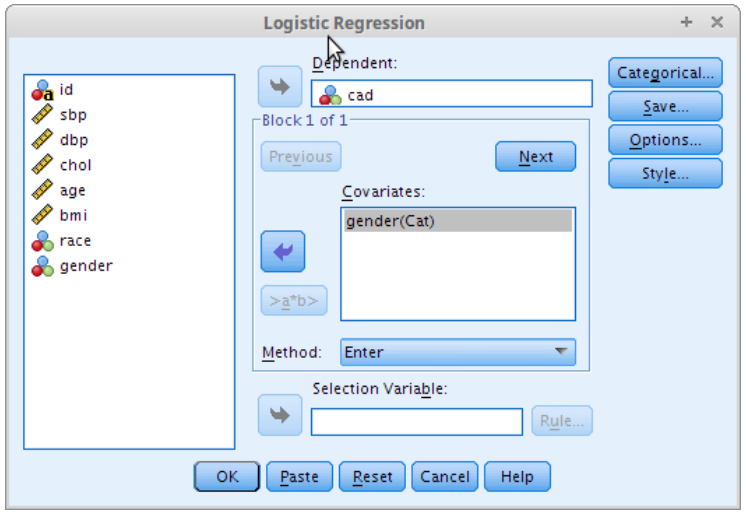
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	sbp	.024	.007	10.290	1	.001	1.024	1.009	1.039
	Constant	-4.684	1.039	20.303	1	.000	.009		

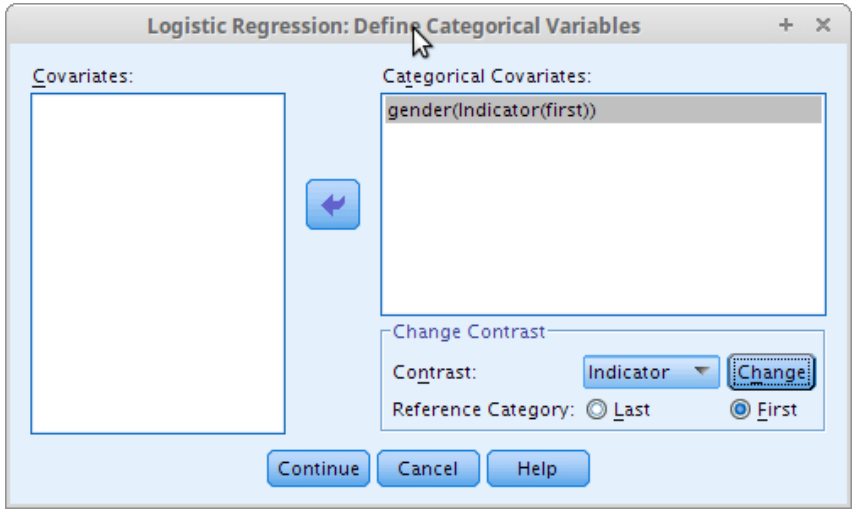
a. Variable(s) entered on step 1: sbp.

- Exp(B) is OR.
- OR(1 unit ↑ in SBP) =1.04(95% CI: 1.01, 1.04)
- Interpretation: 1mmHg increase in SBP increase odds of CAD by 1.02 times

2. Univariable analysis

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





2. Univariable analysis

- Results

Women=0 becomes the reference group.

Categorical Variables Codings

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

Gender *P*-value=0.048 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]	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.

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

2. Univariable analysis

- *OR values and P*-values of IVs

Let's fill in the blanks

Factors		*b*	SE	OR (95% CI)	*P*-value
Systolic Blood Pressure		0.02	0.01	1.02 (1.01, 1.04)	0.001
Diastolic Blood Pressure					
Cholesterol					
Age					
BMI					
Race	**Chinese-vs-Malay** **Indian-vs-Malay**				
Gender	***Man-vs-Woman***	0.75	0.38	2.11 (1.01, 4.44)	0.048

3. Interpretation

- Simple logistic regression of associated factors of coronary artery disease

Factors		*b*	*SE*	OR (95% CI)	*P*-value
Systolic Blood Pressure		0.02	0.01	1.02 (1.01, 1.04)	0.001
Gender	Man vs Woman	0.75	0.38	2.11 (1.01, 4.44)	0.048

1mmHg increase in SBP increase odds of CAD by 1.02 times, while controlling for gender

Man has 2.11 times odds of CAD as compared to woman, while controlling for SBP

To obtain for 10mmHg increase in SBP
OR = exp(c x *b*) = exp(10 x 0.05) = exp(0.5) = 1.22 times

Q&A