St. PAUL INSTITUTE
FOR REPRODUCTIVE HEALTH AND RIGHTS

Binary logistic regression

Mekitie Wondafrash , MD, PhD

St. Paul Institute for Reproductive Health and Rights

Learning objectives

By the end of this session, learners should be able to :

- Understand the concepts of logistic regression.
- Understand the principles and theory underlying logistic regression
- Understand proportions, probabilities, odds, odds ratios, logits, and exponents
- Be able to implement multiple logistic regression analyses using SPSS and accurately interpret the output
- Understand the assumptions underlying logistic regression analyses and how to test them

Introduction to logistic regression

- Logistic regression is a powerful statistical technique used to model the relationship between a **dichotomous dependent variable** (e.g. “yes” or “no,” “success” or “failure,” or “pass” or “fail”) and one or more **independent variables** (predictors).
- We perform logistic regression to predict the **probability** that an observation falls into one of two categories based on the predictor variables.

Logistic regression equation

- Logistic regression is a Generalized Linear Model where the outcome is a two-level categorical variable.
- The outcome, Yi, takes the value 1 with probability pi and the value 0 with probability 1 − pi.
  - logit(pi) = log(pi/ 1 − pi)

Binary (binomial) logistic regression

- Used when you want to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables.
- It is similar to a linear regression model but is suited to models where the dependent variable is dichotomous.
- Logistic regression coefficients can be used to estimate odds ratios for each of the independent variables in the model.

Contents

Key information from logistic regression

Direction
- Negative Odds ratio below 1
- Positive Odds ratio above 1

- Effect size
  - Odds ratio: The odds of the outcome being a case divided by the odds that the outcome is a non-case, for every one-unit increase in x

- Statistical significance
  - P-value
    - $P<0.05$ Statistically significant at the 5 % level
    - $P<0.01$ Statistically significant at the 1 % level

- 95 % Confidence Intervals
  - Interval includes 1:
    - Statistically non significant at the 5 % level
  - Interval does not include 1:
    - Statistically significant at the 5 % level

Odds ratio

- Logistic regression is based on the fact that the outcome has only two possible values: 0 or 1.
- Often, 1 is used to denote a “case” whereas 0 is then a “non-case”.
- Logistic regression is used to predict the “odds” of being a “case” based on the values of the x-variable(s)
- Just as for linear regression analysis, we get a coefficient (log odds) that shows the effect of x on y.
- Odds Ratio is calculated by taking the “exponent” of the coefficient: “exp(B)”

What Is an Odds Ratio?

- An odds ratio indicates how much more likely, with respect to odds, a certain event occurs in one group relative to its occurrence in another group.

Probability of Outcome

	Outcome: Yes	Outcome: No	Total
Group A	20	60	80
Group B	10	90	100
Total	30	150	180

Probability of a **Yes outcome** in Group A = 20/80 **(0.25)**

Probability of a **No outcome** in Group A = 60/80 **(0.75)**

You have a 25% chance of getting the outcome in group A.
What is the chance of getting the outcome in group B?

What is odds ratio?

Odds

Odds of Outcome in Group A

Probability of a Yes outcome in Group A ÷ Probability of a No outcome in Group A

$0.25 \div 0.75 = 0.33$

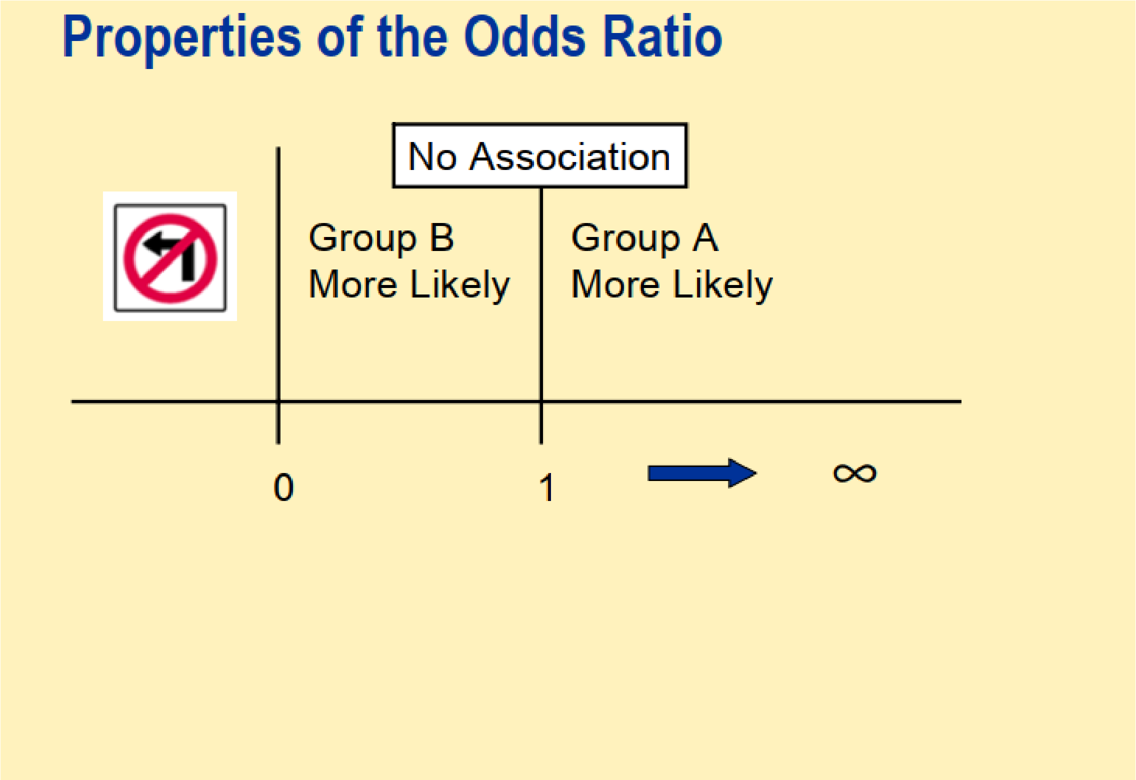
Odds Ratio

Odds Ratio of Group A to Group B

Odds of outcome in Group A ÷ Odds of outcome in Group B

$0.33 \div 0.11 = 3$

What is odds ratio?



- Unlike linear regression, where the null value (i.e. value that denotes no difference) is 0, the null value for logistic regression is 1.
- Also, note that an OR can never be negative – it can range between 0 and infinity.

P-values and confidence intervals and R-squared

P-value and 95% CI

- The p-values and the CI will give you partly different information, but: they are not contradictory.
  - If the p-value is <0.05, the 95 % CI will not include 1
  - If the p-value is above 0.05, the 95 % CI will include 1

R-Squared

- In contrast to linear regression, "R-Squared" or "R^2" is not very usable
- You will, however, get a value for the so-called "Nagelkerke R Square" which is similar to the R-squared

Simple logistic regression

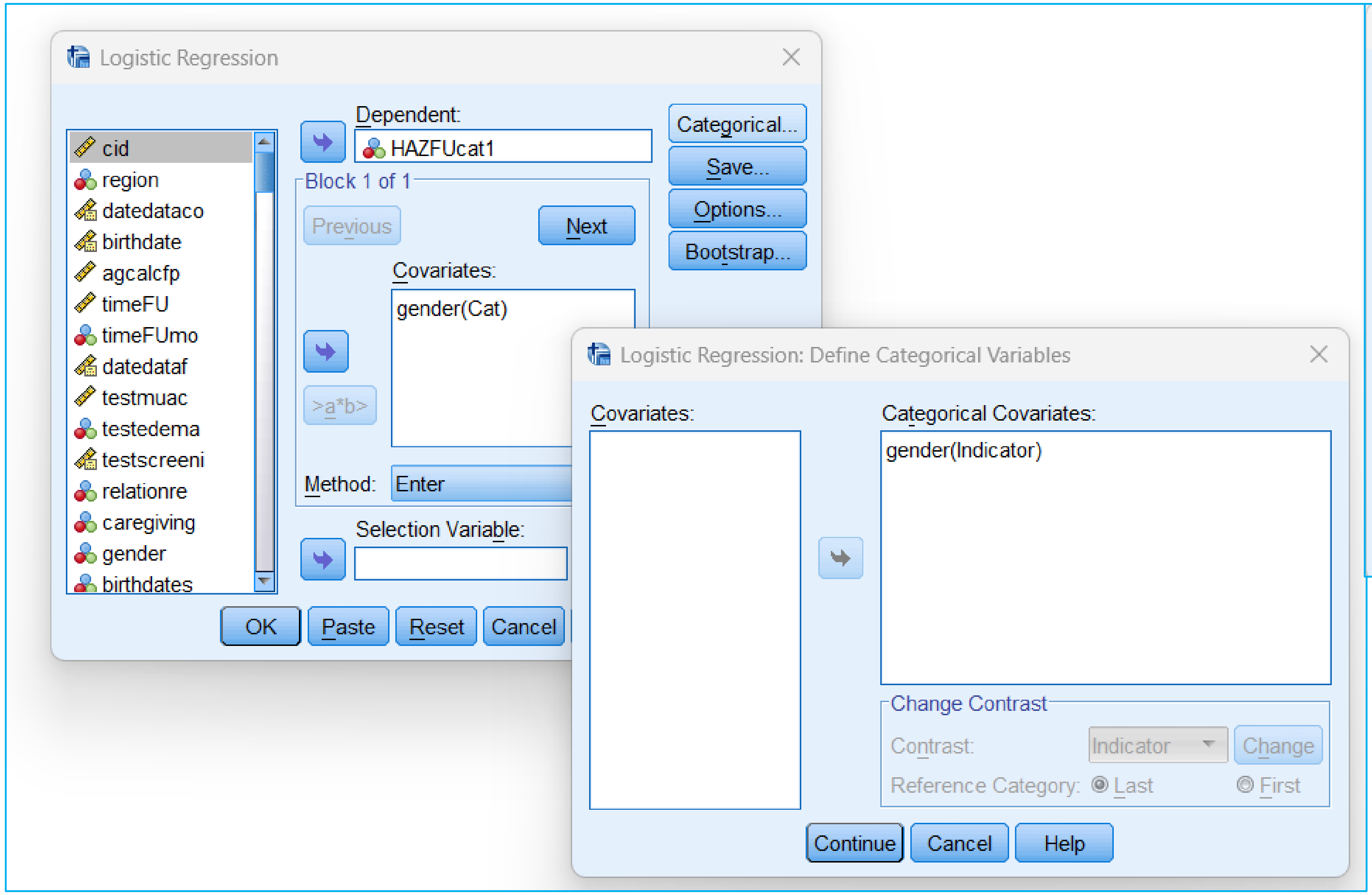
- Number of variables
  - One dependent (y)
  - One independent (x)
- Scale of variable(s)
  - Dependent: binary
  - Independent: categorical (nominal/ordinal) or continuous (ratio/interval)

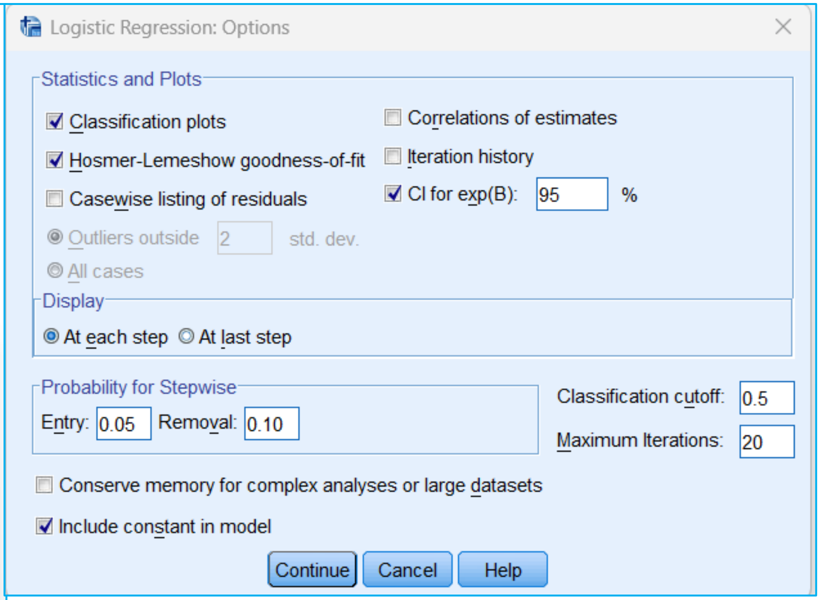
Simple logistic regression

Assumptions

- No normality assumption
- Sample size: the number of cases Vs the number of predictors wish to include in a model (10-15 predictors)
  - N> 30 times the number of predictors
- Multicollinearity: high intercorrelation among predicators
  - No formal test in logistic regression (use linear –regression)

Simple logistic regression





Simple logistic regression

- Stunting (y) : Stunted (0=not stunted; 1=stunted)
- Gender of the child (x) = gender (0=Female; 1=Male)

Variables in the Equation

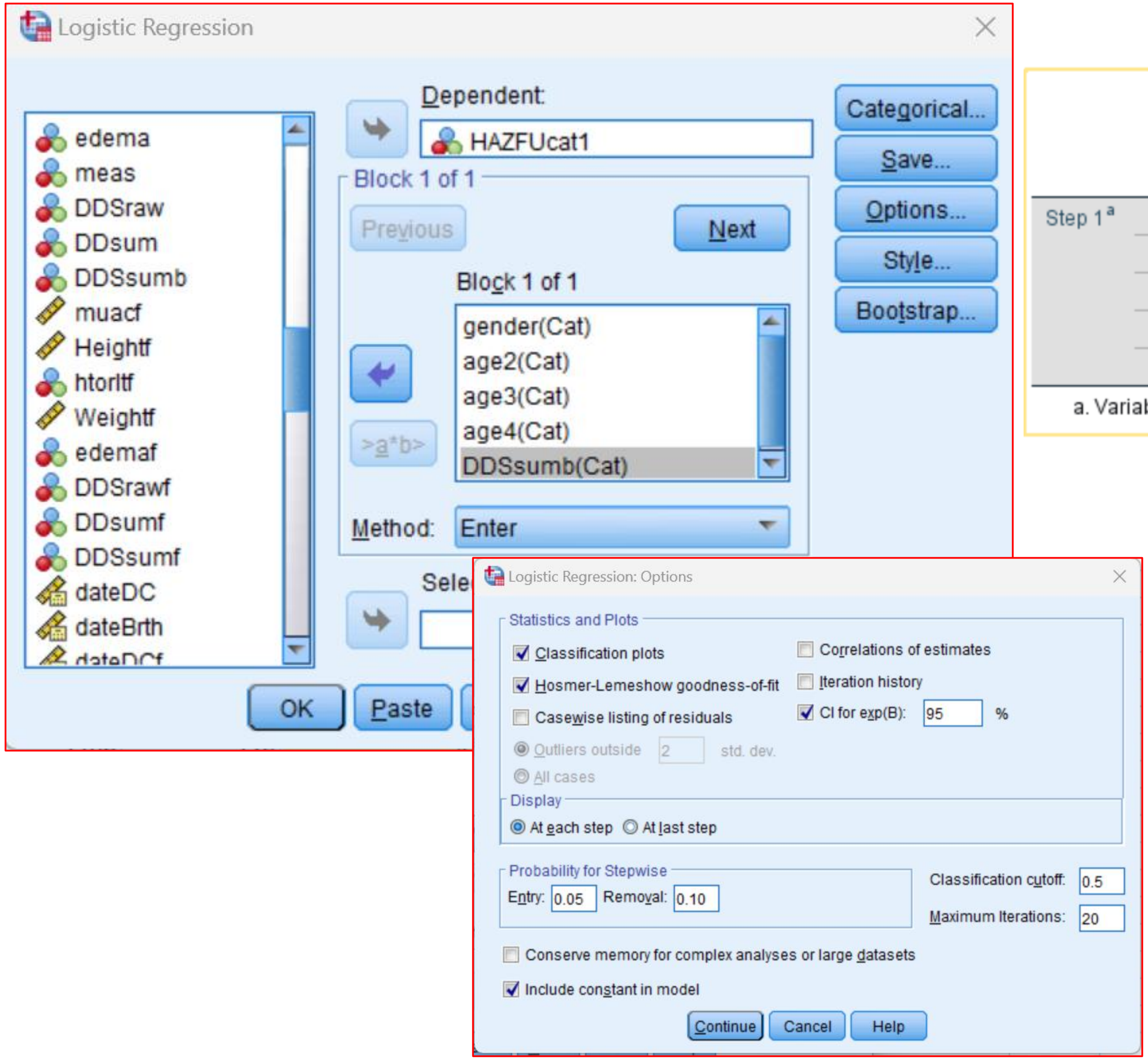
		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Sex of the child(1)	-.417	.143	8.466	1	.004	.659	.498	.873
	Constant	1.151	.115	99.654	1	.000	3.162		

a. Variable(s) entered on step 1: Sex of the child.

- The OR is 0.659, which means that we have a negative association between gender and stunting.
- Males are less likely to be stunted compared to female children

- Sig. shows the value is 0.004> there association between stunting and gender
- 95 % CI for EXP(B) gives us the lower confidence limits (Lower) and the upper confidence limits (Upper).
  - The CI doesn’t include the null value and, thus, the results are statistically significant.

Multiple logistic regression



Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B) Lower	Upper
Step 1[a]	Sex of the child(1)	-.442	.145	9.303	1	.002	.643	.484	.854
	age1(1)	.546	.150	13.206	1	.000	1.727	1.286	2.318
	age3(1)	.708	.246	8.294	1	.004	2.030	1.254	3.288
	DDsum(1)	.068	.180	.142	1	.707	1.070	.752	1.523
	Constant	.097	.328	.087	1	.768	1.102		

a. Variable(s) entered on step 1: Sex of the child, age1, age3, DDsum.

- The OR for gender =**0.643 (95%CI=0.484-0.854)**
- **The OR for age1 is 1.73 and the OR for age3 is 2.03.**
- Those with age1 and age3 are more likely to be stunted (the odds of stunting was higher among those <12mos and 24-36mos)
- P-value and 95 % C.I. for EXP(B) indicate statistical association

Model diagnostics

Goodness of fit

- Determines if the estimated model (i.e. the model with one or more x-variables) predicts the outcome better than the null model (i.e. a model without any x-variables)

- **Estimates of the goodness of fit**
  - Classification tables
  - The Hosmer and Lemeshow test
  - ROC curve

Model diagnostics

Classification tables

- A classification table is similar to the table about sensitivity and specificity
- It is automatically produced by SPSS and appears in the standard output

Sensitivity and specificity			
		Estimated model	
		Non-case	**Case**
"Truth"	**Non-case**	*True negative*	*False positive*
	Case	*False negative*	*True positive*

The Hosmer and Lemeshow test

- A type of a chi-square test.
- It indicates the extent to which the estimated model provides a better fit to the data (i.e. better predictive power) than the null model.
- The test will produce a p-value: if the p-value is above 0.05 the estimated model has adequate fit

Model diagnostics

ROC curve

- Is a graph that shows how well the estimated model predicts cases (sensitivity) and non-cases (specificity)
- What we are interested in here is the "area under the curve" (AUC).
- The AUC ranges between 0.5 and 1.0.

Criteria for AUC

- 0.5-0.6 Fail
- 0.6-0.7 Poor
- 0.7-0.8 Fair
- 0.8-0.9 Good
- 0.9-1.0 Excellent

Model diagnostics

Classification Table

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted: Stunting – Not stunted	Predicted: Stunting – stunted	Percentage Correct
Step 0	Stunting	Not stunted	0	303	.0
		stunted	0	738	100.0
	Overall Percentage				70.9

a. Constant is included in the model.

b. The cut value is .500

Hosmer and Lemeshow test

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	2.252	6	.895

We have a p-value of 0.895. This suggests that the estimated model has adequate fit

- The overall % of cases (stunted) and non-cases (non-stunted) that are correctly classified by the null model is 79.9% which is a lot
- Adding covariates did not make a difference

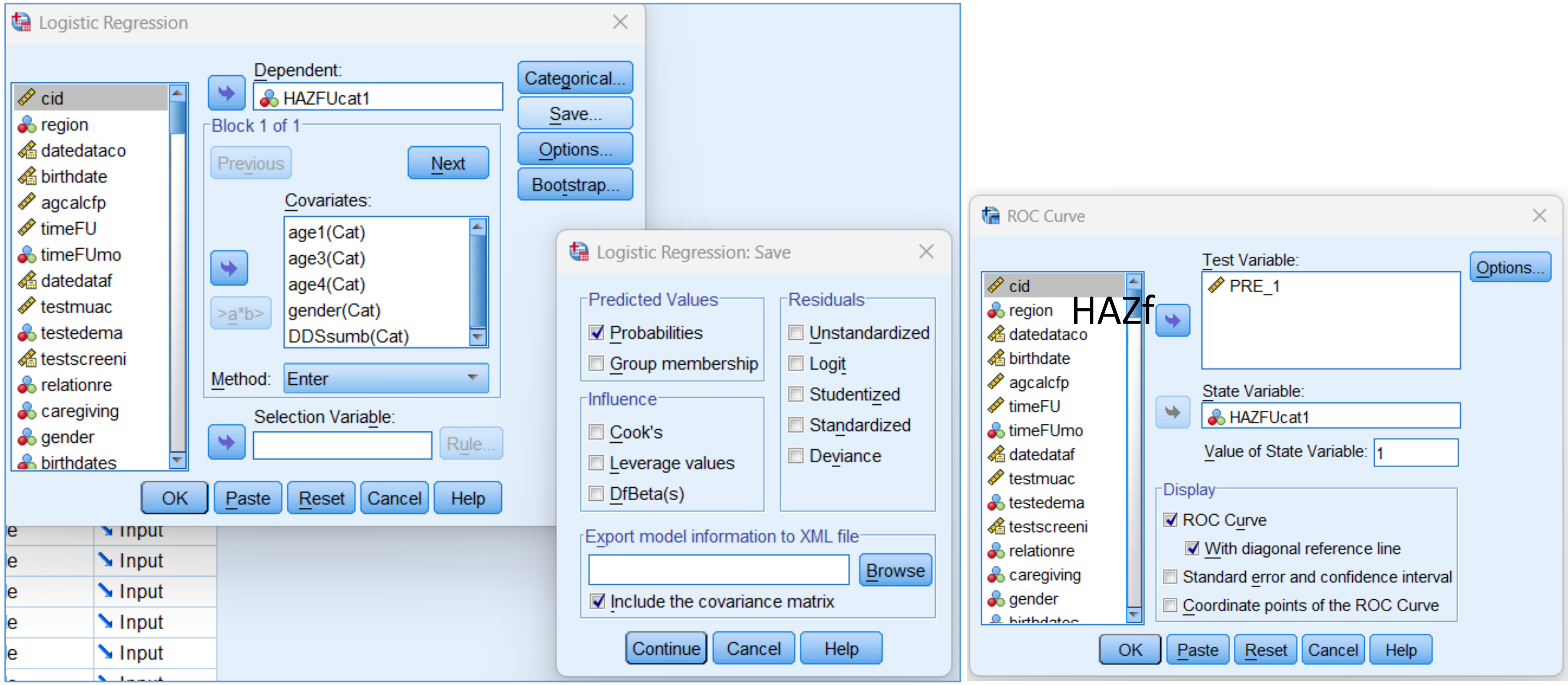
Classification Table[a]

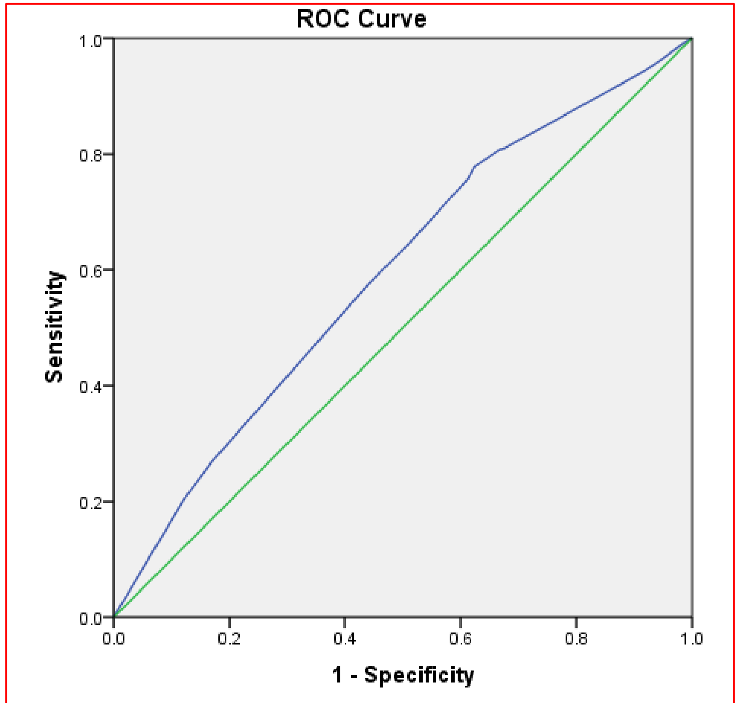
Observed			Predicted: Stunting – Not stunted	Predicted: Stunting – stunted	Percentage Correct
Step 1	Stunting	Not stunted	0	303	.0
		stunted	0	738	100.0
	Overall Percentage				70.9

a. The cut value is .500

Model diagnostics

ROC curve





Diagonal segments are produced by ties.

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
.594	.019	.000	.556	.632

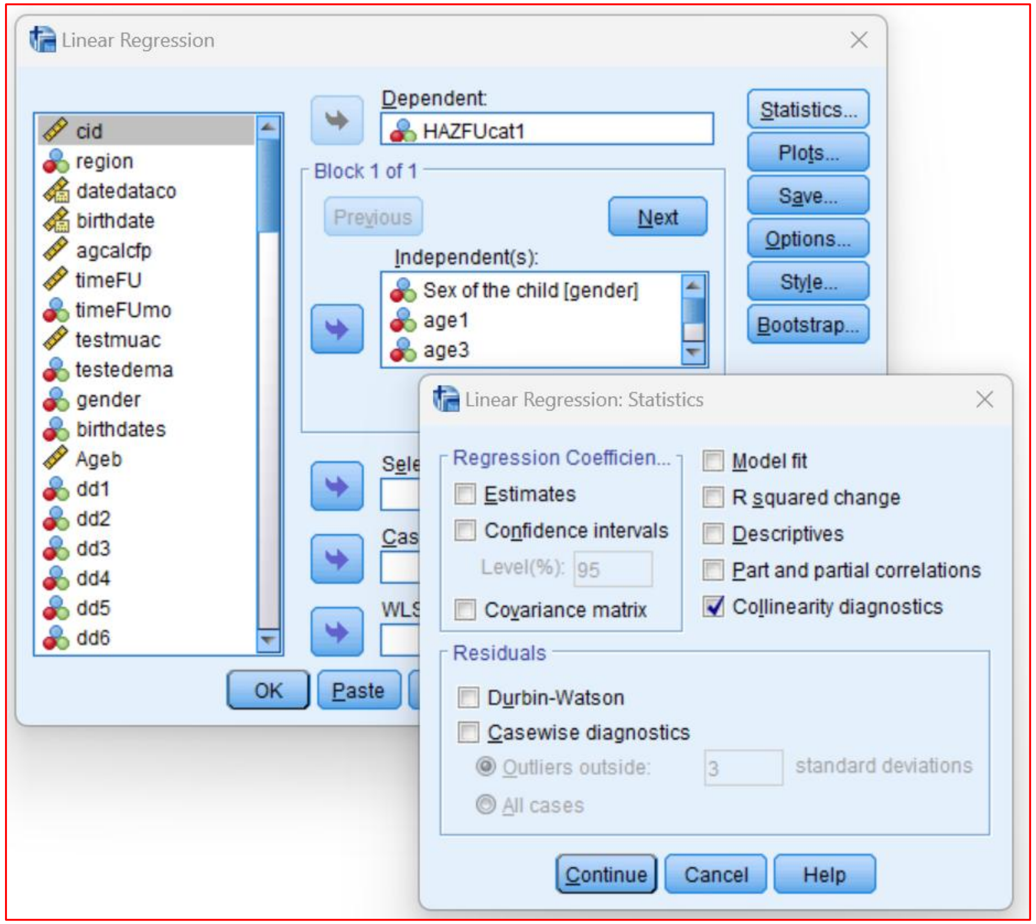
The test result variable(s): Predicted probability has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

a. Under the nonparametric assumption

b. Null hypothesis: true area = 0.5

A value of 0.594~ 0.6suggests rather poor predictive power

Checking for collinearity



Coefficients[a]

Model		Collinearity Statistics Tolerance	VIF
1	Sex of the child	.996	1.004
	age1	.951	1.051
	age4	.957	1.045
	DDsum	.987	1.013

a. Dependent Variable: Stunting

- VIF=1/Tolerance
- <10? < 5?

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Sex of the child	age1	age4	DDsum
1	1	2.326	1.000	.06	.07	.05	.02	.05
	2	1.002	1.524	.00	.00	.16	.66	.00
	3	.827	1.677	.00	.01	.14	.06	.74
	4	.577	2.007	.00	.61	.29	.16	.08
	5	.268	2.948	.94	.31	.36	.10	.13

a. Dependent Variable: Stunting

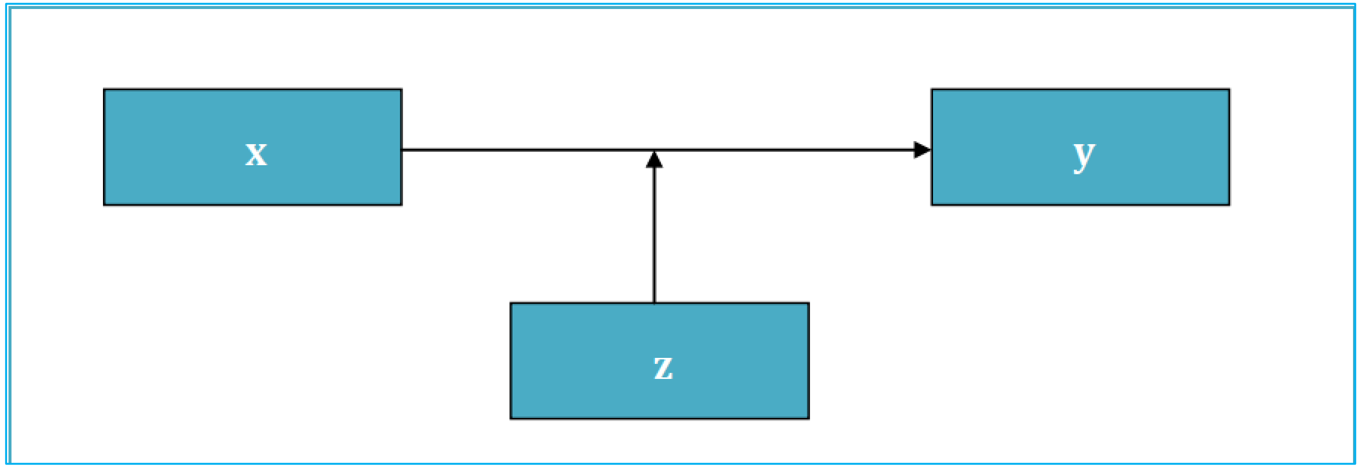
-No two or more predictors have a variance proportion >0.5

Checking for interaction

- Interaction analysis for linear regression
- Interaction analysis for logistic regression

Independent variables

- x=The variable we are mainly interested in with regard to its effect on y (**"main effect term").**
- z=The variable we suspect may modify the effect of x on y (**"main effect term")**
- x*z =The product of x and z – or the x-variable times the z-variable (**"interaction term")**



Checking for interaction

Measurement scale

- Combinations of variables
  - One binary x * one binary z
  - One ordinal/ratio/interval x * one binary z
  - One binary x * one ordinal/ratio/interval z
  - One ordinal/ratio/interval x * one ordinal/ratio/interval z

Direction of association

- If combining two ordinal /ratio/interval variables, they should be in the same direction in relation to the outcome
- Interpretation
  - Keep track of the coding of the variables

Checking for interaction

Example

- Main effect term= age
- Main effect term= DDS
- Interaction term= age*DDS

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	Ageb	.014	.007	3.468	1	.063	1.014	.999	1.028
	DDSsumb(1)	-.281	.402	.489	1	.484	.755	.344	1.659
	age_dds	-.015	.019	.613	1	.434	.985	.949	1.023
	Constant	.939	.374	6.307	1	.012	2.557		

a. Variable(s) entered on step 1: Ageb, DDSsumb, age_dds.

Interpretation

- No statistically significant association between the main effect terms and the outcome (stunting)
- No statistically significant interaction between having age and DDS with regards to stunting.
- Interaction terms shouldn’t be included

Questions?