Correlation coefficient

Elias Zintzaras, M.Sc., Ph.D.

Professor in Biomathematics-Biometry
Department of Biomathematics
School of Medicine
University of Thessaly

Institute for Clinical Research and Health Policy Studies
Tufts University School of Medicine
Boston, MA, USA

Theodoros Mprotsis, MSc, PhD
Teacher & Research Fellow
(http://biomath.med.uth.gr)
University of Thessaly
Email: tmprotsis@uth.gr

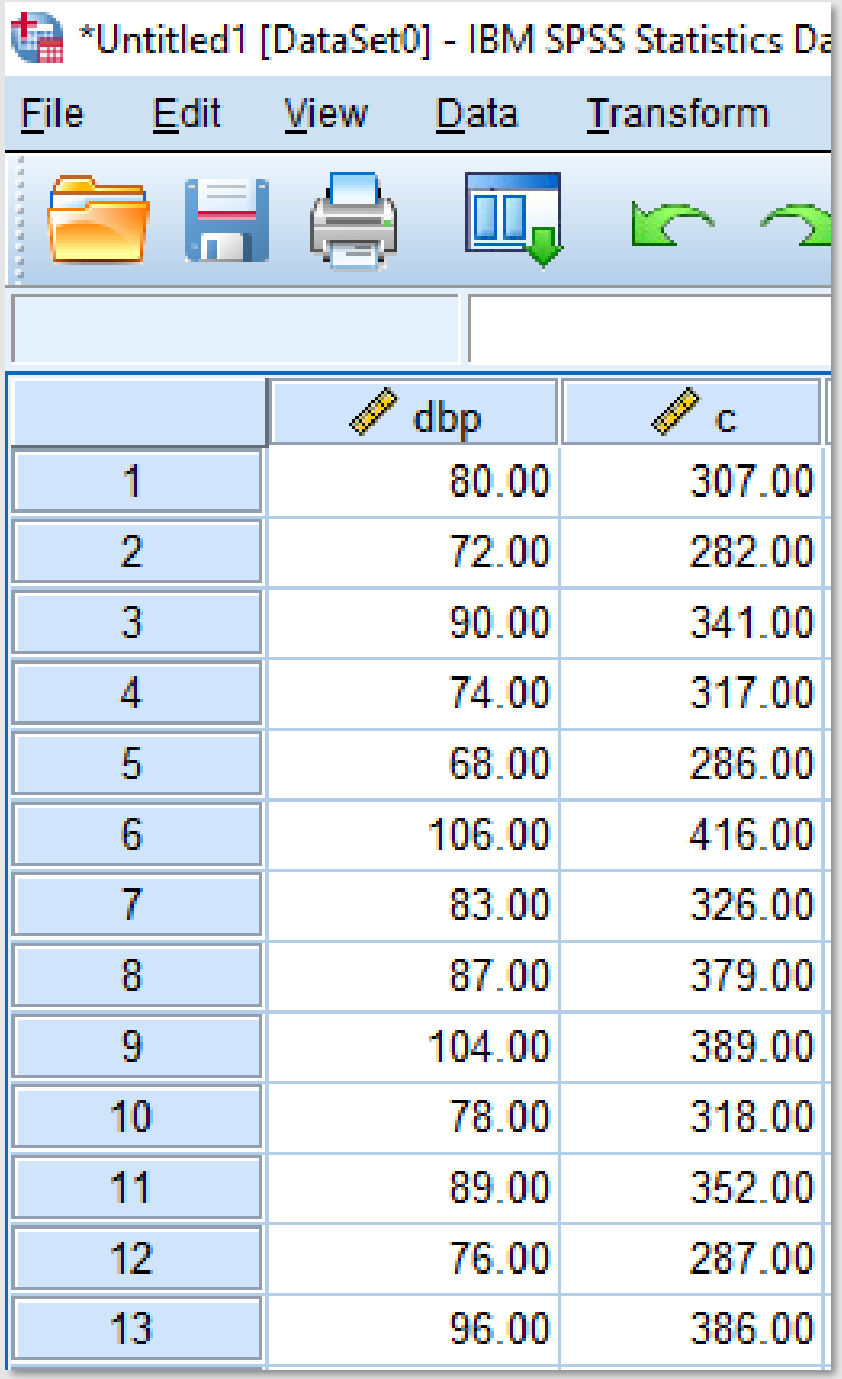
Blood pressure (DBP) and cholesterol levels

Blood pressure (DBP) and cholesterol levels

Suppose 13 people had their blood pressure (DBP) and cholesterol levels (C) measured

We want to test whether there is a **relationship (association)** between DBP and C

Enter the data in the **Data View** and define the variables in the **Variable View**



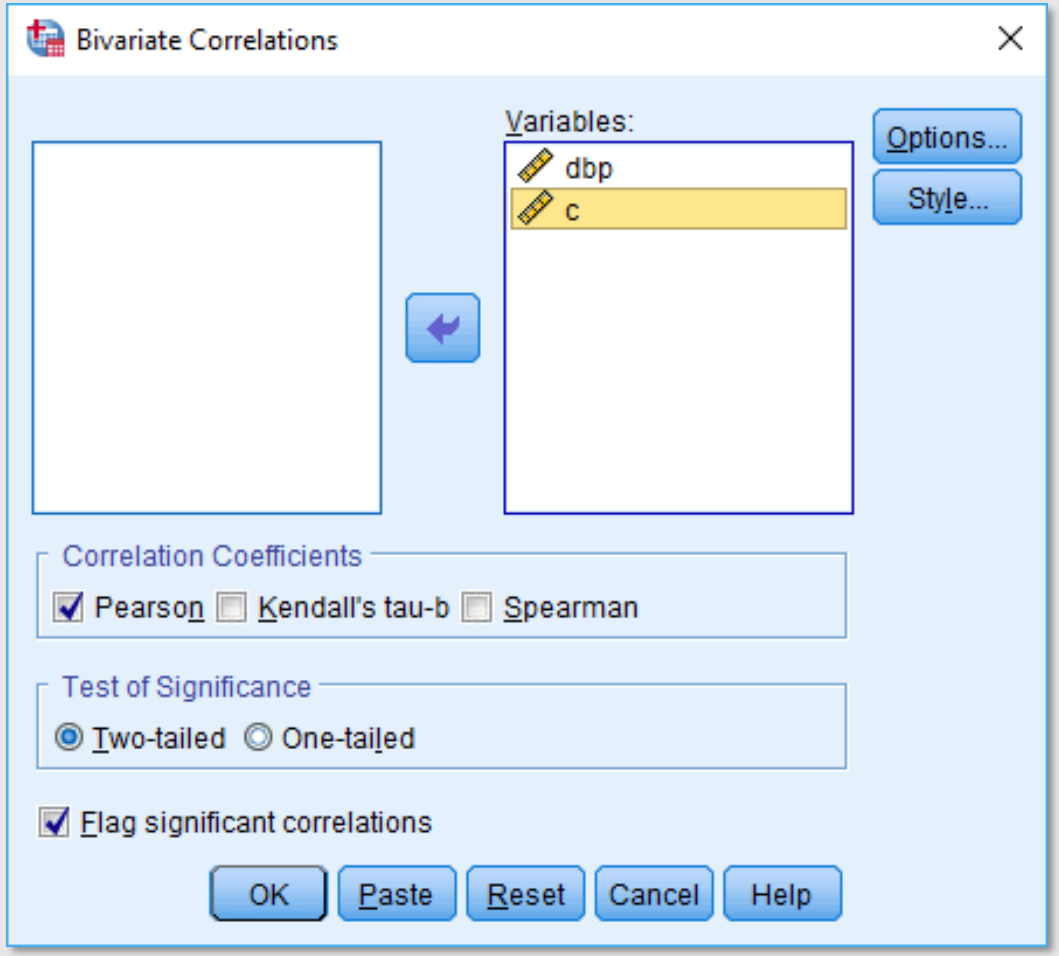
	dbp	c
1	80.00	307.00
2	72.00	282.00
3	90.00	341.00
4	74.00	317.00
5	68.00	286.00
6	106.00	416.00
7	83.00	326.00
8	87.00	379.00
9	104.00	389.00
10	78.00	318.00
11	89.00	352.00
12	76.00	287.00
13	96.00	386.00

Analysis: Bivariate

To test the relationship between the two variables, we need to calculate **the Pearson correlation coefficient**.

From the menu, select **Analyze -> Correlate -> Bivariate**

Drag the two variables from the left box into the **Variables:** box, select **Pearson** in the **Correlation Coefficients** field, and click **OK**



Results and interpretation

In the **Correlations** table, the **Pearson Correlation coefficient** (r) is 0.938, which is statistically significant (p<0.001)

Thus, there is **a strong positive correlation** between the diastolic blood pressure and cholesterol levels. This means that as cholesterol levels increases, blood pressure tend also to increase, and the same is true in reverse.

$r = -1$, perfect negative correlation
$r = 0$, no correlation
$r = 1$, perfect positive correlation
$0.7 < |r| < 1$, strong correlation
$0.5 < |r| < 0.7$, moderate correlation
$0.3 < |r| < 0.5$, weak correlation

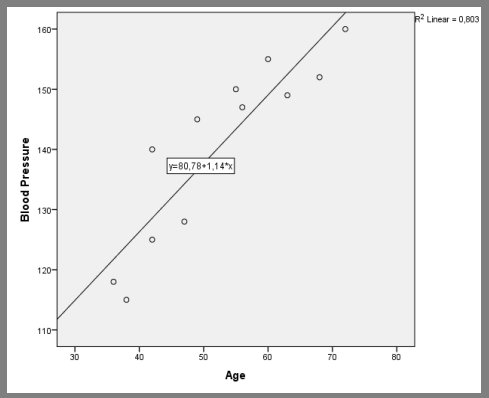
Correlations

Correlations

		dbp	c
dbp	Pearson Correlation	1	.938**
	Sig. (2-tailed)		.000
	N	13	13
c	Pearson Correlation	.938**	1
	Sig. (2-tailed)	.000	
	N	13	13

**. Correlation is significant at the 0.01 level (2-tailed).

Simple linear regression



What is simple linear regression?

- **Simple linear regression** is a statistical method that allows us to study relationships between **two continuous (quantitative)** variables
- One variable, denoted x, is regarded as the **predictor**, **explanatory**, or **independent** variable
- The other variable, denoted *y*, is regarded as the **response**, **outcome**, or **dependent** variable
- In **simple linear regression**, pairs of values for the two variables x and y are used to fit a straight line that best represents the relationship between them
- The null hypothesis is $\mathbf{H_0}$: there is no linear relationship between the independent variable x and the dependent variable y ($\beta_1 = 0$, where β_1 represents the slope of the regression line)

Assumptions

- Your **dependent variable** should be measured at the **continuous** level
- Your **independent variable** should also be measured at the **continuous** level
- There needs to be a **linear relationship** between the two variables (scatterplot)
- There should be **no significant outliers** (boxplot or scatterplot)
- You should have **independence of observations** (Durbin-Watson statistic)
- The data needs to show **homoscedasticity**
- The **residuals (errors)** of the regression line are **approximately normally distributed** (histogram)

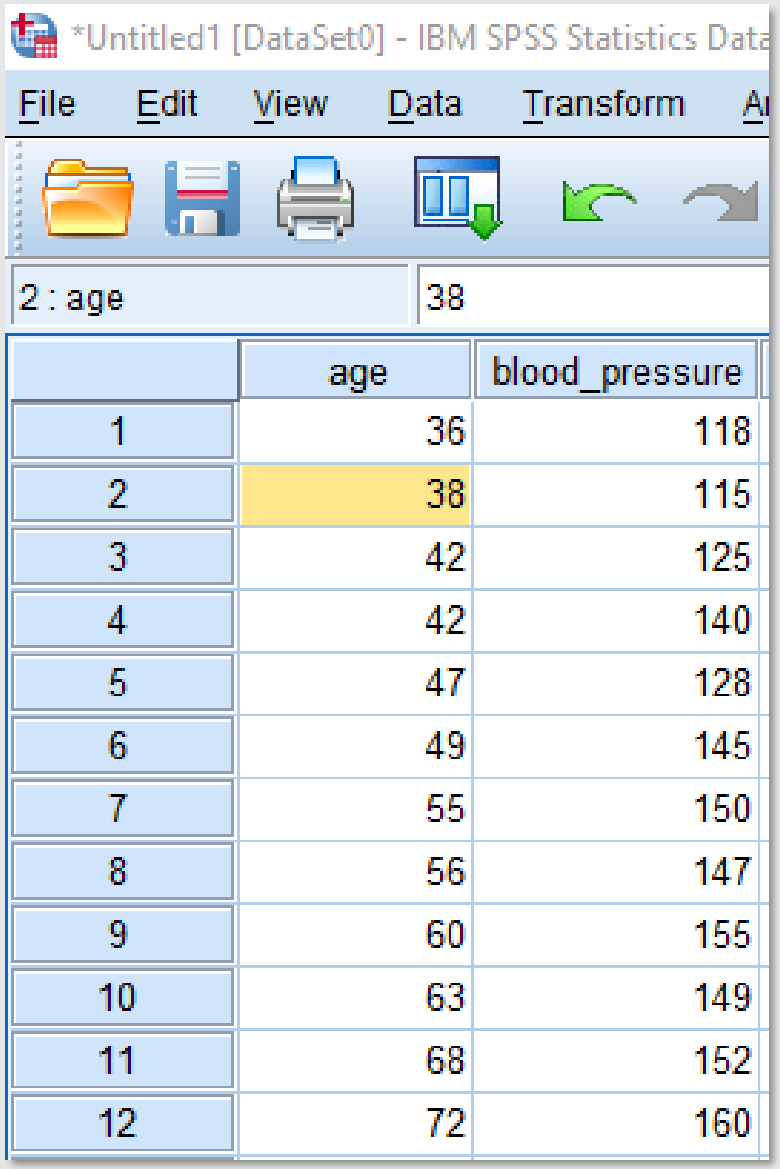
Ages and blood pressure readings from twelve women

The table on the right includes the ages and blood pressure readings from twelve women.

H_0: there is no linear relationship between the independent variable x and the dependent variable y ($\beta_1 = 0$, where β_1 represents the slope of the regression line)

Age (X)	Blood pressure (Y)
36	118
38	115
42	125
42	140
47	128
49	145
55	150
56	147
60	155
63	149
68	152
72	160

Enter the data in the **Data View** and define the variables in the **Variable View**

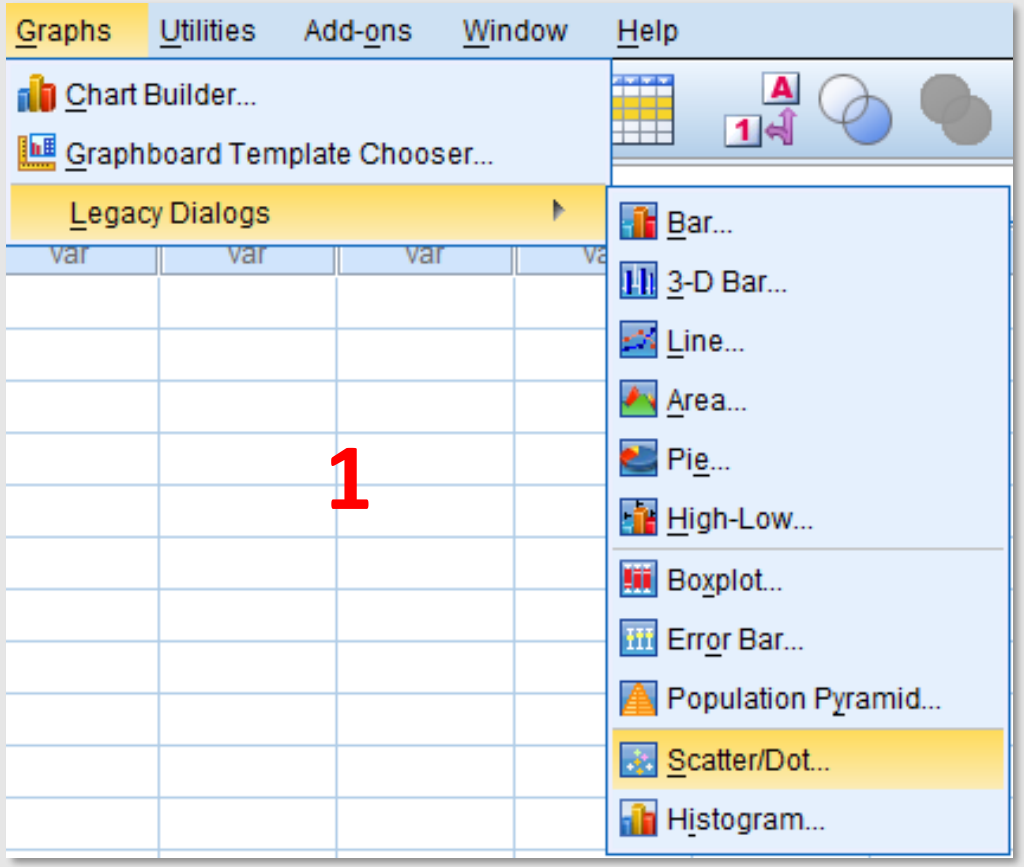


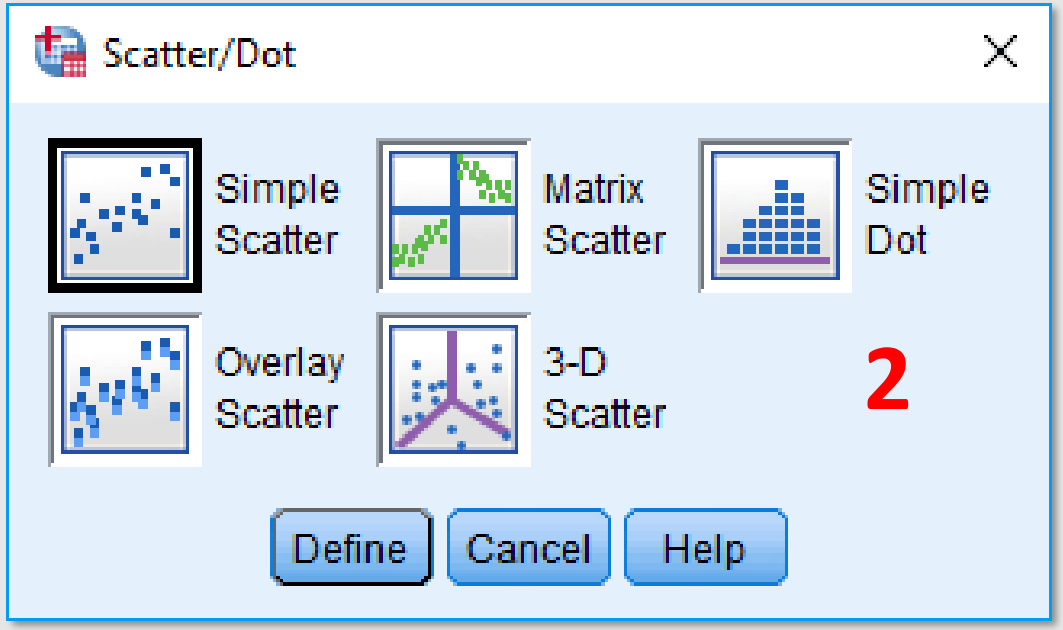
	age	blood_pressure
1	36	118
2	38	115
3	42	125
4	42	140
5	47	128
6	49	145
7	55	150
8	56	147
9	60	155
10	63	149
11	68	152
12	72	160

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
age	Numeric	8	0	Age	None	None	8	Right	Scale
blood_pressure	Numeric	8	0	Blood Pressure	None	None	10	Right	Scale

Scatterplot

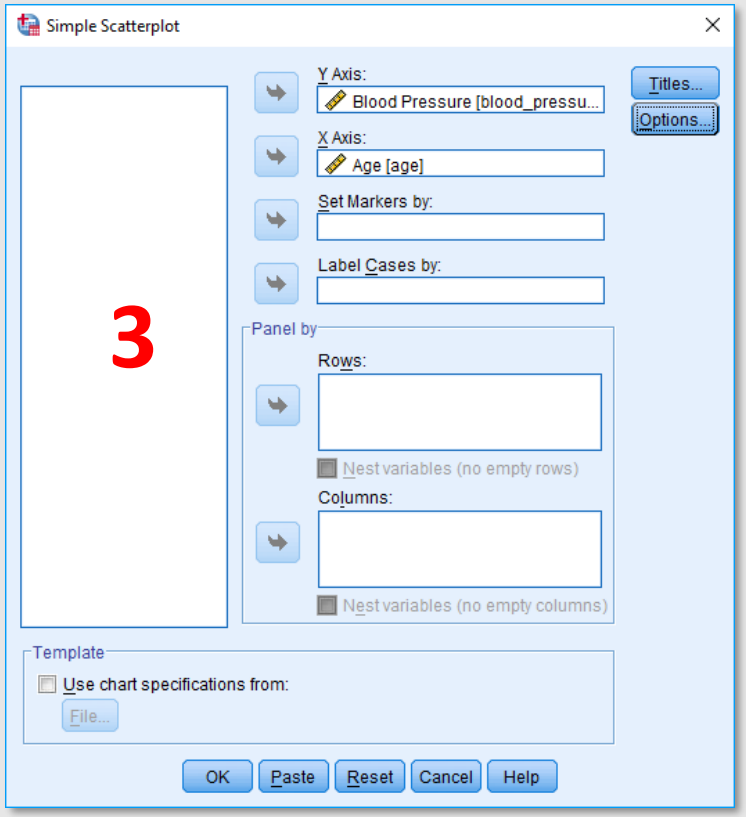
First, plot the **scatterplot** by selecting from the menu **Graphs -> Legacy Dialogs -> Scatter/Dot …**





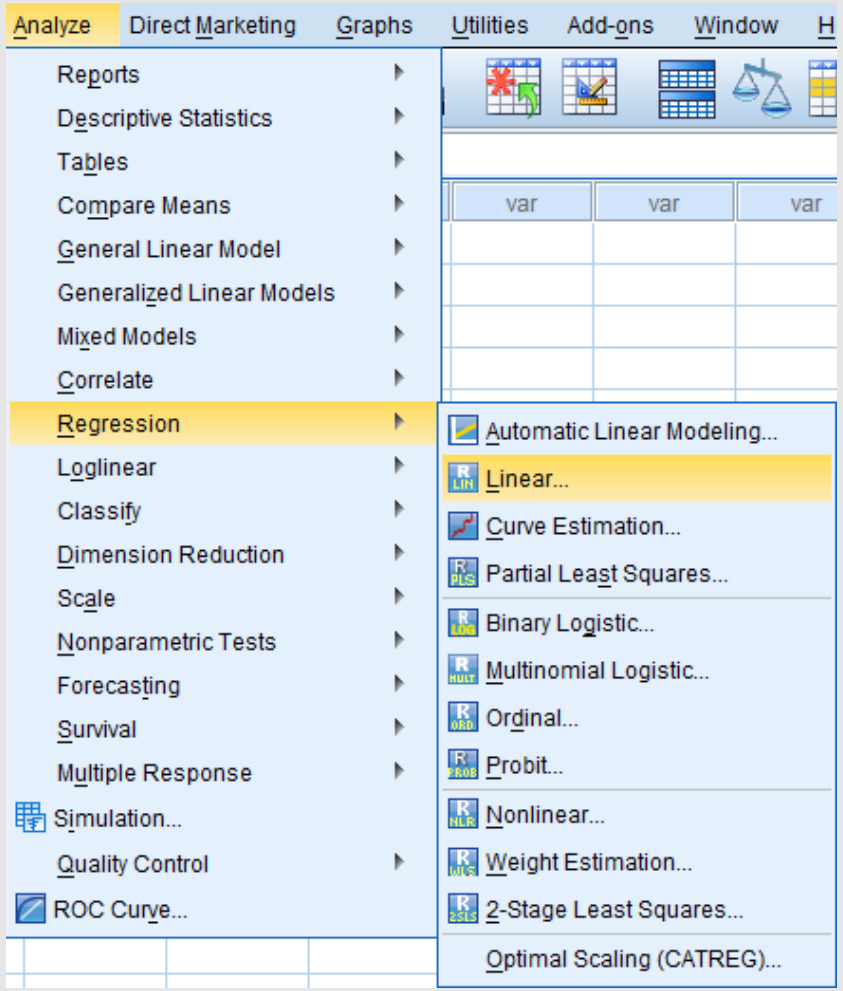
Scatterplot

In the **Simple Scatterplot** window (3), drag the variable **blood_pressure** from the left box into the **Y Axis:** field, and the variable **age** from the left box into the **X Axis:** field. Then, click **OK**

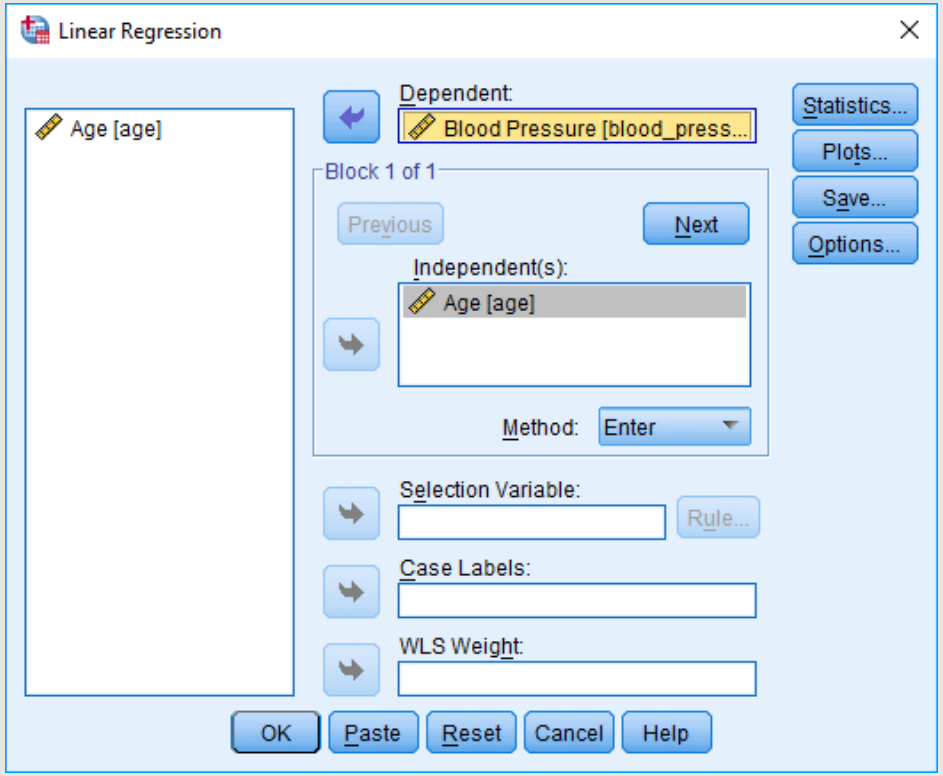


To display the **regression equation**, double click on the plot, and then select the appropriate icon from the toolbar 

Running the analysis



To analyze the data, select **Analyze -> Regression -> Linear** from the menu. Drag the variable **blood_pressure** from the left box into the **Dependent:** field, and the variable **age** from the left box into the **Independents(s):** box



R, R^2

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,896[a]	,803	,783	7,018

a. Predictors: (Constant), Age

- In the **Model Summary** table the **R** value represents the simple correlation and is 0.869, which indicates a high degree of correlation
- To **R Square** indicates how much of the **total variation** in the dependent variable **can be explained** from the independent variable. In this **regression model**, **80.3%** can be explained, which is very large

$r = -1$, perfect negative correlation
$r = 0$, no correlation
$r = 1$, perfect positive correlation
$0.7 < |r| < 1$, strong correlation
$0.5 < |r| < 0.7$, moderate correlation
$0.3 < |r| < 0.5$, weak correlation

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2008,200	1	2008,200	40,778	,000[b]
	Residual	492,467	10	49,247		
	Total	2500,667	11			

a. Dependent Variable: Blood Pressure

b. Predictors: (Constant), Age

- The **ANOVA** table, reports how well the **regression equation** fits the data (i.e., predicts the dependent variable)
- This table indicates that the regression model predicts the dependent variable significantly well ($p<0.001$) (i.e., it is a good fit for the data)

Regression model equation

The **Coefficients** table provides us with the necessary information to predict blood pressure from age

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	80,778	9,544		8,464	,000
	Age	1,138	,178	,896	6,386	,000

a. Dependent Variable: Blood Pressure

- The value +1.138 is the slope of the line. It represents the effect of the independent variable (age) on the dependent variable (blood pressure)
- Each additional year of age is associated with an increase of 1.138 mm Hg in blood pressure
- Thus, for an increase in age of 10 years, the estimated mean blood pressure increases by 11.38 mm Hg

The regression model formula is

$$\boldsymbol{y = a + b * x}$$

where

- $\boldsymbol{y}$ is the dependent variable (blood pressure)
- $\boldsymbol{x}$ is the independent variable (age)
- $\boldsymbol{a}, \boldsymbol{b}$ are the parameters of the regression model, with $\boldsymbol{a}$ being the intercept and $\boldsymbol{b}$ being the slope

$$\textbf{blood pressure} = \mathbf{80.778} + \mathbf{1.138} \times \textbf{Age}$$

Reporting the results

We found a significant relationship ($p<0.001$) between age and blood pressure, with an R^2 of 0.803. This suggests that age accounts for approximately 80.3% of the variance in blood pressure. Additionally, the result indicate that each additional year of age is associated with an increase of 1.138 mm Hg in blood pressure ($t = 6.386$, $p < 0.001$).

Practical exercise

A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

Analyze the relationship between age at birth (the independent variable) and birth weight (the dependent variable) using **simple linear regression**

Infant ID #	Gestational Age (weeks)	Birth Weight (grams)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005