PREDICTING THE FUTURE: INTRODUCTION TO REGRESSION ANALYSIS

NURULJANNAH BT NOR AZMI

What is regression analysis?

We often hear of new, complex “machine learning” methods that:

- Allow us to generate human language.
- Very accurately predict changes in the stock market.
- Recognize that an image contains a person or specific object.

What is regression analysis?

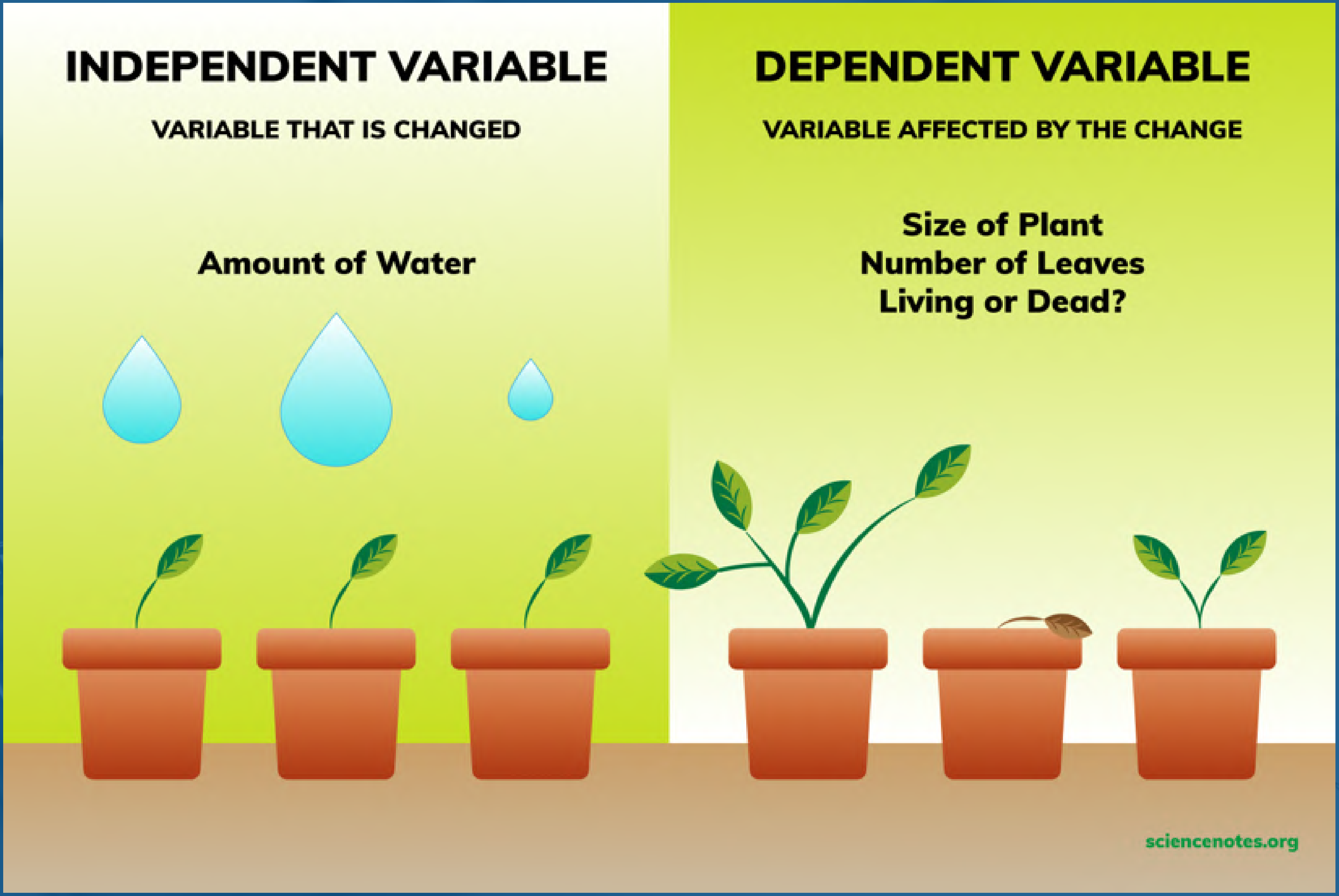
- Regression model analysis is utilized in various applications.
- Adrien-Marie Legendre introduced the concept of regression models in 1805.
- Since then, regression-based modeling has remained fundamental in applied statistics!

What is regression analysis?

Regression analysis comprises a set of statistical techniques aimed at estimating the relationship between:

Dependent variable (outcome variable)

One OR More independent variables (predictor variables)

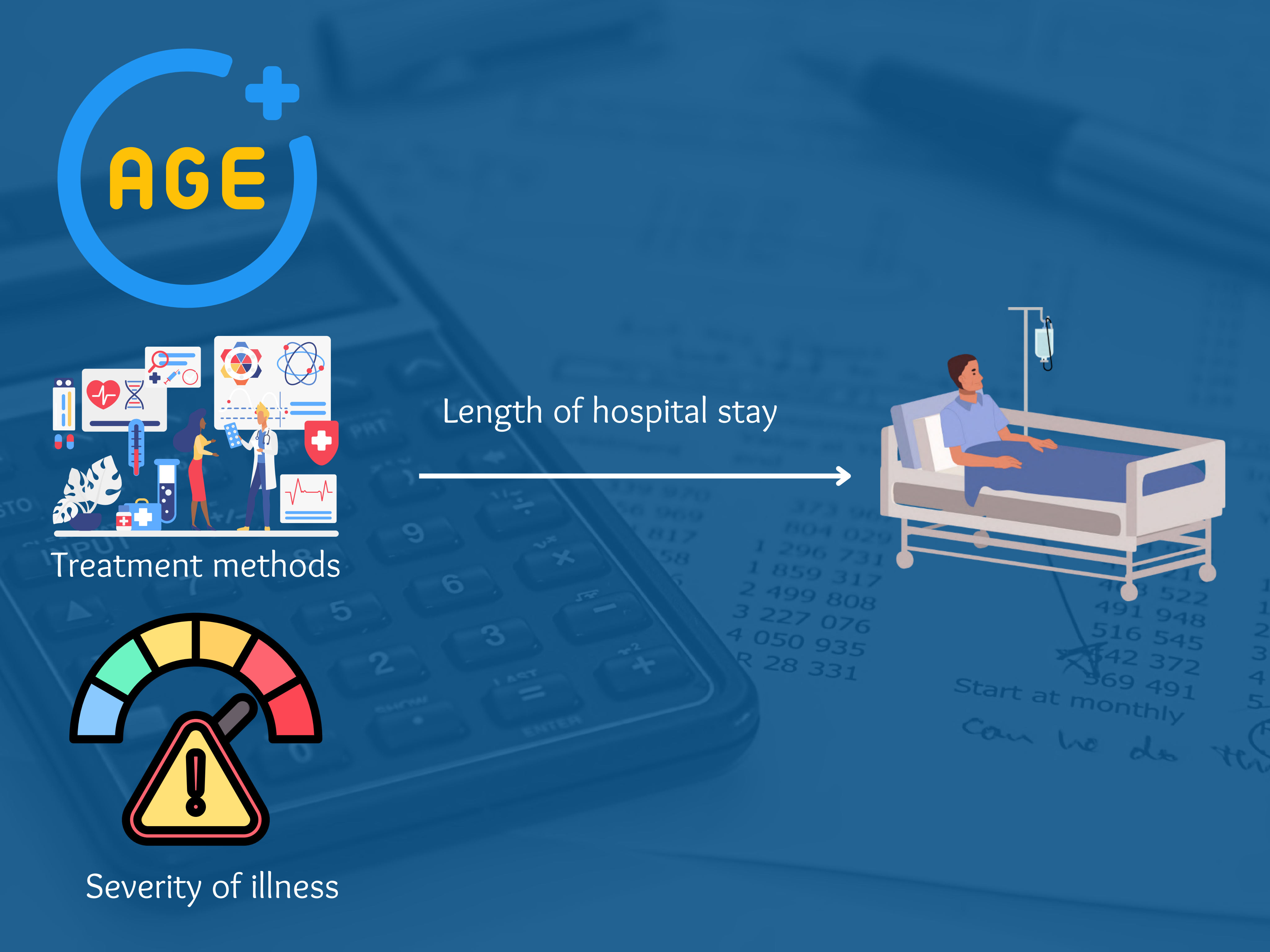

INDEPENDENT VARIABLE
VARIABLE THAT IS CHANGED
Amount of Water
DEPENDENT VARIABLE
VARIABLE AFFECTED BY THE CHANGE
Size of Plant
Number of Leaves
Living or Dead?
sciencenotes.org

When to apply Regression Analysis?

Regression analysis can address a broad range of questions, such as:

1. Is the relationship between two variables linear?
2. Which variable contributes the most to the outcome measurement?
3. How accurately can we predict future values?
4. Is our outcome variable caused by another variable?


AGE
DIET PLAN

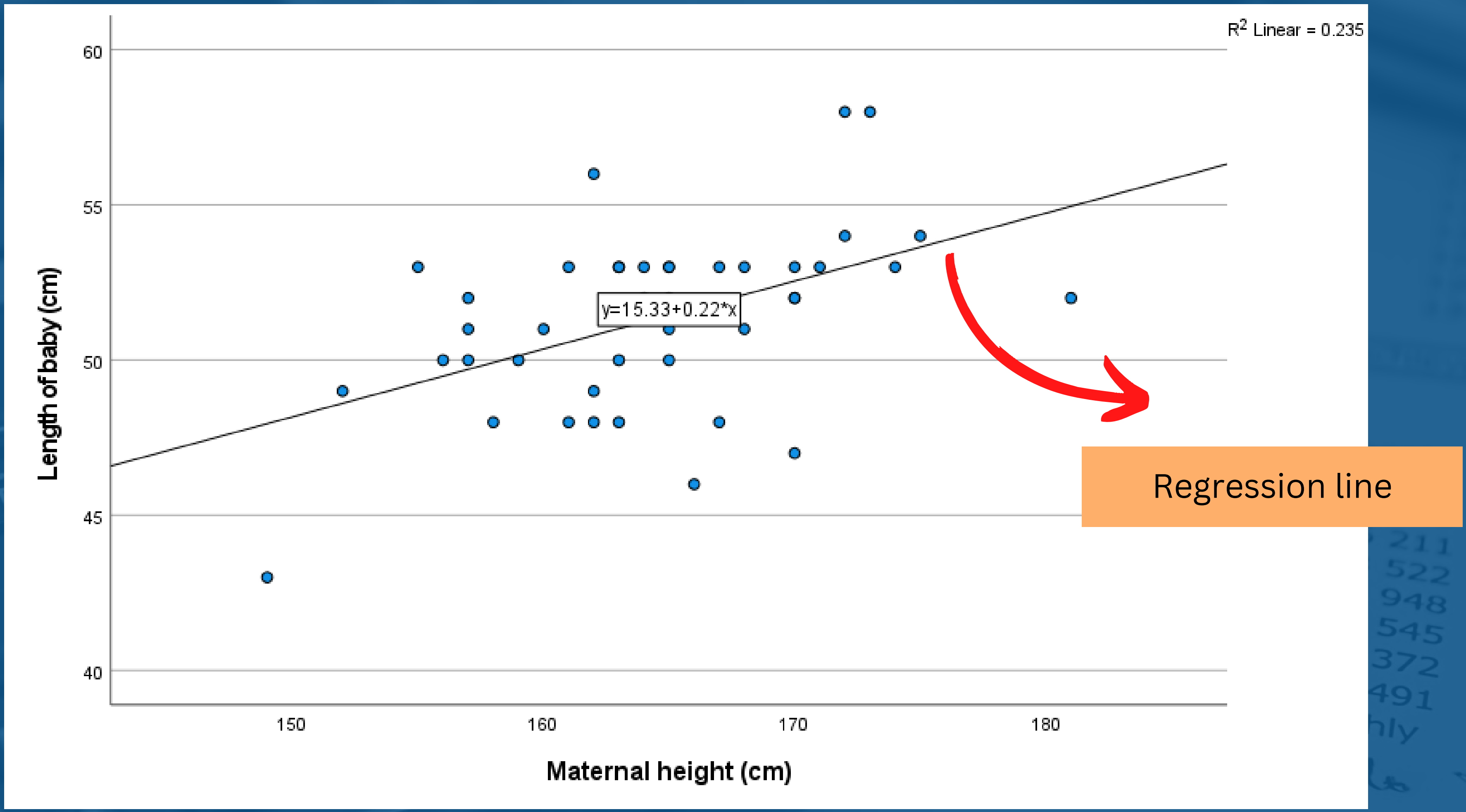

AGE
Treatment methods
Severity of illness
Length of hospital stay

Estimating Coefficient(s)

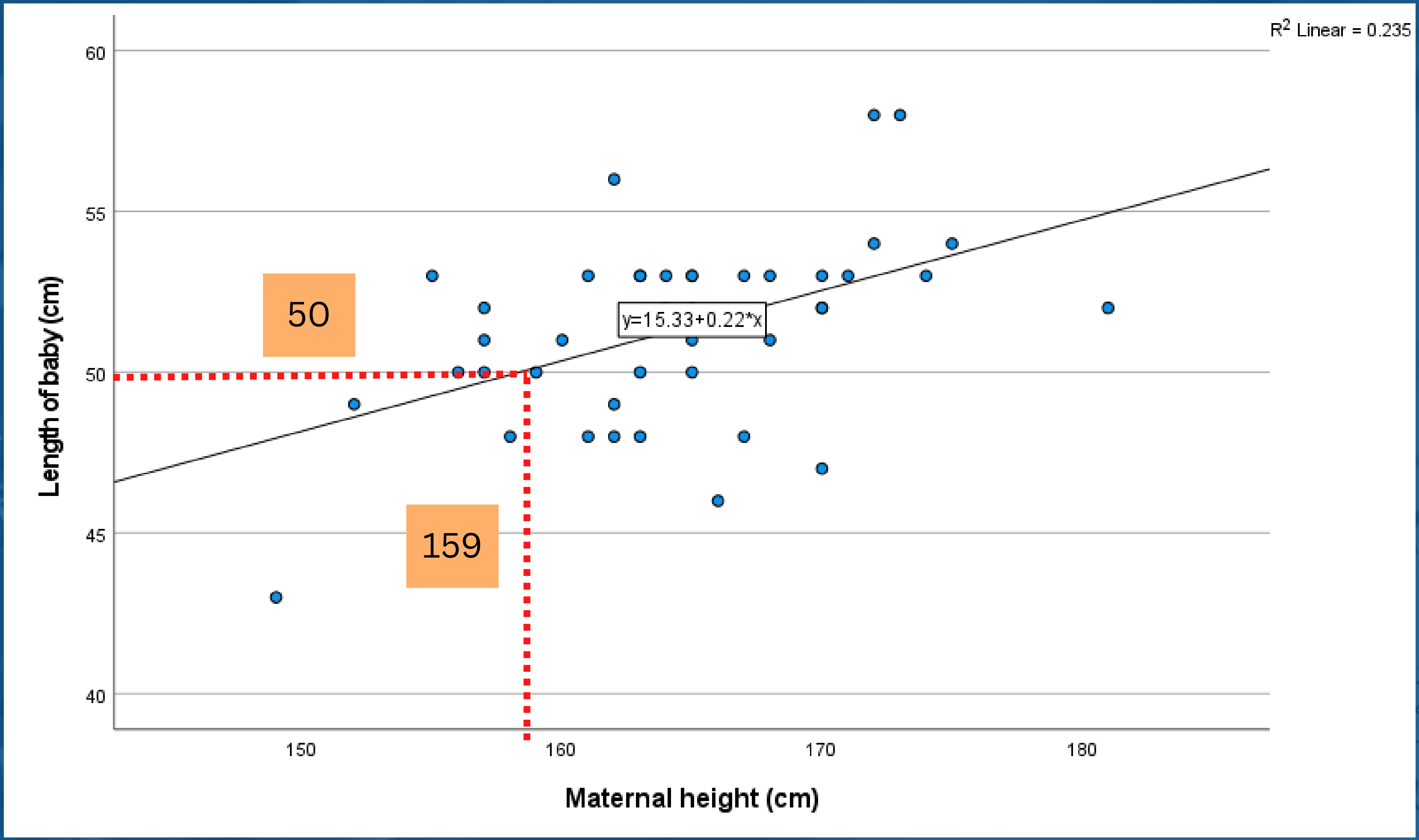
Assuming we are analyzing a basic model consisting of a single predictor, one outcome variable, and one coefficient, we can formally represent this model as follows:

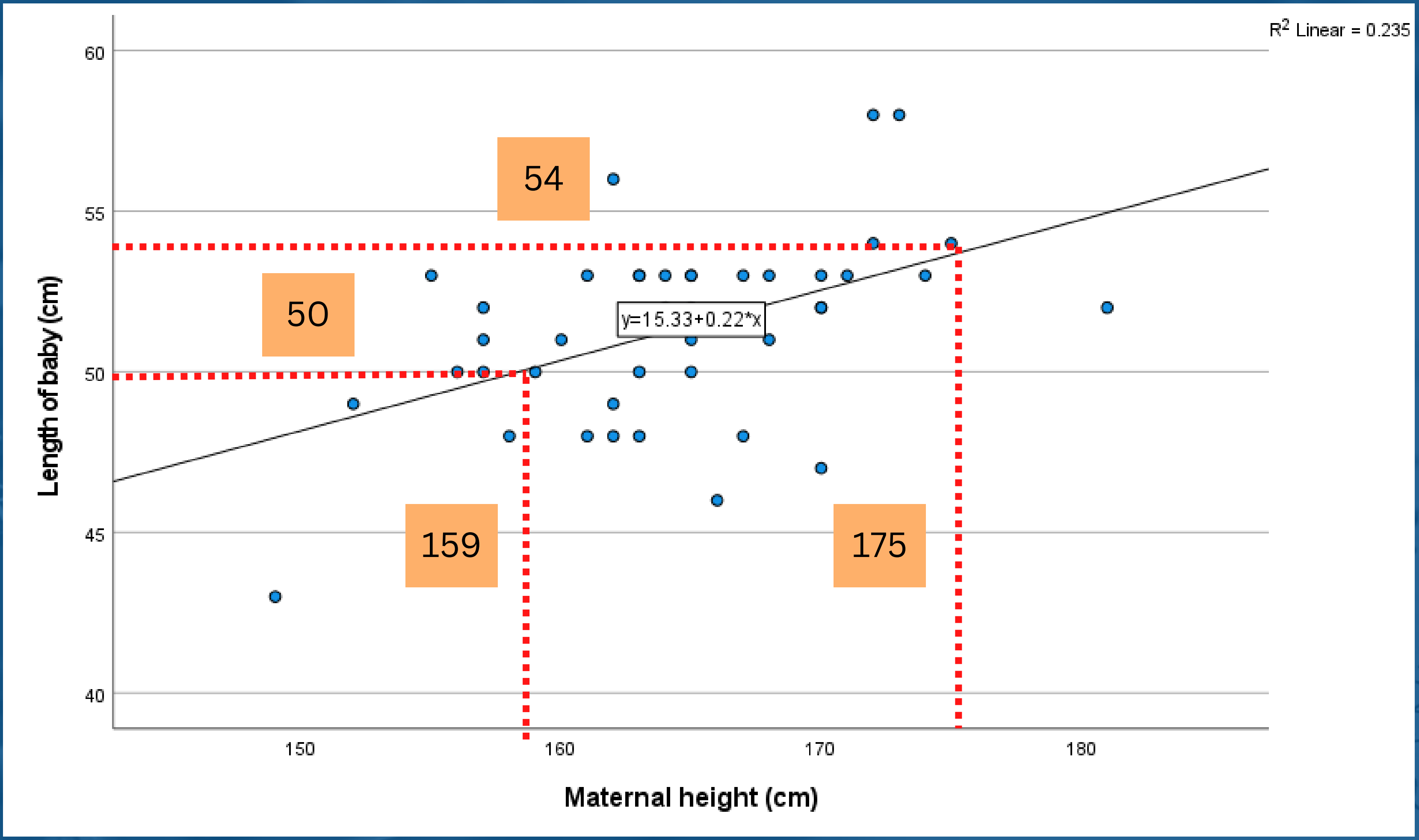
$$Y = \beta_0 + \beta_1 X_1 + \text{error}$$

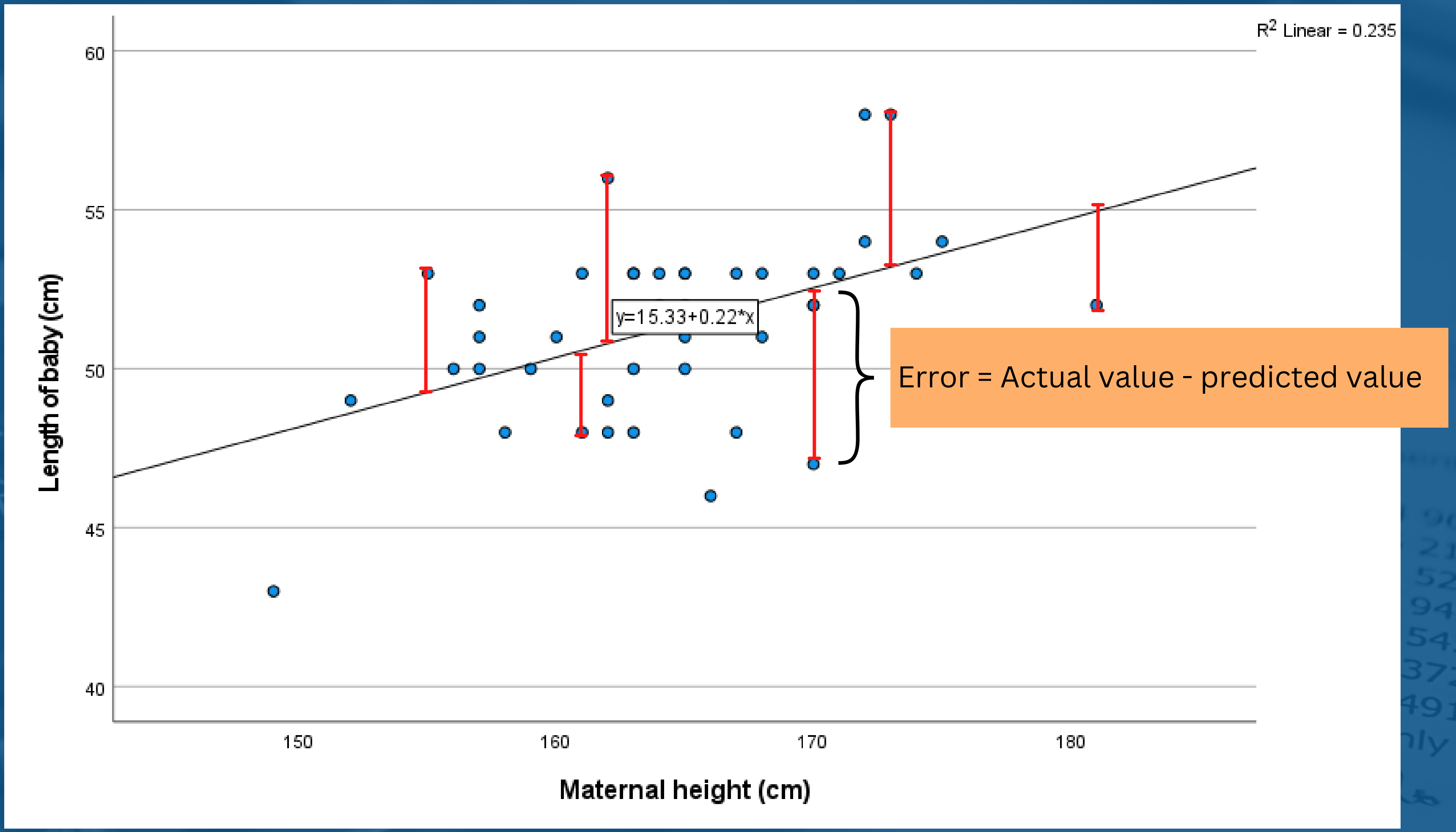
Y = outcome
β_0 = an intercept of the regression line
β_1 = a slope of the line / regression coefficient for independent variable
X_1 = independent variable
error = residual



R2 Linear = 0.235
60
55
50
45
40
Length of baby (cm)
y=15.33+0.22*x
150
160
170
180
Maternal height (cm)
Regression line


R2 Linear = 0.235
y=15.33+0.22*x
50
159
Length of baby (cm)
Maternal height (cm)
60
55
50
45
40
150
160
170
180


R2 Linear = 0.235
60
55
50
45
40
Length of baby (cm)
150
160
170
180
Maternal height (cm)
y=15.33+0.22*x
54
50
159
175


R^2 Linear = 0.235
60
55
50
45
40
Length of baby (cm)
y=15.33+0.22*x
Error = Actual value - predicted value
150
160
170
180
Maternal height (cm)

SIMPLE LINEAR REGRESSION

Simple Linear Regression is used to estimate the relationship between two quantitative variables.

Dependent variable : numerical

Independent variable : numerical

When to apply Simple Linear Regression?

You can use simple linear regression when you want to identify:

1. How strong the relationship between two variables.
2. To predict a value of one variable for a given value of the other.

How much the value Y (dependent variable) varies with one unit of change in value X (independent variable)

SIMPLE LINEAR REGRESSION - ONLY ONE INDEPENDENT VARIABLE

Independent variable (x) → Dependent variable (y)

Mother's height → Length of baby

MULTIPLE LINEAR REGRESSION - MORE THAN ONE INDEPENDENT VARIABLES

Independent variables (x) → Dependent variable (y)

Mother's height
Mother's weight
Age
→ Length of baby

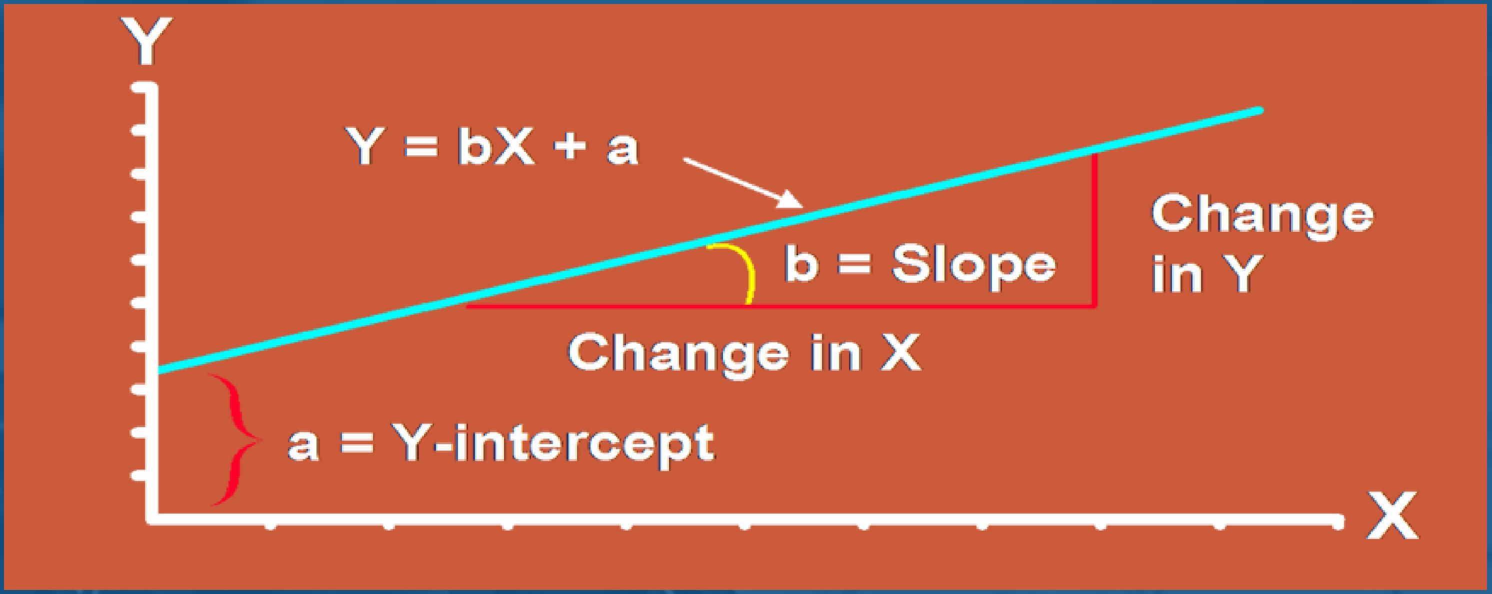
INTRODUCTION

If independent variables are combination of numerical and categorical or categorical only - **General Linear Regression**

Dependent (outcome) : numerical

Independent (predictor) : 2 or more combination of numerical and categorical or categorical only

Regression Equation



y = a + bx

x : independent variable
y: dependent variable

a: an intercept of the regression line (value of Y when X=0)
b: a slope of the line (an amount of change in Y for a unit change in X)

Coefficient of Determination (r^2)

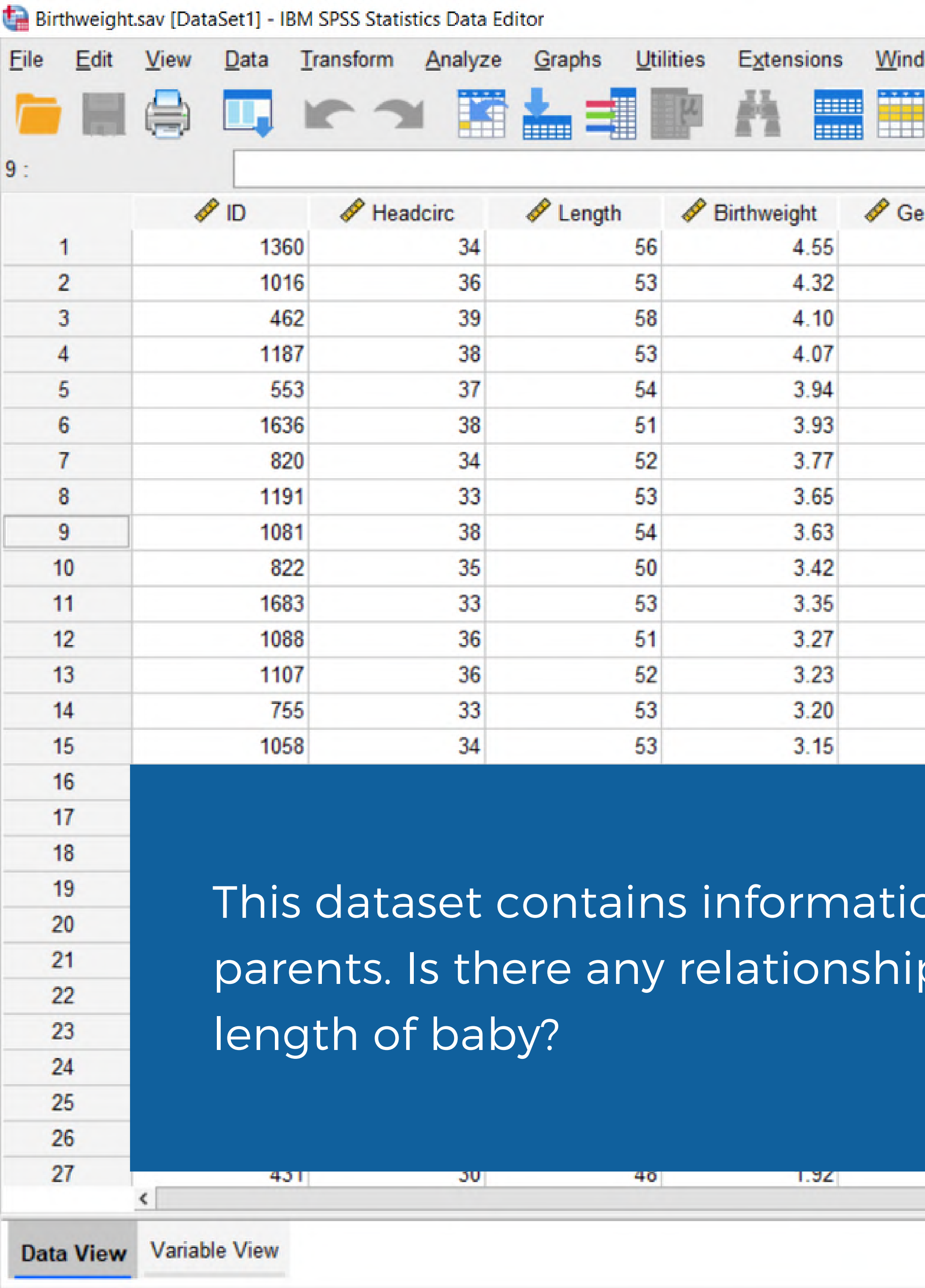
- Ranges from 0 to 1.
- It provides a measure of how well future outcomes are likely to be predicted by the model.
- How much the independent of ‘x’ is explained by dependent ‘y’.

ASSUMPTIONS OF THE MODEL

L	I	N	E
Relationship between the independent and dependent variable is linear (Linearity)	Independent observation	Residuals should be approximately normally distributed	Homoscedasticity (Equal variances)

CHECKING MODEL ASSUMPTIONS

Assumptions	How to check?
1.Relationship between the independent and dependent variable is linear (Linearity).	Scatter plot between independent and dependent variable
2.Independent observation	Done during design stage
3.Residuals should be approximately normally distributed	Histogram with overlaid normal curve of residuals
4.Homoscedasticity (Equal variances)	Scatter plot between residuals and predicted values (XP - YR)



Birthweight.sav [DataSet1] - IBM SPSS Statistics Data Editor

	ID	Headcirc	Length	Birthweight
1	1360	34	56	4.55
2	1016	36	53	4.32
3	462	39	58	4.10
4	1187	38	53	4.07
5	553	37	54	3.94
6	1636	38	51	3.93
7	820	34	52	3.77
8	1191	33	53	3.65
9	1081	38	54	3.63
10	822	35	50	3.42
11	1683	33	53	3.35
12	1088	36	51	3.27
13	1107	36	52	3.23
14	755	33	53	3.20
15	1058	34	53	3.15

EXAMPLE

Open dataset: birthweight.sav

This dataset contains information on new born babies and their parents. Is there any relationship between maternal height and length of baby?

EXAMPLE

A study was conducted to determine the relationship between mother's height and the length of baby, with the researcher aiming to forecast the baby's length using the mother's height as a predictor.

Mother's height — Length of baby

Numerical — Numerical

Simple Linear Regression

List down all the variables

Identify the types of variables

Identify the right statistical analysis

STEPS IN SIMPLE LINEAR REGRESSION

Step 1: State your research hypothesis

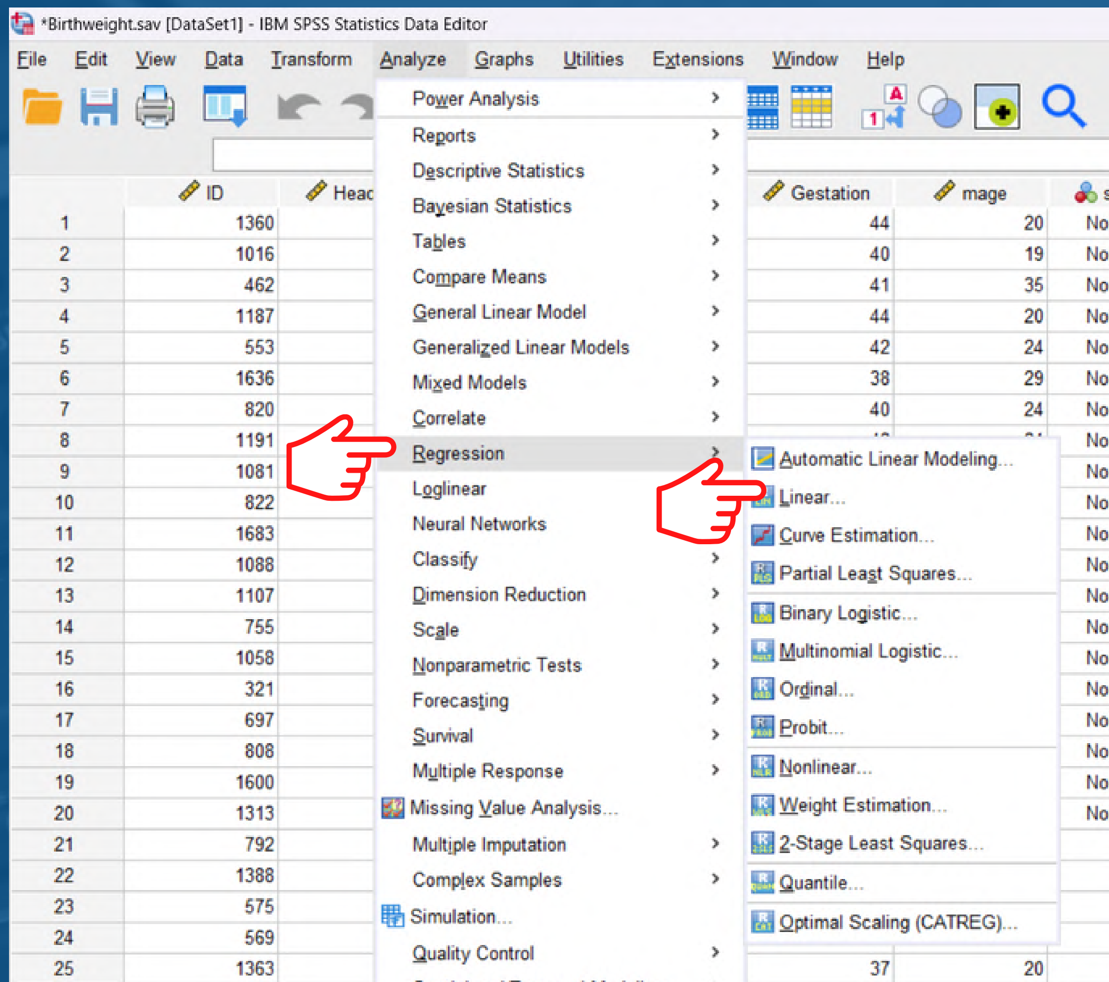
Null hypothesis and Alternative hypothesis

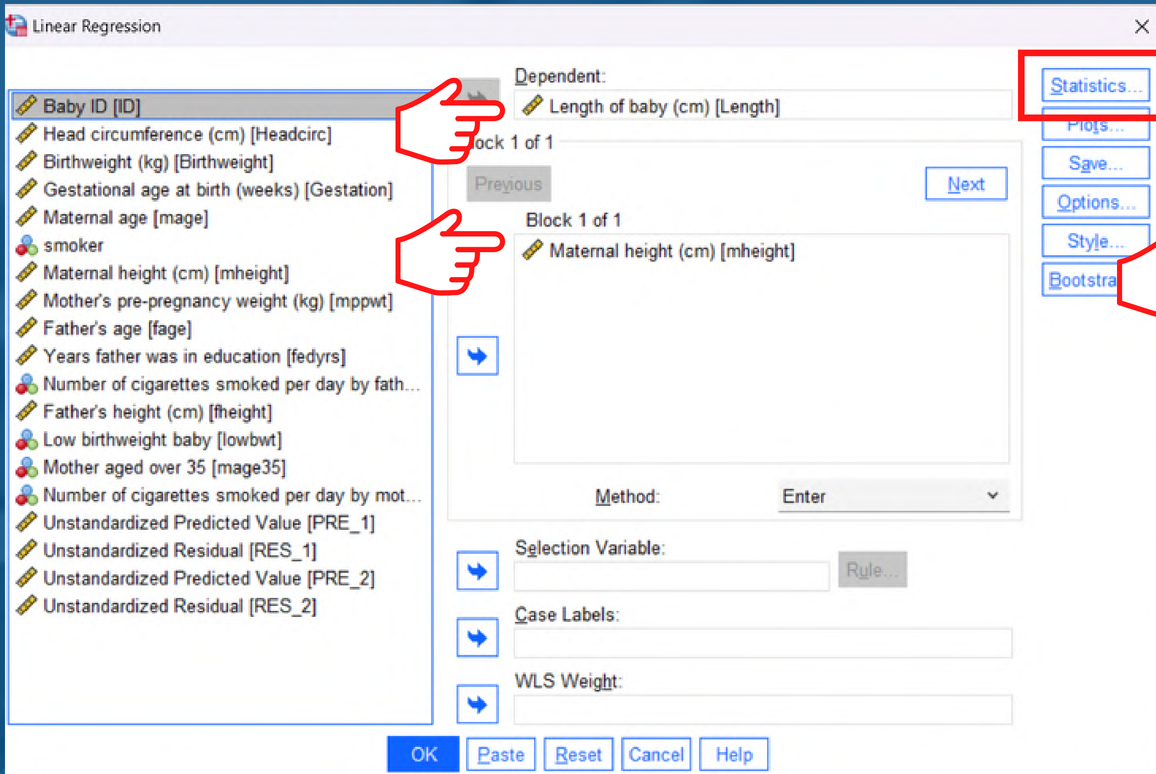
Ho : There is no relationship between mother’s height and the length of baby

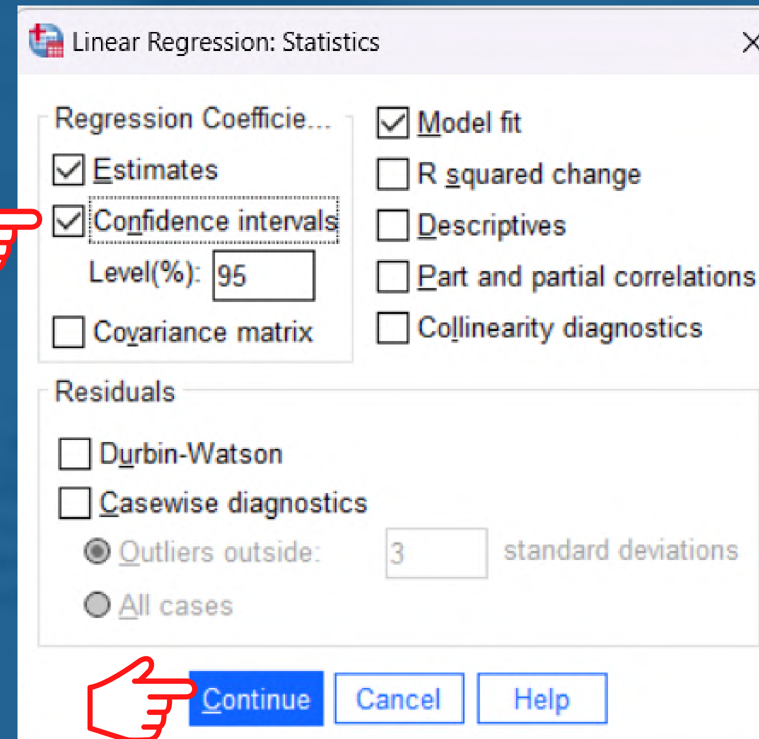
H_A : There is a relationship between mother’s height and the length of baby

Step 2: Run Simple Linear Regression

Go to: Analyze > Regression > Linear







Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	15.334	10.271		1.493	.143	-5.425	36.093
	Maternal height (cm)	.219	.062	.485	3.507	.001	.093	.345

a. Dependent Variable: Length of baby (cm)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.485[a]	.235	.216	2.599

a. Predictors: (Constant), Maternal height (cm)

b. Dependent Variable: Length of baby (cm)

23.5% of the variation in length of baby is explained by mother's height according to the linear regression model (r^2= 0.235).

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	83.110	1	83.110	12.302	.001[b]
	Residual	270.223	40	6.756		
	Total	353.333	41			

a. Dependent Variable: Length of baby (cm)

b. Predictors: (Constant), Maternal height (cm)

This table indicates that the regression model predicts the dependent variable significantly well (refer to p-value).

Here, $p < 0.05$, which is less than 0.05, and indicates that the regression model statistically significantly predicts the outcome variable.

Result presentation for Simple Linear Regression

Table 1: Simple linear regression

Variable	SLR[a]	
	b* (95% CI)	***p*-value**
Mother's height	0.22 (0.09,0.35)	0.001

[a] Simple linear regression

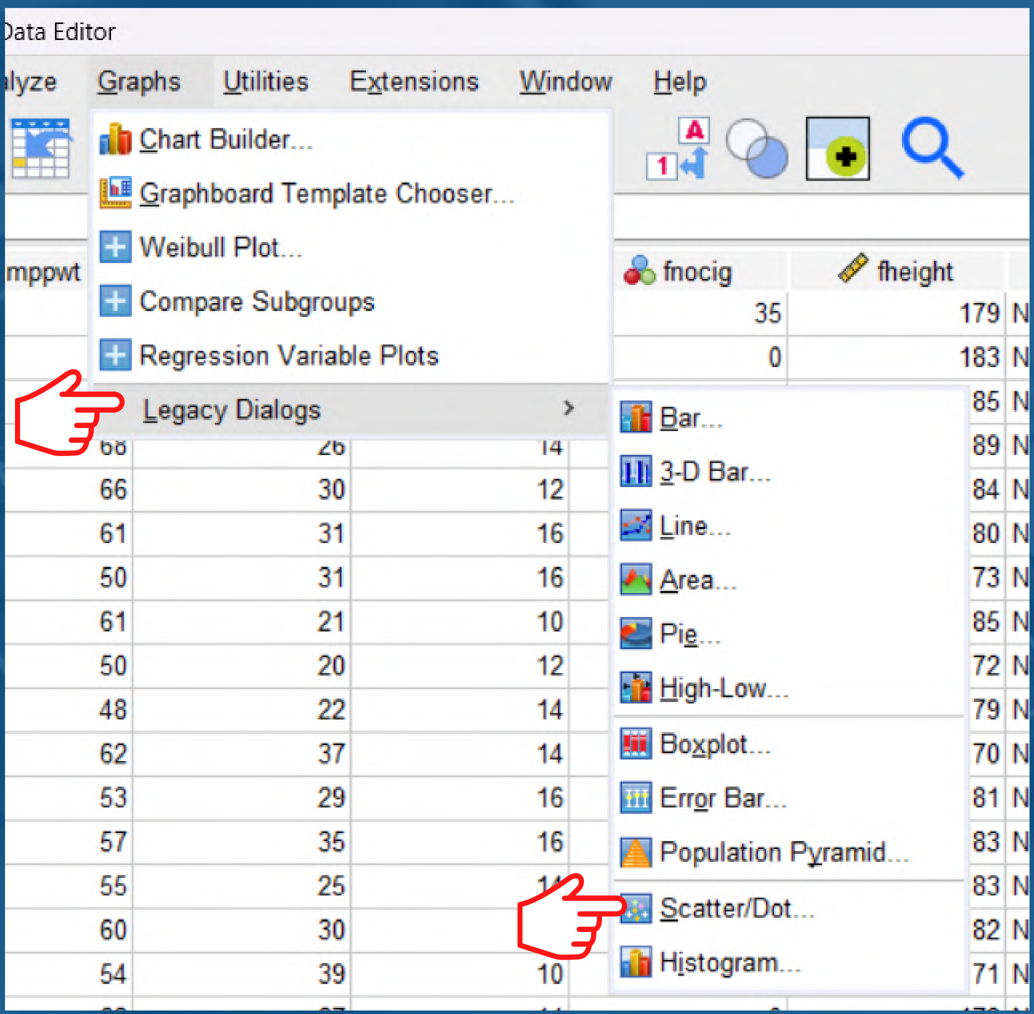
b* = crude regression coefficient

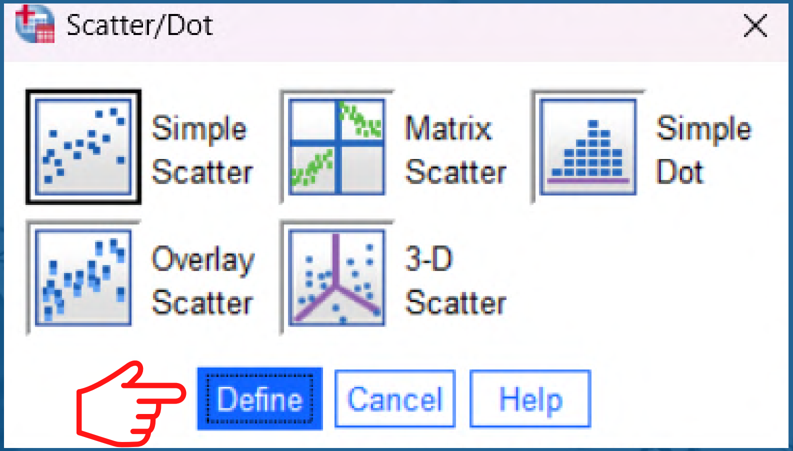
Step 3: Checking assumptions

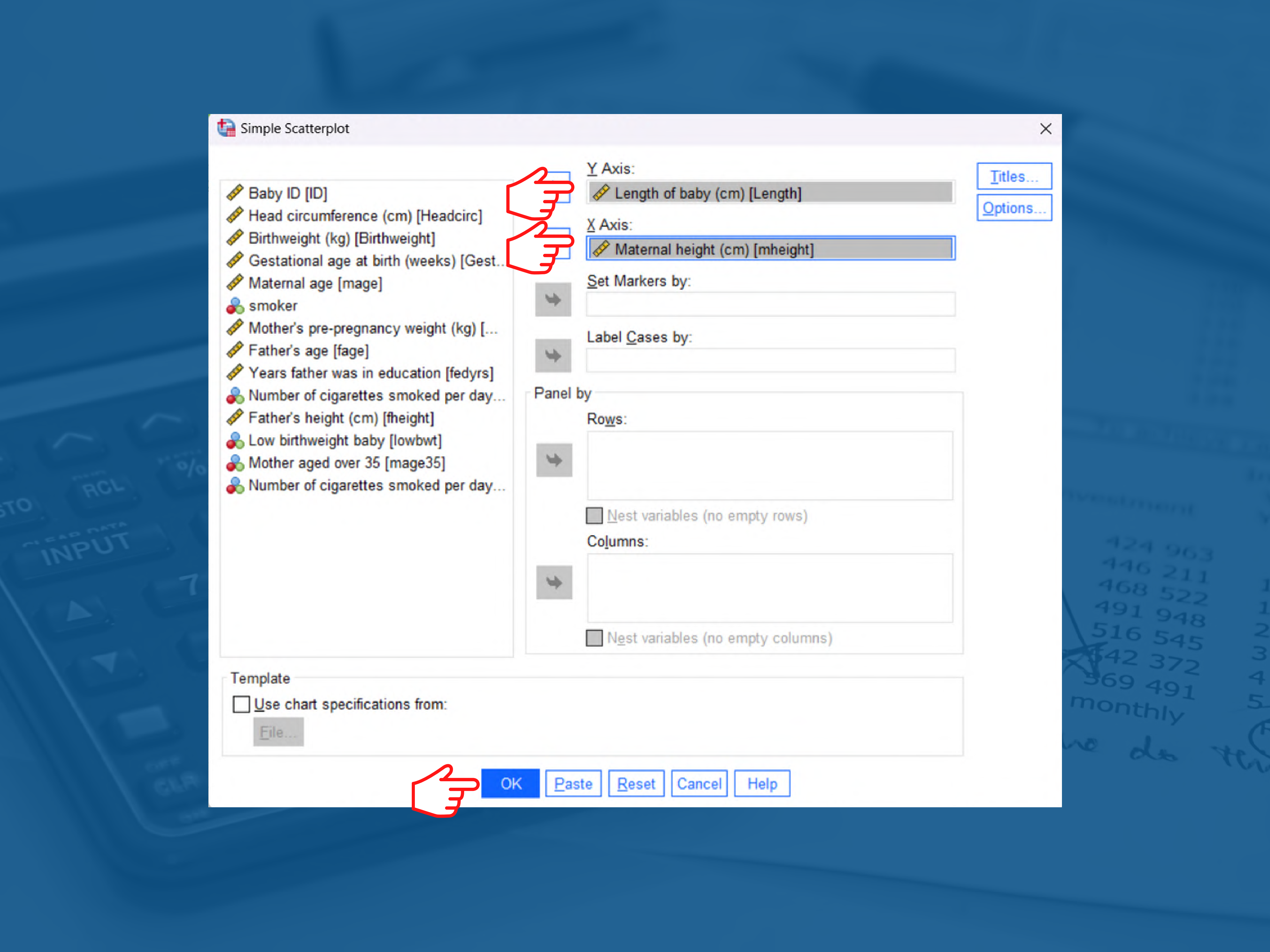
Assumptions	How to check?
1.Independent observation	Done during design stage
2.Relationship between the independent and dependent variable is linear (Linearity)	Scatter plot between independent and dependent variable
3.Homoscedasticity (Equal variances)	Scatter plot between residuals and predicted values (XP - YR)
4.Residuals should be approximately normally distributed	Histogram with overlaid normal curve of residuals

Checking assumption : Linearity

Go to: Graph > Legacy Dialogs > Scatter/Dot

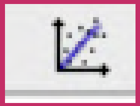


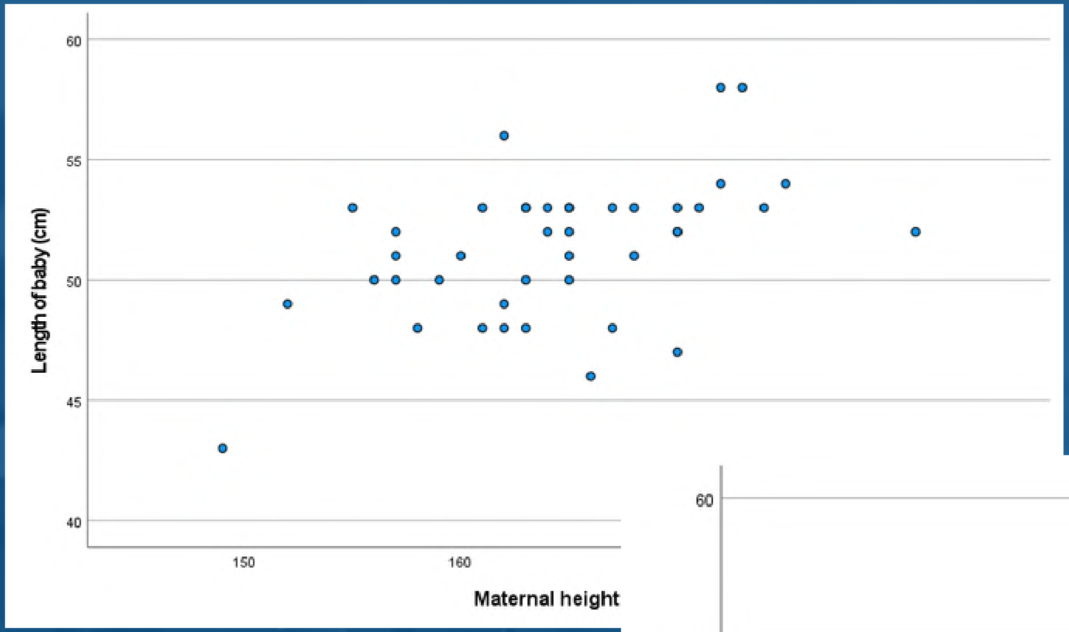


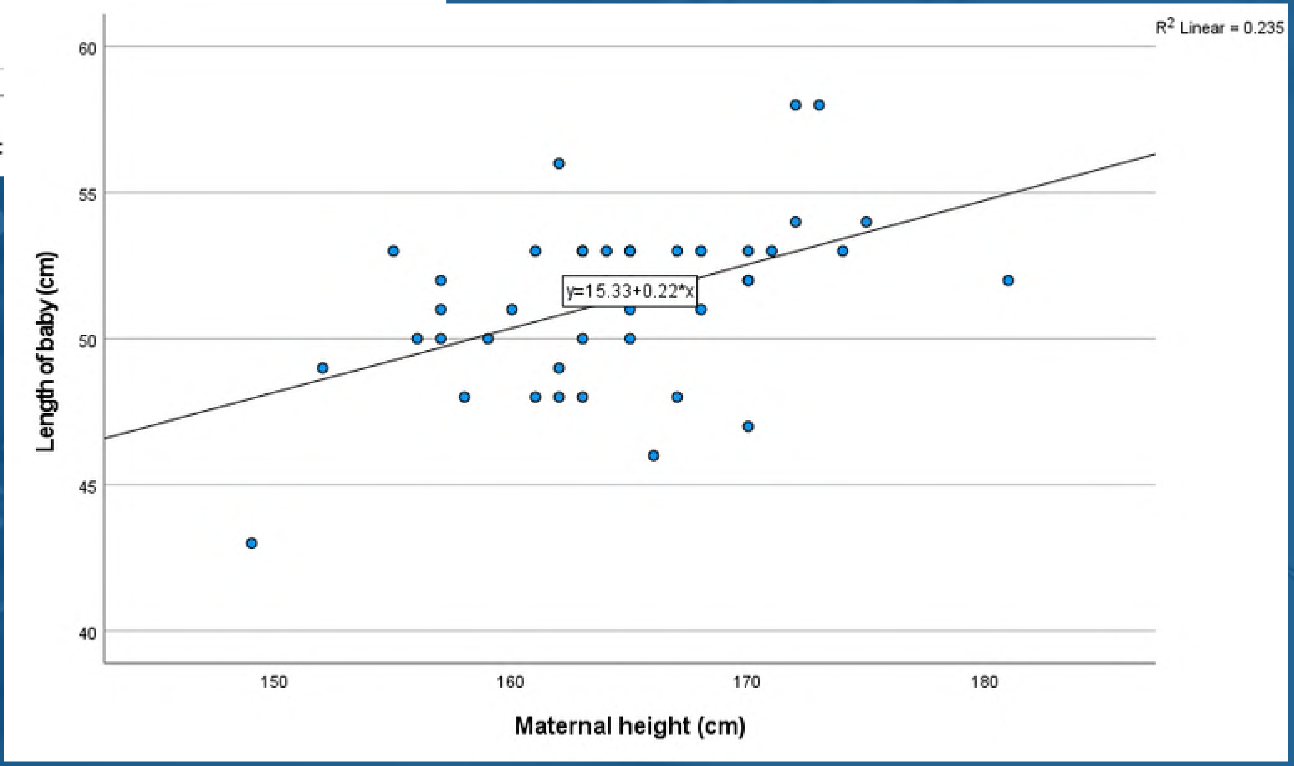


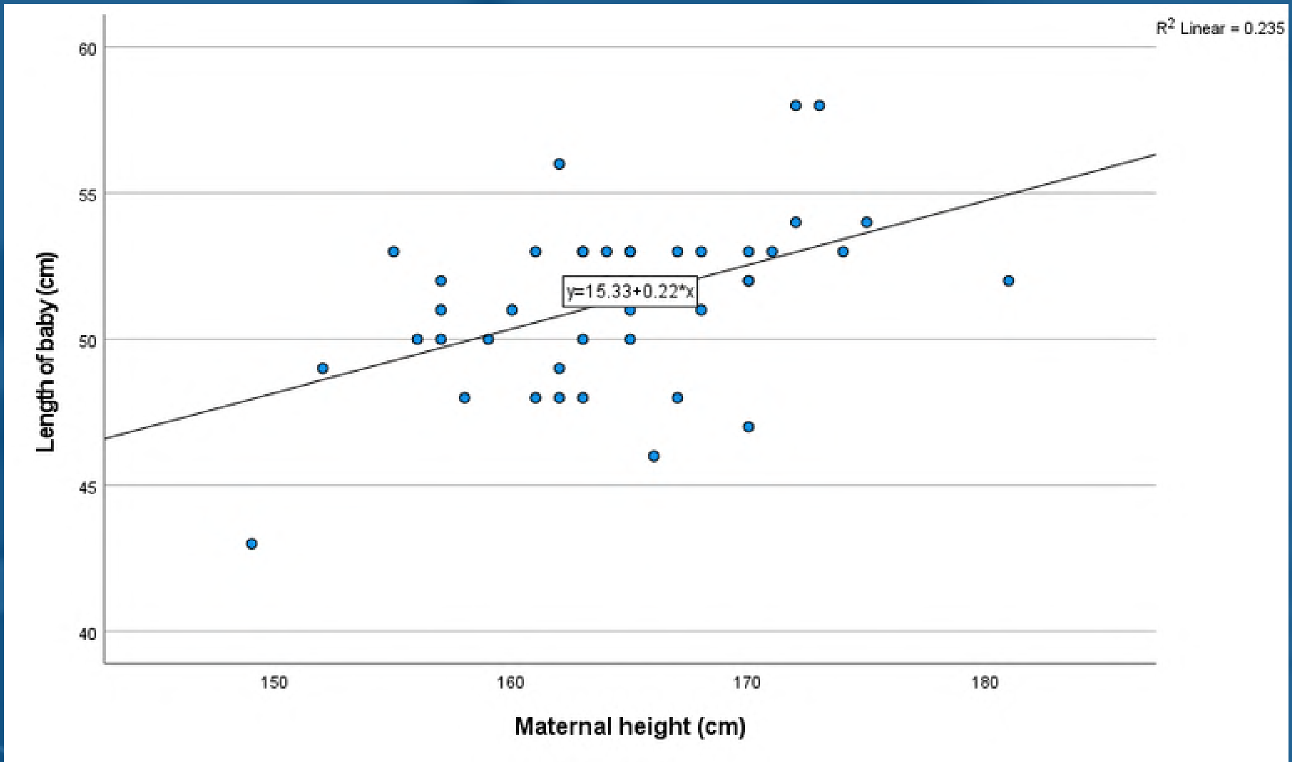
Simple Scatterplot
Baby ID [ID]
Head circumference (cm) [Headcirc]
Birthweight (kg) [Birthweight]
Gestational age at birth (weeks) [Gest...
Maternal age [mage]
smoker
Mother's pre-pregnancy weight (kg) [...
Father's age [fage]
Years father was in education [fedyrs]
Number of cigarettes smoked per day...
Father's height (cm) [fheight]
Low birthweight baby [lowbwt]
Mother aged over 35 [mage35]
Number of cigarettes smoked per day...
Y Axis:
Length of baby (cm) [Length]
X Axis:
Maternal height (cm) [mheight]
Set Markers by:
Label Cases by:
Panel by
Rows:
Nest variables (no empty rows)
Columns:
Nest variables (no empty columns)
Titles...
Options...
Template
Use chart specifications from:
File...
OK
Paste
Reset
Cancel
Help

Double click the plot and click

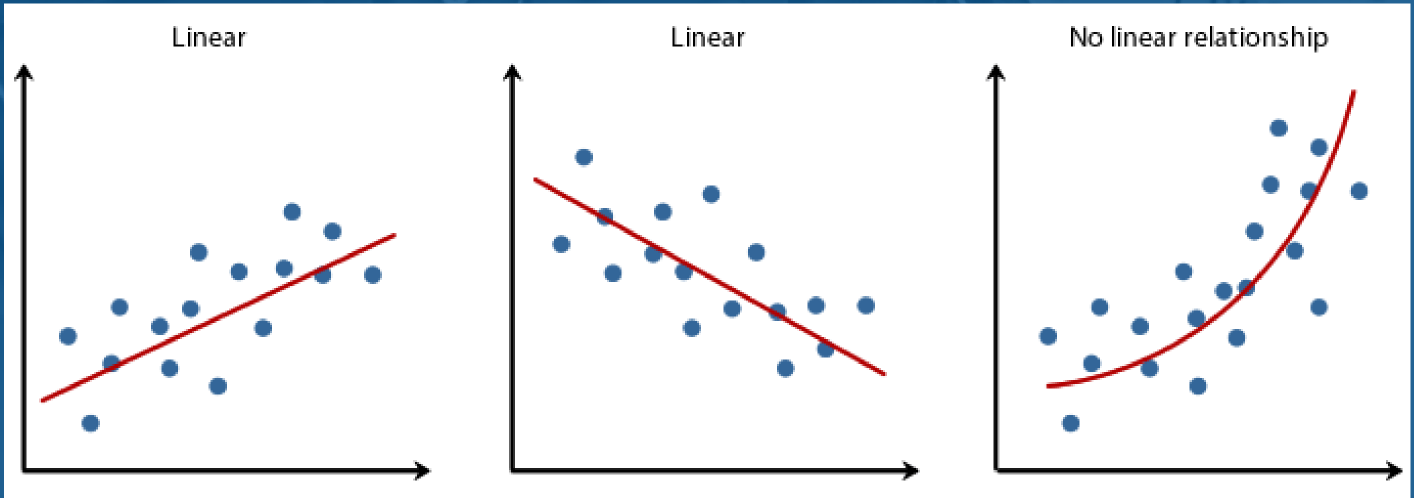








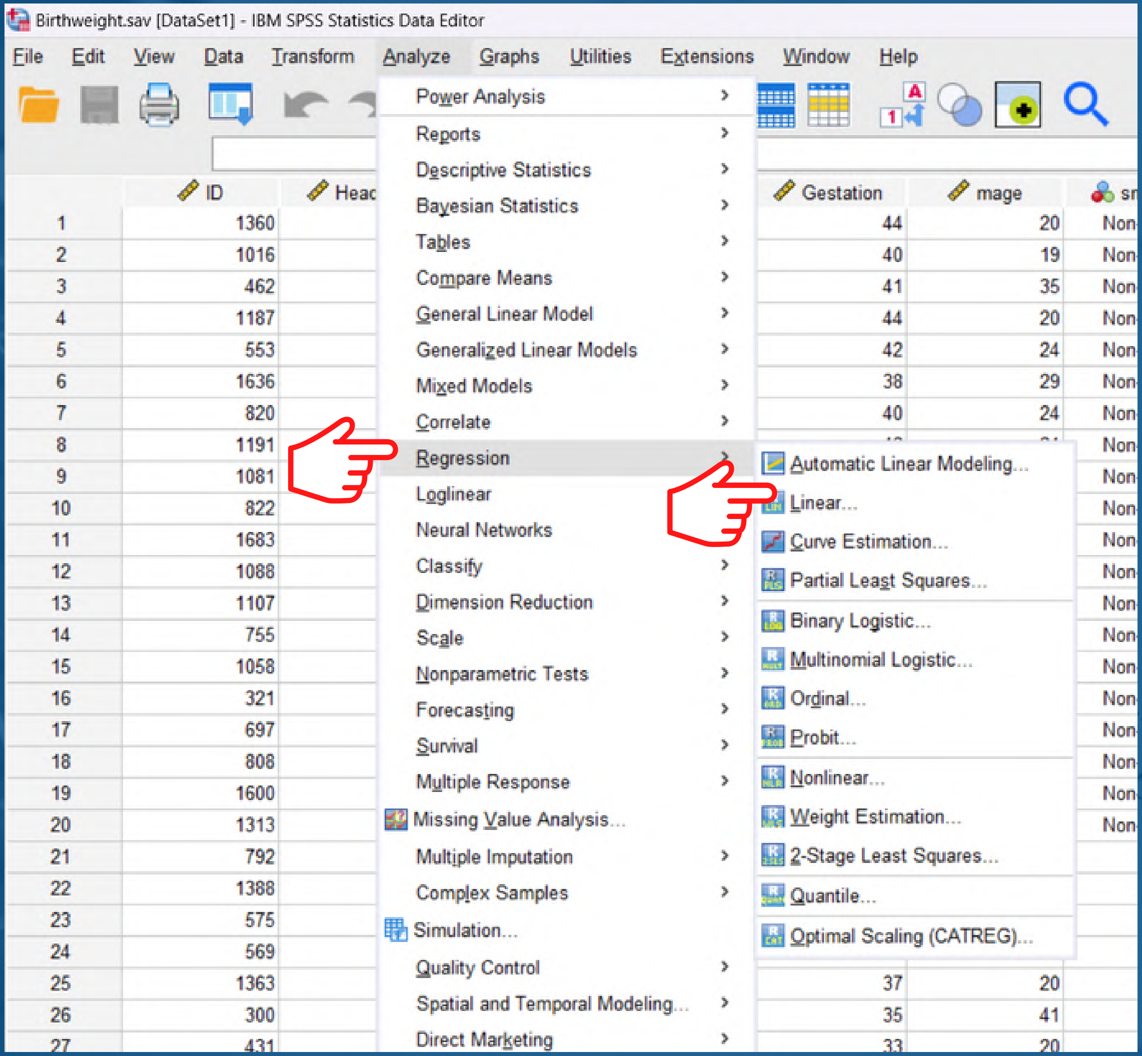
Linear relationship of two continuous variables. Linearity assumption is met.

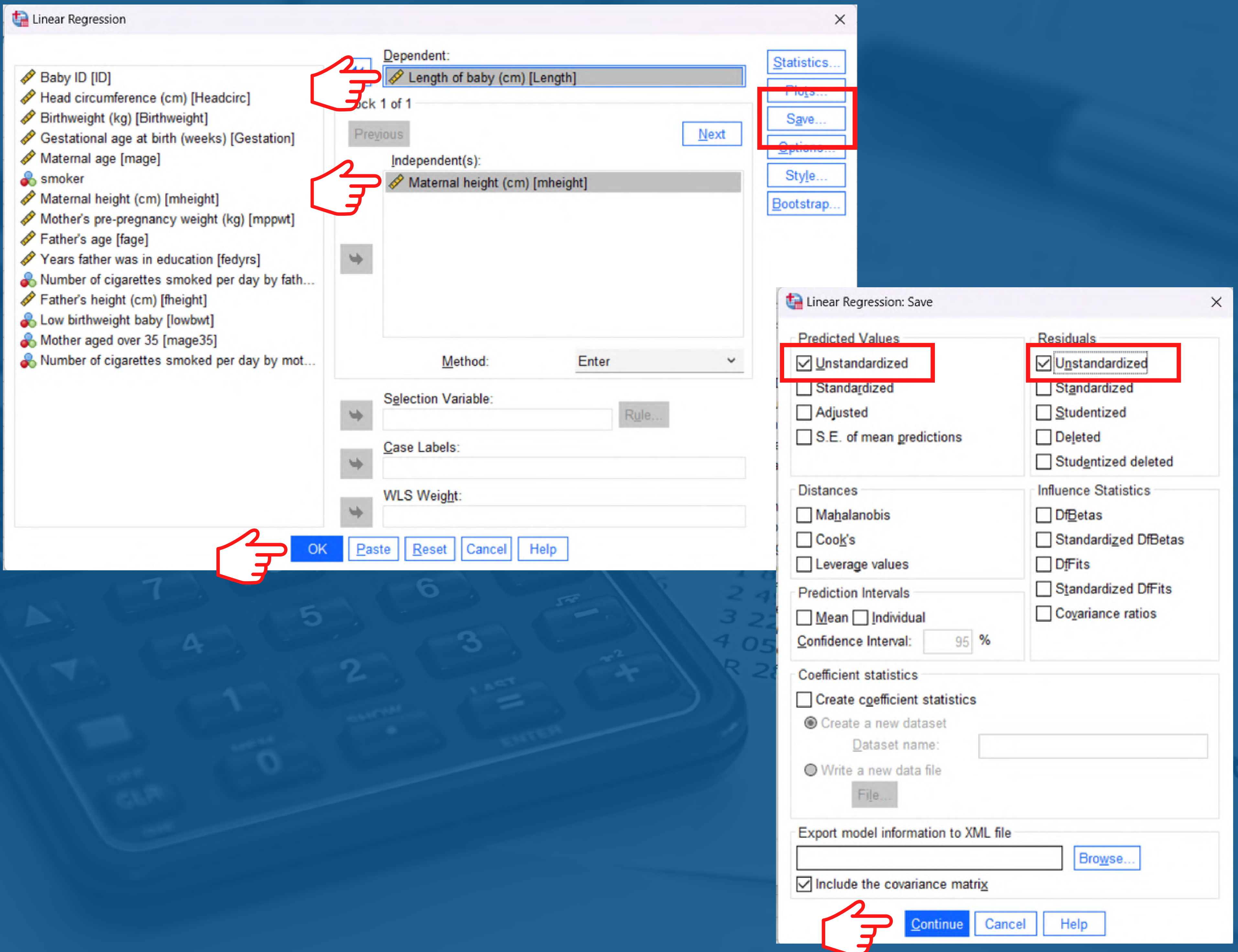


Example of linear and non-linear relationship

Checking assumption: Homoscedasticity (equal variances)

Go to: Analyze > Regression > Linear




Linear Regression
Baby ID [ID]
Head circumference (cm) [Headcirc]
Birthweight (kg) [Birthweight]
Gestational age at birth (weeks) [Gestation]
Maternal age [mage]
smoker
Maternal height (cm) [mheight]
Mother's pre-pregnancy weight (kg) [mppwt]
Father's age [fage]
Years father was in education [fedyrs]
Number of cigarettes smoked per day by fath...
Father's height (cm) [fheight]
Low birthweight baby [lowbwt]
Mother aged over 35 [mage35]
Number of cigarettes smoked per day by mot...
Dependent:
Length of baby (cm) [Length]
Previous
Next
Independent(s):
Maternal height (cm) [mheight]
Method:
Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Save...
Style...
Bootstrap...
OK
Paste
Reset
Cancel
Help
Linear Regression: Save
Predicted Values
Unstandardized
Standardized
Adjusted
S.E. of mean predictions
Residuals
Unstandardized
Standardized
Studentized
Deleted
Studentized deleted
Distances
Mahalanobis
Cook's
Leverage values
Influence Statistics
DfBetas
Standardized DfBetas
DfFits
Standardized DfFits
Covariance ratios
Prediction Intervals
Mean
Individual
Confidence Interval:
95
%
Coefficient statistics
Create coefficient statistics
Create a new dataset
Dataset name:
Write a new data file
File...
Export model information to XML file
Browse...
Include the covariance matrix
Continue
Cancel
Help

Predicted value

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	15.334	10.271		1.493	.143
	Maternal height (cm)	.219	.062	.485	3.507	.001

a. Dependent Variable: Length of baby (cm)

	ID	Length	mheight	PRE_1	RES_1
1	1360	56	162	50.79650	5.20350

y = a + bx

Length of baby = 15.33 + (0.22*mother's height)

Predicted length of baby whose mother's height is 162cm is:

Length of baby = 15.33 + (0.22*162) = 50.97

Residuals

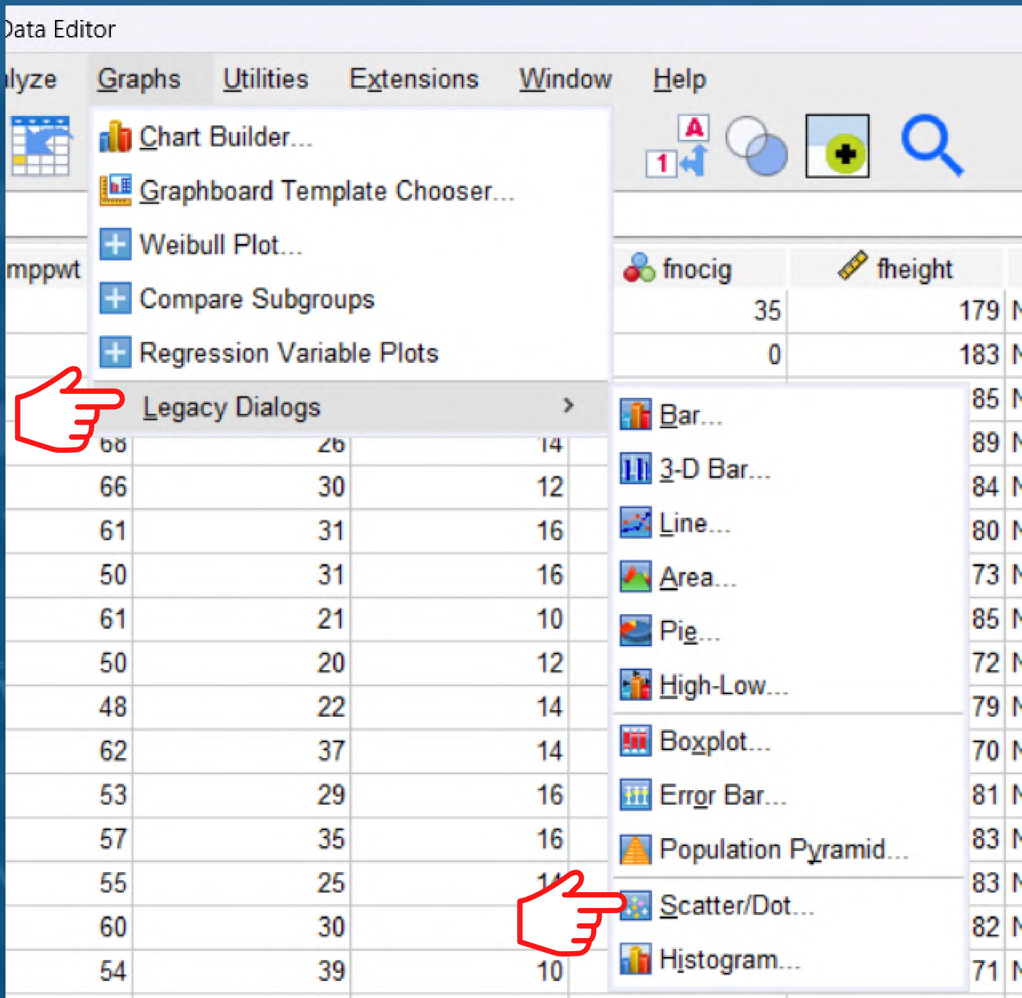
	ID	Length	mheight	PRE_1	RES_1
1	1360	56	162	50.79650	5.20350

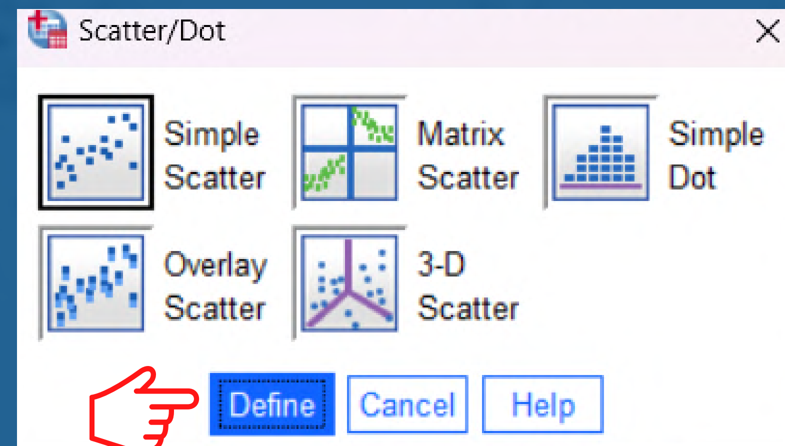
The difference between the observed value of the dependent variable and the value predicted by the regression line.

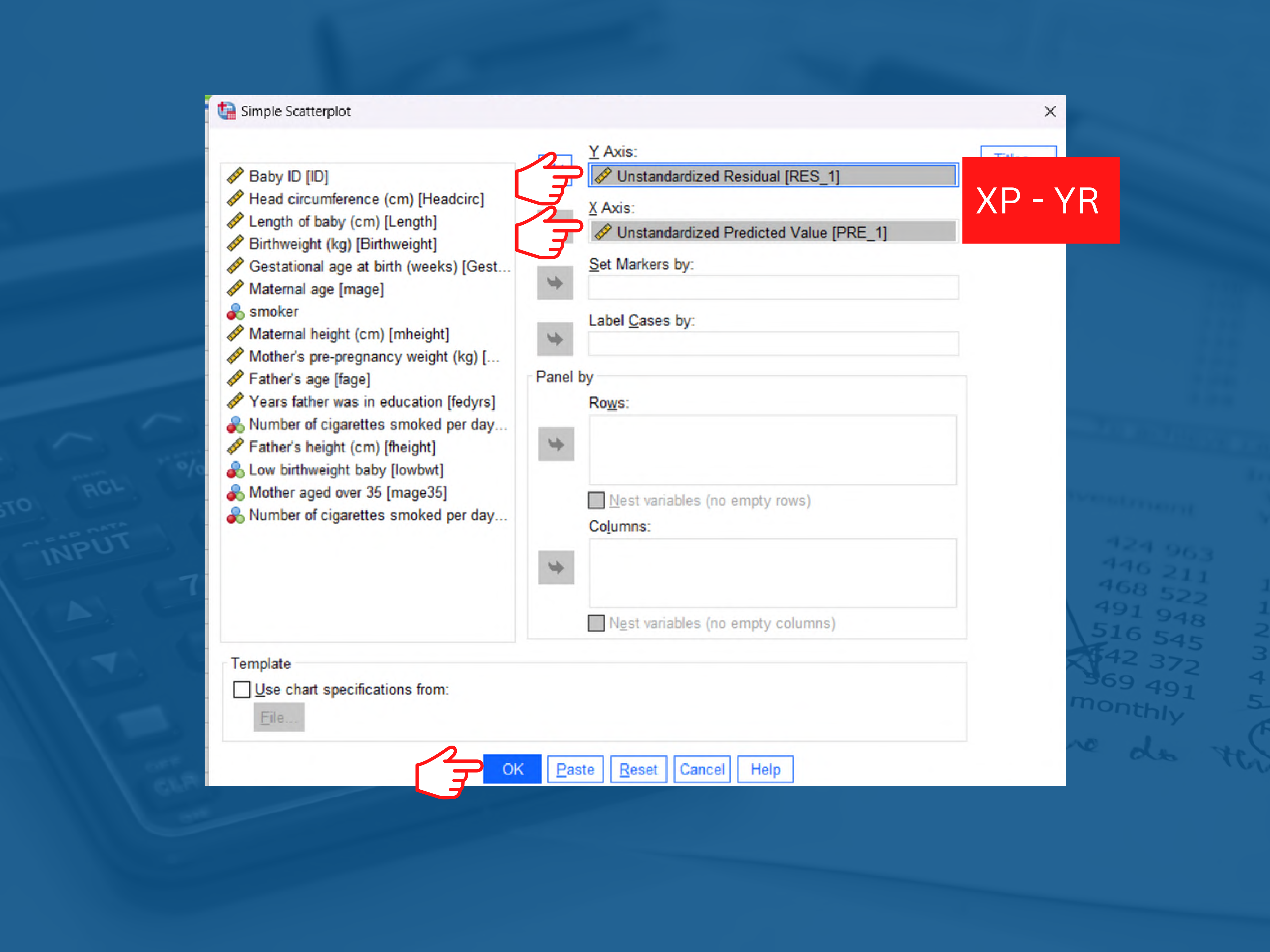
Residual = observed length - predicted length

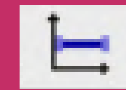
Residual = 56 - 50.8 = 5.2

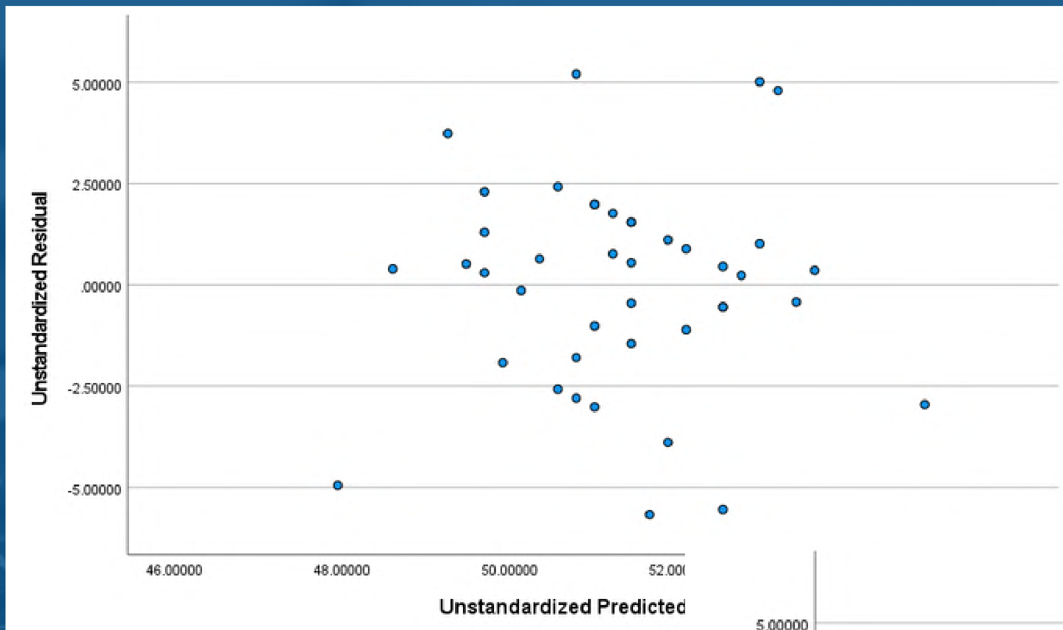
Go to: Graph > Legacy Dialogs > Scatter/Dot

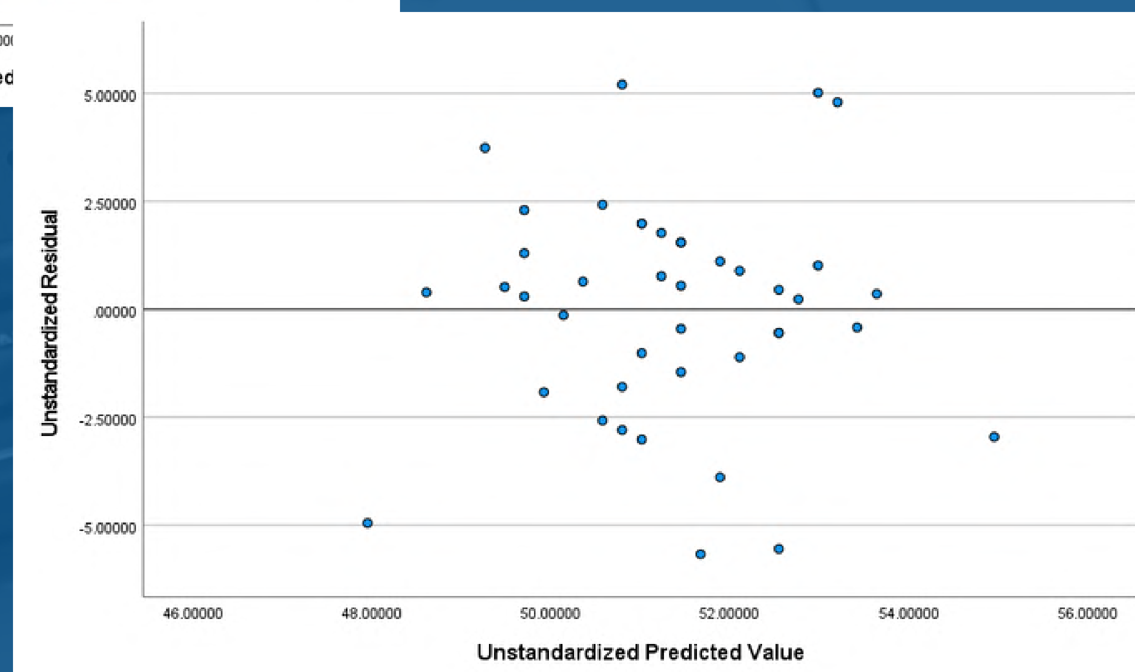


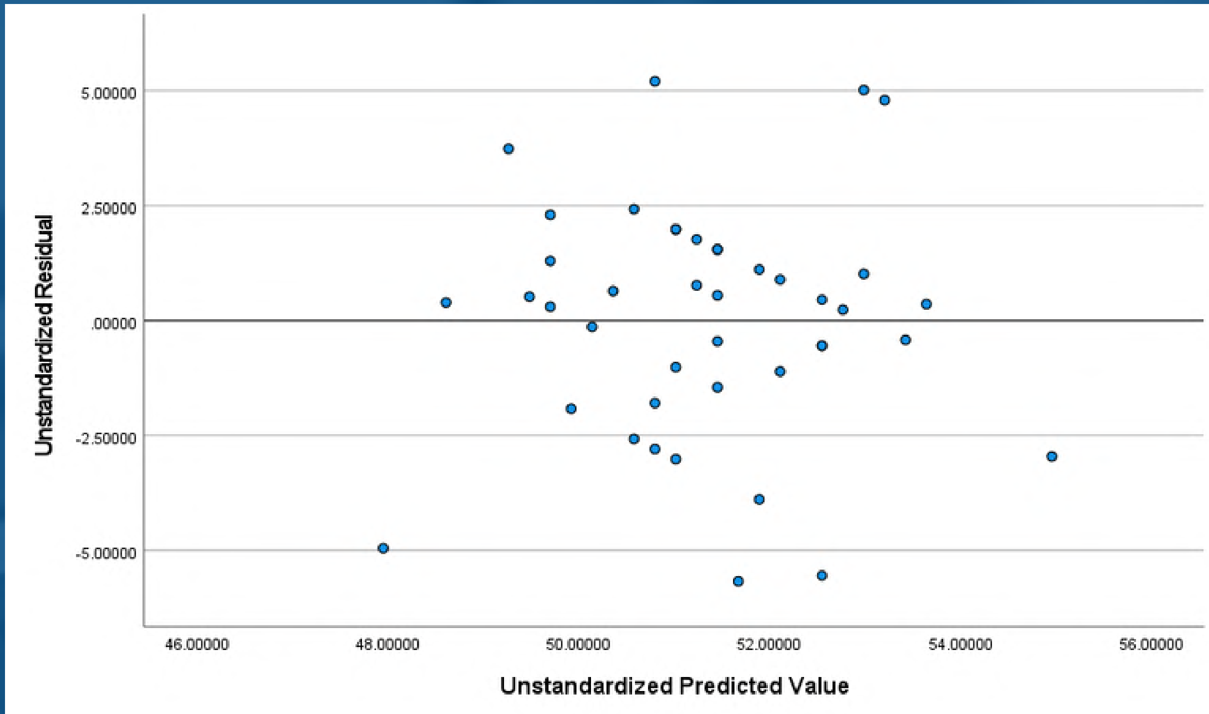



Simple Scatterplot
Baby ID [ID]
Head circumference (cm) [Headcirc]
Length of baby (cm) [Length]
Birthweight (kg) [Birthweight]
Gestational age at birth (weeks) [Gest...
Maternal age [mage]
smoker
Maternal height (cm) [mheight]
Mother's pre-pregnancy weight (kg) [...
Father's age [fage]
Years father was in education [fedyrs]
Number of cigarettes smoked per day...
Father's height (cm) [fheight]
Low birthweight baby [lowbwt]
Mother aged over 35 [mage35]
Number of cigarettes smoked per day...
Y Axis:
Unstandardized Residual [RES_1]
X Axis:
Unstandardized Predicted Value [PRE_1]
XP - YR
Set Markers by:
Label Cases by:
Panel by
Rows:
Nest variables (no empty rows)
Columns:
Nest variables (no empty columns)
Template
Use chart specifications from:
File...
OK
Paste
Reset
Cancel
Help

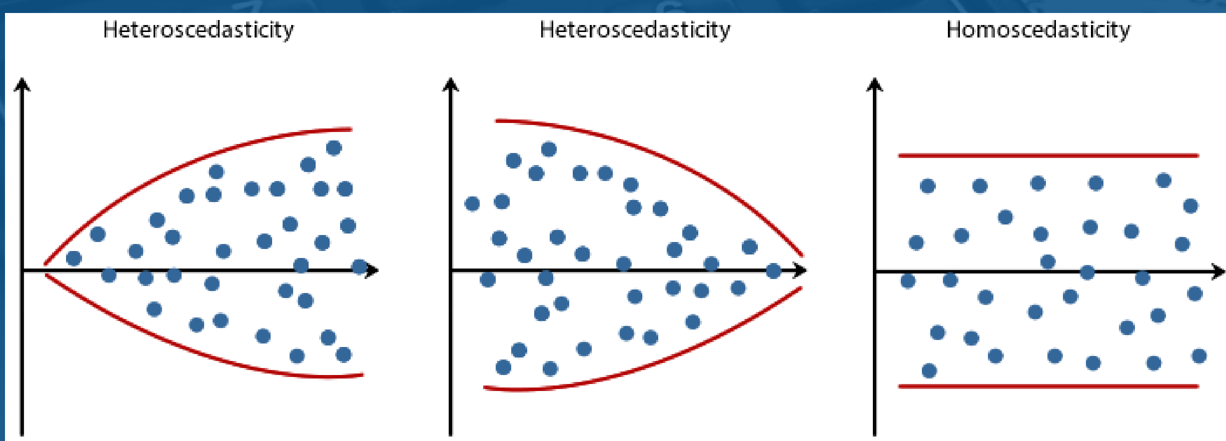
Double click the plot and click 







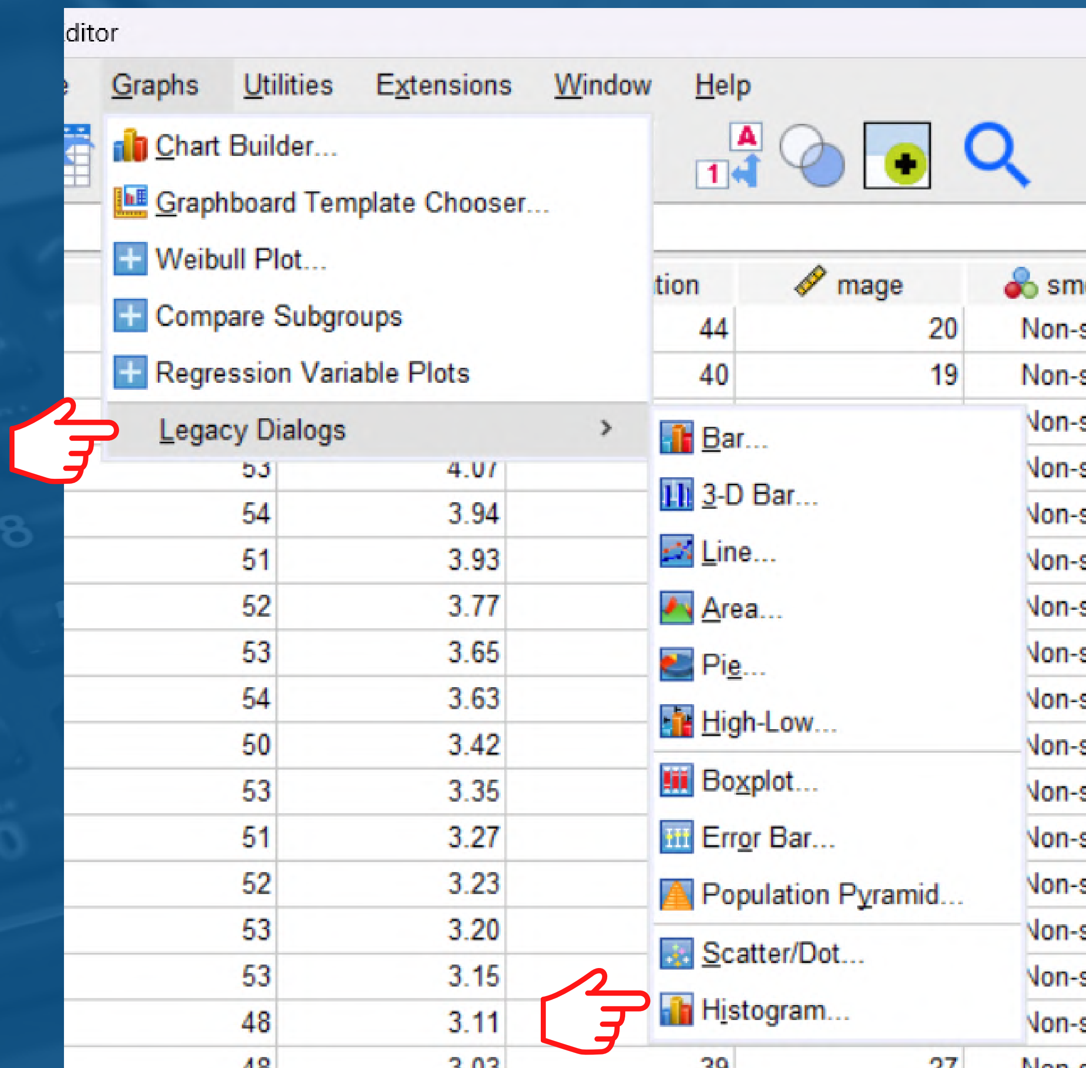
There is no pattern in the scatter. Homoscedasticity assumption is met.

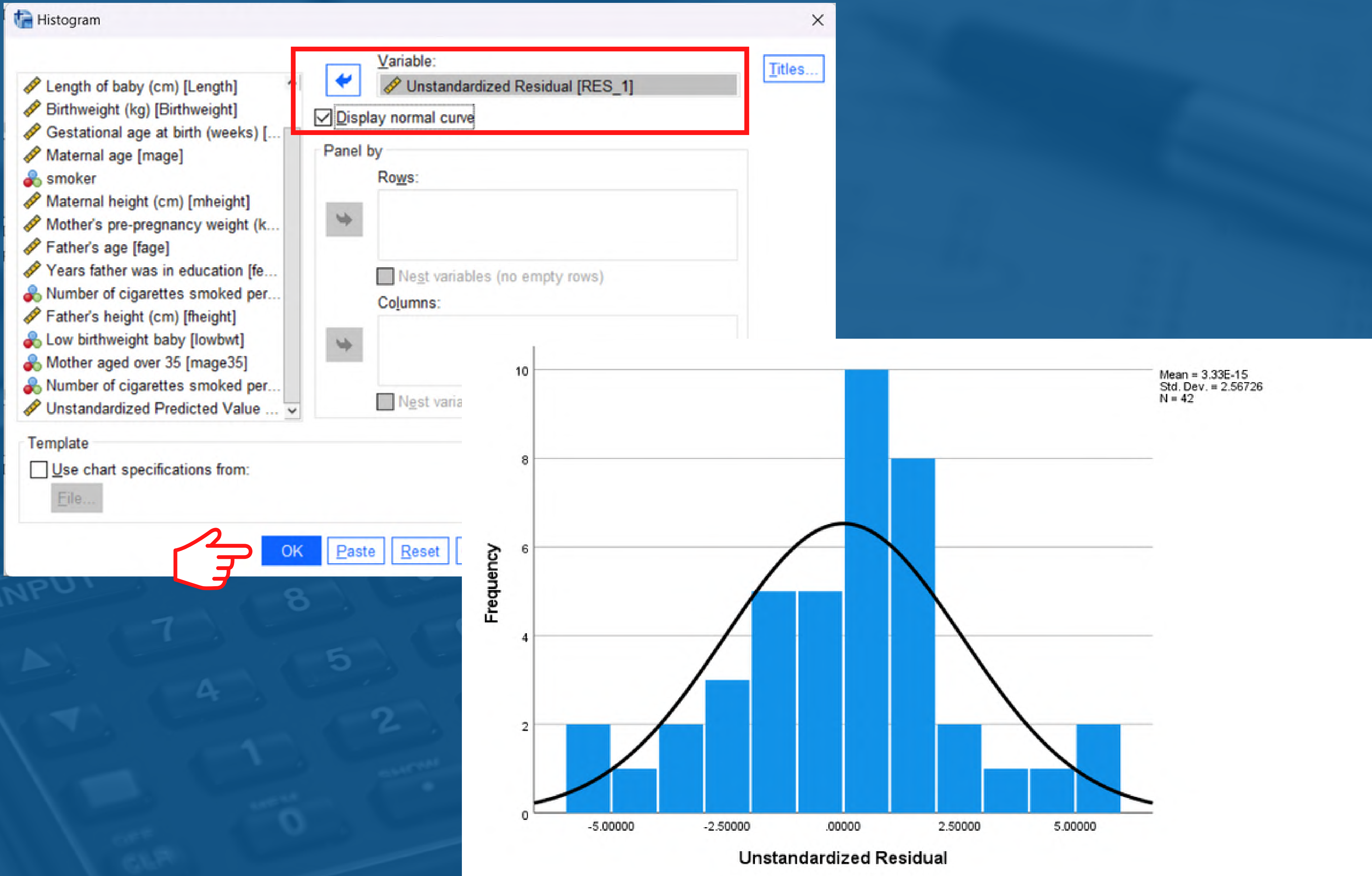


Example of heteroscedasticity and homoscedaticity

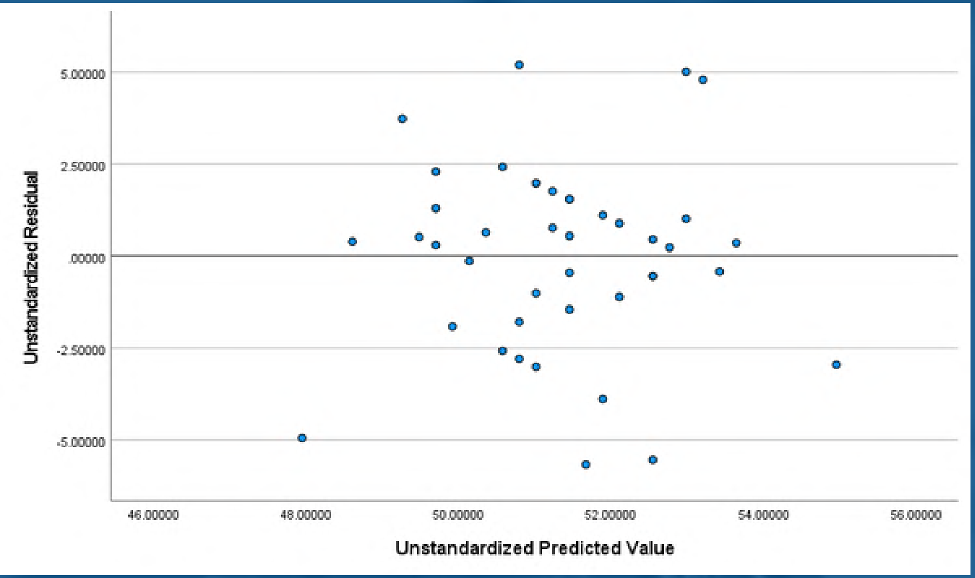
Checking assumption: Normality distribution of residuals

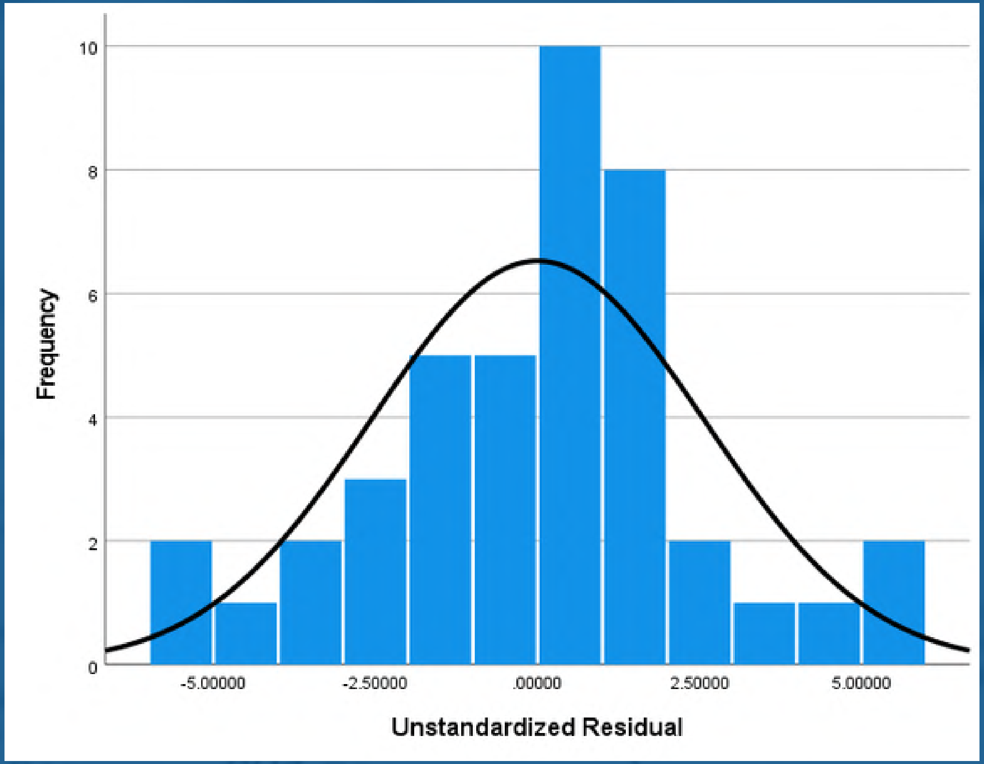
Go to: Graphs > Legacy Dialogs > Histogram

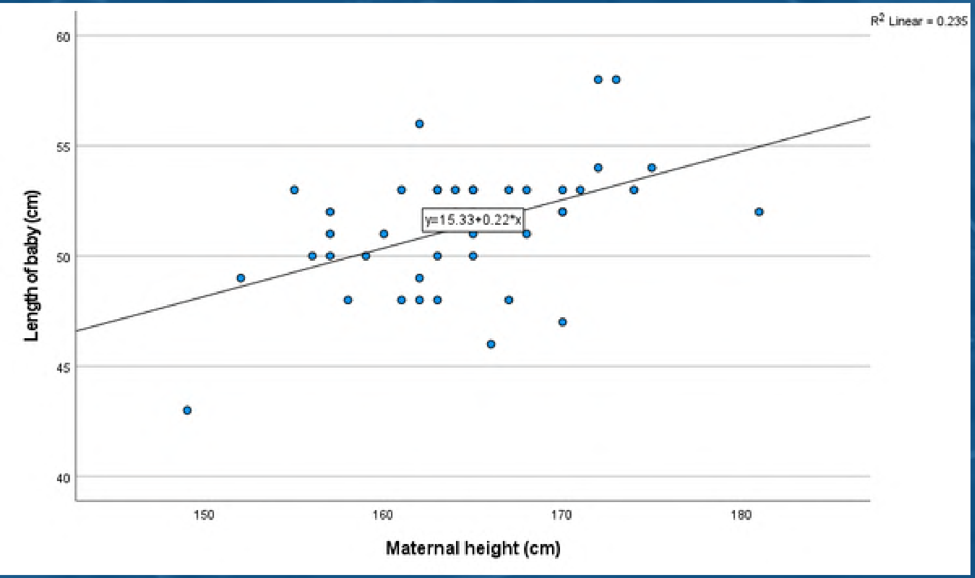




Residuals are normally distributed. Assumption is met.







Assumptions for homoscedasticity, linearity and normally distributed are met.

Step 4: Result Interpretation & Conclusion

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	15.334	10.271		1.493	.143	-5.425	36.093
	Maternal height (cm)	.219	.062	.485	3.507	.001	.093	.345

a. Dependent Variable: Length of baby (cm)

Interpretation:

Increasing the mother's height by 1 cm will result in a 0.2 cm increase in the length of the baby (b=0.22, 95% CI 0.09, 0.35, p=0.001).

Regression equation:

$y = a + bx$

length of baby = 15.33 + (0.22*mother's height)

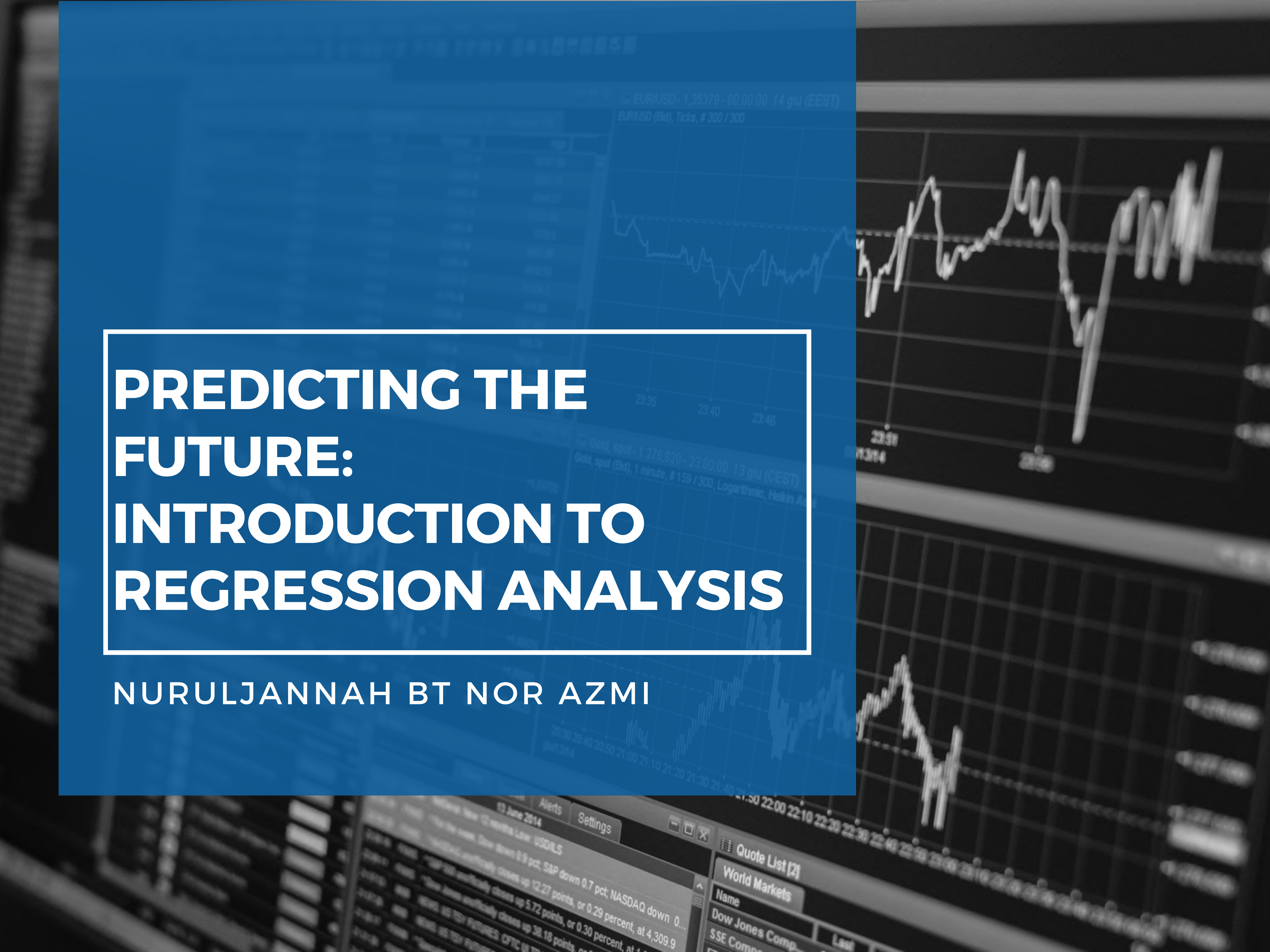
REGRESSION ANALYSIS

DEPENDENT VARIABLE	INDEPENDENT VARIABLE	STATISTICAL TEST
Numerical	Numerical	Multiple Linear Regression
Categorical (dichotomous)	Numerical and categorical	Multiple/Binary Logistic Regression
Categorical (polytomous - nominal)	Numerical and categorical	Multinomial Logistic Regression
Categorical (ordinal)	Numerical and categorical	Ordinal Logistic Regression

MDM NURULJANNAH BT NOR AZMI

EMAIL: nuruljannah@mahsa.edu.my


THANK
YOU



PREDICTING THE FUTURE: INTRODUCTION TO REGRESSION ANALYSIS

NURULJANNAH BT NOR AZMI

MULTIPLE LINEAR REGRESSION

Multiple linear regression is used to estimate the relationship between **two or more independent variables** and **one dependent variable.**

Dependent (outcome) : numerical

Independent (predictor) : 2 or more numerical variables



MULTIPLE LINEAR REGRESSION

If independent variables are combination of numerical and categorical or categorical only - **General Linear Regression**

Dependent (outcome) : numerical

Independent (predictor) : 2 or more combination of numerical and categorical or categorical only

SIMPLE LINEAR REGRESSION - ONLY ONE INDEPENDENT VARIABLE

Independent variable (x) → Dependent variable (y)

Mother's height → Length of baby

MULTIPLE LINEAR REGRESSION - MORE THAN ONE INDEPENDENT VARIABLES

Independent variables (x) → Dependent variable (y)

Mother's height
Mother's weight
Age
→ Length of baby

Multiple Linear Regression Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots\ldots\ldots \beta_n X_n$$

Y = outcome

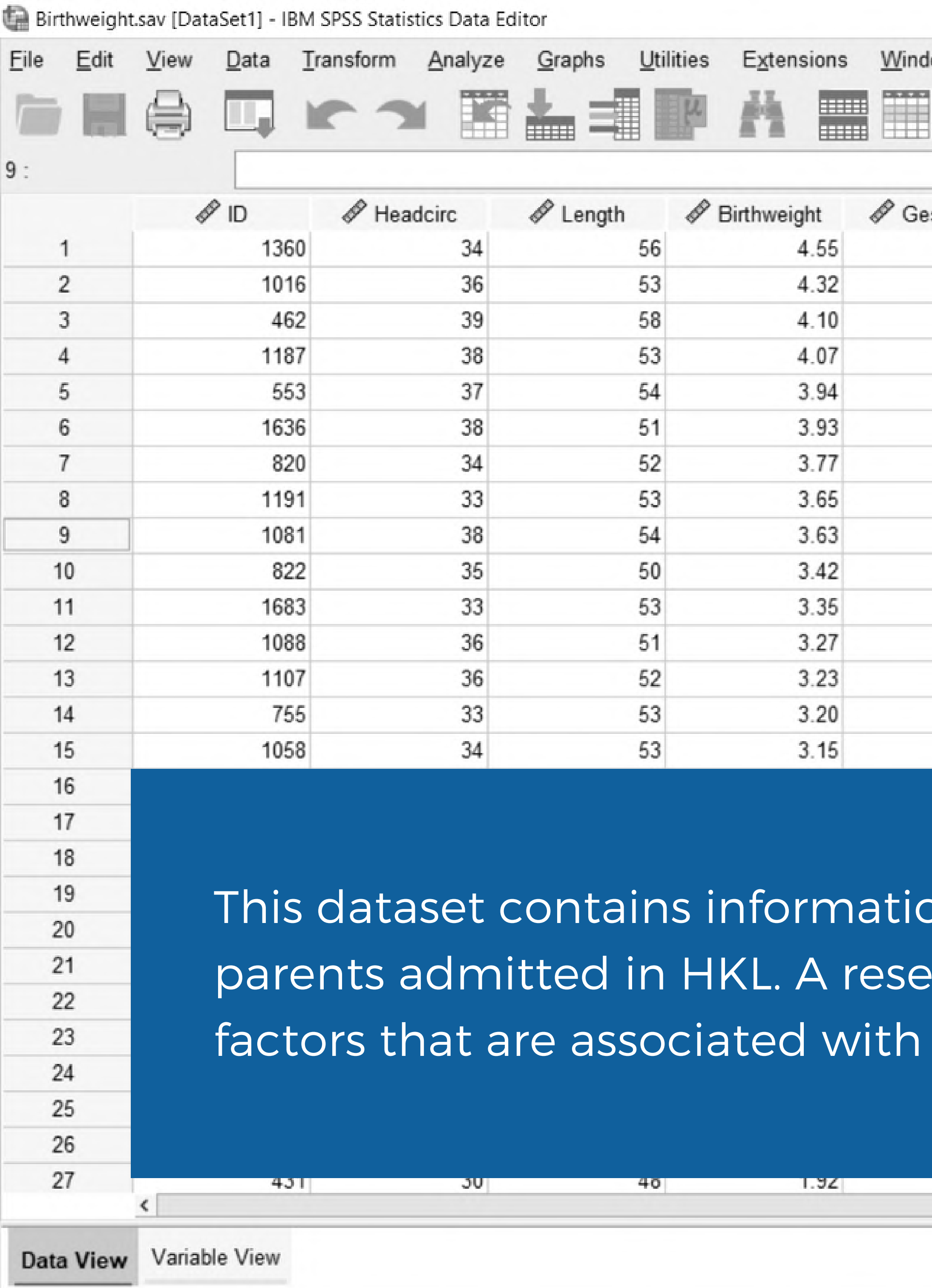
β_0 = intercept

β_1 .... β_n = regression coefficient for independent variable

X_1 .... X_n = independent variable

STEPS IN MULTIPLE LINEAR REGRESSION

1. Descriptive statistics
2. Simple linear regression (Univariable analysis)
3. Multiple linear regression (Multivariable analysis)
4. Checking multicollinearity & interaction (Preliminary final model)
5. Checking assumptions (final model)
6. Interpretation & presentation



	ID	Headcirc	Length	Birthweight	Ges
1	1360	34	56	4.55	
2	1016	36	53	4.32	
3	462	39	58	4.10	
4	1187	38	53	4.07	
5	553	37	54	3.94	
6	1636	38	51	3.93	
7	820	34	52	3.77	
8	1191	33	53	3.65	
9	1081	38	54	3.63	
10	822	35	50	3.42	
11	1683	33	53	3.35	
12	1088	36	51	3.27	
13	1107	36	52	3.23	
14	755	33	53	3.20	
15	1058	34	53	3.15	

EXAMPLE

Open dataset:
birthweight.sav

This dataset contains information on new born babies and their parents admitted in HKL. A researcher wants to determine the factors that are associated with the length of baby.

EXAMPLE

RQ: What are the factors that associated with the length of baby?

Length of baby (DV)

Factors (IV)

- Mother's age
- Mother's height
- Mother's weight

List down all the variables

Numerical

Numerical

Identify the types of variables

Multiple Linear Regression

Identify the right statistical analysis

STEP 1: DESCRIPTIVE STATISTICS

1.Data exploration and cleaning.

2.For categorical data, run the data by using Frequencies in SPSS.

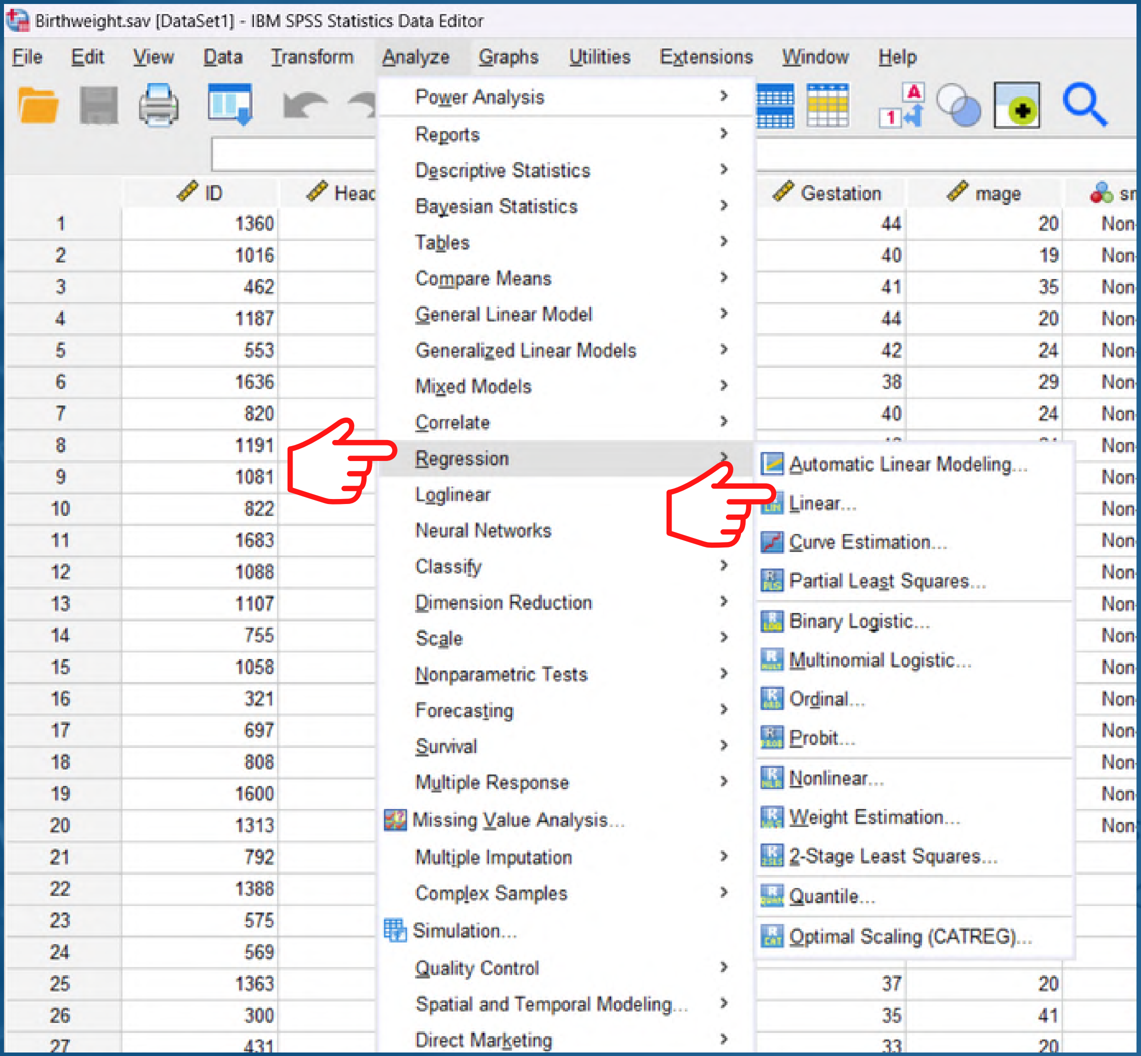
3.For numerical data, run the data by using Descriptives/Explore in SPSS.

STEP 2: SIMPLE LINEAR REGRESSION (UNIVARIABLE ANALYSIS)

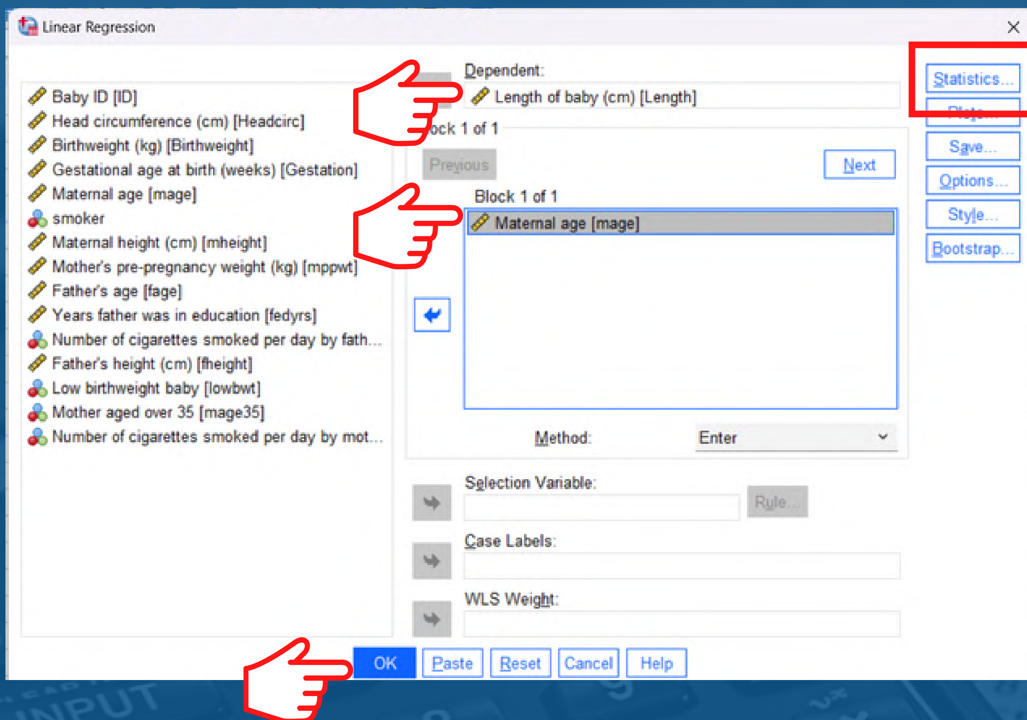
1. Do Simple Linear Regression analysis for each independent variable:
   - Mother's age
   - Mother's height
   - Mother's weight

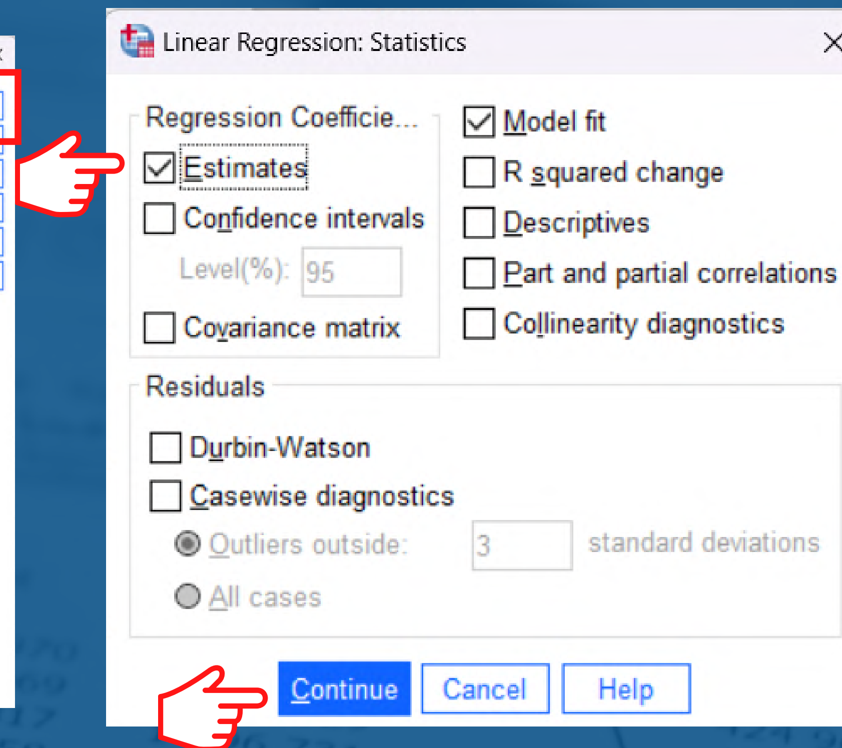
2. At the end, choose variables with p-value < 0.25 and/or clinically important.

Go to: Analyze > Regression > Linear



Length of baby vs Mother's age





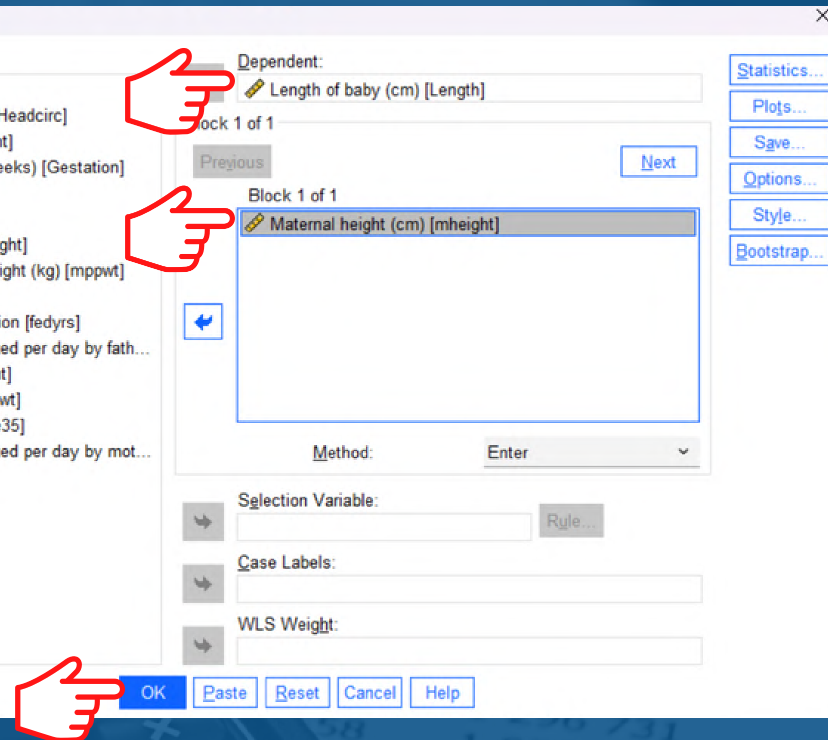
Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	60.461	1.400		43.182	<.001	57.631	63.291
	Maternal age	-.318	.048	-.727	-6.691	<.001	-.414	-.222

a. Dependent Variable: Length of baby (cm)

- There is a significant relationship between mother's age and the length of baby.

Length of baby vs Mother's height



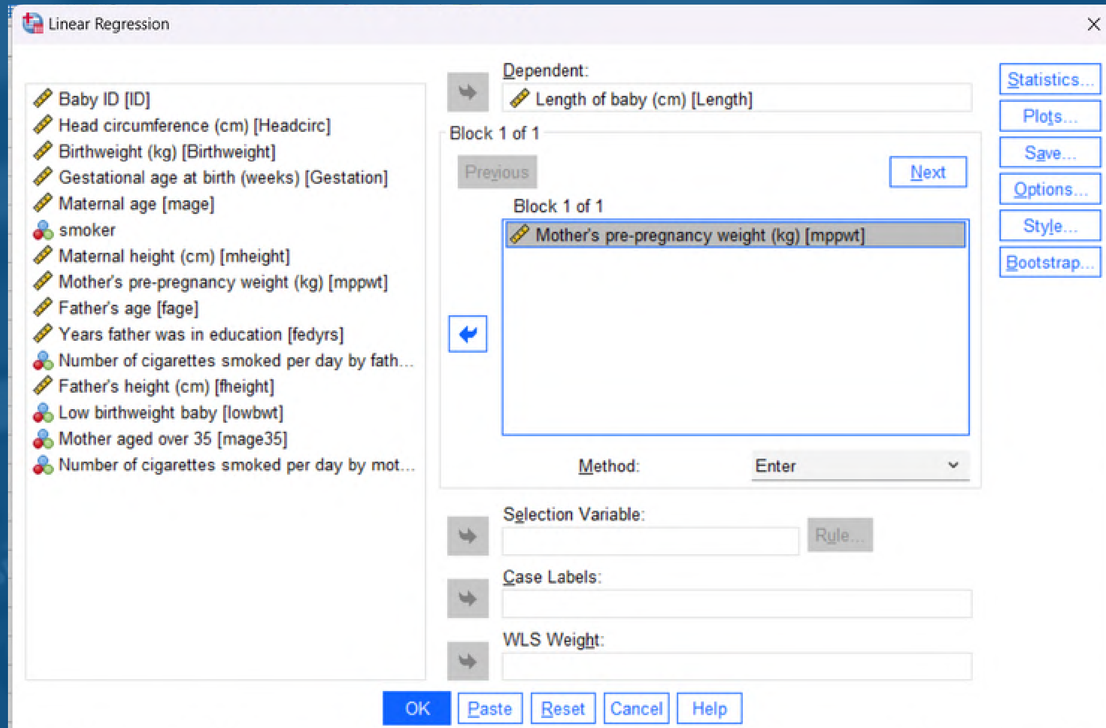
Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	15.334	10.271		1.493	.143	-5.425	36.093
	Maternal height (cm)	.219	.062	.485	3.507	.001	.093	.345

a. Dependent Variable: Length of baby (cm)

- There is a significant relationship between mother's height and the length of baby.

Length of baby vs Mother's weight



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	41.996	3.427		12.255	<.001	35.070	48.922
	Mother's pre-pregnancy weight (kg)	.162	.059	.398	2.745	.009	.043	.282

a. Dependent Variable: Length of baby (cm)

- There is a significant relationship between mother's pre-pregnancy weight and the length of baby.

Table 1: Associated factors of the length of baby by Simple Linear Regression

Variable	Simple Linear Regression	
	b* (95% CI)	*p*-value
Mother's age	-0.32 (-0.41,-0.22)	<0.001
Mother's height (cm)	0.22 (0.09,0.35)	0.001
Mother's pre-pregnancy weight (kg)	0.16 (0.04,0.28)	0.009

STEP 3: MULTIPLE LINEAR REGRESSION (MULTIVARIABLE ANALYSIS)

1.Variables selection can be done by using following methods:

- Forward
- Backward
- Stepwise

2. Perform all the methods and select the model with all variables significant as the preliminary main effect model.

STEP 4: CHECKING MULTICOLLINEARITY

1.Multicollinearity occurs when independent variables in a regression model are correlated.

2. This correlation is a problem because independent variables should be independent.

3.If the degree of correlation between variables is high enough, it can cause problems when you fit the model and interpret the results.

4. There is a high chance of getting inaccurate p-values and wide confidence interval of regression coefficient.

STEP 4: CHECKING MULTICOLLINEARITY

5. Multicollinearity can be checked by using Variance Inflation Factor (VIF).

6. If VIF is more than 10, then there is a multicollinearity amongst independent variables.

STEP 4: CHECKING INTERACTION

1. An interaction effect occurs when the effect of one variable depends on the value of another variable.

2. The interaction terms need to be biologically meaningful.

3. The interaction term needs to be computed in SPSS and then added to the model as an independent variable. If you have more than one interaction term, add to the model one by one.

4. If the interaction term is statistically significant, include the term in the model.

STEP 5: CHECKING ASSUMPTIONS

Assumptions	How to check?
1.Independent observation	Done during design stage
2.Overall linearity	Scatter plot between residuals and predicted values (XP - YR)
3.Homoscedasticity (Equal variances)	Scatter plot between residuals and predicted values (XP - YR)
4.Linearity of each independent variable	Scatter plot residual vs each independent variable (XI - YR)
5.Residuals should be approximately normally distributed	Histogram with overlaid normal curve of residuals

STEP 6: INTERPRETATION AND PRESENTATION

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B Lower Bound	95.0% Confidence Interval for B Upper Bound
1	(Constant)	35.959	8.019		4.484	<.001	19.738	52.180
	Maternal age	-.282	.045	-.644	-6.303	<.001	-.372	-.191
	Maternal height (cm)	.143	.046	.316	3.094	.004	.049	.236

a. Dependent Variable: Length of baby (cm)

Run the final model. All the assumptions were checked and MET.

STEP 6: INTERPRETATION AND PRESENTATION

Table 2: Factors associated with the length of baby (n=42)

Variable	Simple Linear Regression		Multiple Linear Regression	
	b[a] (95% CI)	*p*-value	b[b] (95% CI)	*p*-value
Mother's age	-0.32 (-0.41,-0.22)	<0.001	-0.28 (-0.37,-0.19)	<0.001
Mother's height (cm)	0.22 (0.09,0.35)	0.001	0.14 (0.05,0.24)	0.004
Mother's pre-pregnancy weight (kg)	0.16 (0.04,0.28)	0.009	-	-

[a] Crude regression coefficient
[b] Adjusted regression coefficient
All model assumptions are fulfilled.
No multicollinearity problem detected and there were no interaction among the independent variables.
Coefficient of determination (R^2) = 0.621

Final model equation:
Length of baby = 35.96 – (0.28*mother's age) + (0.14*mother's height)

STEP 6: INTERPRETATION AND PRESENTATION

- There is a significant linear negative relationship between mother's age and the length of baby. For every one-year increase in the mother's age, the baby's length is 0.28 cm lower. (adjusted b = -0.28; 95% CI -0.37,-0.19; $p<0.001$)
- There is a significant linear positive relationship between mother's height and the length of baby. For every 1 cm increase in the mother's height, the baby's length increases by 0.14 cm. (adjusted b = 0.14; 95% CI 0.05,0.24; $p=0.004$)
- 62.1% of the variation in the length of baby is explained by mother's age and height according to the multiple linear regression model ($R^2 = 0.621$).

REGRESSION ANALYSIS

DEPENDENT VARIABLE	INDEPENDENT VARIABLE	STATISTICAL TEST
Numerical	Numerical	Multiple Linear Regression
Categorical (dichotomous)	Numerical and categorical	Multiple/Binary Logistic Regression
Categorical (polytomous - nominal)	Numerical and categorical	Multinomial Logistic Regression
Categorical (ordinal)	Numerical and categorical	Ordinal Logistic Regression

MDM NURULJANNAH BT NOR AZMI

EMAIL: nuruljannah@mahsa.edu.my


THANK
YOU