St. PAUL INSTITUTE
FOR REPRODUCTIVE HEALTH AND RIGHTS

Introduction to regression analysis

Mekitie Wondafrash (MD, PhD)
St. Paul Institute for Reproductive Health and Rights
Ethiopia

Learning objectives

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

- Understand the different types of regression based on the type of outcome variables
- Understand the requirements for carrying out regression analysis
- Define confounders, mediators, and moderators in regression analysis

Introduction to regression

- Regression analysis is a way of predicting an outcome variable from one predictor variable (simple regression) or several predictor variables (multiple regression).
- Regression analysis is used to:
  - Predict the value of a dependent variable based on the value of at least one independent variable
  - Explain the impact of changes in an independent variable on the dependent variable

What type of regression should be used?

Outcome (y)	Type of regression
Nominal with two categories, i.e. dichotomous (binary)	Logistic regression
Nominal with more than two categories, i.e. polytomous	Multinomial regression
Ordinal	Ordinal regression
Continuous (ratio/interval)	Linear regression

- The outcome (y) determines the type of regression
- x-variable(s) can take on any form – they can be categorical or continuous
- Simple/bivariate regression – only one x variable
- Multiple regression –two or more x variables

X, Y and Z variables in regression

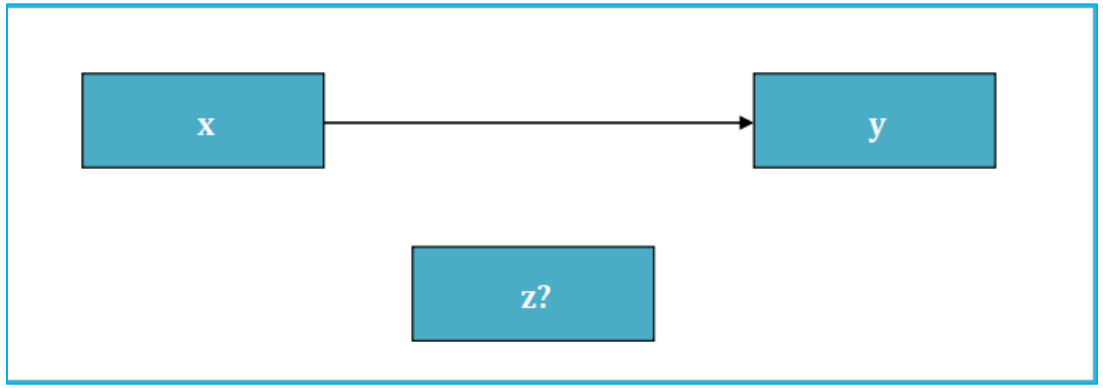
- Variables play different roles in the analysis.
- Researchers often use various terms to distinguish between these roles.

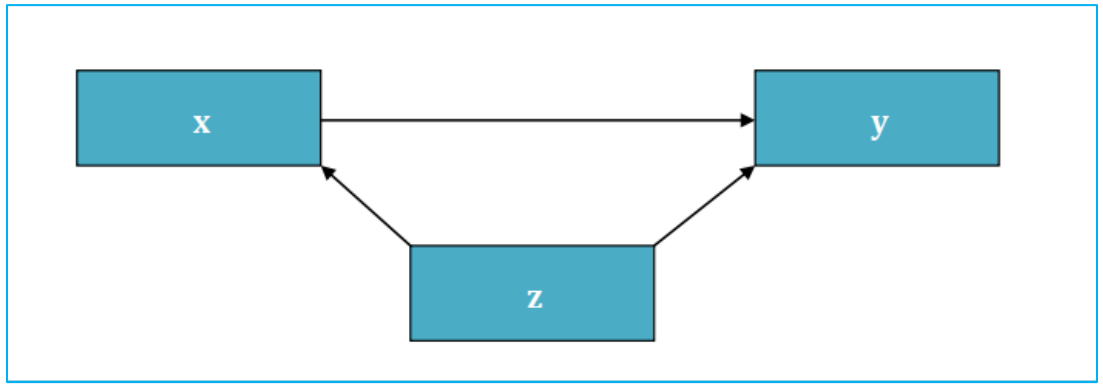
  X= Independent variable; Exposure; Predictor

  Y= Dependent variable; Outcome

  Z =Covariate; Confounder; Mediator; Moderator

Confounding, mediating and moderating variables





- **Confounder** : Both x and y are affected by z
- **Mediator** : A part of the association between x and y goes through z
- **Moderator** : Z affects the association between x and y

Preparation for data analysis

Important preparations

- Data distribution/ parametric (normally distributed)
  - Check normality by histogram, QQ-plot, and test for normality
  - Usually try two or three, as each gives some different information
  - Statistical tests/ visual inspection

Dummy variables

- In regression analysis – regardless of the type - we can only include x-variables that are continuous (ratio/interval) or binary (i.e. they consist of only two values).
- A binary variable is sometimes called "dichotomous", "binomial" or "dummy".
- Categorical variable with more than two values needs to be changed to dummy variables to "trick" the regression analysis to correctly analyze those variables.

Example

	Categories	*Dummy*
Educational attainment	1=Compulsory	1=Compulsory, 0=Other
	2=Upper secondary	1=Upper secondary, 0=Other
	3=University	1=University, 0=Other

Choosing a reference category in regression

- Use the normative category- characteristics that represent most people (e.g. attended formal schooling vs not attended formal school)
- Use the largest category
- Avoid categories with small sample sizes
- Consider the ordinal nature of categories: selecting the first or last category as the reference can be logical

Preparation for data analysis

- When running a statistical test, two hypotheses are being tested
  - Null Hypothesis: the default, or ‘boring’ state
    - Typically ‘no change’, ‘no difference’, or ‘no relationship’
  - Alternative Hypothesis: something else is happening
- Construct the two hypotheses based on your question from the data exploration step

Simple versus multiple regression model

Multiple regression

- Each x-variable's effect on y is estimated while taking into account the other x-variables' effects on y.
- Other x-variables are "held constant", "adjusted for", or "controlled for".
- It is a way of dealing with the issue of "confounding" variables, and to some extent also "mediating" variables

- Run a simple regression for each of the x-variables before including them in multiple regression to :
  - See the effect on B-coefficients by adding more x-variables (comparison)
  - Including more x-variables mostly reduces the strength of associations (especially if the x-variables overlapped in their effect on y)

Multivariate analysis

- It is the 3rd step in data analysis
- It is the process of examining the effects of two or more independent variables on the dependent variable simultaneously
- It allows us to:
  - Control for alternative effects and thus assess the extent of spuriousness
  - Conducts definitive test of our hypotheses
  - Gains a more sophisticated view of social reality

Questions?