Correlation coefficient

Συντελεστής συσχέτισης

Ζιντζαράς Ηλίας, M.Sc., Ph.D.

Καθηγητής Βιομαθηματικών-Βιομετρίας
Εργαστήριο Βιομαθηματικών
Τμήμα Ιατρικής
Πανεπιστήμιο Θεσσαλίας

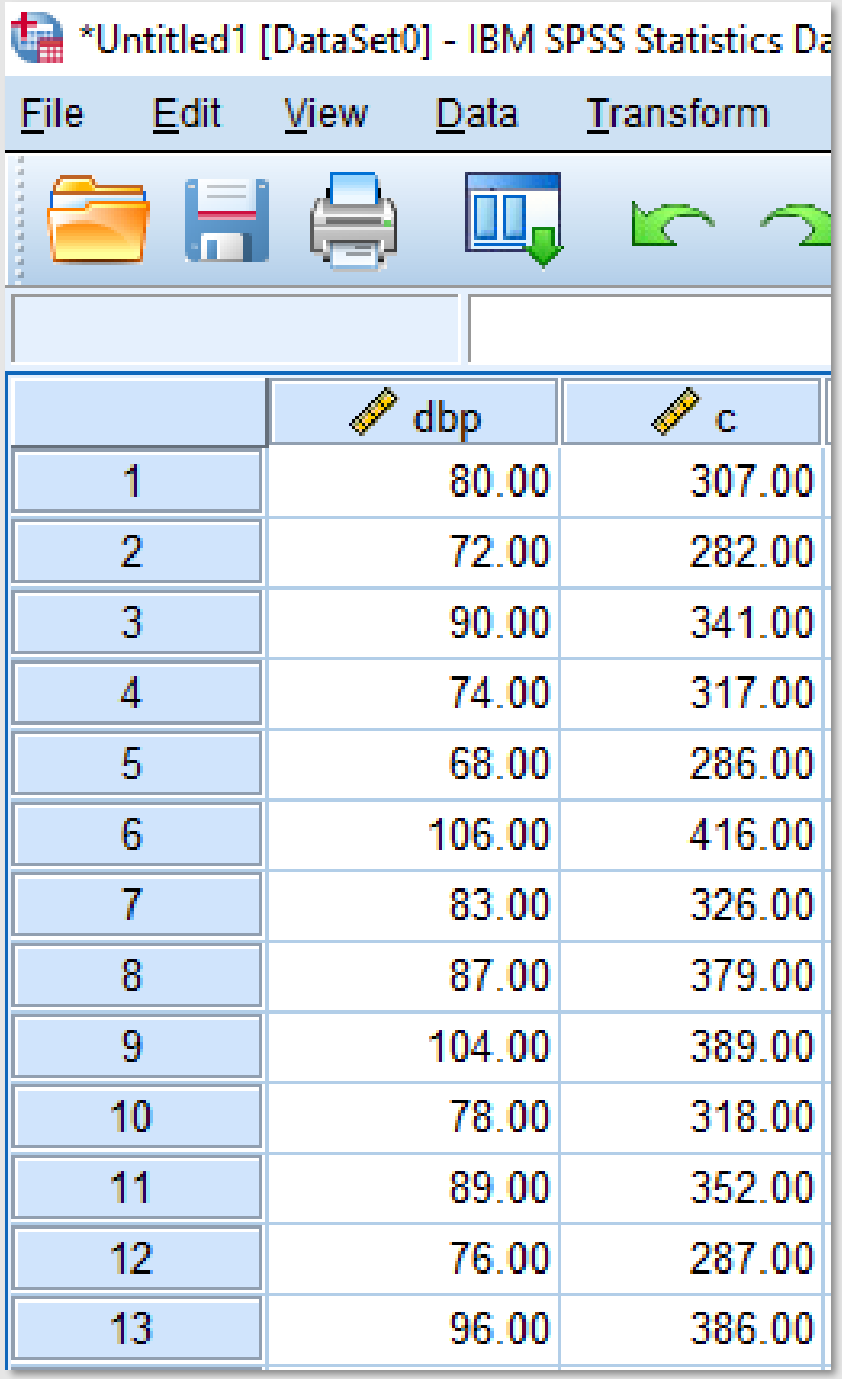
Institute for Clinical Research and Health Policy Studies
Tufts University School of Medicine
Boston, MA, USA

Θεόδωρος Μπρότσης, MSc, PhD
Εντεταλμένος Διδάσκων
(http://biomath.med.uth.gr)
Πανεπιστήμιο Θεσσαλίας
Email: tmprotsis@uth.gr

Πίεση αίματος και επίπεδα χοληστερόλης

Έστω ότι σε 13 άτομα μετρήθηκε η πίεση στο αίμα (DBP) και τα επίπεδα χοληστερόλης (C)

Θέλουμε να ελέγξουμε εάν υπάρχει σχέση μεταξύ DBP και C



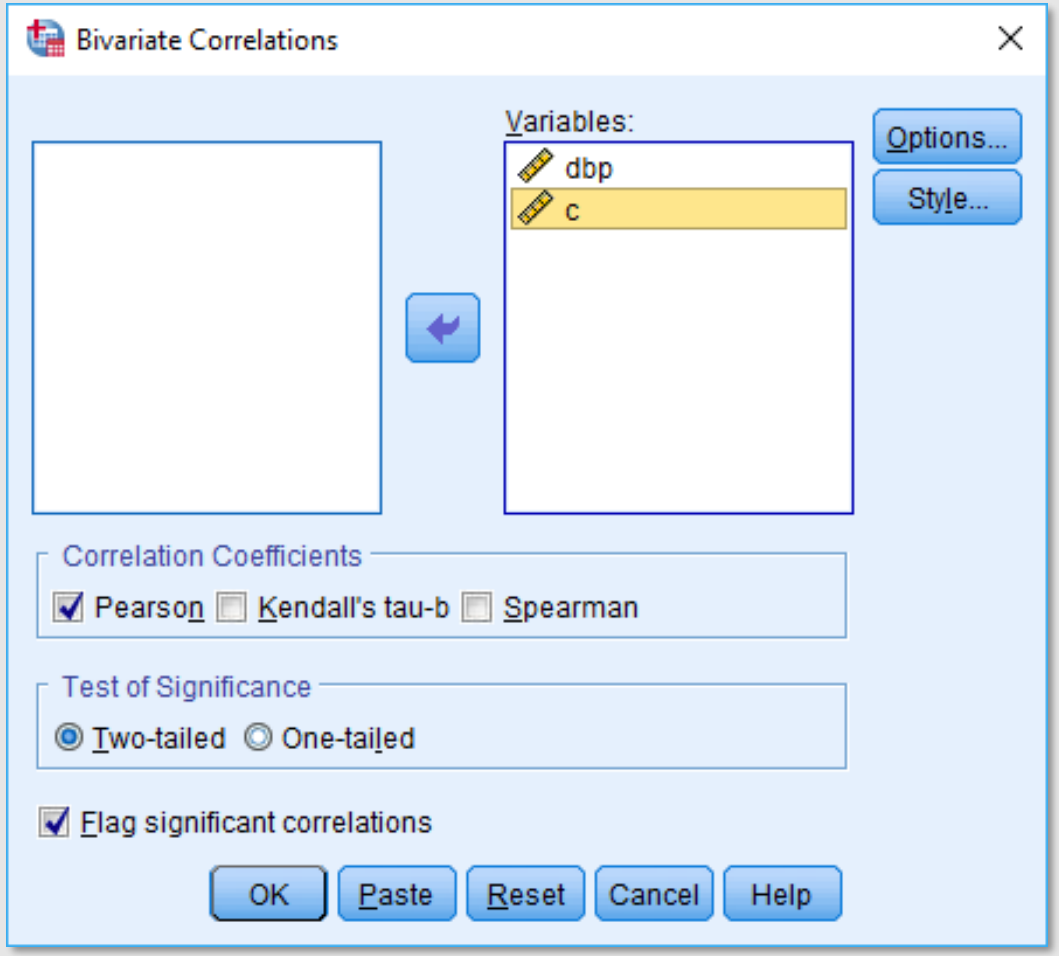
	dbp	c
1	80.00	307.00
2	72.00	282.00
3	90.00	341.00
4	74.00	317.00
5	68.00	286.00
6	106.00	416.00
7	83.00	326.00
8	87.00	379.00
9	104.00	389.00
10	78.00	318.00
11	89.00	352.00
12	76.00	287.00
13	96.00	386.00

Ανάλυση: Bivariate

Για να υπολογίσουμε το συντελεστή συσχέτισης r (Pearson Correlation) επιλέγουμε από το μενού διαδοχικά

Analyze -> Correlate -> Bivariate

Στην συνέχεια από την Λίστα Μεταβλητών, σέρνουμε τις μεταβλητές (DBP και C) στο πεδίο **Variables:**, επιλέγουμε **Pearson** και πατάμε OK



Ερμηνεία

Στα αποτελέσματα του output βρίσκουμε ότι $\boldsymbol{r = 0.938}$ που είναι σημαντικό σε $P = 0.000$ [Sig. (2-tailed)], δηλ. $P < 0.05$

Οπότε, υπάρχει στατιστικά σημαντική θετική σχέση μεταξύ χοληστερόλης και πίεσης

$r = -1$, τέλεια αρνητική συσχέτιση
$r = 0$, Μηδενική (δεν υπάρχει συσχέτιση)
$r = 1$, Τέλεια θετική συσχέτιση
$0.7 < |r| < 1$, ικανοποιητική ως πολύ ισχυρή
$0.5 < |r| < 0.7$, μέτρια έως ικανοποιητική
$0.3 < |r| < 0.5$, Ασθενής έως μέτρια

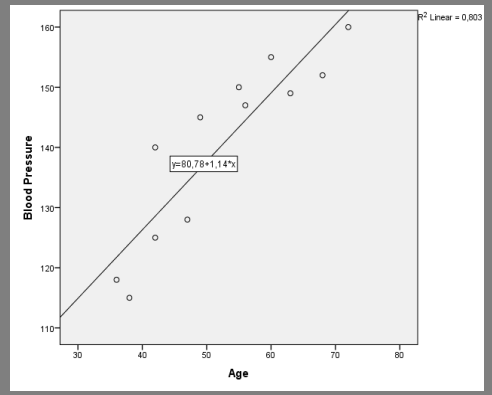
Correlations

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).

Απλή γραμμική παλινδρόμηση (simple linear regression)



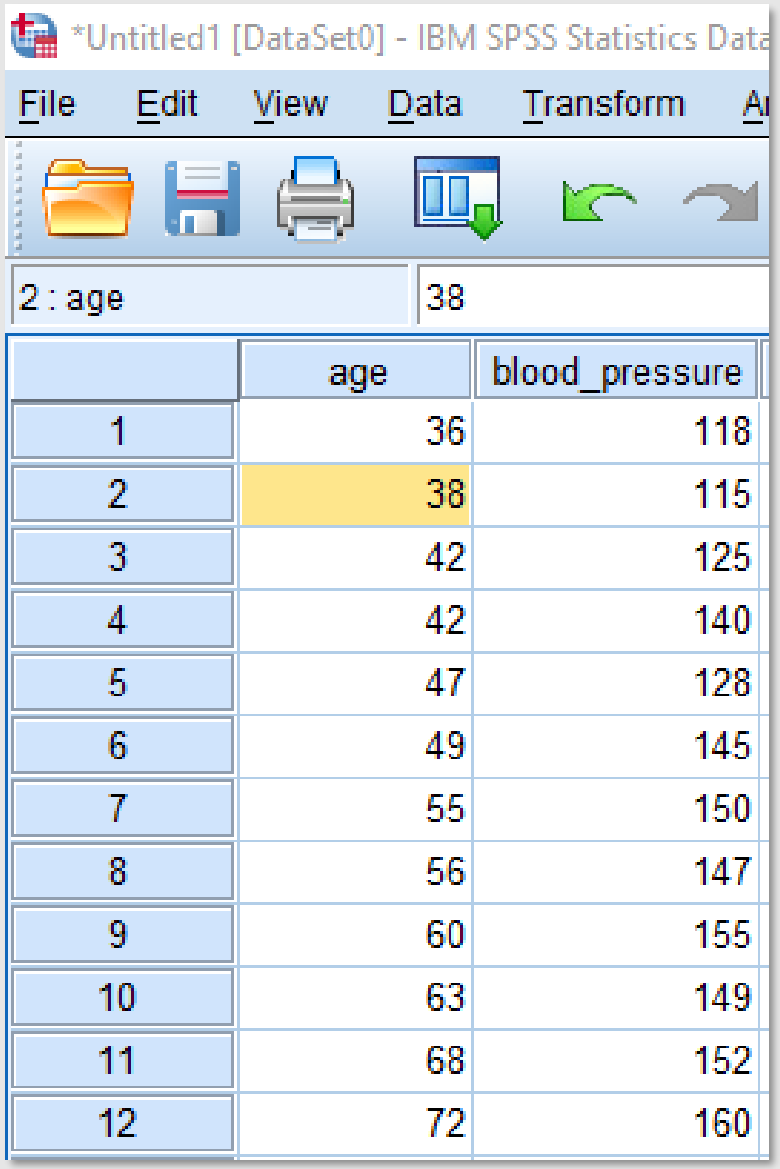
Απλή γραμμική παλινδρόμηση

- Διερευνά τη σχέση μεταξύ δυο (scaled) μεταβλητών Χ, Υ (π. χ. Χ: ηλικία και Υ: πίεση αίματος)
- Η μεταβλητή Χ καλείται ανεξάρτητη
- Η μεταβλητή Υ καλείται εξαρτημένη
- Τα ζεύγη των τιμών των δυο μεταβλητών (x, y) προσαρμόζονται σε μία ευθεία
- Ψάχνουμε τους συντελεστές της ευθείας και αν γίνεται καλή προσαρμογή
- Εξετάζουμε την H_0 (μηδενική υπόθεση): δεν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών Χ και Υ

Παράδειγμα: Πίεση αίματος και ηλικία

Από $n = 12$ γυναίκες λαμβάνουμε τις ακόλουθες τιμές πίεσης αίματος και της αντίστοιχης ηλικίας σε έτη

Ηλικία (X)	Πίεση αίματος (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

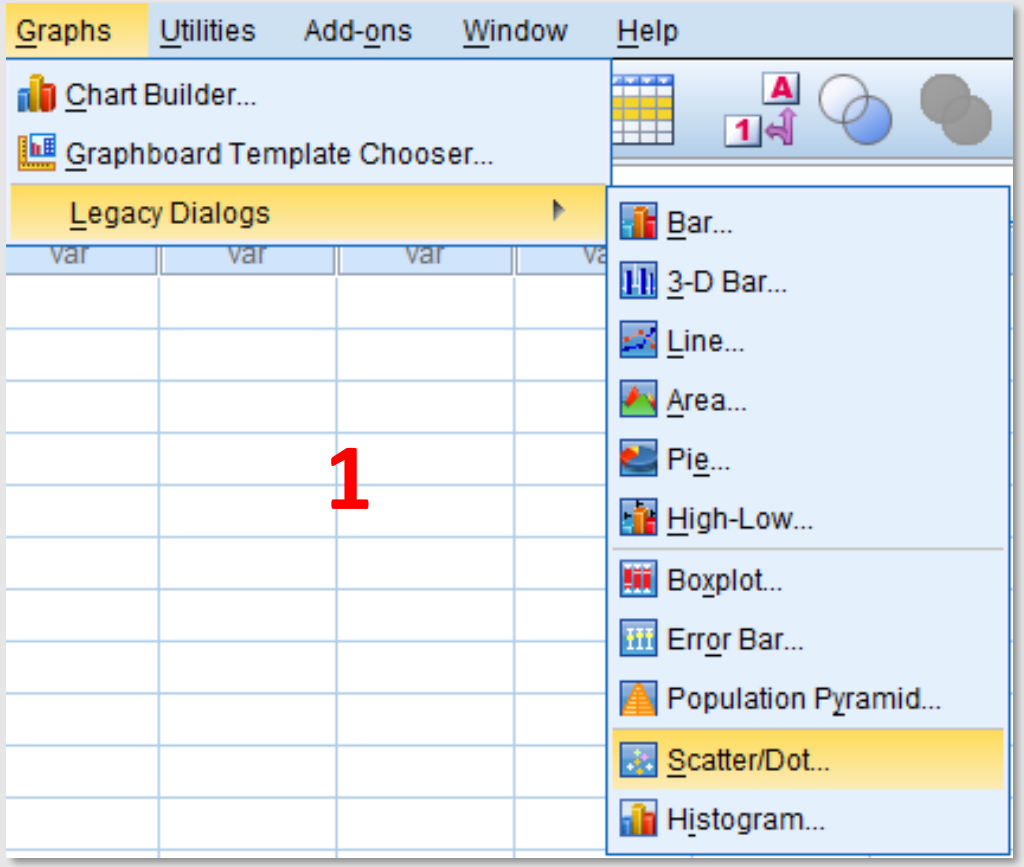


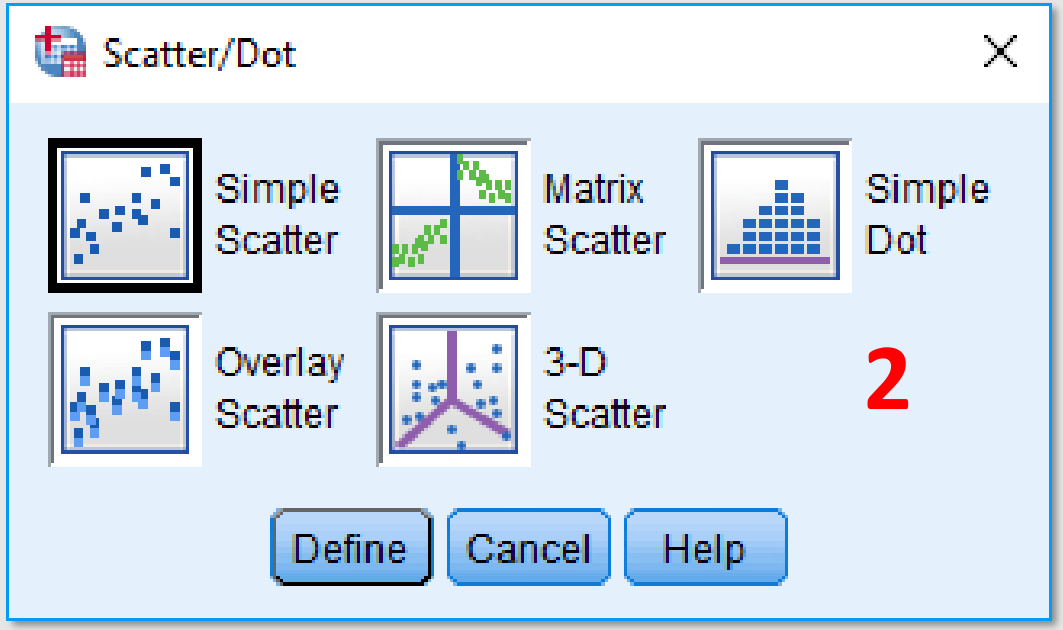
	age	blood_pressure
1	36	118
2	38	115
3	42	125
4	42	140
5	47	128
6	49	145
7	55	150
8	56	147
9	60	155
10	63	149
11	68	152
12	72	160

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

Διάγραμμα συσχέτισης

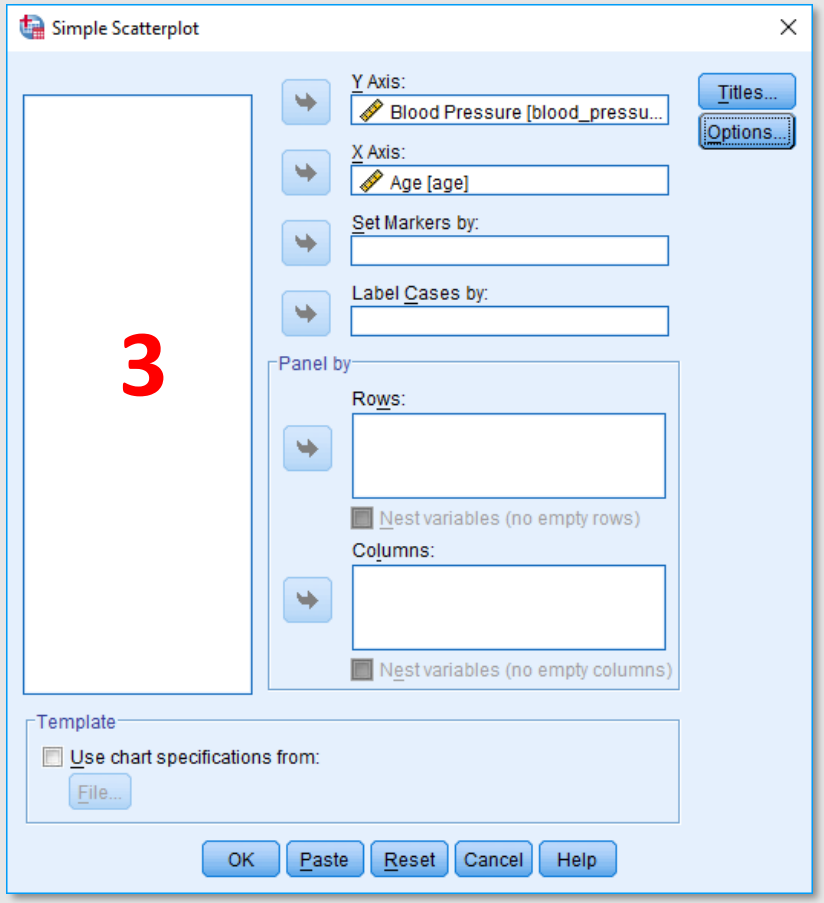
Δημιουργούμε αρχικά το Scatter Plot επιλέγοντας από το μενού **Graphs -> Legacy Dialogs -> Scatter/Dot …**





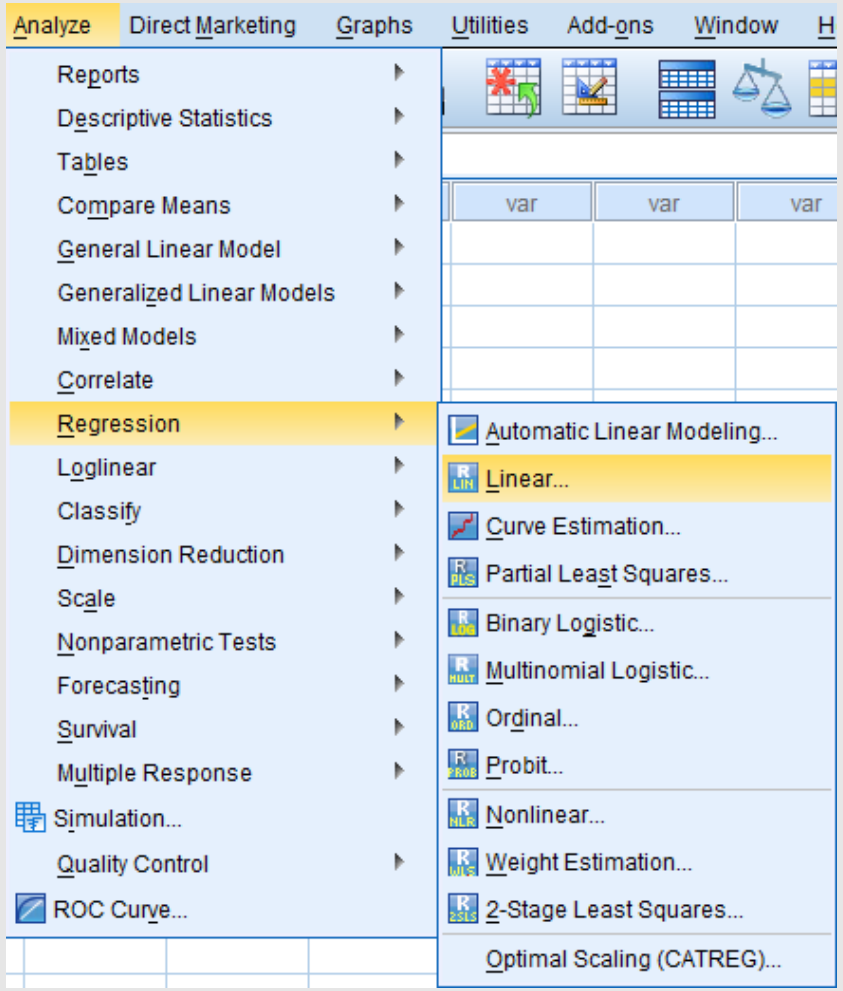
Διάγραμμα συσχέτισης

… **Scatter/Dot**

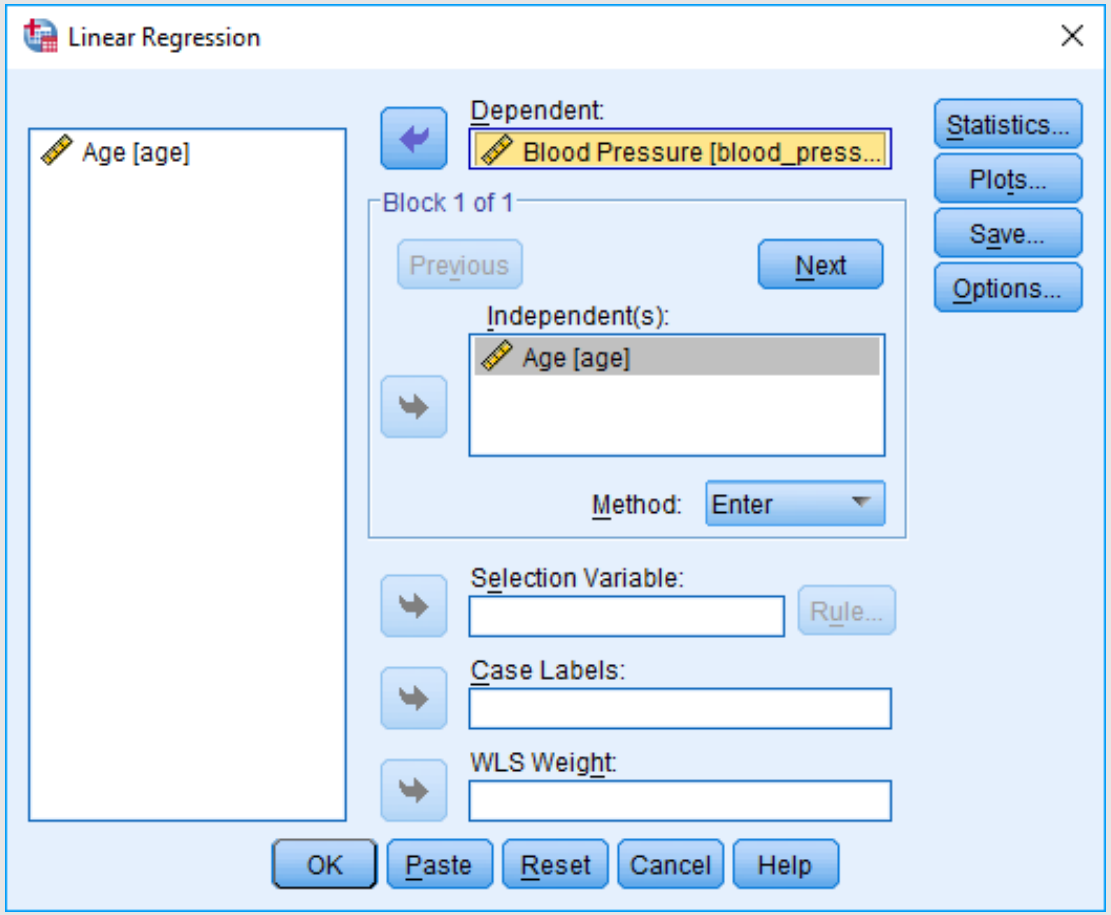


Η γραμμή εξίσωσης εμφανίζεται με δύο κλικ πάνω στο διάγραμμα και επιλογή του εικονιδίου

Ανάλυση: Απλή γραμμική παλινδρόμηση



Για να αναλύσουμε τα δεδομένα επιλέγουμε **Analyse -> Regression -> Linear**



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

- Στο **Model Summary** η τιμή **R** αναφέρεται στην απόλυτη τιμή του συντελεστή γραμμικής συσχέτισης.
- Το **R Square** ονομάζεται συντελεστής προσδιορισμού. Ο συντελεστής αυτός φανερώνει **το ποσοστό της μεταβλητότητας** των δεδομένων που εξηγείται από το γραμμικό μοντέλο. Το συγκεκριμένο μοντέλο εξηγεί το $\mathbf{80.3}\%$ της μεταβλητότητας των δεδομένων.

$r = -1$, τέλεια αρνητική συσχέτιση
$r = 0$, Μηδενική (δεν υπάρχει συσχέτιση)
$r = 1$, Τέλεια θετική συσχέτιση
$0.7 < |r| < 1$, ικανοποιητική ως πολύ ισχυρή
$0.5 < |r| < 0.7$, μέτρια έως ικανοποιητική
$0.3 < |r| < 0.5$, Ασθενής έως μέτρια

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

- Ο αριθμός $\boldsymbol{SSR = 2008.200}$ δείχνει την διακύμανση που εξηγείται από το μοντέλο, ενώ ο $\boldsymbol{SST = 2500.667}$ την συνολική διακύμανση.
- Η διαφορά τους είναι η διακύμανση που δεν εξηγείται από το μοντέλο $(SSE = SST - SSR = 492.467)$
- $r^2 = \frac{SSR}{SST} = \frac{2008.200}{2500.667} = 0.803$

- Καθώς το $p - value\ (0.000) < 0.001$ απορρίπτουμε την μηδενική υπόθεση, οπότε υπάρχει γραμμική σχέση μεταξύ των μεταβλητών, στατιστικά σημαντική
- Το μοντέλο παλινδρόμησης προβλέπει καλά την εξαρτημένη μεταβλητή από την ανεξάρτητη (p<0.001) (καλή προσαρμογή των δεδομένων)

Εξίσωση γραμμής παλινδρόμησης

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

Η τιμή $+\mathbf{1.138}$ είναι η κλίση της ευθείας (slope). Επίσης φανερώνει την επίδραση της ανεξάρτητης στην εξαρτημένη μεταβλητή

Για κάθε αύξηση της ανεξάρτητης μεταβλητής (age) κατά μία μονάδα η εκτιμώμενη μέση τιμή της εξαρτημένης μεταβλητής (blood pressure) αυξάνεται κατά 1.138 μονάδες.

Έτσι, για μία αύξηση της ηλικίας κατά 10 έτη, η αύξηση της εκτιμώμενης μέσης πίεσης είναι 11.38 μονάδες.

Το μοντέλο είναι της μορφής

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

όπου

- $\boldsymbol{y}$ είναι η εξαρτημένη μεταβλητή (Blood pressure)
- $\boldsymbol{x}$ η ανεξάρτητη (age)
- $\boldsymbol{a, b}$ οι παράμετροι του μοντέλου τις οποίες εκτιμάμε

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

Αναφορά αποτελεσμάτων

- Η **ηλικία** σε σχέση με την **πίεση** περιγράφεται με την εξής γραμμή παλινδρόμησης:

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

- Η γραμμή παλινδρόμησης που έχει εκτιμηθεί είναι στατιστικά σημαντική $(F(1, 10) = 40.778, p < 0,001)$
- Το **80.3%** της διακύμανσης της πίεσης ερμηνεύεται από την διακύμανση της ηλικίας
- Επίσης, το εκτιμώμενο μοντέλο υποδεικνύει ότι όταν η **ηλικία** είναι αυξημένη κατά **μία μονάδα**, τότε η **πίεση** αναμένεται να είναι αυξημένη κατά **1.138 μονάδες**
- Για μία αύξηση της ηλικίας κατά 10 έτη, η αύξηση της εκτιμώμενης μέσης πίεσης είναι 11.38 μονάδες

Πρακτική άσκηση

Διεξήχθη μια μικρή μελέτη στην οποία συμμετείχαν 17 βρέφη για να διερευνηθεί η συσχέτιση μεταξύ της ηλικίας κύησης κατά τη γέννηση, μετρούμενη σε εβδομάδες, και του βάρους γέννησης, μετρούμενο σε γραμμάρια.

Αναλύστε τη σχέση μεταξύ της ηλικίας κατά τη γέννηση (η ανεξάρτητη μεταβλητή) και του βάρους γέννησης (η εξαρτημένη μεταβλητή) χρησιμοποιώντας **απλή γραμμική παλινδρόμηση**

Α/Α	Ηλικία κύησης (εβδομάδες)	Βάρος γέννησης (γραμμάρια)
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