



# 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](mailto: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

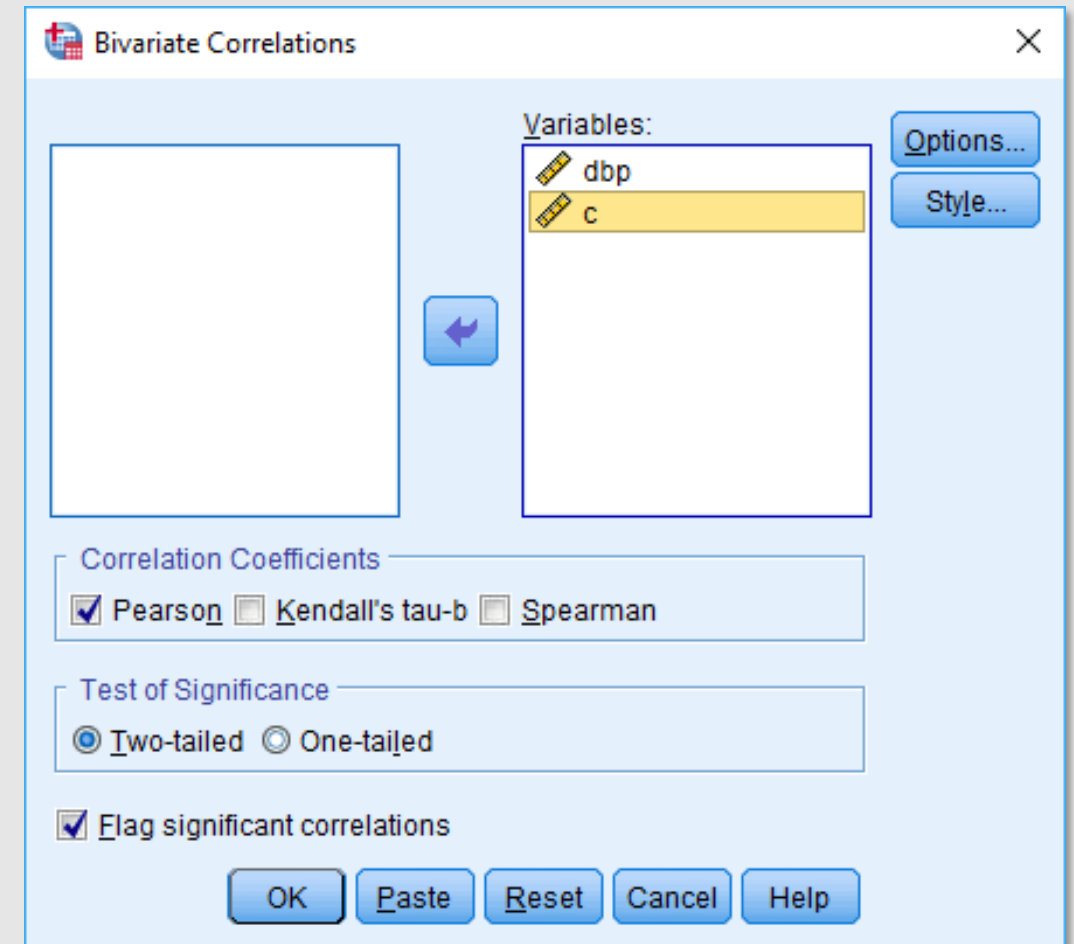


# Ανάλυση: Bivariate

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

**Analyze -> Correlate -> Bivariate**

Στην συνέχεια από την Λίστα Μεταβλητών, επιλέγουμε τις μεταβλητές (DBP και C), για τις οποίες θέλουμε να υπολογίσουμε τον συντελεστή συσχέτισης, και κατόπιν την επιλογή **Pearson**





# Ερμηνεία

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

Οπότε, υπάρχει στατιστικά σημαντική σχέση μεταξύ πίεσης και χοληστερόλης

$r = -1$ , 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

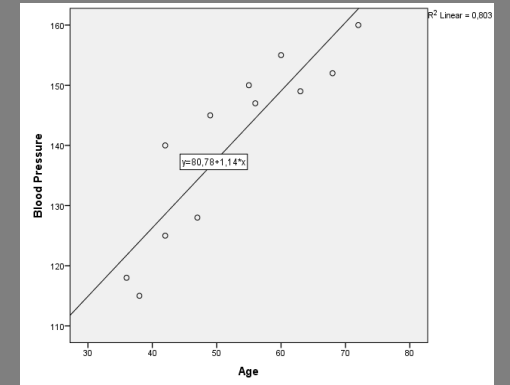
## → Correlations

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  $X, Y$  ( π. χ.  $X$ : ηλικία και  $Y$ : πίεση αίματος)
- Η μεταβλητή  $X$  καλείται ανεξάρτητη
- Η μεταβλητή  $Y$  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  $X$  και  $Y$



## 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



\*Untitled1 [DataSet0] - IBM SPSS Statistics Data Editor

File	Edit	View	Data	Transform	Alt
------	------	------	------	-----------	-----

2 : age	38
---------	----

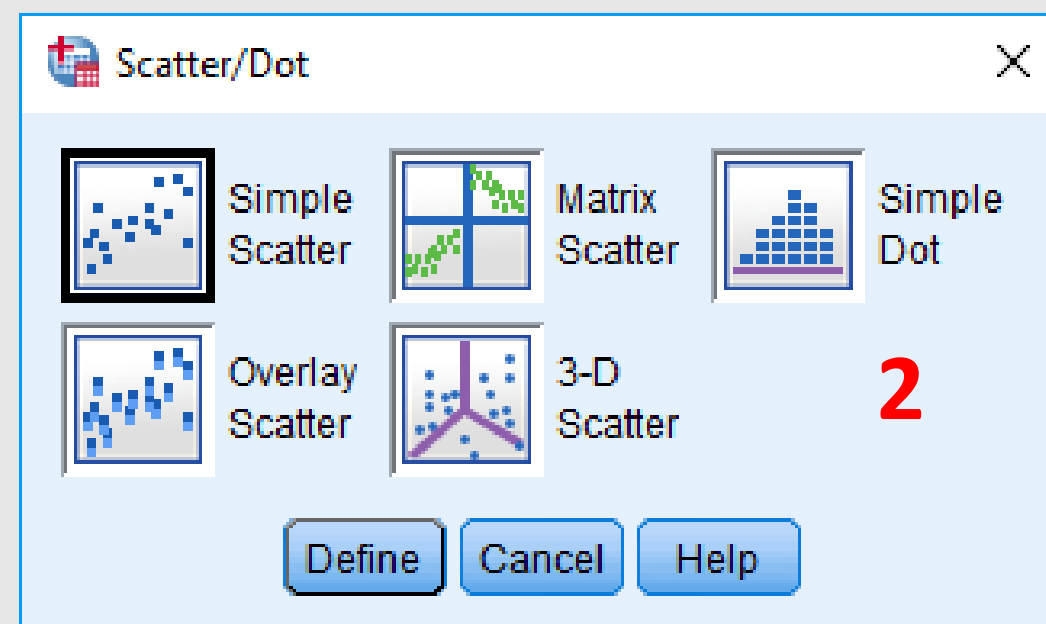
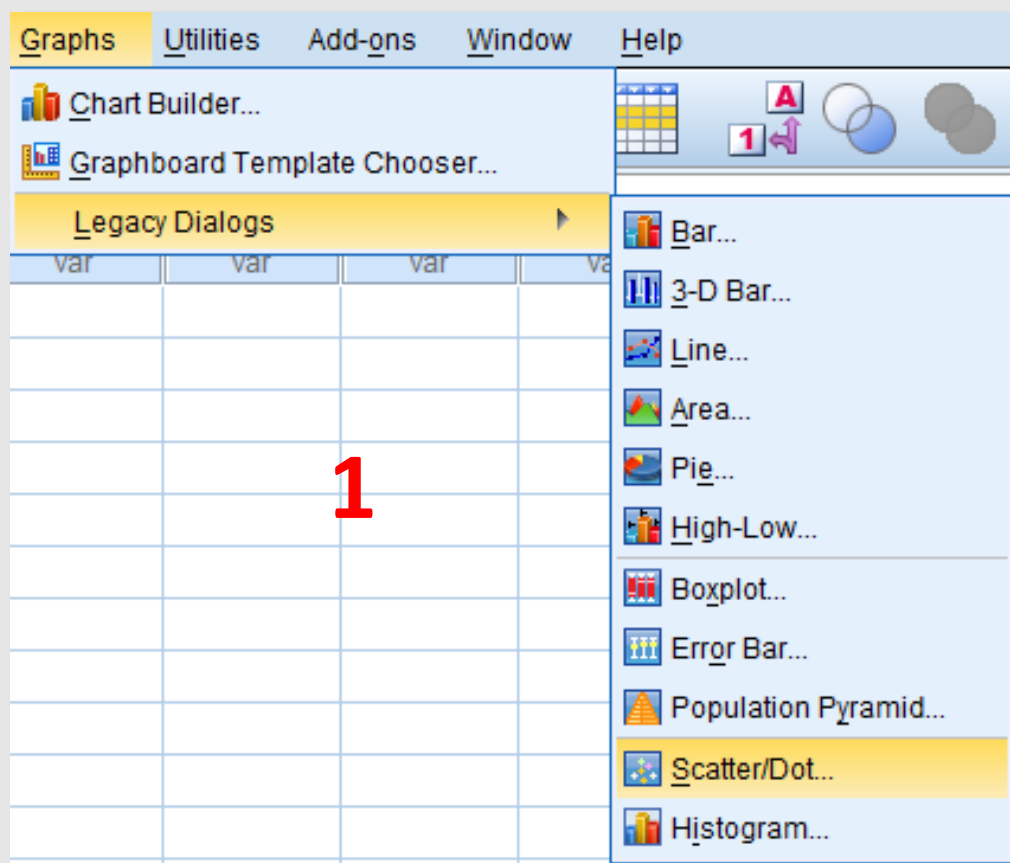
	age	blood_pressure
1	36	118
2	38	115
3	42	125
4	42	140
5	47	128
6	49	145
7	55	150
8	56	147
9	60	155
10	63	149
11	68	152
12	72	160

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

Simple Scatterplot

Y Axis: Blood Pressure [blood\_pressu...]

X Axis: Age [age]

Set Markers by:

Label Cases by:

Panel by:

Rows:

Columns:

Template

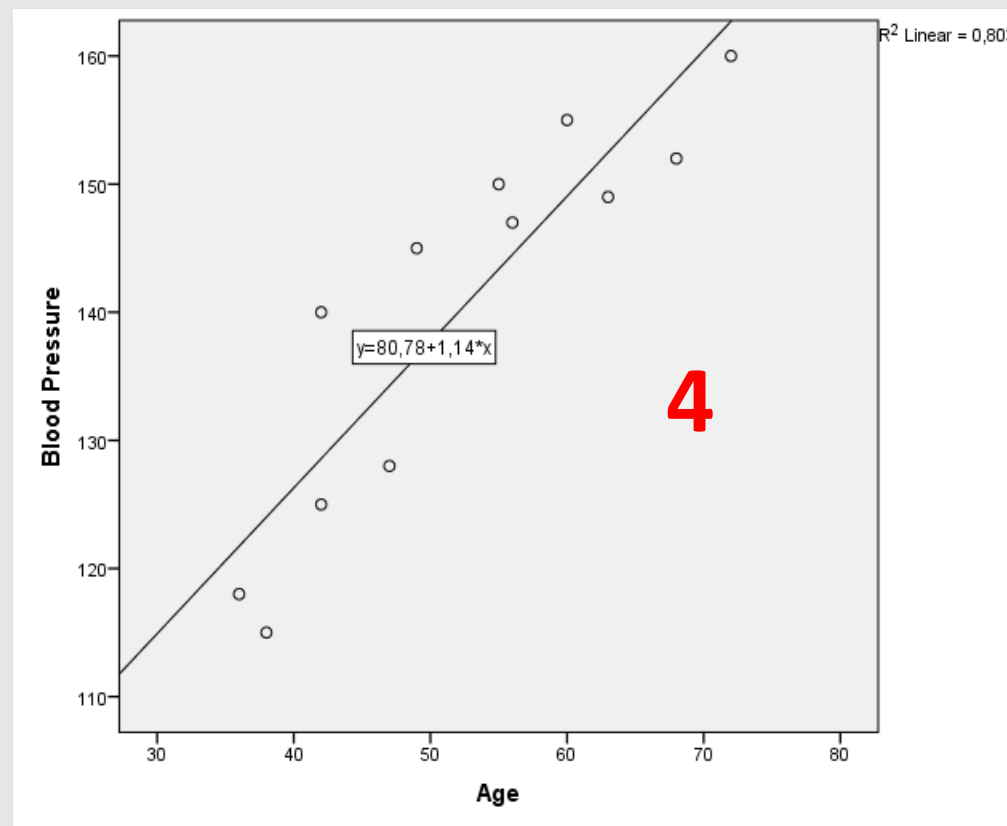
☐ Use chart specifications from:

File...

OK Paste Reset Cancel Help

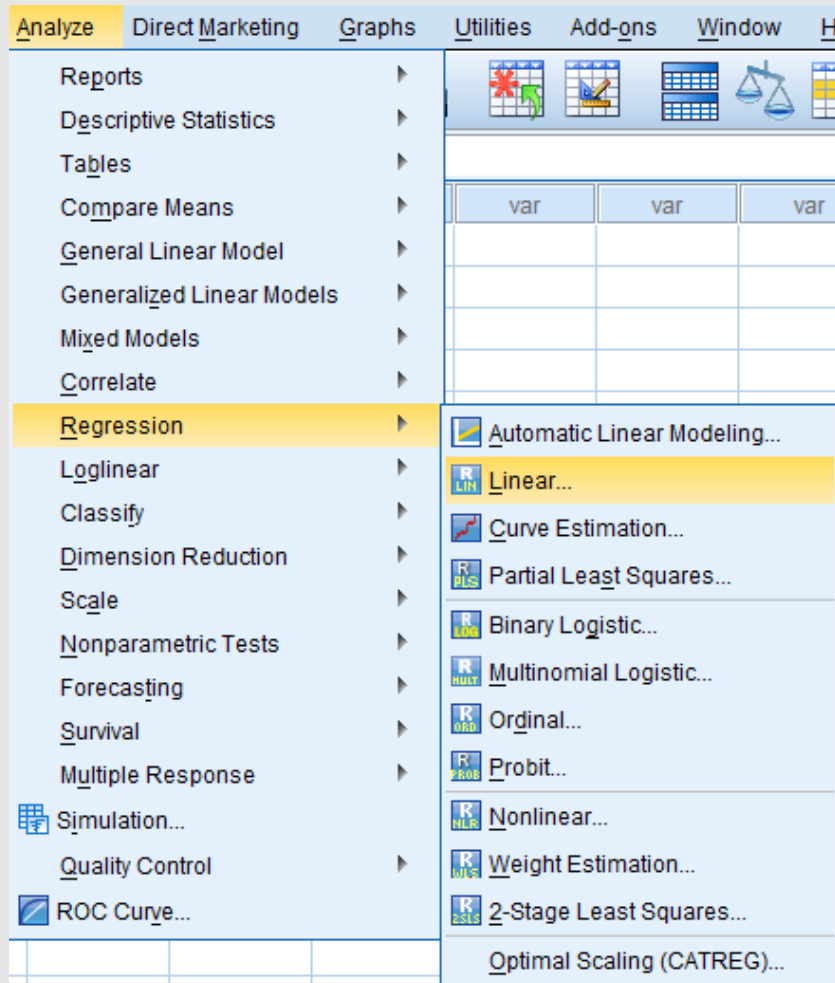
3

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

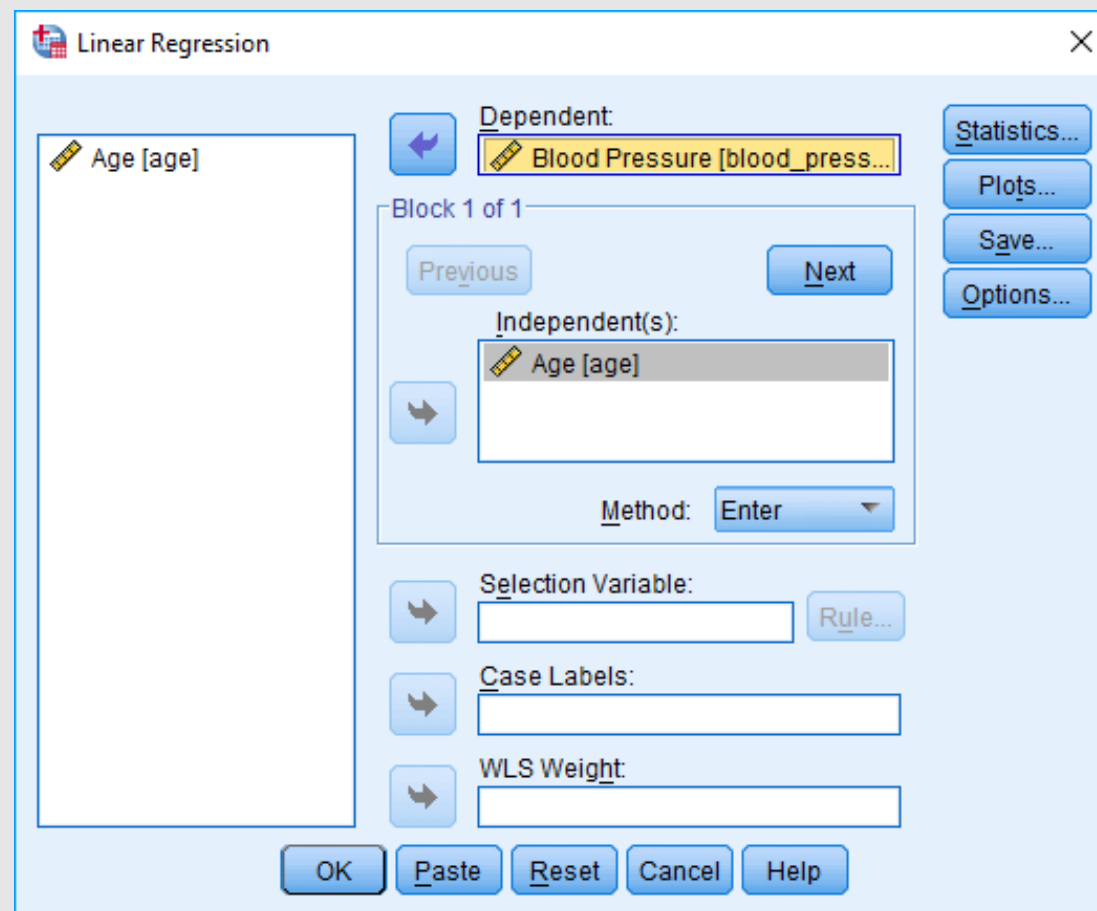




# Ανάλυση: Απλή γραμμική παλινδρόμηση



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



# $R, R^2, Adjusted R Square$

**Model Summary**

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,896 <sup>a</sup>	,803	,783	7,018

a. Predictors: (Constant), Age

$r = -1$ , 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

- 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 **80.3%** της μεταβλητότητας των δεδομένων.
- Ο προσαρμοσμένος συντελεστής (**Adjusted R Square**) λαμβάνει υπόψη και το μέγεθος του δείγματος.

**ANOVA<sup>a</sup>**

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

a. Dependent Variable: Blood Pressure

b. Predictors: (Constant), Age

- Καθώς το  $p - value$  ( $0.000$ )  $< 0.001$  απορρίπτουμε την μηδενική υπόθεση, οπότε υπάρχει γραμμική σχέση μεταξύ των μεταβλητών, στατιστικά σημαντική
- Ο αριθμός  **$SSR = 2008.200$**  δείχνει την διακύμανση που εξηγείται από το μοντέλο, ενώ ο  **$SST = 2500.667$**  την συνολική διακύμανση.
- Η διαφορά τους είναι η διακύμανση που δεν εξηγείται από το μοντέλο ( $SSE = SST - SSR = 492.467$ )
- $$r^2 = \frac{SSR}{SST} = \frac{2008.200}{2500.667} = 0.803$$



# Εξίσωση γραμμής παλινδρόμησης

Coefficients<sup>a</sup>

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 **+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

$$y = a + b * x$$

όπου

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

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



## Αναφορά αποτελεσμάτων

- Η **ηλικία** σε σχέση με την **πίεση** περιγράφεται με την εξής γραμμή παλινδρόμησης:  
$$\text{Blood pressure} = 80.778 + 1.138 \times \text{Age}$$
- 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