Mohamed Khider University of Biskra
Faculty of SNVSTU

Practical Protocol 02 – Linear Regression
Biostatistics, M1 Biology

Dr. Ben Gherbal Hanane
hanane.bengherbal@univ-biskra.dz

Suppose we have a sample of n individuals. For each individual i ($i = 1, \ldots, n$), two variables are measured: y_i, corresponding to the observed value of the quantitative variable Y, and x_i, representing the value of the explanatory variable X.

The objective is to analyze the relationship between these two variables, particularly the influence of X on Y, using SPSS software. To make the steps understandable, we illustrate them through the following example.

Exercise

A study was conducted to analyze the effect of phosphate concentration on chlorophyll production of a certain type of algae. The obtained results are presented in the following table:

Concentration	0	100	200	400	600
Chlorophyll	305	378	458	540	565

To model these data, we propose the following linear model:

$$Y = \alpha X + \beta$$

- Plot the scatter diagram of the points (X_i, Y_i). What can you conclude about the proposed model?
- Estimate the parameters α and β and determine the regression line.
- Compute the linear correlation coefficient. What can you conclude?

Solution Using SPSS

Step 1: Define Variables

Launch SPSS and define the two variables in **Variable View**: X (concentration) and Y (chlorophyll).

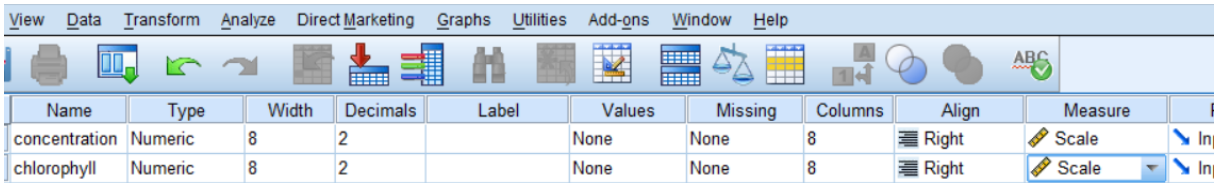

Figure 1: Variable View

Enter the data in **Data View**:

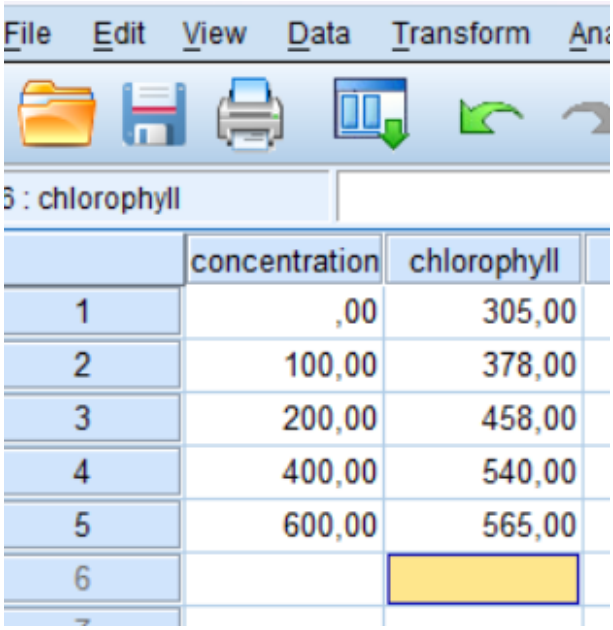

Figure 2: Data View

Step 2: Scatter Plot

First, we draw the scatter plot representing the relationship between the explanatory variable X and the response variable Y:

Graphs → **Scatter/Dot** → **Simple Scatter** → **Define**.

Place:

- X (concentration) on the X-axis
- Y (chlorophyll) on the Y-axis

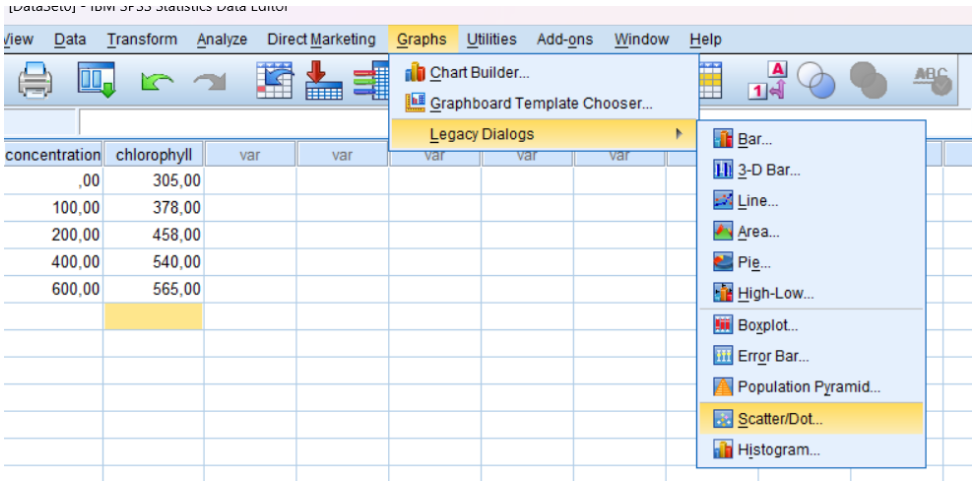

Figure 3: Scatter Plot

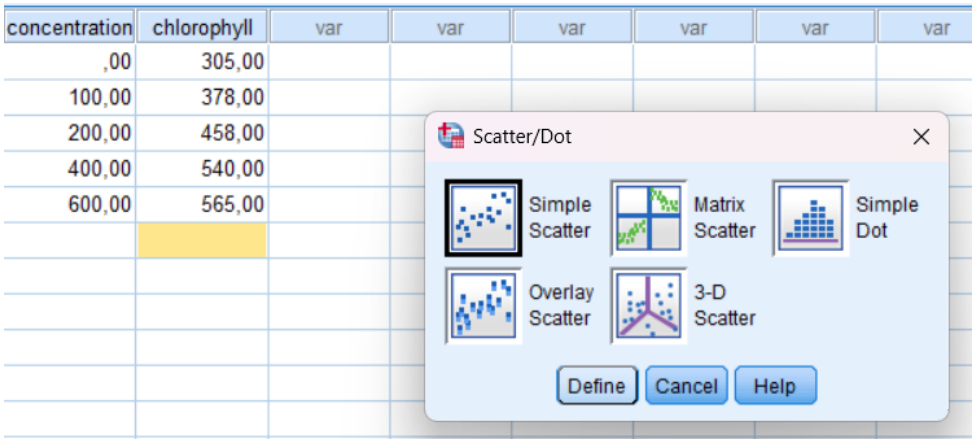

Figure 4: Scatter Plot

Place the explanatory variable X (concentration) in the "X-Axis" field and the response variable Y (chlorophyll) in the "Y-Axis" field.

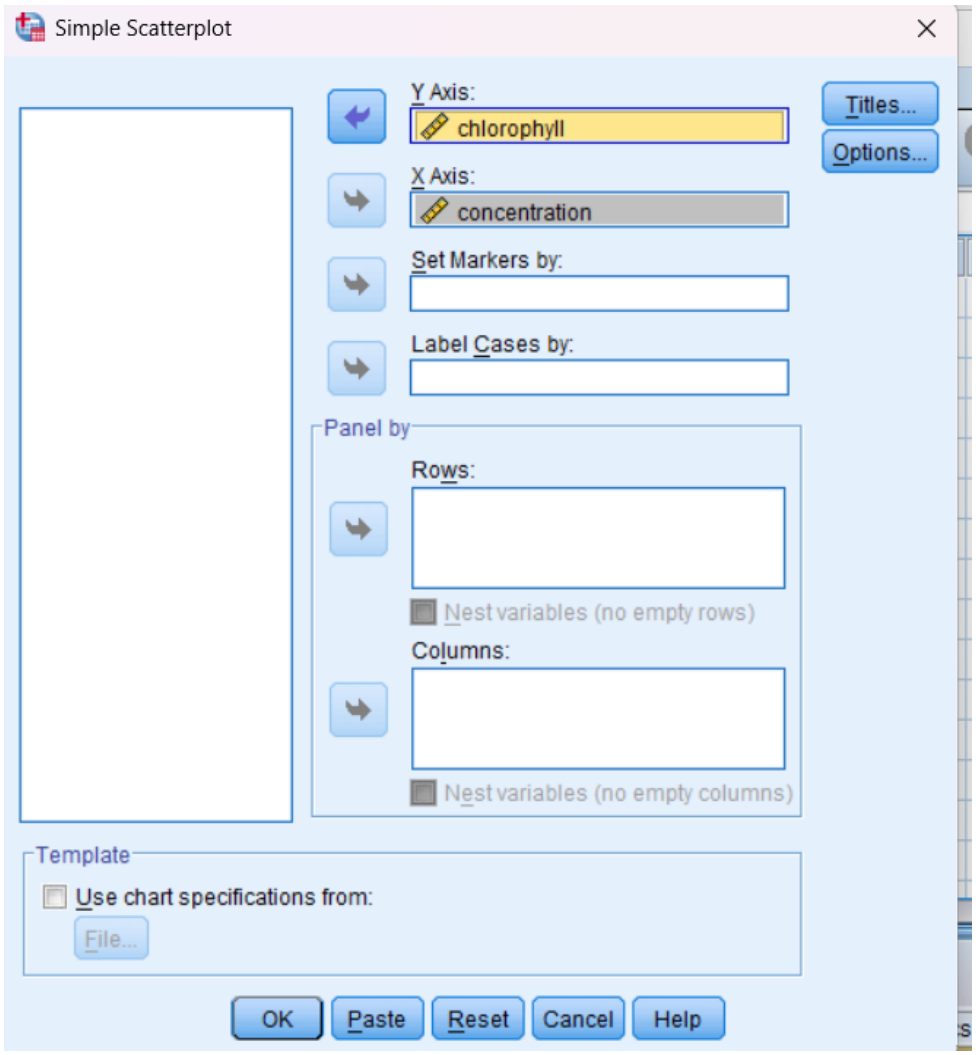

Figure 5: Scatter Plot

After clicking **OK**, the following window appears:

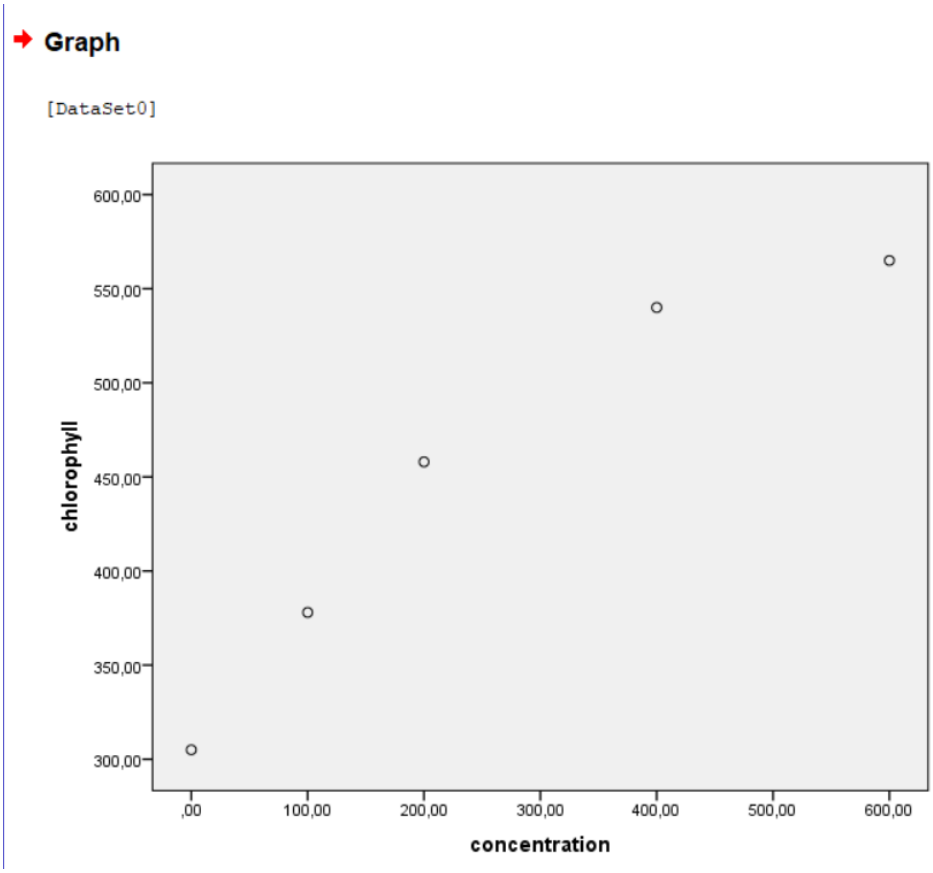

Figure 6: Scatter Plot

From observation of the graph, the points are almost aligned in a straight-line pattern and follow an increasing trend. This suggests a positive linear relationship between phosphate concentration and chlorophyll production.

We double-click on the graph; a side window appears, from which we choose:
Elements → **Add Fit Line at Total** → **Linear**.

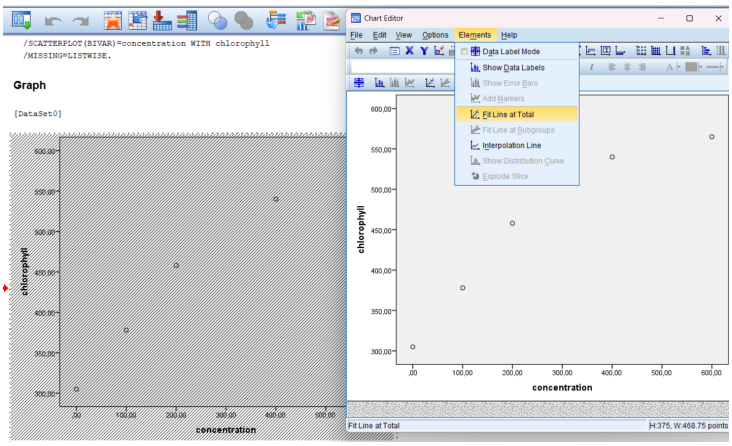

Figure 7: Linear Fit

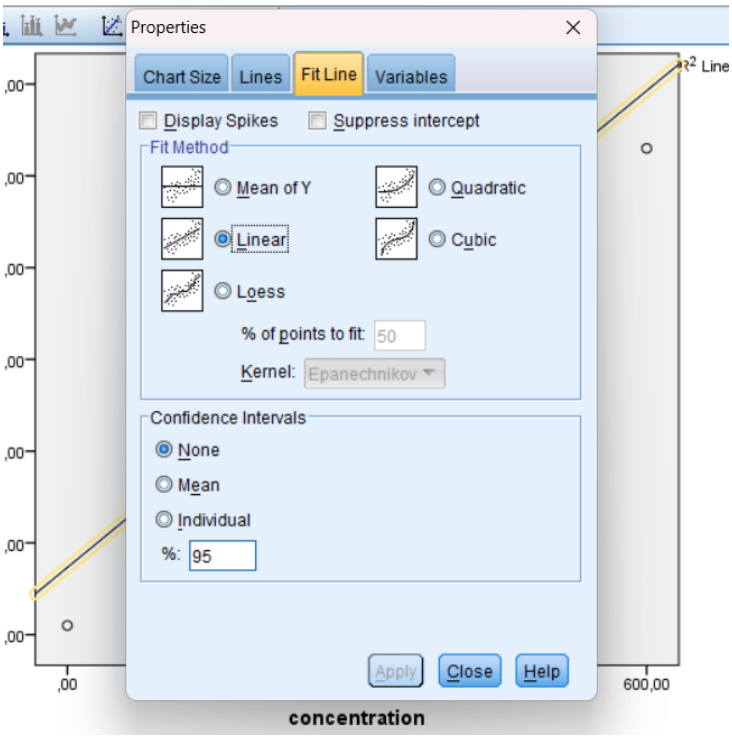

Figure 8: Linear Fit

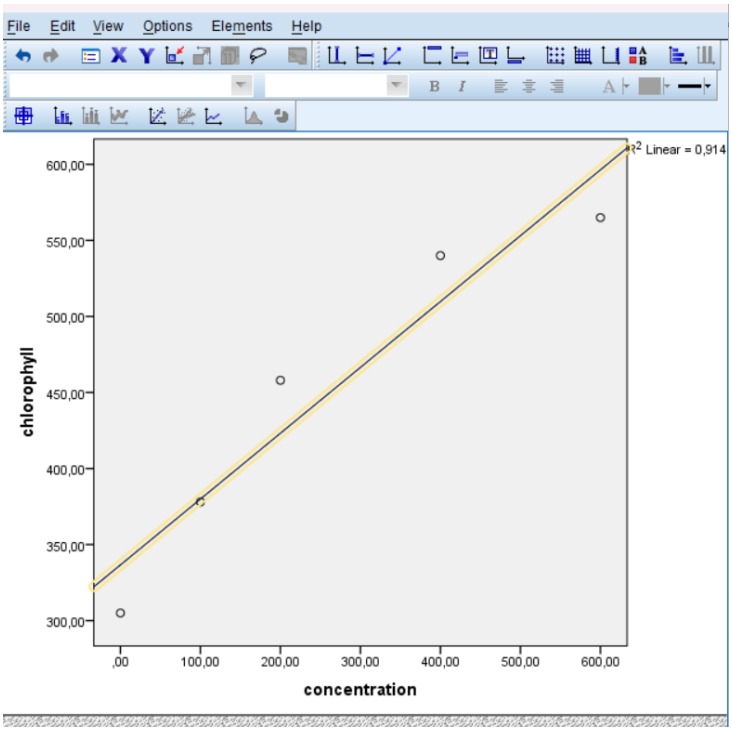

Figure 9: Linear Fit

The arrangement of the points around a straight line allows us to conclude that the linear model

$$Y = \alpha X + \beta$$

is appropriate for the data. The scatter plot visually confirms that the use of a simple linear regression model is relevant, which is consistent with the high coefficient of determination ($R^2 \approx 0.914$) displayed at the top of the graph.

Step 3: Linear Regression

To estimate the parameters of the regression line and evaluate the goodness of fit of the model, a linear regression test is performed according to the following steps:
Analyze → Regression → Linear

- Independent variable: X
- Dependent variable: Y

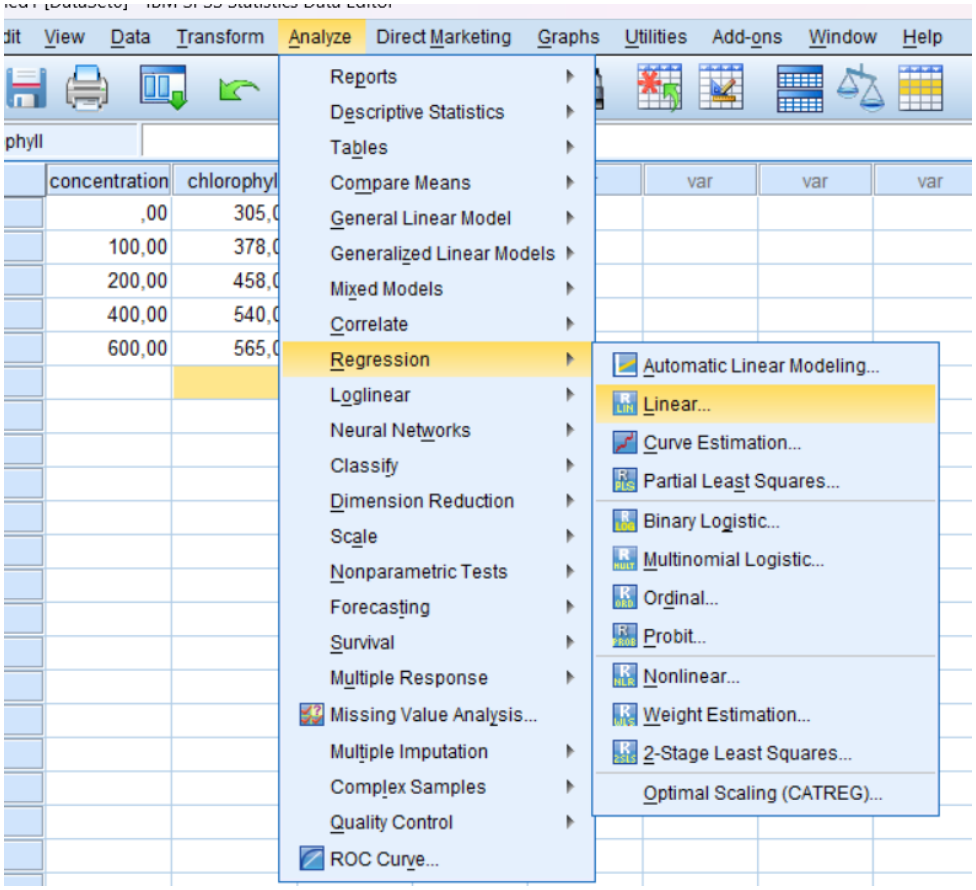

Figure 10: Regression Setup

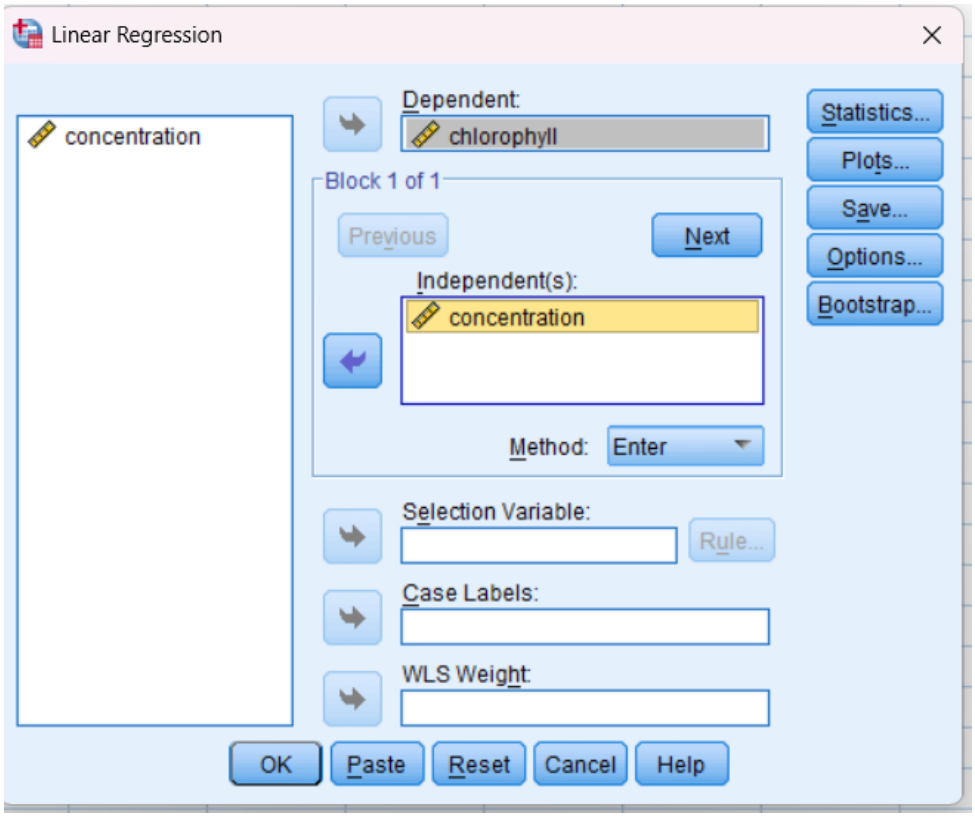

Figure 11: Regression Setup

Under **Statistics**, select: Estimates, Model fit, Descriptives.

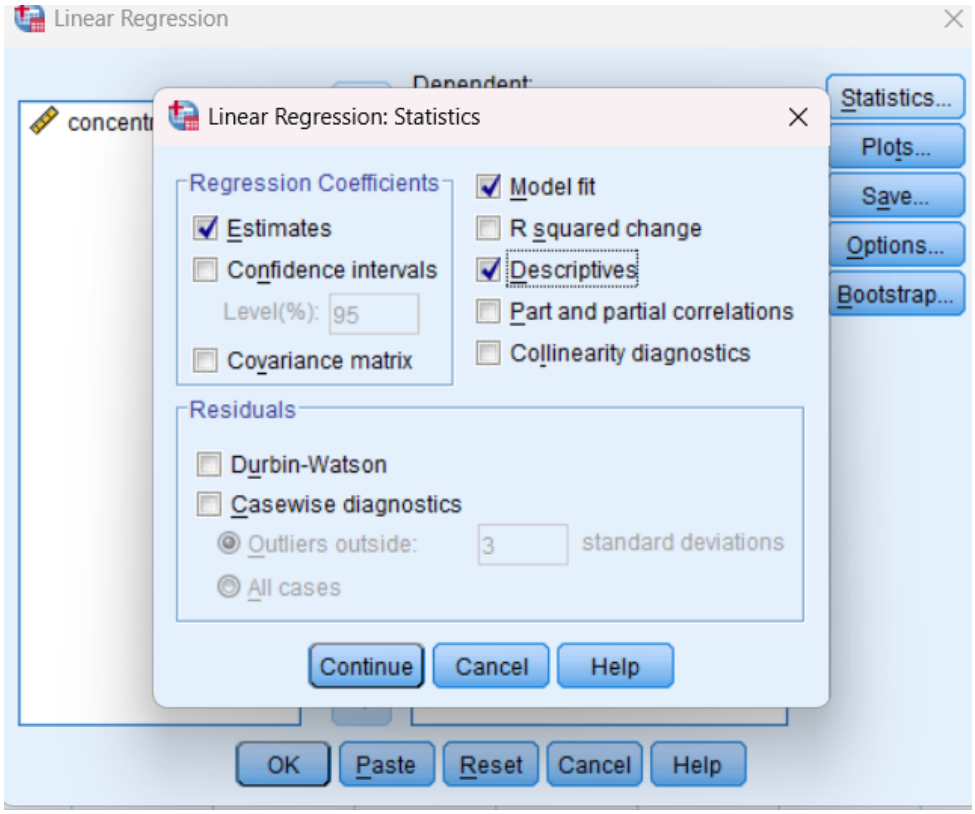

Figure 12: Regression Setup

Under **Plots**: ZPRED on X-axis, ZRESID on Y-axis.

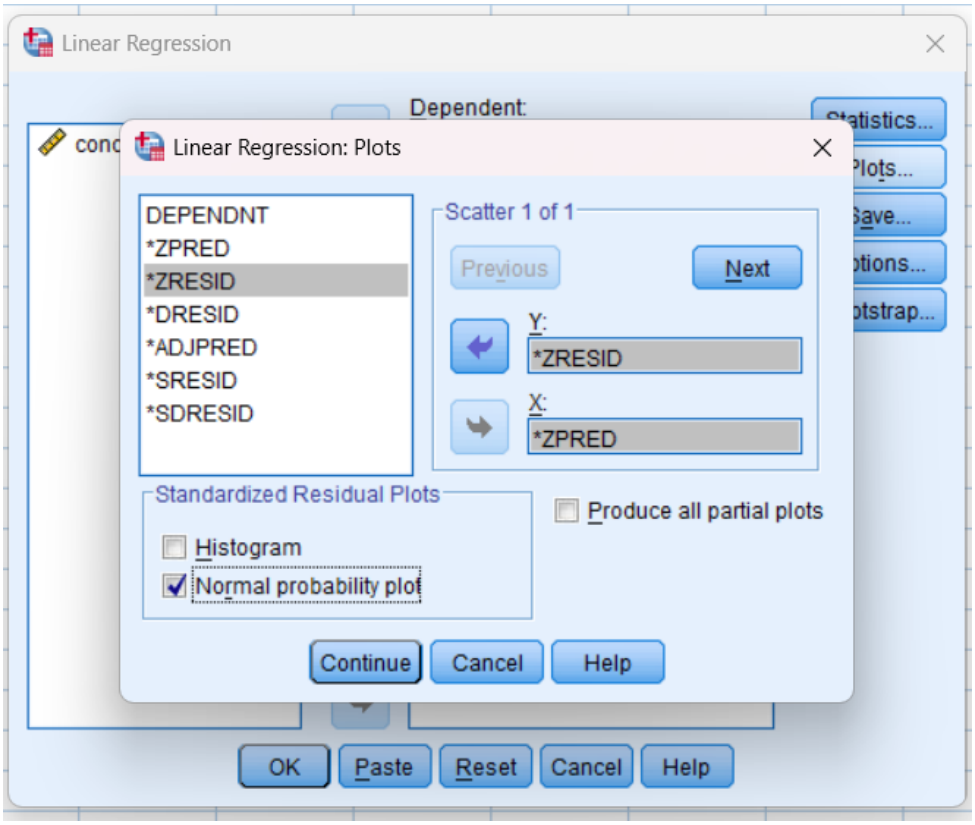

Figure 13: Regression Setup

Interpretation of Results

The first two tables provide a general overview of the descriptive statistics (means and standard deviations) for the two variables. The second table presents the correlation between the two variables, which is equal to 0.956. This means that there is a strong linear relationship between the two variables: approximately 95.6% of the variation in Y is explained by the variation in X.

Thus, the p-value (0.006), which is very close to zero, indicates that the model is statistically significant.

Regression

[DataSet0]

Descriptive Statistics

	Mean	Std. Deviation	N
chlorophyll	449,2000	109,08116	5
concentration	260,0000	240,83189	5

Correlations

		chlorophyll	concentration
Pearson Correlation	chlorophyll	1,000	,956
	concentration	,956	1,000
Sig. (1-tailed)	chlorophyll	.	,006
	concentration	,006	.
N	chlorophyll	5	5
	concentration	5	5

Figure 14: Regression Setup

The three tables displayed in the following image contain the following information: The first table presents the correlation coefficient r, which is equal to 0.956.

In the second table, we observe the value of the test statistic ($t \equiv D$), which is the same as the one we calculated manually during the lecture.

Since the value of sig = 0.011 is less than 0.05, this indicates that there is a strong linear relationship between the two variables. Based on the following decision rule:

$$\begin{cases} \text{sig} < \alpha & \text{Reject } H_0\text{, i.e., there is a linear correlation.} \\ \text{sig} \geq \alpha & \text{Accept } H_0\text{, i.e., there is no linear correlation.} \end{cases}$$

with

$$\begin{cases} H_0 : R = 0 & \text{there is no linear correlation,} \\ H_1 : R \neq 0 & \text{there is a linear correlation.} \end{cases}$$

The third table provides the estimated values of the coefficients β (intercept) and α (slope) of the linear model, which are respectively 336.638 and 0.433.

Thus, the equation of the linear relationship between the two variables is written as:

$$Y = 0.433X + 336.638.$$

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,956[a]	,914	,885	37,01895

a. Predictors: (Constant), concentration

b. Dependent Variable: chlorophyll

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	43483,593	1	43483,593	31,731	,011[b]
	Residual	4111,207	3	1370,402		
	Total	47594,800	4			

a. Dependent Variable: chlorophyll

b. Predictors: (Constant), concentration

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	336,638	25,950		12,973	,001
	concentration	,433	,077	,956	5,633	,011

a. Dependent Variable: chlorophyll

Figure 15: Regression Setup