

The background of the slide features a collage of financial data. On the left, there's a blue-tinted area with a white border containing the title. Behind this and across the rest of the slide are various financial charts and tables. A prominent line chart on the right shows price fluctuations with a dashed horizontal line. Below it, another chart shows a similar pattern. At the bottom, a table titled 'Quote List (2)' lists market data for 'World Markets', including 'Dow Jones Comp.' and 'SSE Comp.'. Other visible text includes 'EURUSD - 1.35379 - 00.00.00 14 Jun (EEST)', 'Gold, spot - 1,276,820 - 23.00.00 13 Jun (CEST)', and '13 June 2014'.

# 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?**



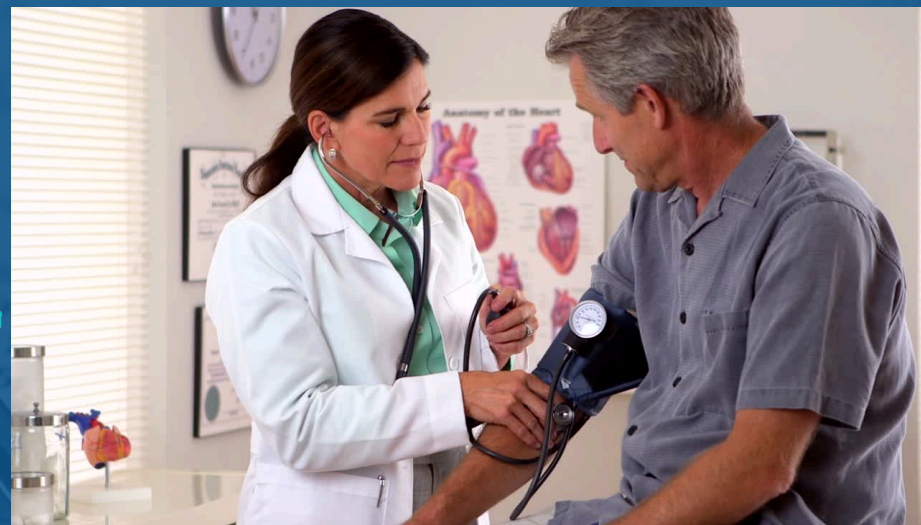
# When to apply Regression Analysis?

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?



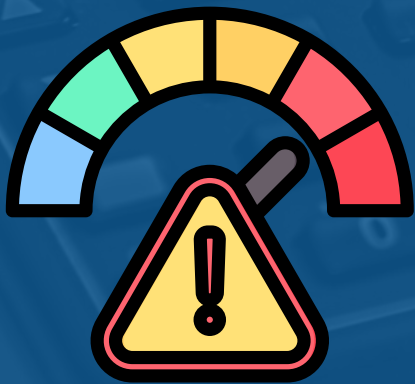
AGE +



42 372  
369 491  
Start at monthly  
Can we do this



Treatment methods



Severity of illness

Length of hospital stay







# 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)

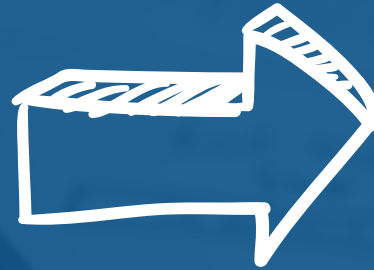
---



## SIMPLE LINEAR REGRESSION - ONLY ONE INDEPENDENT VARIABLE

Independent variable (x)

Mother's height



Dependent variable (y)

Length of baby

---

## MULTIPLE LINEAR REGRESSION - MORE THAN ONE INDEPENDENT VARIABLES

Independent variables (x)

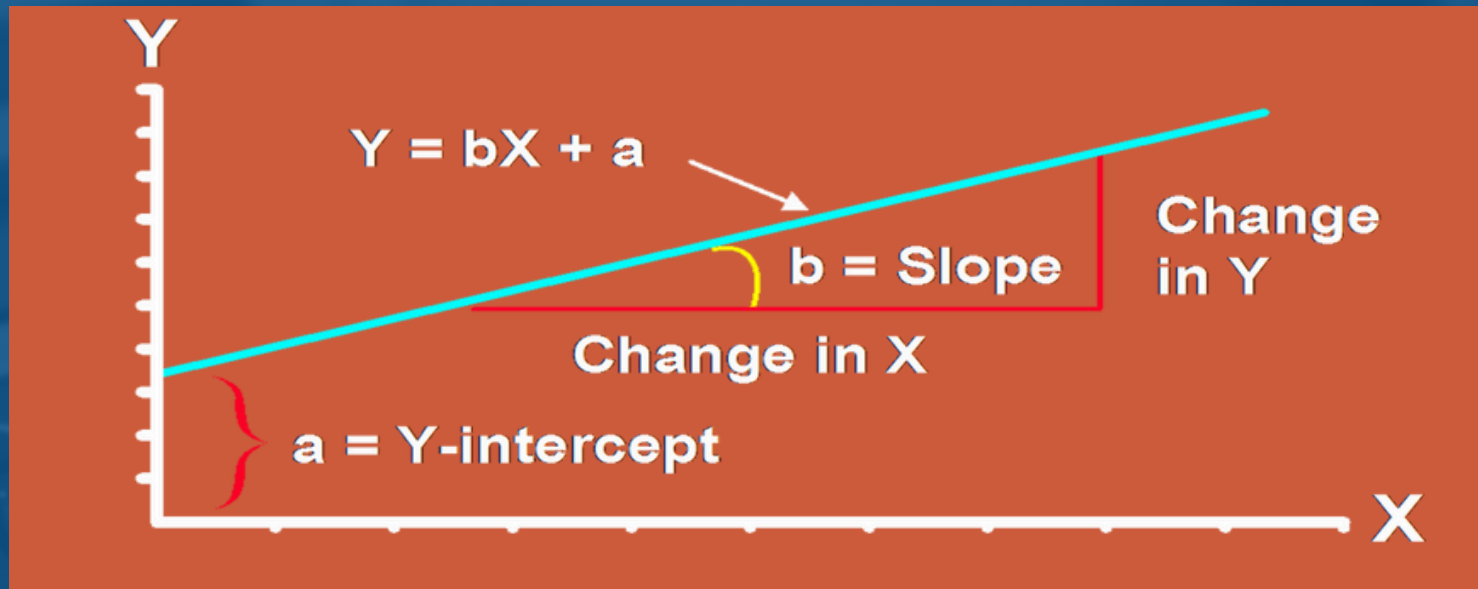
Mother's height  
Mother's weight  
Age



Dependent variable (y)

Length of baby

# 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)



# Simple Linear Regression Model

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:

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

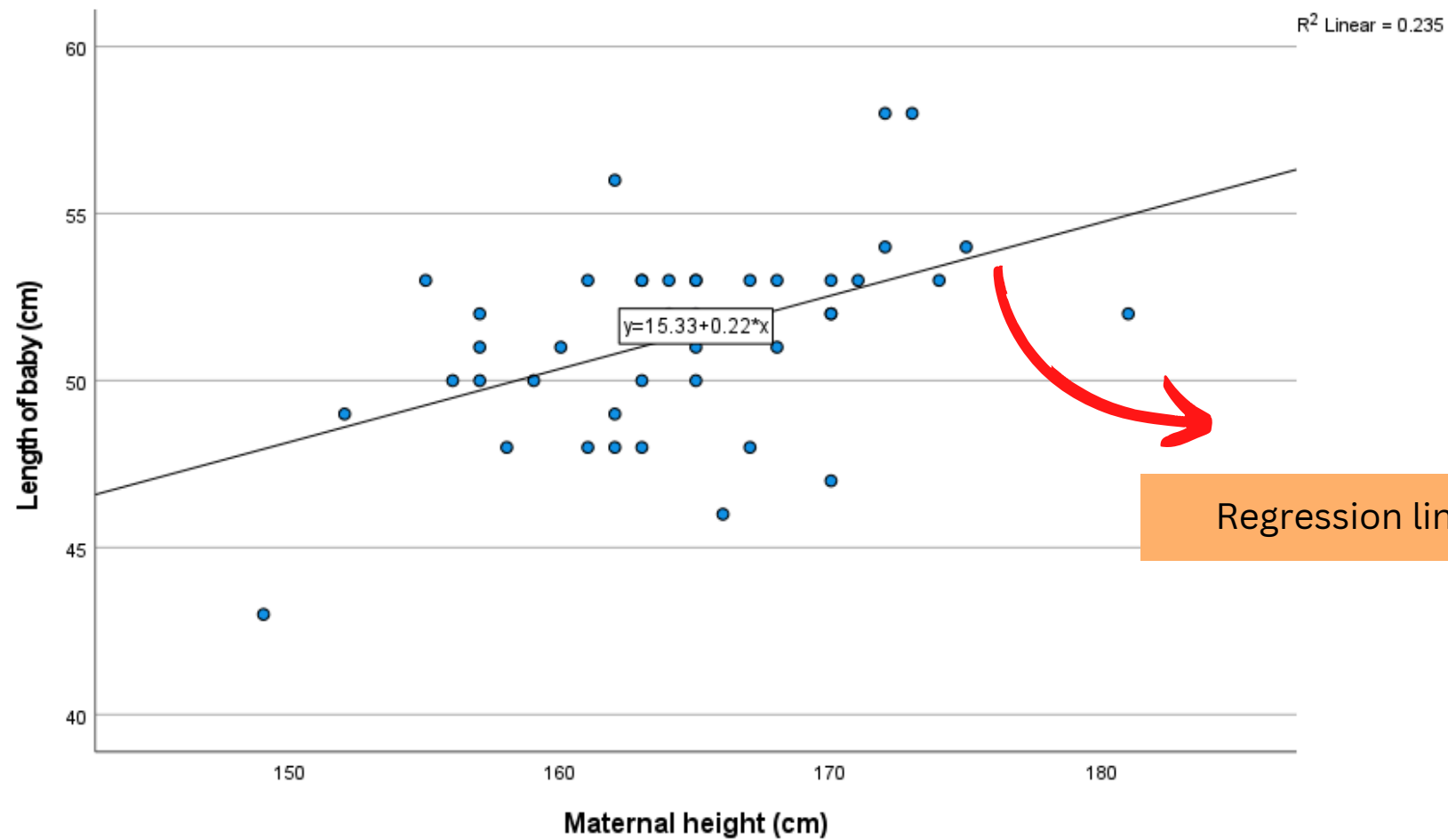
$Y$  = outcome

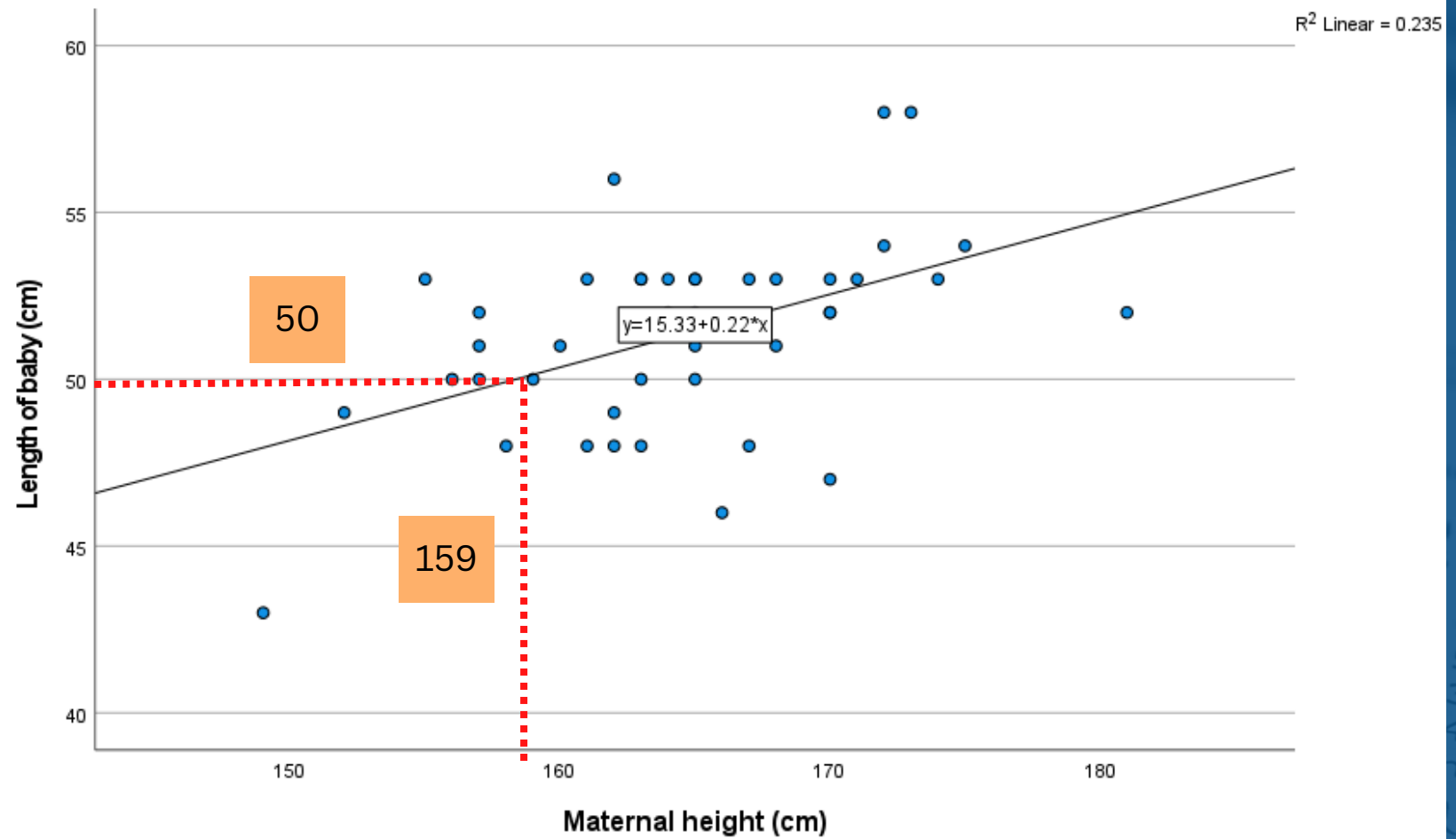
$\beta_0$  = an intercept of the regression line

$\beta_1$  = a slope of the line / regression coefficient for independent variable

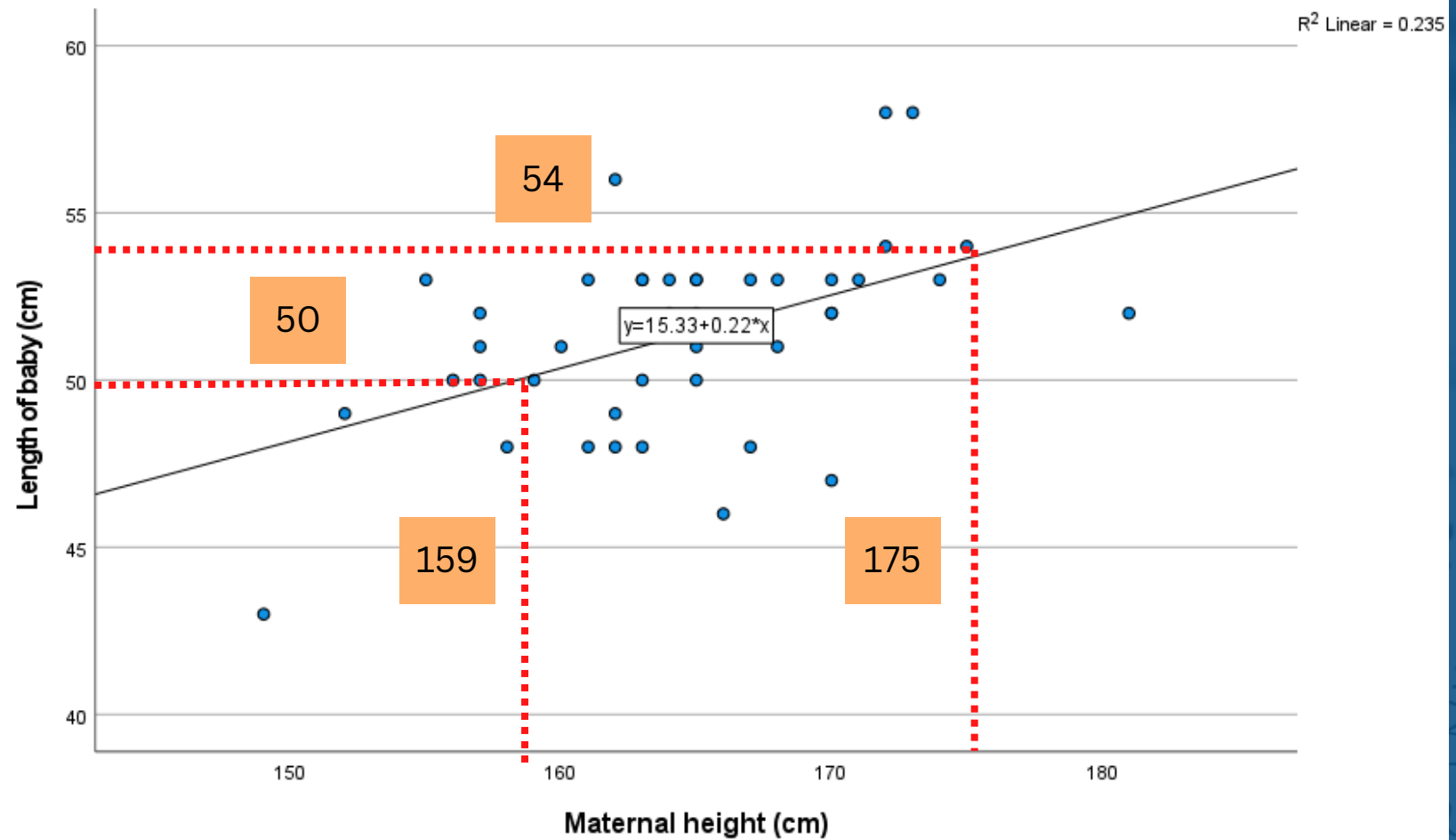
$X_1$  = independent variable

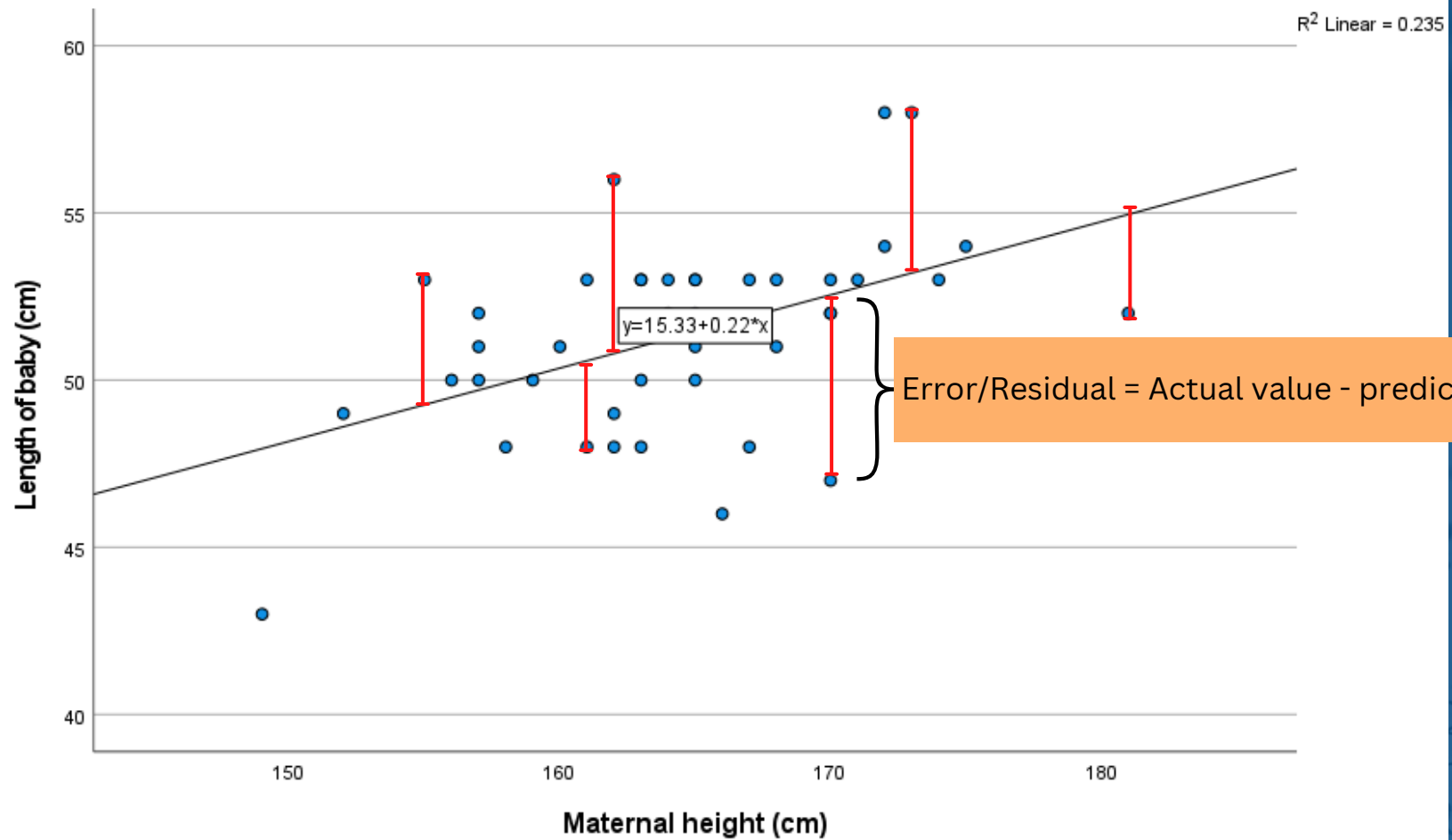
error = residual











# ASSUMPTIONS OF THE MODEL

---

**L**

Relationship between the independent and dependent variable is linear.  
(Linearity).

**I**

Independent  
observation

**N**

Residuals should be approximately  
normally  
distributed

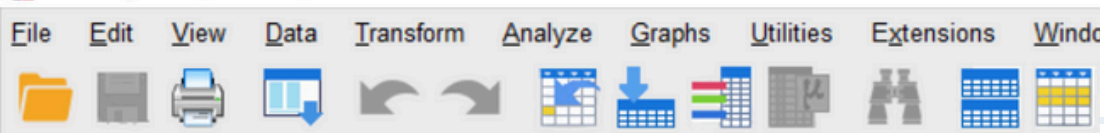
**E**

Homoscedasticity  
(Equal variances)



# CHECKING MODEL ASSUMPTIONS

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



9 :

	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	
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27	431	30	48	1.92	

# 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

The diagram illustrates the process of identifying variables and the statistical method used. It features a background image of a calculator and a document with handwritten notes. The text 'Mother's height' is positioned above the word 'Numerical', which is above a large orange arrow pointing to 'Simple Linear Regression'. Similarly, 'Length of baby' is above 'Numerical', which is also above the same large orange arrow. To the right of these elements are three pink boxes containing the steps: 'List down all the variables', 'Identify the types of variables', and 'Identify the right statistical analysis'.

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

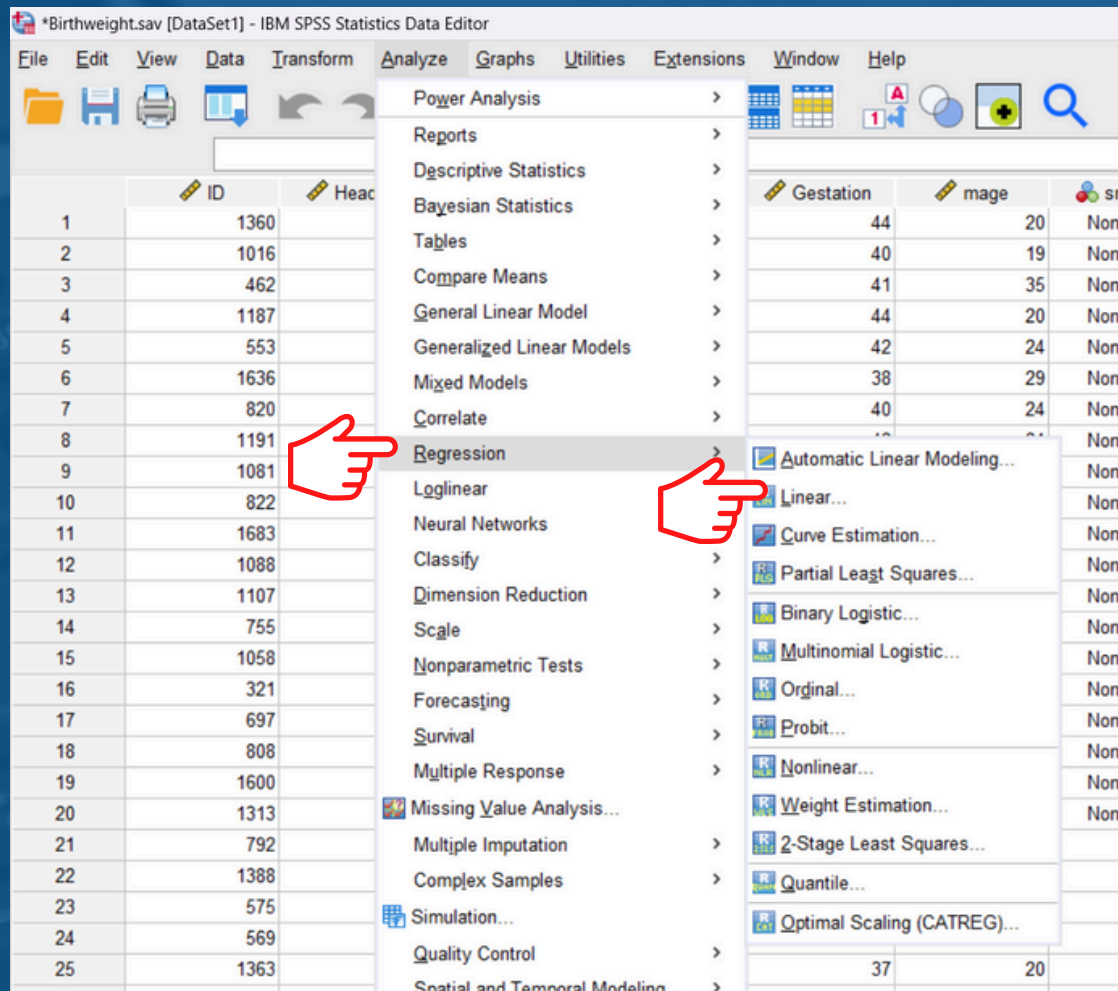
$H_0$  : There is no relationship between mother's height and the length of baby

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



## Step 2: Run Simple Linear Regression

Go to: Analyze > Regression > Linear



Linear Regression

Dependent: Length of baby (cm) [Length]

Block 1 of 1

Maternal height (cm) [mheight]

Method: Enter

Selection Variable: Rule...

Case Labels:

WLS Weight:

OK Paste Reset Cancel Help

Statistics... Projs... Save... Options... Style... Bootstrap

Linear Regression: Statistics

Regression Coefficients

☒ Model fit

☒ Estimates

☒ Confidence intervals

Level(%): 95

☐ Covariance matrix

☐ R squared change

☐ Descriptives

☐ Part and partial correlations

☐ Collinearity diagnostics

Residuals

☐ Durbin-Watson

☐ Casewise diagnostics

☒ Outliers outside: 3 standard deviations

☐ All cases

Continue Cancel Help

### Coefficients<sup>a</sup>

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	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 <sup>b</sup>				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.485 <sup>a</sup>	.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 ( $r^2 = 0.235$ ).

### Coefficient of Determination ( $r^2$ )

- Ranges from 0 to 1
- Provides a measure of how well future outcomes are likely to be predicted by the model.

ANOVA <sup>a</sup>						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	83.110	1	83.110	12.302	.001 <sup>b</sup>
	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 <sup>a</sup>	
	b* (95% CI)	p-value
Mother's height	0.22 (0.09,0.35)	0.001

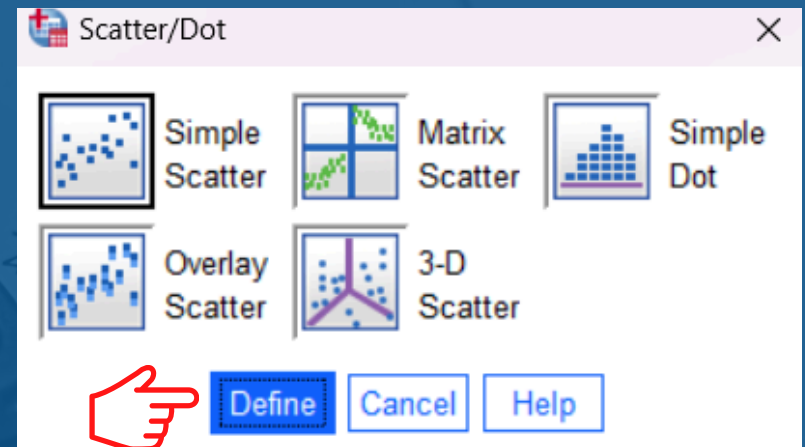
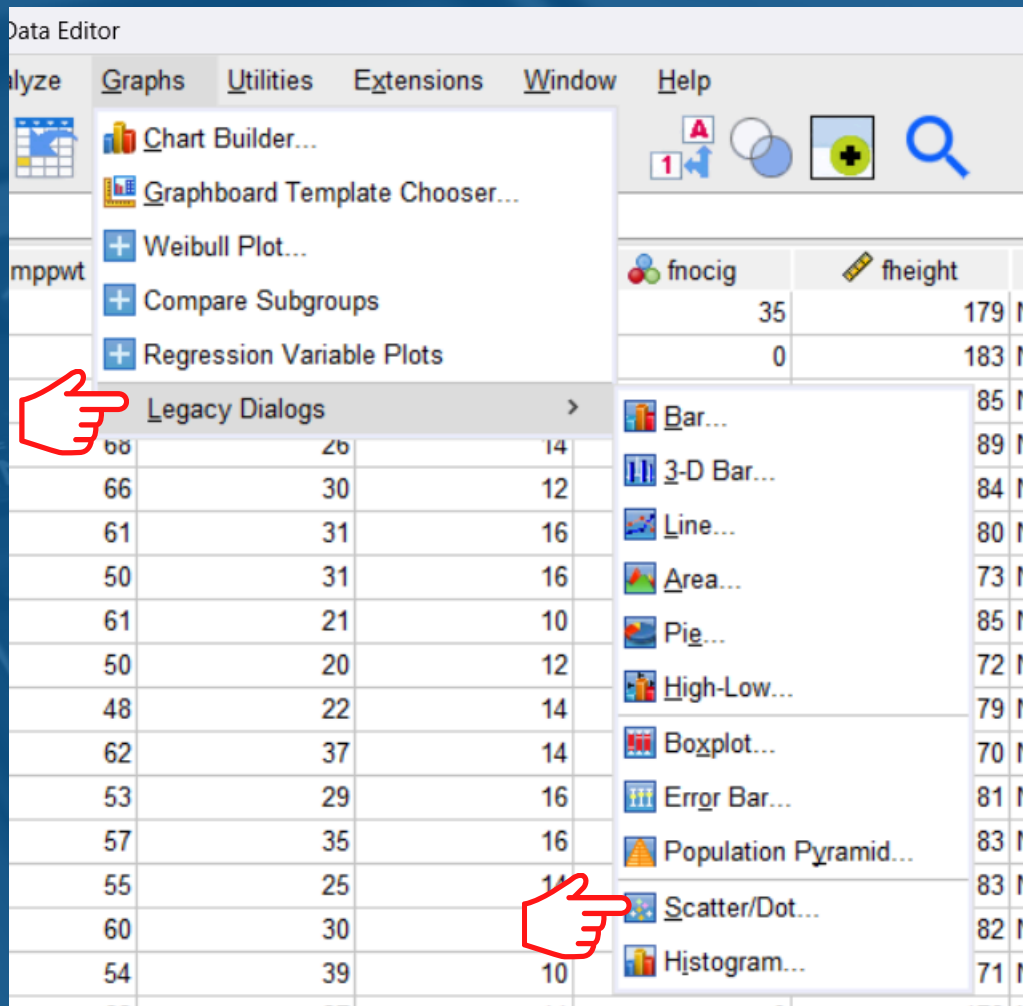
<sup>a</sup> Simple linear regression  
b\* = crude regression coefficient

## Step 3: Checking assumptions

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

# Checking assumption : Linearity

Go to: Graph > Legacy Dialogs > Scatter/Dot



Simple Scatterplot

Y Axis: Length of baby (cm) [Length]

X Axis: Maternal height (cm) [mheight]

Set Markers by:

Label Cases by:

Panel by

Rows:

Columns:

Template

☐ Use chart specifications from:

File...

OK Paste Reset Cancel Help

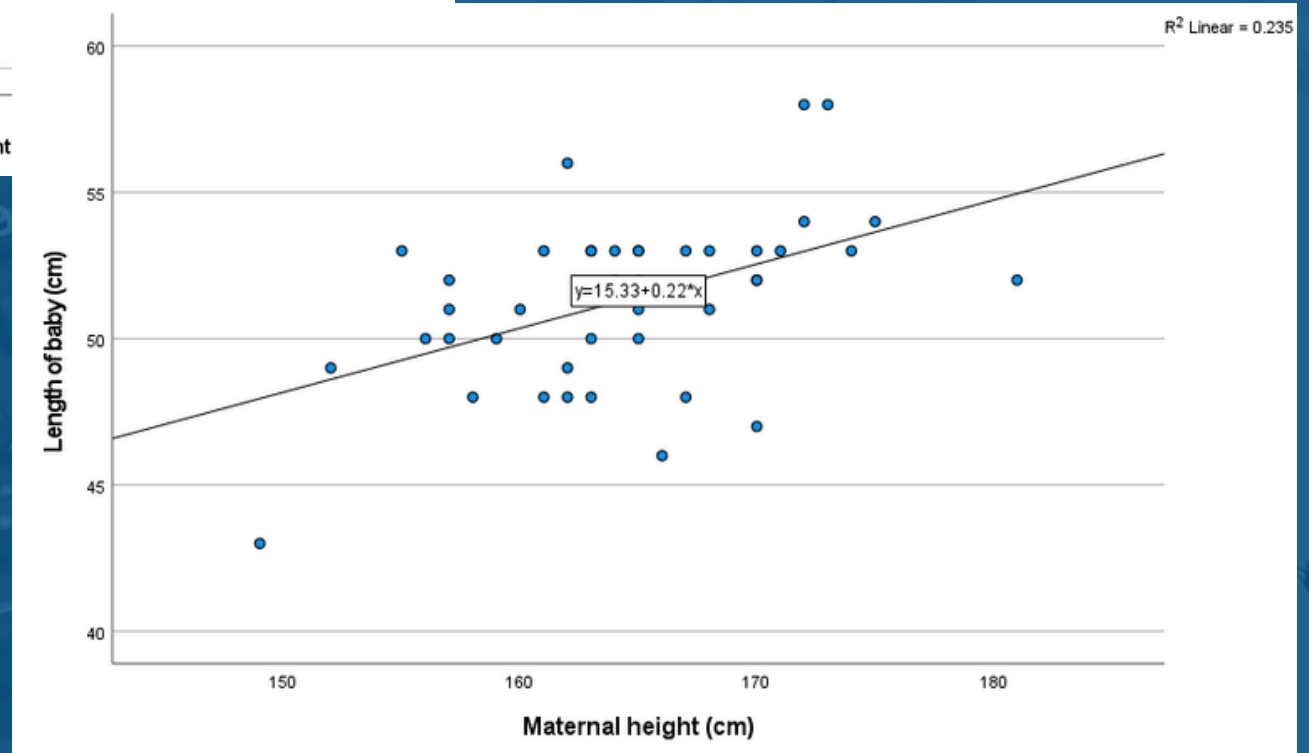
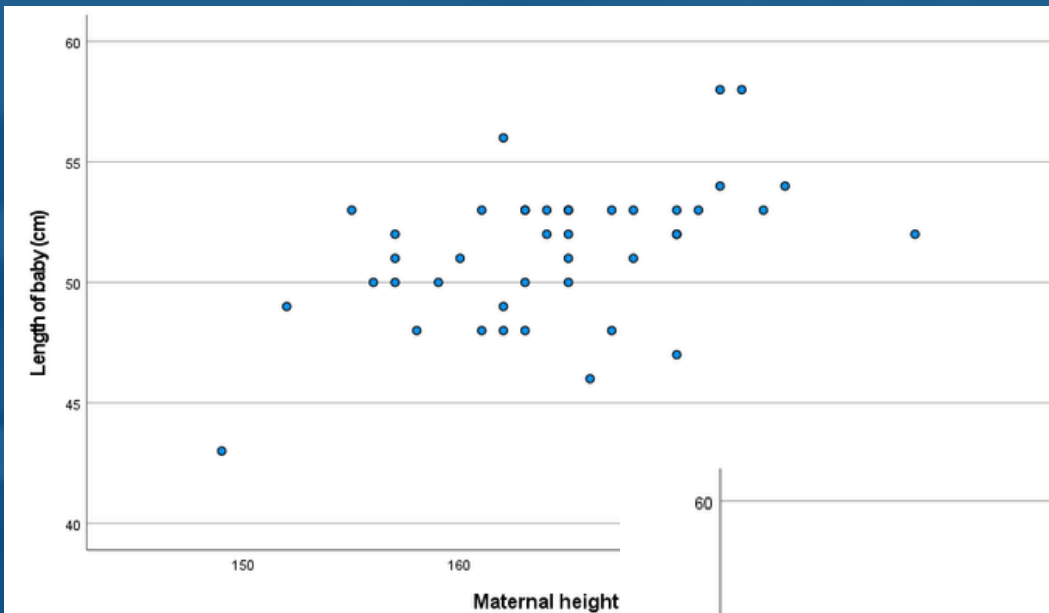
Titles... Options...

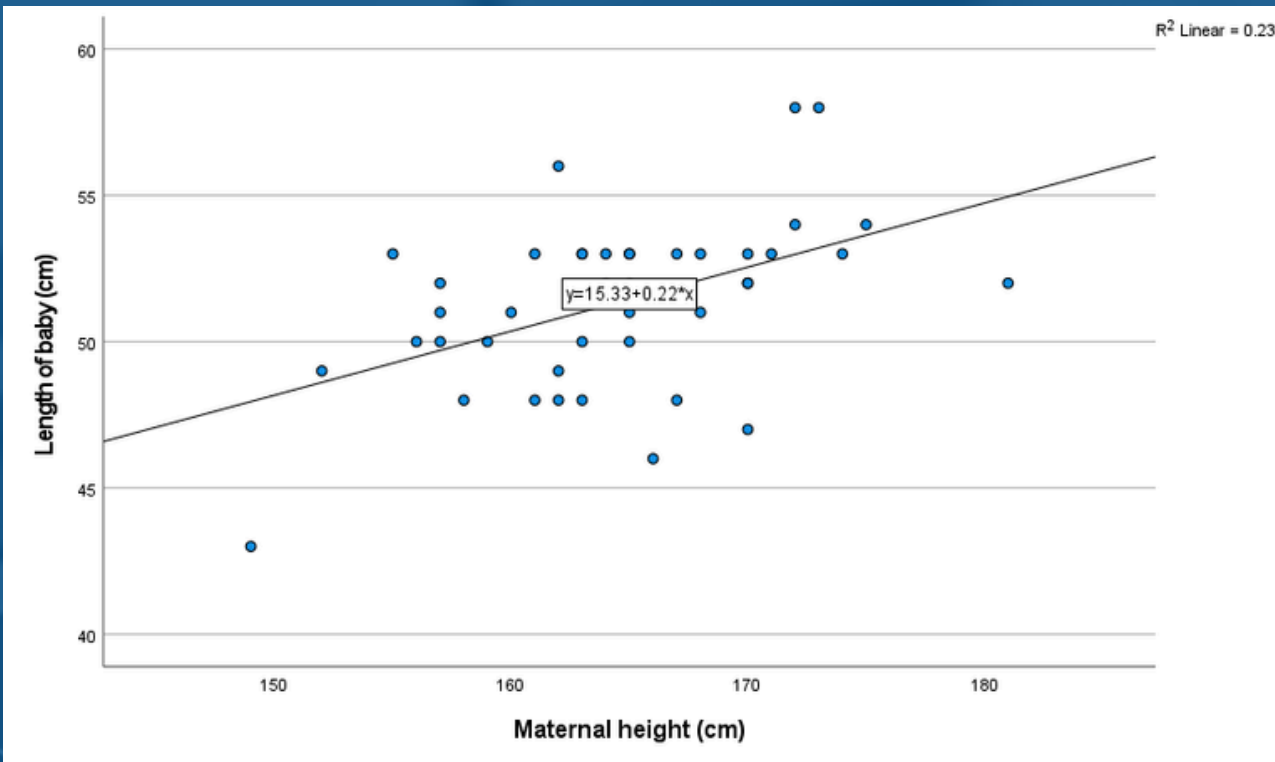
Baby ID [ID]  
Head circumference (cm) [Headcirc]  
Birthweight (kg) [Birthweight]  
Gestational age at birth (weeks) [Gest...]  
Maternal age [mage]  
smoker  
Mother's pre-pregnancy weight (kg) [...]  
Father's age [fage]  
Years father was in education [fedysr]  
Number of cigarettes smoked per day...  
Father's height (cm) [fheight]  
Low birthweight baby [lowbwt]  
Mother aged over 35 [mage35]  
Number of cigarettes smoked per day...

Nest variables (no empty rows)  
Nest variables (no empty columns)

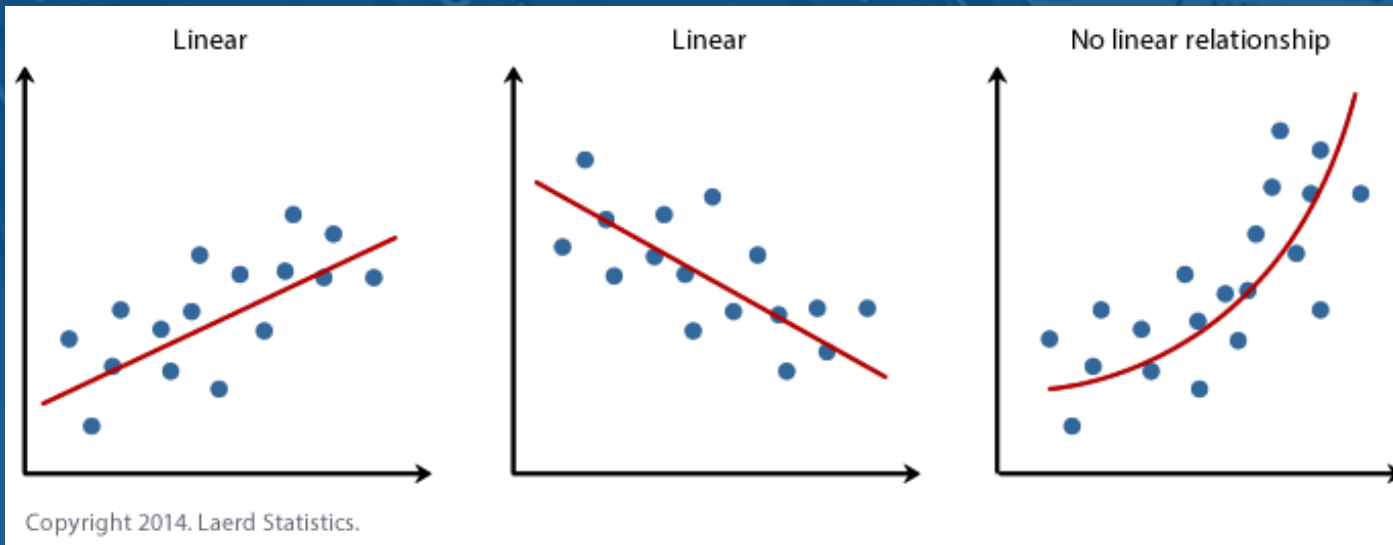


Double click the plot and click





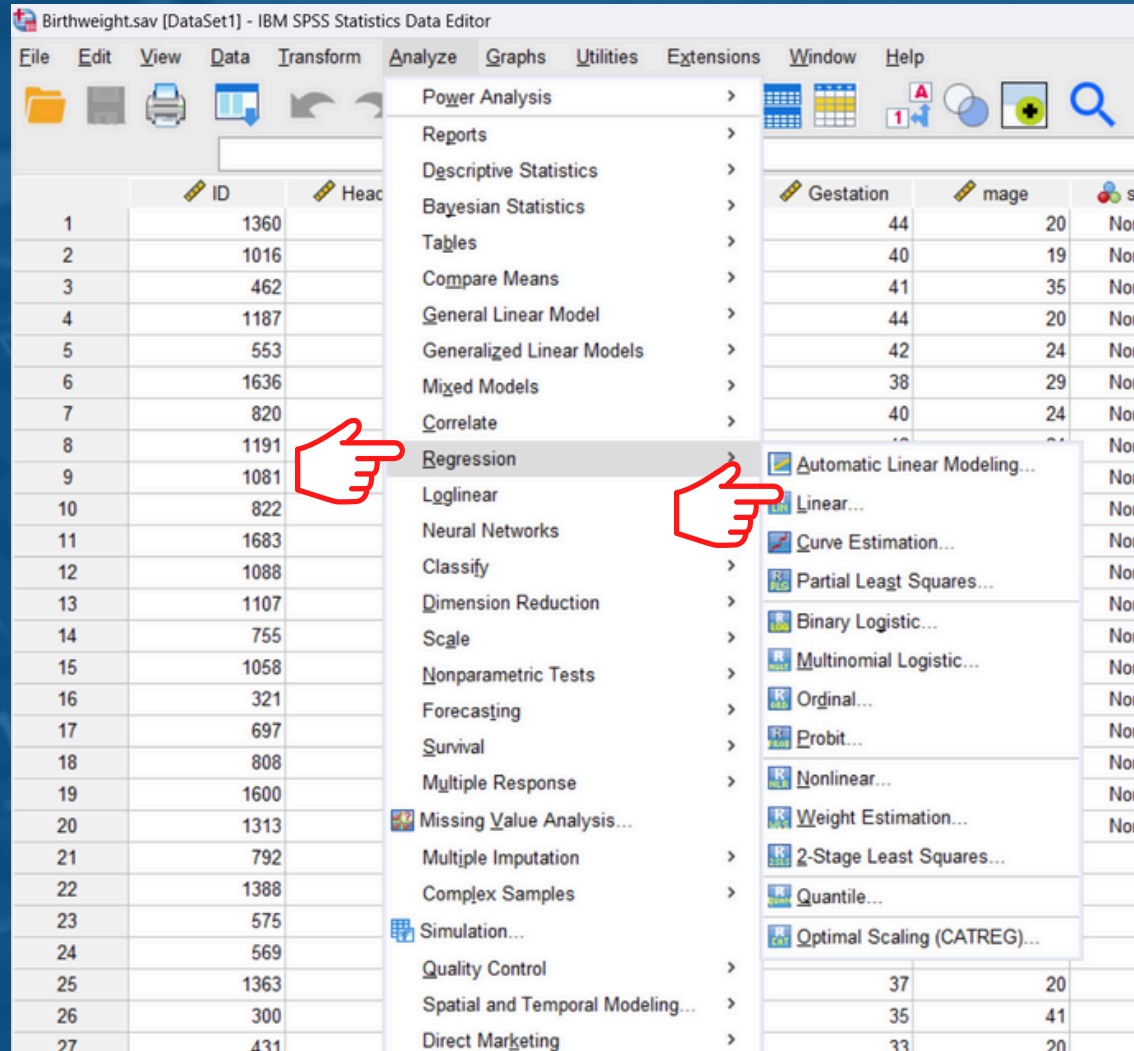
Linear relationship of two continuous variables. Linearity assumption is met.



Example of linear and non-linear relationship

# Checking assumption: Homoscedasticity (equal variances)

Go to: Analyze > Regression > Linear



Linear Regression

Dependent: Length of baby (cm) [Length]

Independent(s): Maternal height (cm) [mheight]

Method: Enter

Selection Variable: Rule...

Case Labels:

WLS Weight:

OK Paste Reset Cancel Help

Statistics... Plots... Save... Options... Style... Bootstrap...

Previous Next

Block 1 of 1

smoker

Low birthweight baby [lowbwt]

Mother aged over 35 [mage35]

Number of cigarettes smoked per day by mot...

Years father was in education [fedyr]

Father's height (cm) [fheight]

Father's age [fage]

Mother's pre-pregnancy weight (kg) [mppwt]

Maternal height (cm) [mheight]

Maternal age [mage]

Gestational age at birth (weeks) [Gestation]

Birthweight (kg) [Birthweight]

Head circumference (cm) [Headcirc]

Baby ID [ID]

Linear Regression: Save

Predicted Values

☒ Unstandardized

☐ Standardized

☐ Adjusted

☐ S.E. of mean predictions

Distances

☐ Mahalanobis

☐ Cook's

☐ Leverage values

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

Residuals

☒ Unstandardized

☐ Standardized

☐ Studentized

☐ Deleted

☐ Studentized deleted

Influence Statistics

☐ DfBetas

☐ Standardized DfBetas

☐ DfFits

☐ Standardized DfFits

☐ Covariance ratios



# Predicted value

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
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

$$y = a + bx$$

$$\text{Length of baby} = 15.33 + (0.22 * \text{mother's height})$$

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

$$\text{Length of baby} = 15.33 + (0.22 * 162) = 50.97$$

# Residuals

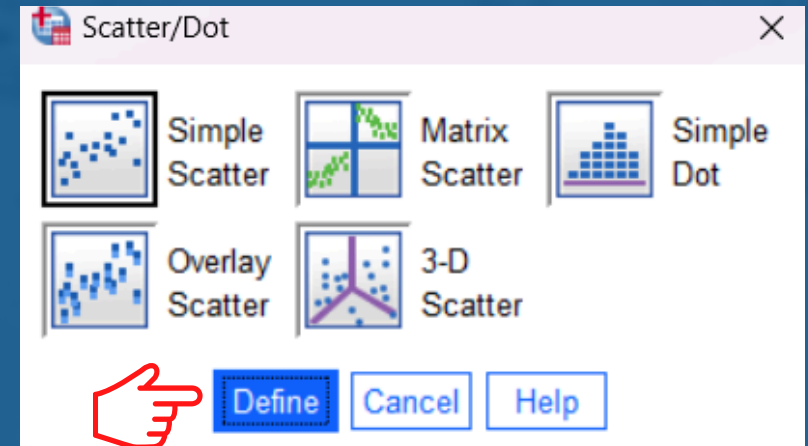
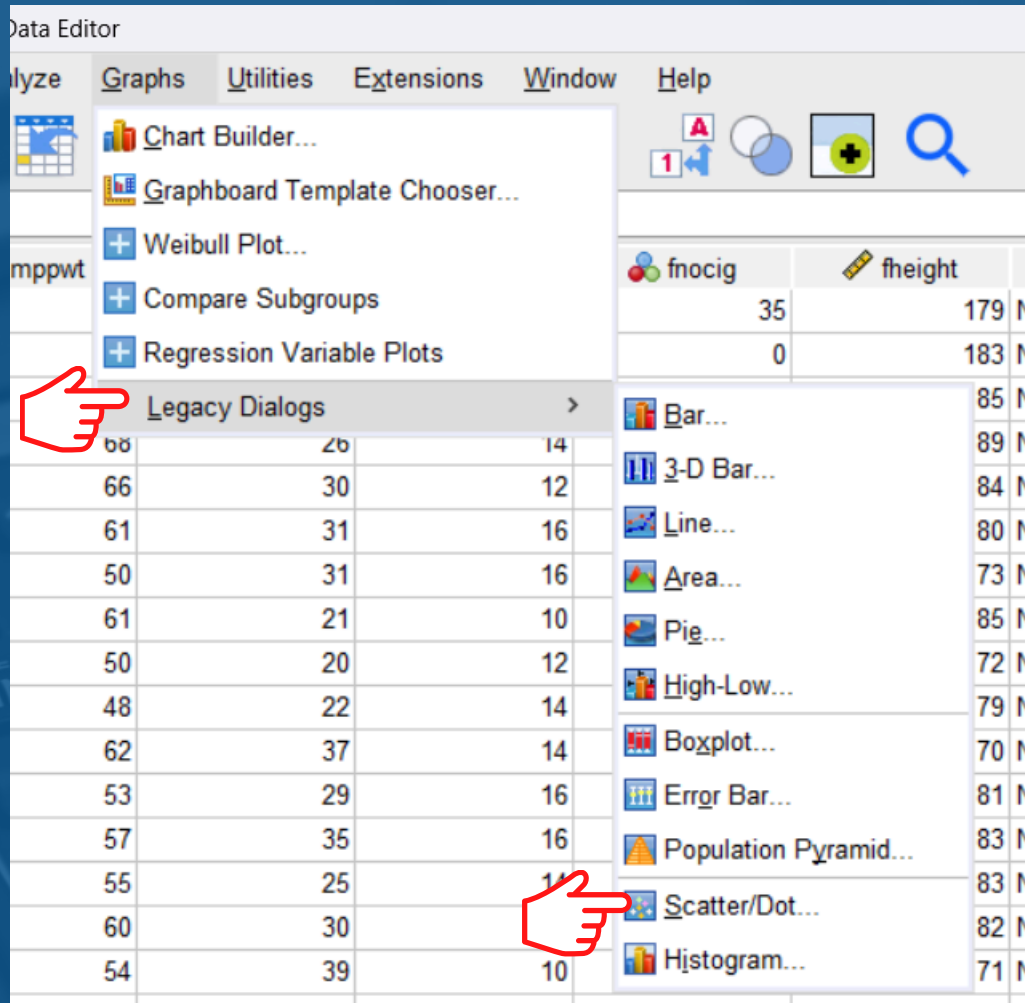
	ID	Length	mheight	PRE_1	RES_1
1	1360	56	162	50.79650	5.20350

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

Y Axis:  
Unstandardized Residual [RES\_1]

X Axis:  
Unstandardized Predicted Value [PRE\_1]

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:

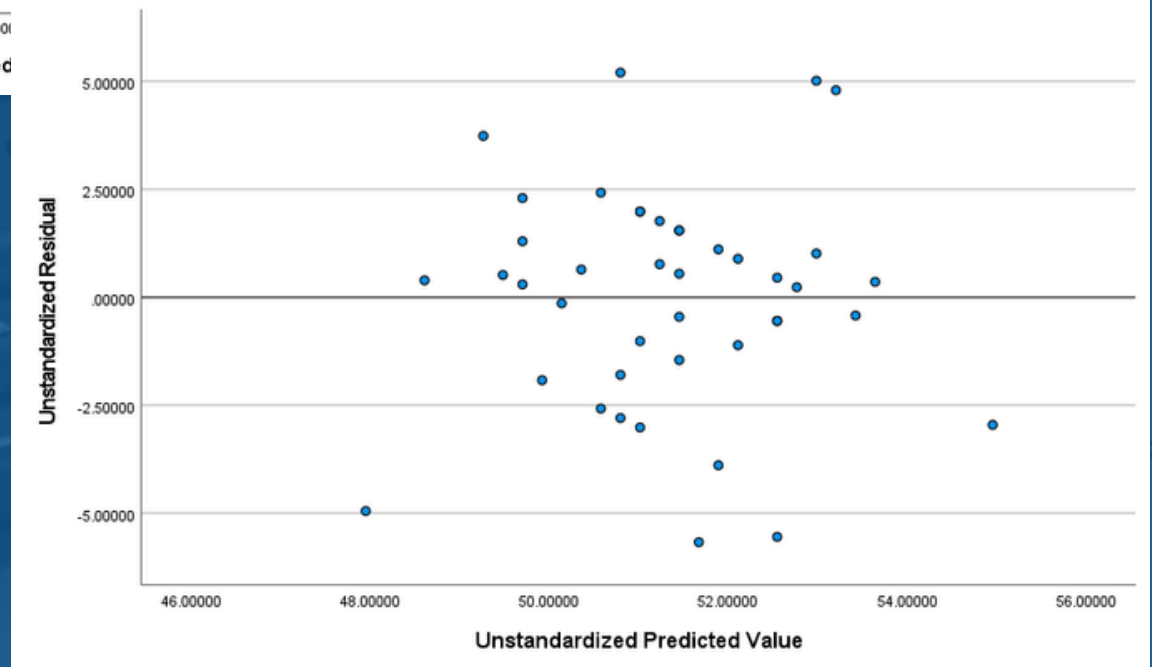
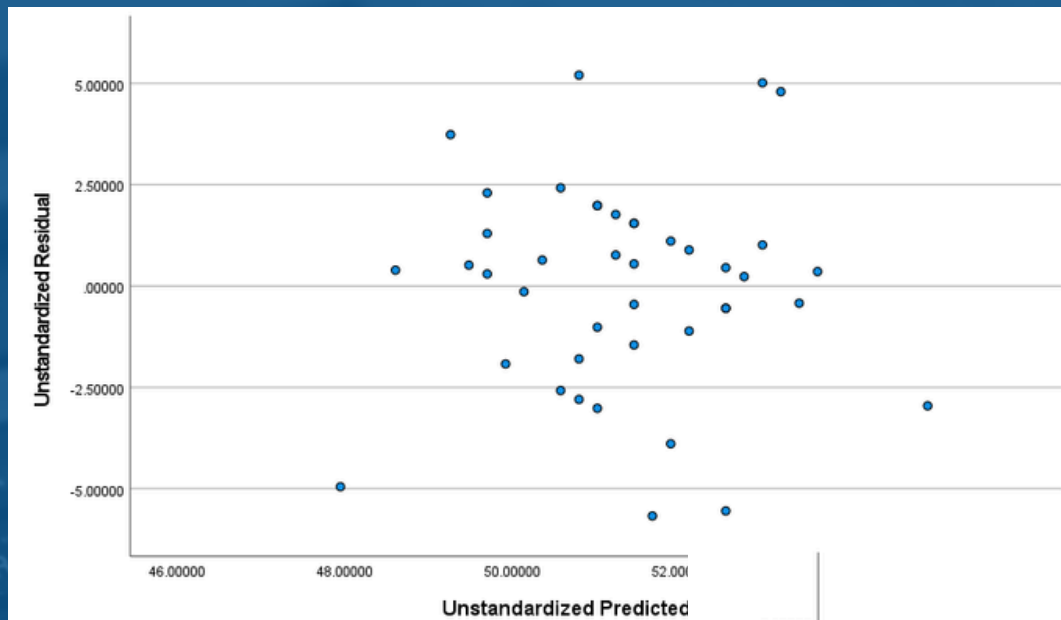
File...

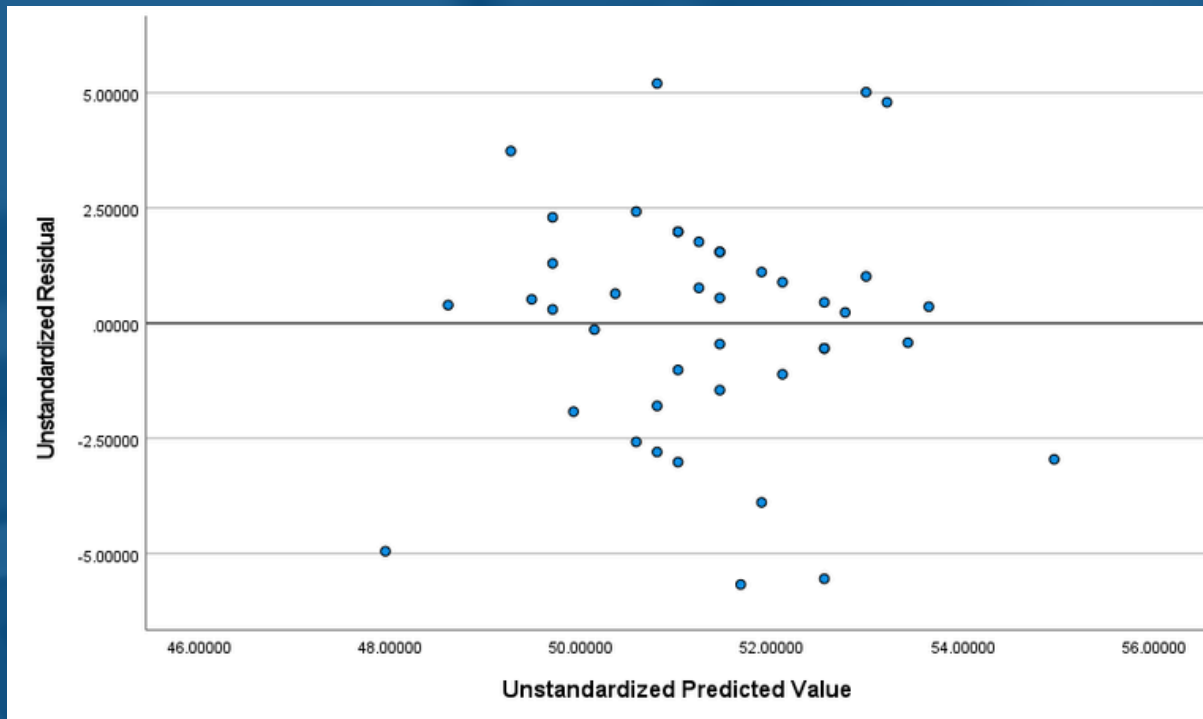
OK Paste Reset Cancel Help

XP - YR

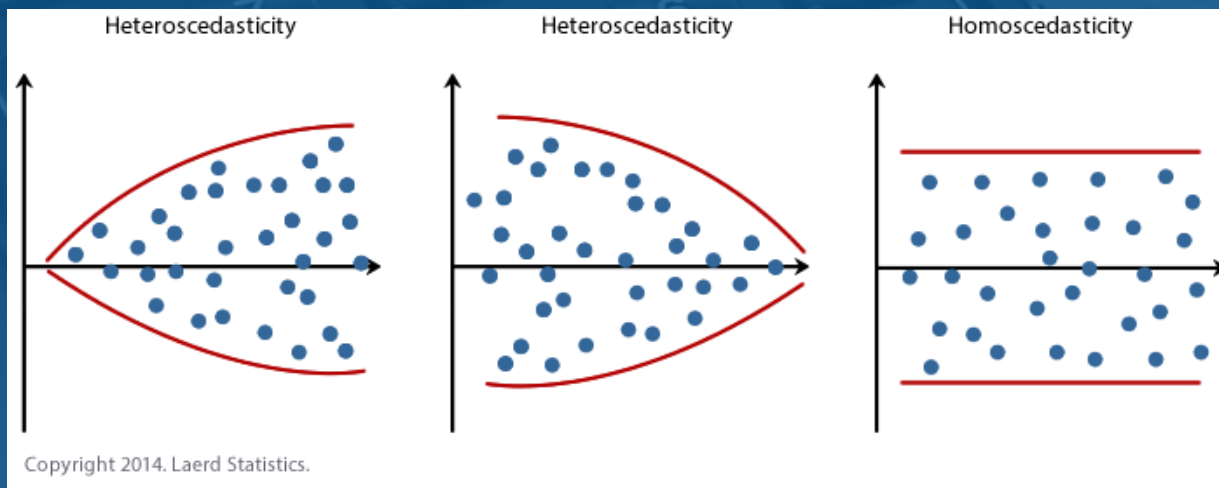


Double click the plot and click





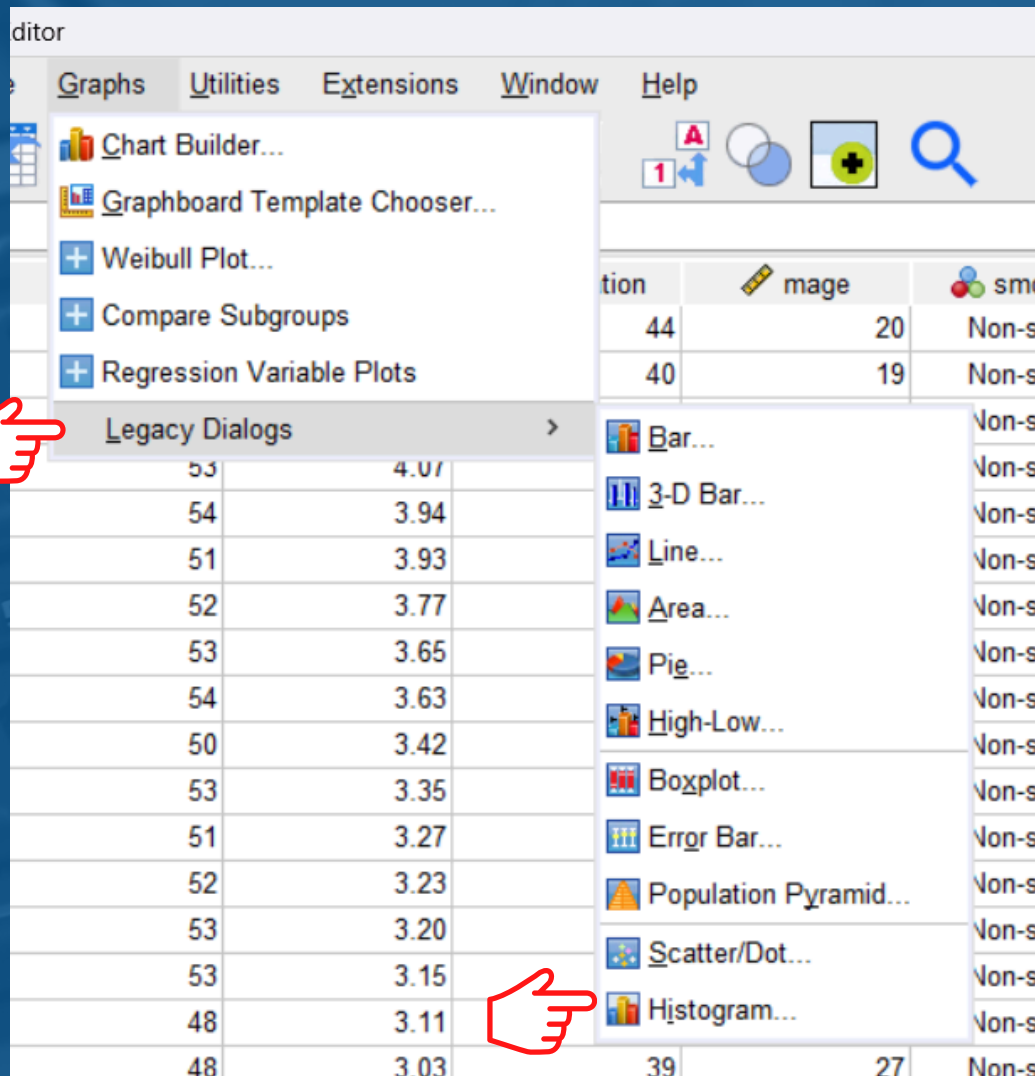
There is no pattern in the scatter.  
Homoscedasticity assumption is met.

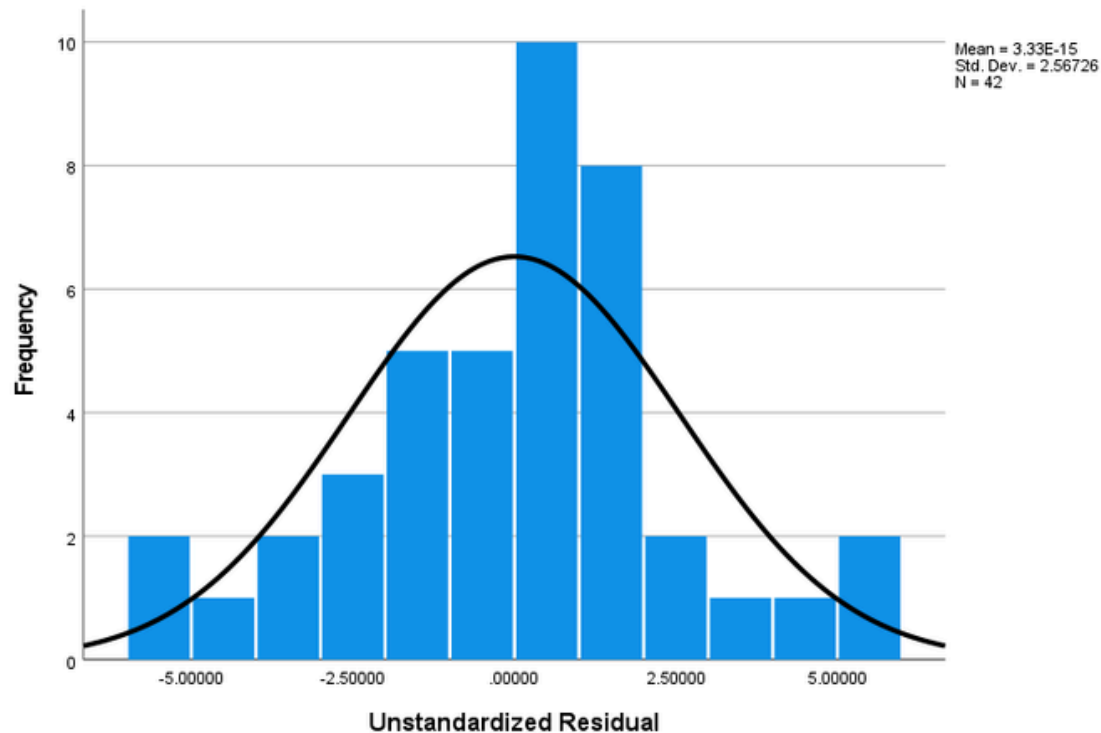
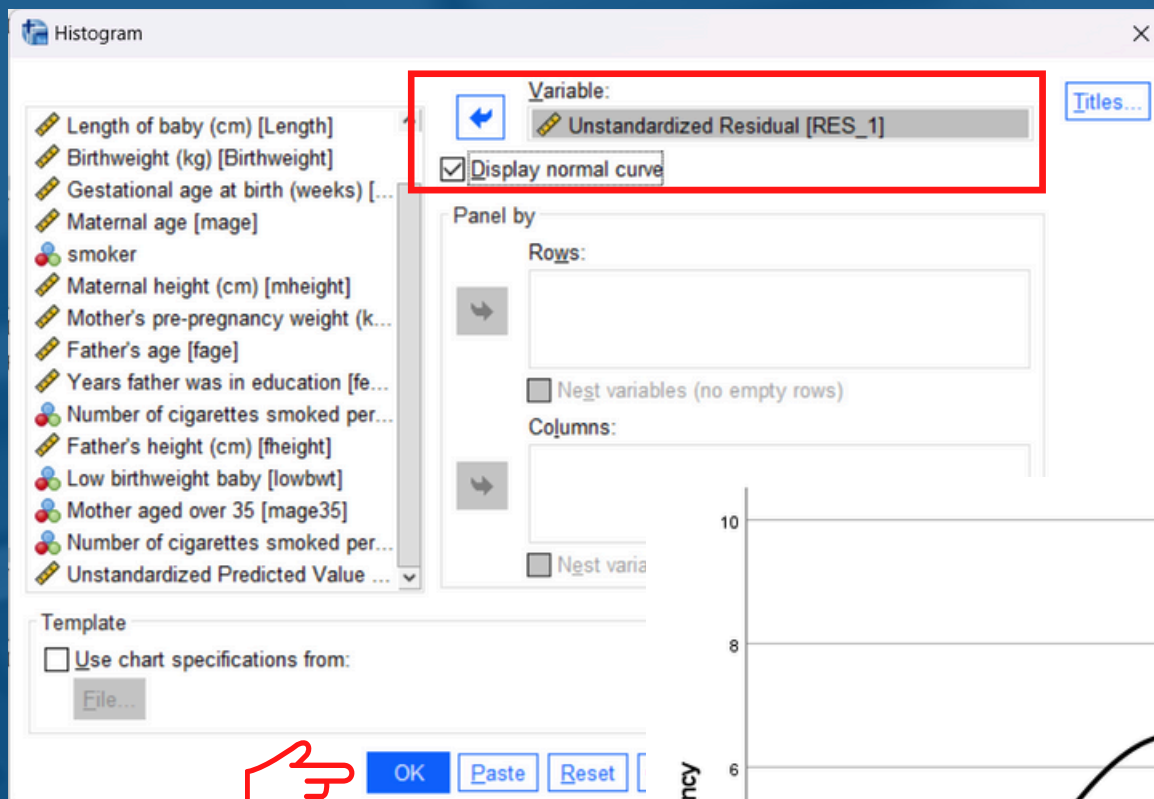


Example of heteroscedasticity and homoscedasticity

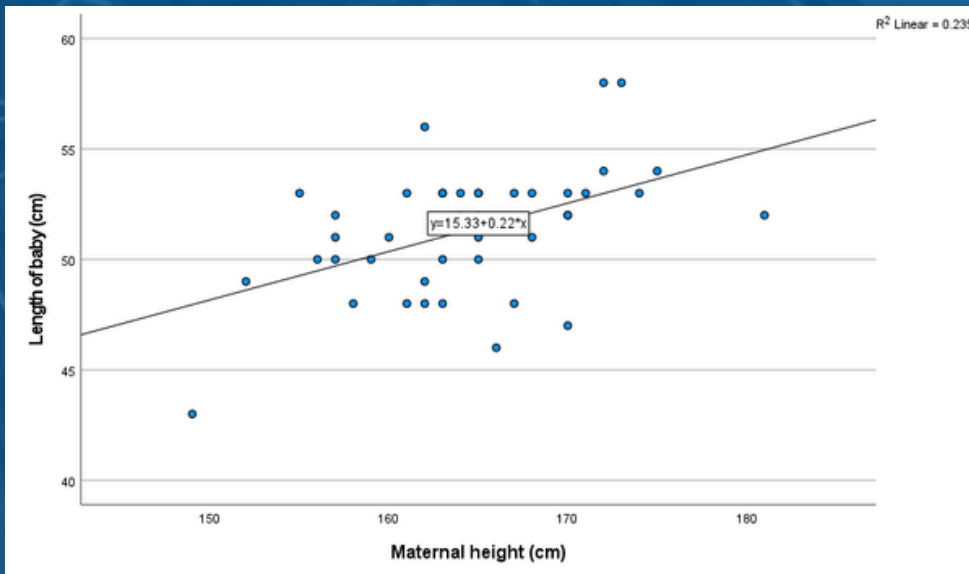
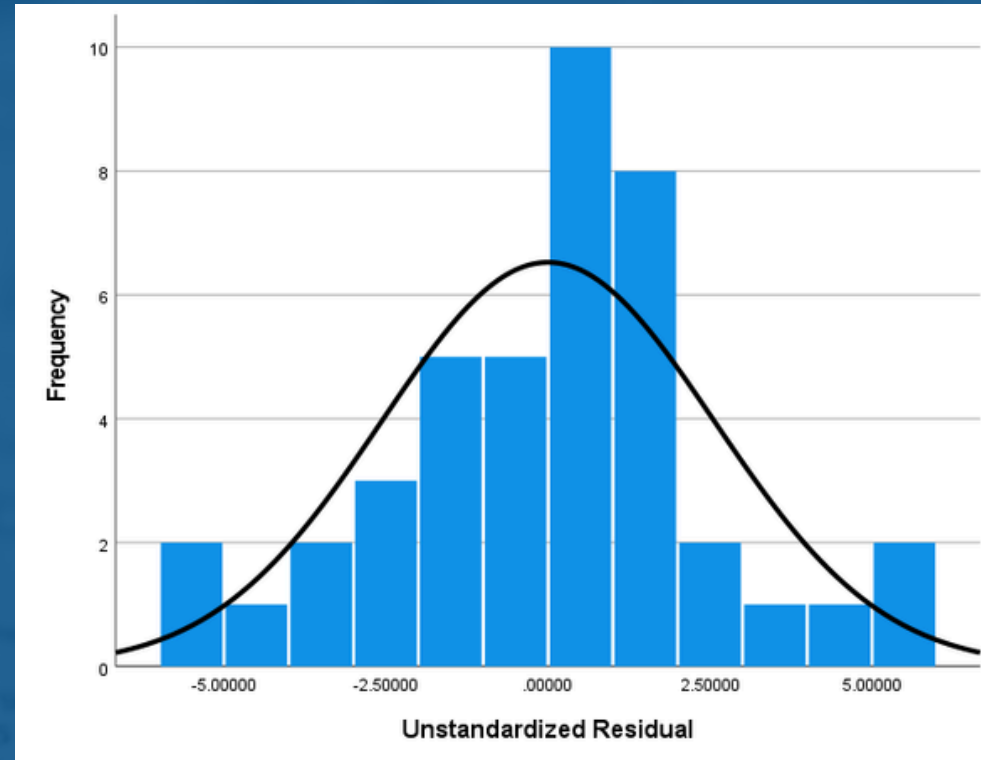
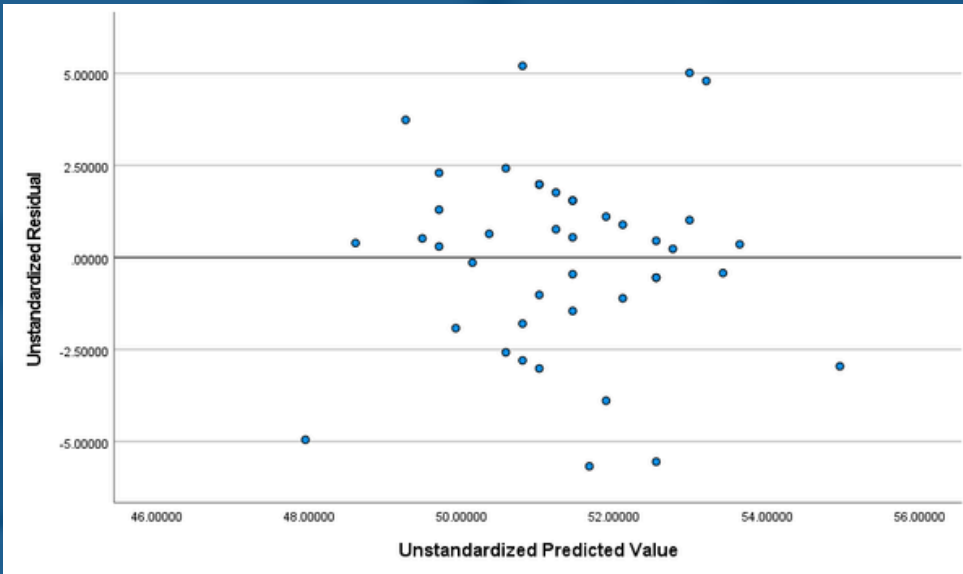
# Checking assumption: Normality distribution of residuals

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 <sup>a</sup>								
		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
Model		B	Std. Error	Beta	t	Sig.	Lower Bound	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:

$$y = a + bx$$

$$\text{length of baby} = 15.33 + (0.22 * \text{mother's height})$$



# 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](mailto:nuruljannah@mahsa.edu.my)





**THANK  
YOU**