ΒΙΟΣΤΑΤΙΣΤΙΚΗ

Εξάρτηση ποσοτικών χαρακτηριστικών: η έννοια της εξάρτησης δύο ποσοτικών μεταβλητών

Μ.Ι Κάσδαγλη, Phd

Εργ. Υγιεινής, Επιδημιολογίας και Ιατρικής Στατιστικής,
Ιατρική Σχολή Πανεπιστημίου Αθηνών

Διερεύνηση σχέσεων

	Έκβαση (ανταπόκριση)	
Έκθεση / Παρέμβαση	Ποσοτική	Ποιοτική
Ποιοτική		
Ποσοτική		

Διερεύνηση σχέσεων

	Έκβαση (ανταπόκριση)	
Έκθεση / Παρέμβαση	Ποσοτική	Ποιοτική
Ποιοτική	• **t-test**	• **X^2**
Ποσοτική		

Διερεύνηση σχέσεων

	Έκβαση (ανταπόκριση)	
Έκθεση / Παρέμβαση	Ποσοτική	Ποιοτική
Ποιοτική	• **t-test** • **Γραμμική εξάρτηση**	• **χ^2** • Λογαριθμιστική (logistic)
Ποσοτική	• Συσχέτιση • **Γραμμική εξάρτηση**	• Λογαριθμιστική (logistic)

Σε μία έρευνα συλλέγονται, μεταξύ άλλων, τα εξής δεδομένα για τους συμμετέχοντες:

- Ηλικία (έτη)
- Φύλο
- Συστολική πίεση (mm Hg)
- Κάπνισμα (ναι/όχι)

Πιθανά ερωτήματα

– Διαφέρουν τα επίπεδα διαστολικής πίεσης μεταξύ ανδρών και γυναικών;

- Πίεση: ποσοτική μεταβλητή
- Φύλο: ποιοτική μεταβλητή με 2 επίπεδα

Στατιστική δοκιμασία ? **T-Test**

Πιθανά ερωτήματα

- Διαφέρουν τα επίπεδα διαστολικής πίεσης μεταξύ ανδρών και γυναικών;
  - Πίεση: ποσοτική μεταβλητή
  - Φύλο: ποιοτική μεταβλητή με 2 επίπεδα

  Στατιστική δοκιμασία ? **T-Test**

- Διαφέρει η συχνότητα καπνιστών μεταξύ ανδρών και γυναικών; (Υπάρχει σχέση μεταξύ καπνίσματος και φύλου;)
  - Κάπνισμα: ποιοτική μεταβλητή με 2 επίπεδα
  - Φύλο: ποιοτική μεταβλητή με 2 επίπεδα

  Στατιστική δοκιμασία ? **X^2**

Πιθανά ερωτήματα

– Υπάρχει σχέση μεταξύ ηλικίας και διαστολικής πίεσης;
  - Ηλικία: ποσοτική μεταβλητή
  - Συστολική πίεση: ποσοτική μεταβλητή

Στατιστική δοκιμασία ?

Συσχέτιση ή Γραμμική Εξάρτηση

Συσχέτιση και εξάρτηση

Η διάκριση μεταξύ συσχέτισης και παλινδρόμησης (εξάρτησης) είναι περισσότερο εννοιολογική και λιγότερο στατιστική

- Εάν μας ενδιαφέρει η ένταση της σχέσης των δύο μεταβλητών, αρκεί η συσχέτιση

- Εάν μας ενδιαφέρει η μελέτη της εξάρτησης της μιας μεταβλητής από την άλλη (εξαρτημένη μεταβλητή-ανεξάρτητη μεταβλητή) τότε επιλέγουμε την εξάρτηση (παλινδρόμηση)

Εξάρτηση

ΠΡΟΥΠΟΘΕΣΗ:

- Το εξαρτημένο μέγεθος θα πρέπει να κατανέμεται κανονικά (για κάθε συγκεκριμένη τιμή του ανεξάρτητου)
- **Αυθαίρετη** επιλογή του ενός μεγέθους (π.χ. ηλικία: εξασφάλιση επιλογής από όλες τις ηλικιακές ομάδες)

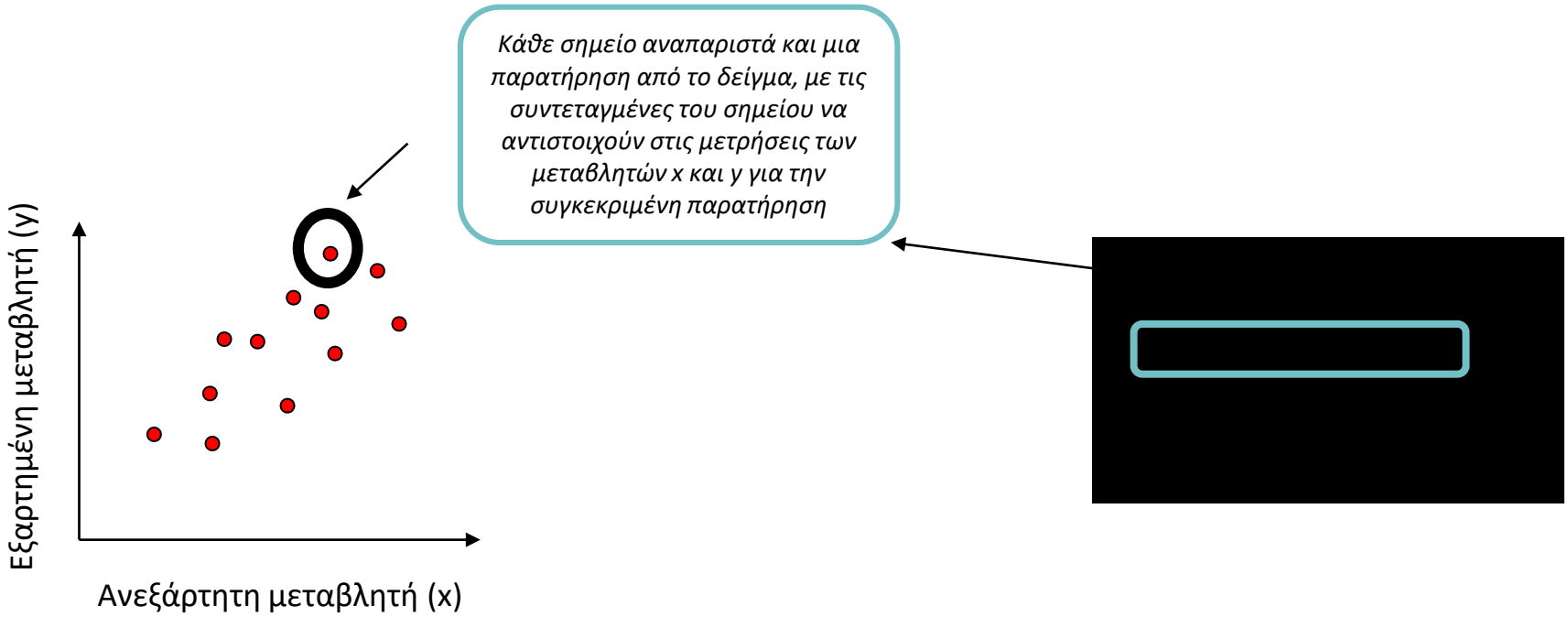
ΜΕ ΤΗΝ ΕΞΑΡΤΗΣΗ ΕΞΕΤΑΖΟΥΜΕ ΤΗ

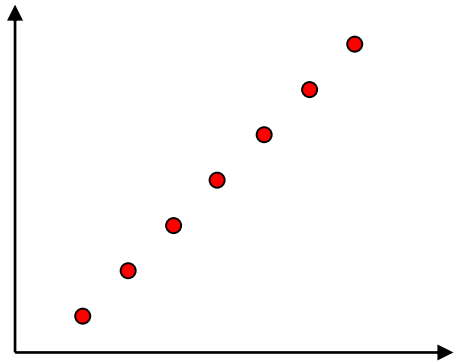
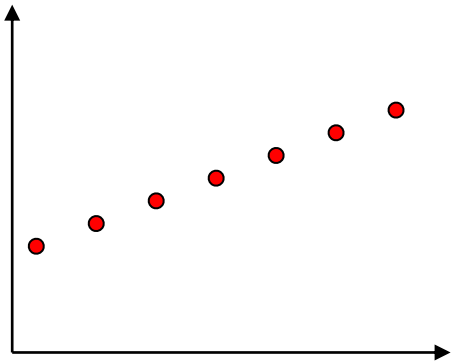
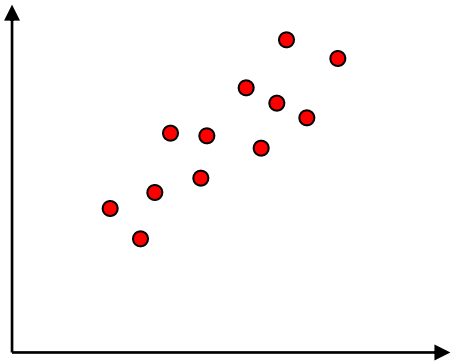
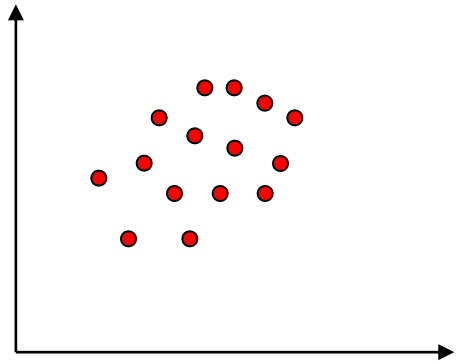
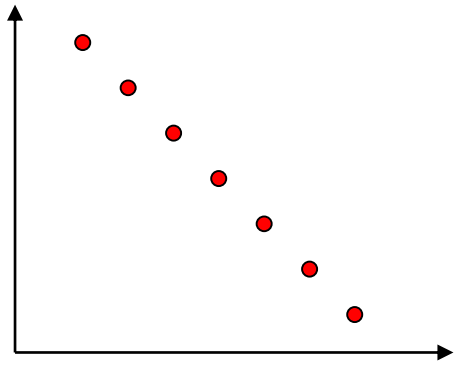
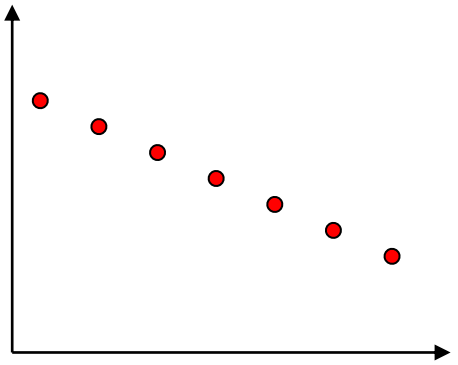
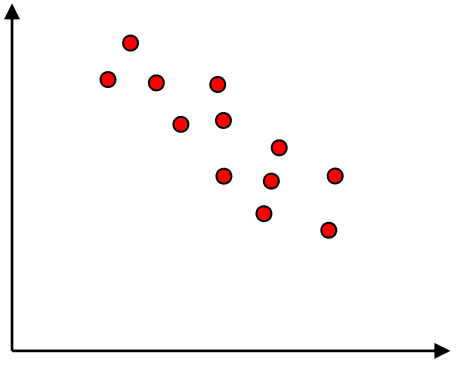
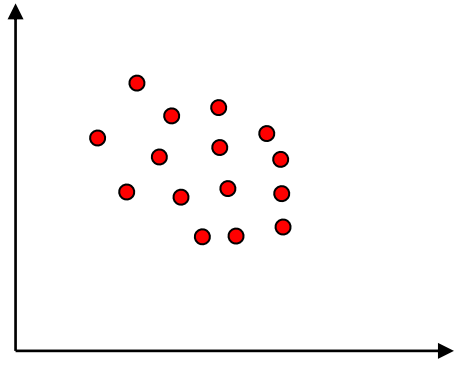
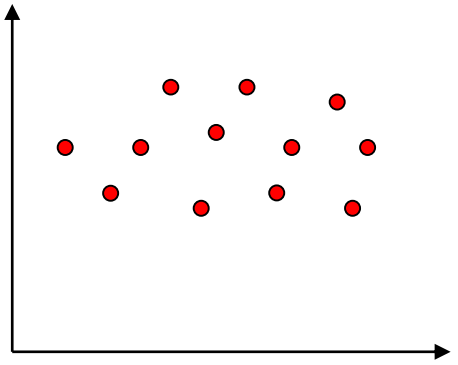
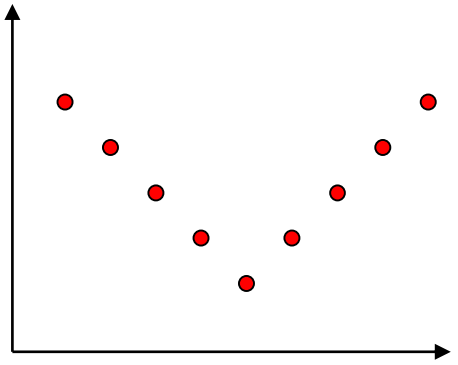
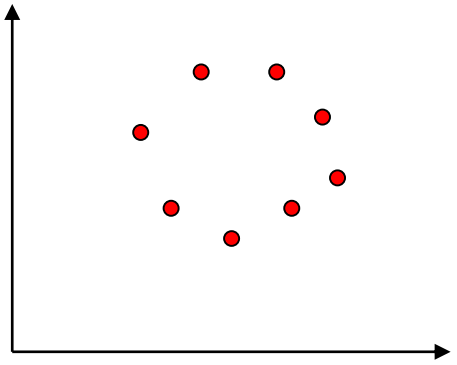
ΓΡΑΜΜΙΚΗ ΣΧΕΣΗ

ΤΩΝ ΔΥΟ ΜΕΤΑΒΛΗΤΩΝ

Απλή γραμμική εξάρτηση (simple linear regression)

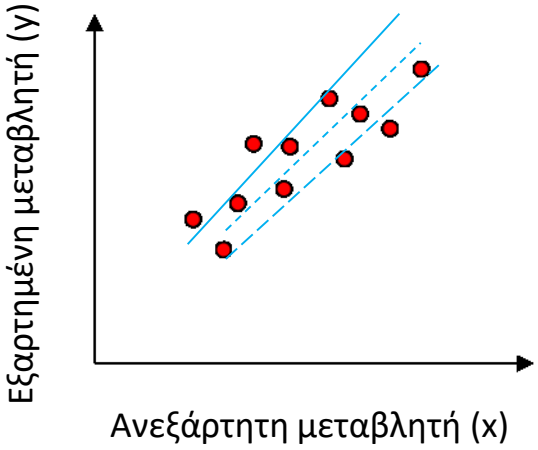
- Στην απλή εξάρτηση διερευνάται η **ΓΡΑΜΜΙΚΗ** σχέση μιας εξαρτημένης μεταβλητής με μία μόνο ανεξάρτητη μεταβλητή.
- Γραφική αναπαράσταση: Στικτόγραμμα (Scatter plot)



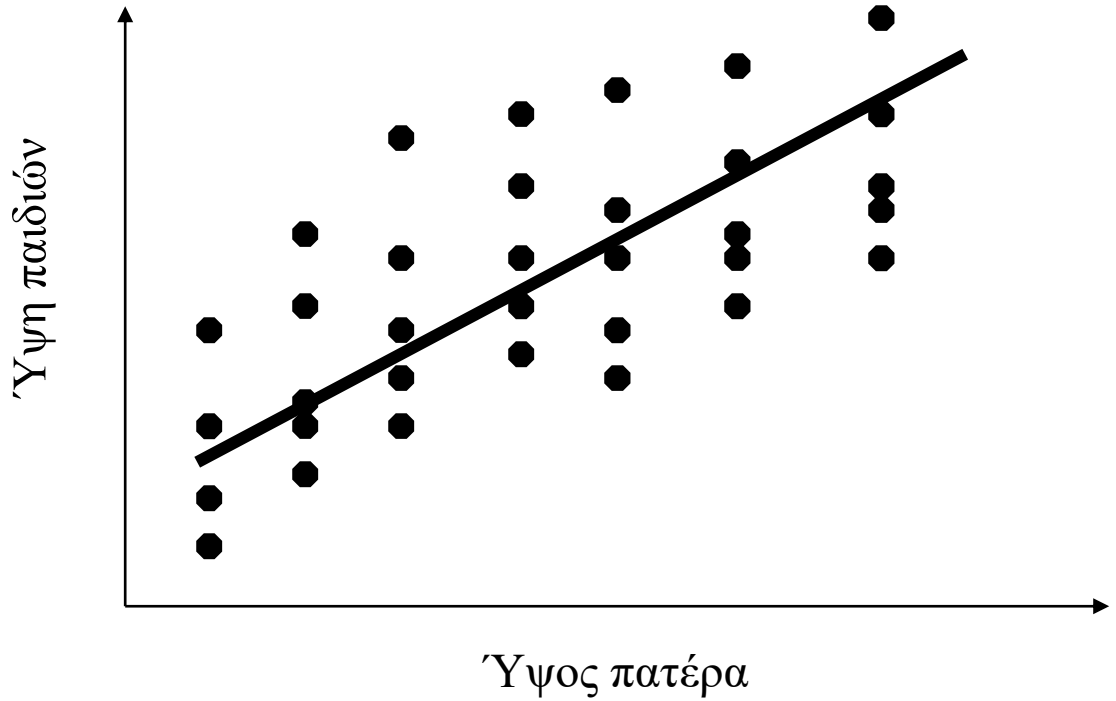
Απλή γραμμική εξάρτηση (simple linear regression)

- Ποια είναι η λογική της;
  - Προσπαθούμε να εκτιμήσουμε την ευθεία που χαρακτηρίζει «καλύτερα» τη σχέση μεταξύ της εξαρτημένης και ανεξάρτητης μεταβλητής
  - Με βάση αυτή τη γραμμή μπορούμε να βρούμε κάθε τιμή της εξαρτημένης που αντιστοιχεί σε συγκεκριμένη τιμή της ανεξάρτητης.



Προϋποθέσεις γραμμικής εξάρτησης

1. Γραμμική σχέση



2. Ομοιόμορφη διασπορά παρατηρήσεων γύρω από τη γραμμή
3. Κανονική κατανομή εξαρτημένης

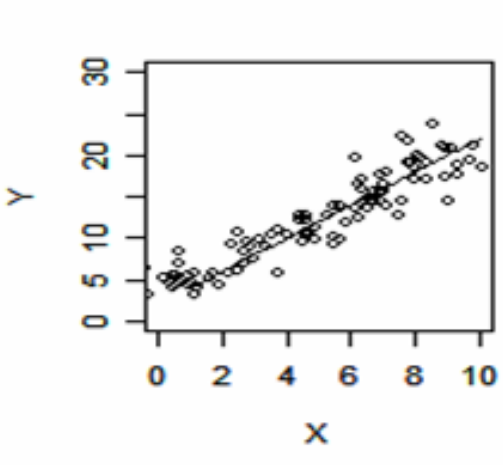
Η εκτίμηση γίνεται ώστε να είναι βέλτιστη γραμμική σχέση μεταξύ των δυο παραμέτρων.

ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

- Μέθοδος για την διερεύνηση των μεταβολών των τιμών μιας μεταβλητής (*εξαρτημένη*) συναρτήσει των μεταβολών των τιμών της άλλης (*ανεξάρτητη*).

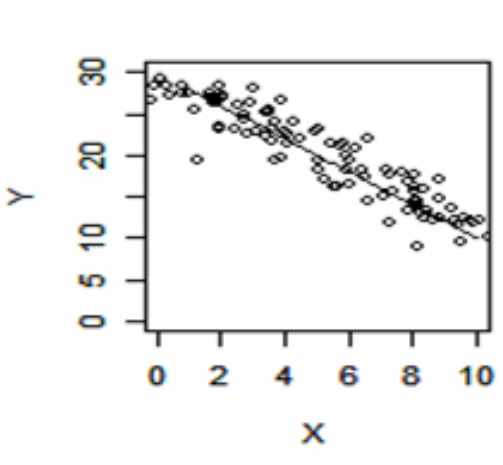
  «Η διερεύνηση γραμμικής σχέσης εξάρτησης μεταξύ 2 μεταβλητών, εκ των οποίων η μια καλείται εξαρτημένη και η άλλη ανεξάρτητη».

- Δηλαδή, Υ (εξαρτημένη) από την Χ (ανεξάρτητη), συνδέονται με τη σχέση:



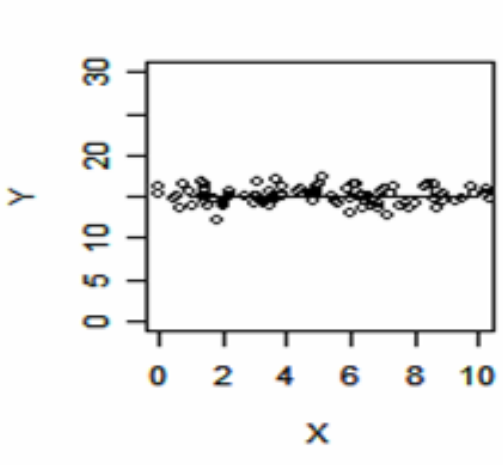
b>0

θετική σχέση



b<0

αρνητική σχέση



b=0

απουσία σχέσης

Ερμηνεία

- Ο b εκφράζει τη **κατά μέσο όρο** μεταβολή (αύξηση ή μείωση ανάλογα με το πρόσημο) της εξαρτημένης μεταβλητής (Y) όταν η ανεξάρτητη (X) μεταβληθεί (αυξηθεί) κατά μία μονάδα.
- Ο συντελεστής εξάρτησης b μπορεί να είναι αρνητικός (αρνητική εξάρτηση) ή θετικός αριθμός (θετική εξάρτηση) ή να ισούται προς 0 (απουσία εξάρτησης).
- Έχει σαν μονάδες το λόγο των μονάδων της εξαρτημένης προς τις **μονάδες** της ανεξάρτητης μεταβλητής (μονάδες Y ανά μονάδες X)

- Ο a εκφράζει την αναμενόμενη τιμή της εξαρτημένης μεταβλητής (Y) όταν η ανεξάρτητη (X) παίρνει την τιμή **0**. (Y= a+b*0= a)
  - *Δεν εμπλέκεται στην εξάρτηση της Y με τη X.*
  - *Χρησιμοποιείται όταν θέλουμε να υπολογίσουμε την αναμενόμενη τιμή της Y για συγκεκριμένη τιμή της X.*

Συντελεστής εξάρτησης b

Η Υ (εξαρτημένη) από την Χ (ανεξάρτητη), συνδέονται με τη σχέση: Y = a + b* X

$\hat{Y}_1$: η αναμενόμενη τιμή της y για x=X_1

$\hat{Y}_2$: η αναμενόμενη τιμή της y για x=X_1+1

$$\hat{Y}_1 = a + b * X_1$$

$$\hat{Y}_2 = a + b * (X_1 + 1)$$

Πόσο διαφέρουν οι Y_1 και Y_2, όταν η ανεξάρτητη μεταβλητή αυξηθεί κατά 1;

$$\hat{Y}_2 - \hat{Y}_1 = [a + b(X_1+1)] - [a + bX_1]$$

$$\hat{Y}_2 - \hat{Y}_1 = [a + bX_1 + b] - [a + bX_1]$$

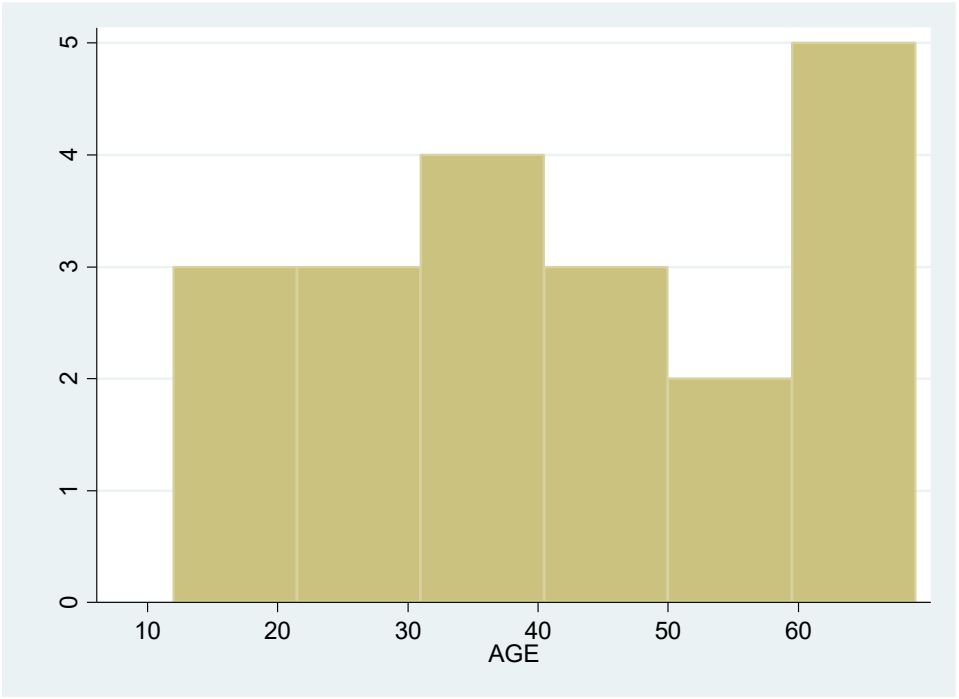
$$\hat{Y}_2 - \hat{Y}_1 = \cancel{a} + \cancel{bX_1} + b - \cancel{a} - \cancel{bX_1}$$

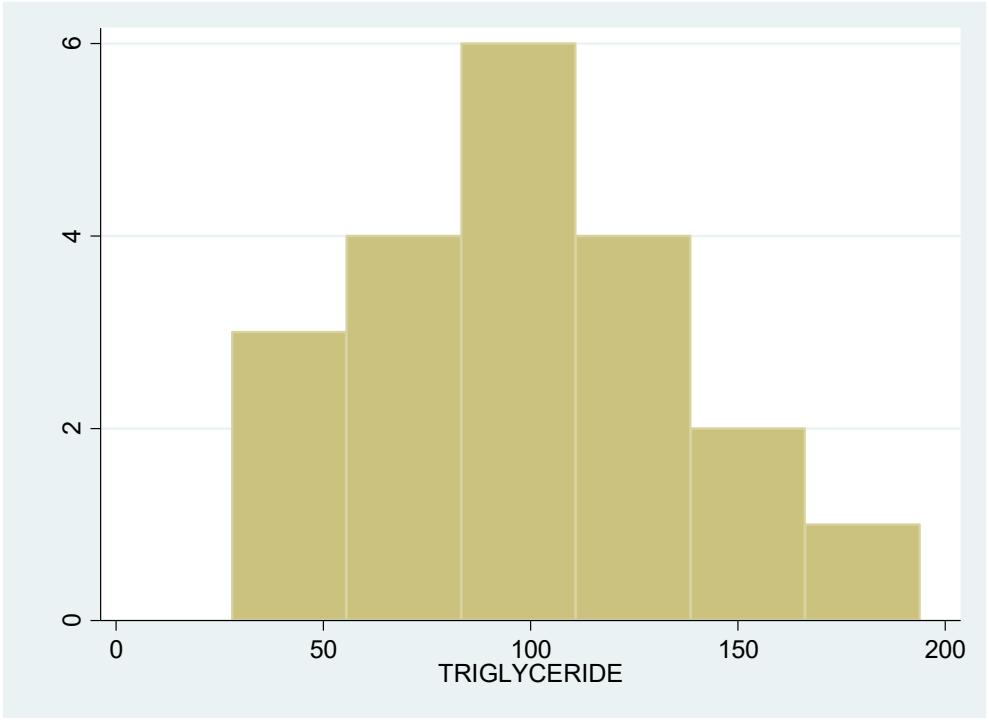
$$\hat{Y}_2 - \hat{Y}_1 = b$$

Ηλικία	Τριγλυκερίδια ορού (mg/100ml)
12	28
12	52
18	106
24	87
26	90
27	67
33	99
35	80
38	130
40	50
44	83
46	95
48	111
51	124
57	83
62	119
63	194
67	165
68	152
69	91

Παράδειγμα

- Ηλικία και τριγλυκερίδια σε δείγμα 20 υγιών ανδρών
- Ποια η σχέση μεταξύ ηλικίας και τριγλυκεριδίων;


0
1
2
3
4
5
10
20
30
40
50
60
AGE


0
2
4
6
0
50
100
150
200
TRIGLYCERIDE

Ποια θα θεωρηθεί εξαρτημένη μεταβλητή και ποια ανεξάρτητη;

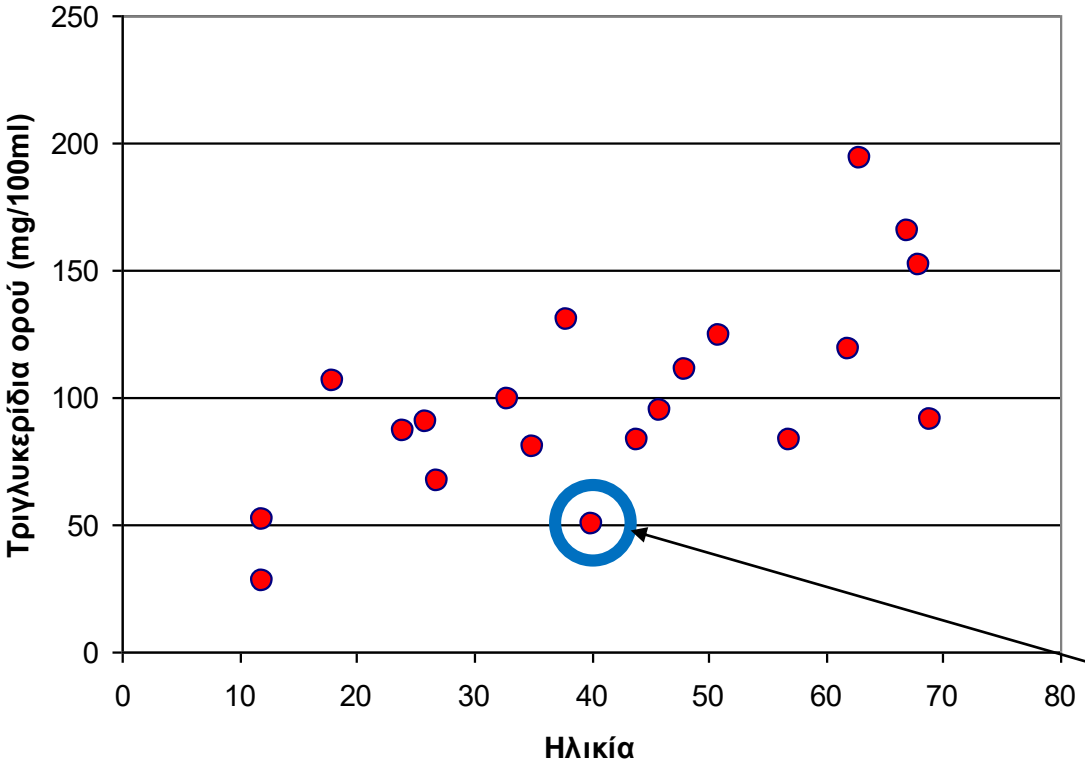
- Εννοιολογικά: τριγλυκερίδια εξαρτώνται από την ηλικία και όχι η ηλικία από τα τριγλυκερίδια

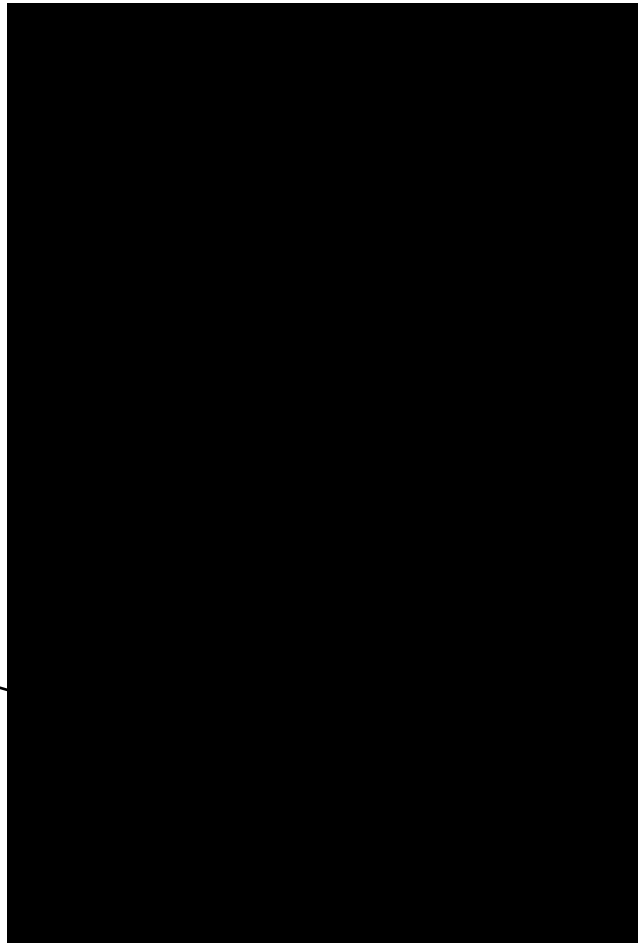
- Με βάση την κατανομή τους: η ηλικία δεν έχει κανονική κατανομή ενώ τα τριγλυκερίδια έχουν κατά προσέγγιση κανονική κατανομή

→ Εξαρτημένη (Υ) : ΤΡΙΓΛΥΚΕΡΙΔΙΑ

→ Ανεξάρτητη (Χ): ΗΛΙΚΙΑ

Scatter plot (στικτόγραμμα) ηλικίας (Χ) και τριγλυκεριδίων (Υ)





- Τι τιμή αναμένετε να έχει ο b;
  - Περίπου 0, >0 ή <0; → αναμένουμε b>0

Απλή γραμμική παλινδρόμηση (εξάρτηση) για τη διερεύνηση της σχέσης ηλικίας και τριγλυκεριδίων :

Y= a+b*X => **Τριγλυκερίδια= a + b * Ηλικία**

$$b = +1.466 \ \text{mgr/100 ml/year}$$

Μονάδες Y Μονάδες X

- Όταν η ηλικία αυξηθεί κατά ένα έτος η τιμή των τριγλυκεριδίων του ορού αναμένεται να αυξηθεί **κατά μέσο όρο** κατά 1,466 mgr/100ml

- Εναλλακτικά, σε άτομα που η ηλικία τους διαφέρει κατά 1 έτος, η τιμή των τριγλυκεριδίων του ορού αναμένεται να διαφέρει κατά μέσο όρο κατά 1,466 mgr/ml

- **Είναι όμως σημαντική η σχέση αυτή?**

H_0: β=0
H_1: β≠ 0

Στατιστική αξιολόγηση του b

Η ποσότητα $\frac{b}{SE_b}$ ακολουθεί την t κατανομή με n-2 ΒΕ

Ελέγχουμε την τιμή στον πίνακα με τις οριακές τιμές (συνήθως με την τιμή που αντιστοιχεί στο 5% επίπεδο σημαντικότητας) σε n-2 βαθμούς ελευθερίας

Αν η ποσότητα $\frac{b}{SE_b}$ (κατά απόλυτη τιμή) είναι μεγαλύτερη από την οριακή τιμή του πίνακα, τότε απορρίπτουμε την Η0.

95% όρια αξιοπιστίας (CI) του b

- Τα **95% CI** του συντελεστή εξάρτησης b υπολογίζονται από τον τύπο:

$$b \pm t_{0.05,(n-2)} SE(b)$$

- Τα 95% CI μας επιτρέπουν να αξιολογήσουμε αν ο b διαφέρει στατιστικά σημαντικά από το 0 (αν υπάρχει γραμμική σχέση)
  - Αν τα όρια δεν περιλαμβάνουν το 0 → διαφέρει στατιστικά σημαντικά π.χ. (-3, -1) ή (0.25, 0.81) (ύπαρξη γραμμικής σχέσης θετικής ή αρνητικής ανάλογα το πρόσημο του b)
  - Αν τα όρια περιλαμβάνουν το 0 → μη στατιστικά σημαντική σχέση π.χ. (-0.92, 1,12) (απουσία γραμμικής σχέσης)

Αξιολόγηση του b στο παράδειγμα ηλικίας-τριγλυκεριδίων (B)

- b=1,466 και SE(b)=0,374
- 95% CI είναι:

$$1{,}466 \pm 2{,}10 * 0{,}374 \rightarrow (0{,}681 - 2{,}251)$$

- Τα όρια δεν περιλαμβάνουν το 0 → στατιστικά σημαντική σχέση ηλικίας-τριγλυκεριδίων στο 5% επίπεδο σημαντικότητας

p<0,001 **Στατιστικά πολύ σημαντική εξάρτηση**

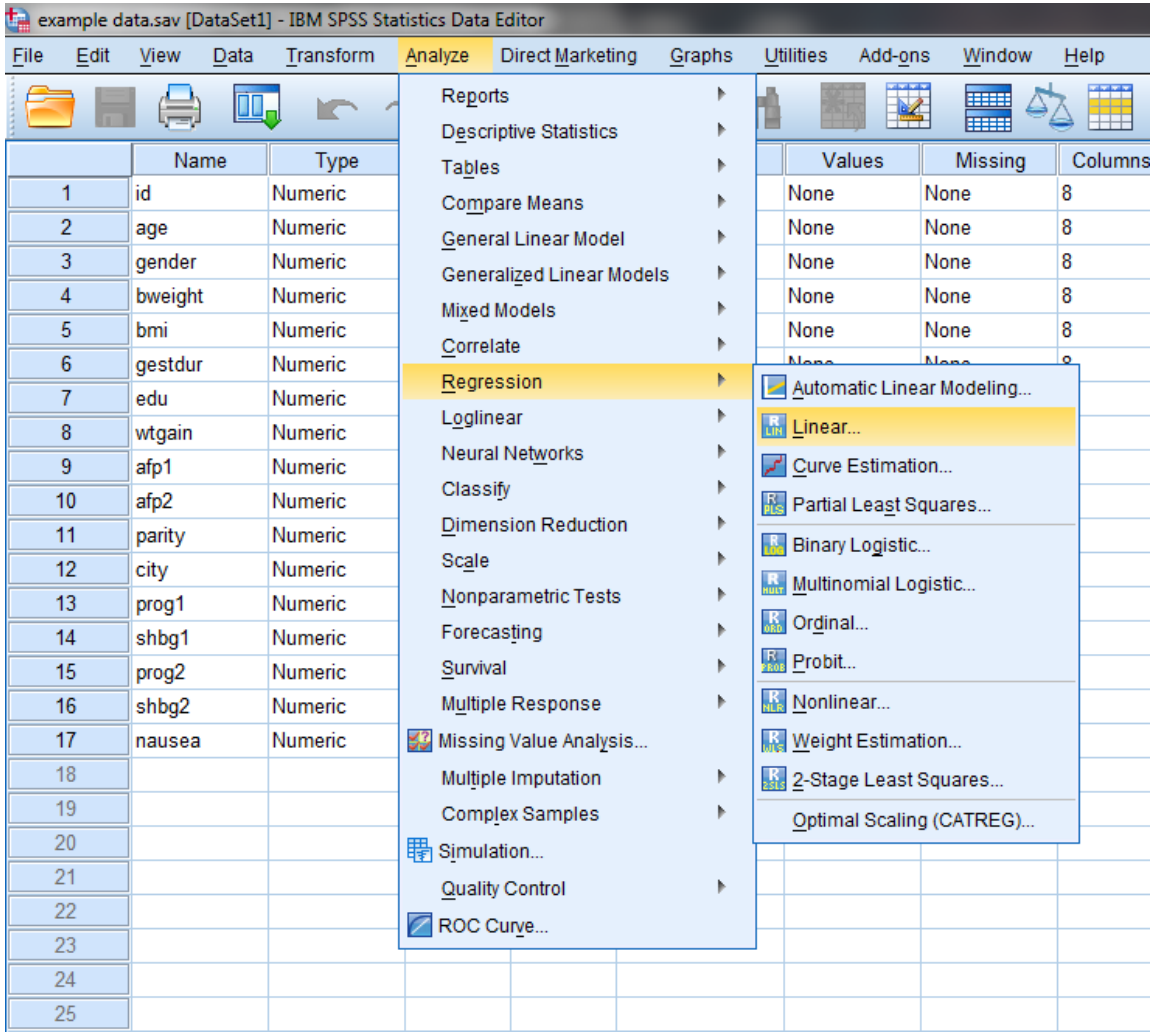
Ερμηνεία

- Όταν η ηλικία αυξηθεί κατά 1 έτος η τιμή των τριγλυκεριδίων του ορού αναμένεται να αυξηθεί **κατά μέσο όρο** κατά 1,466 mgr/100 ml (ενώ με 95% πιθανότητα η αύξηση αυτή μπορεί να κυμαίνεται από 0,68 μέχρι 2,25 mgr/100 ml)

- **Ανά δεκαετία:** b=1,466*10=14,66 mgr/100 ml/έτος

  95% CI: 6,81-22,51 → Σε άτομα που η ηλικία τους διαφέρει κατά 10 έτη, η τιμή των τριγλυκεριδίων του ορού αναμένεται να διαφέρει κατά μέσο όρο κατά 14,66 χιλιοστόγραμμα ανά 100 χιλιοστόλιτρα

Παράδειγμα SPSS: εξάρτηση βάρους γέννησης από διάρκεια κύησης



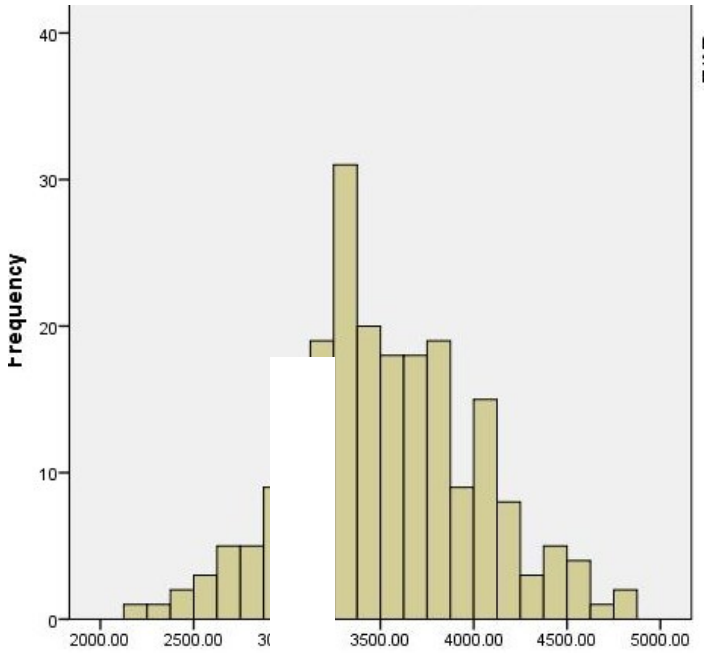
ΒΗΜΑΤΑ

- Επιλογή στατιστικής δοκιμασίας ανάλογα το είδος των μεταβλητών

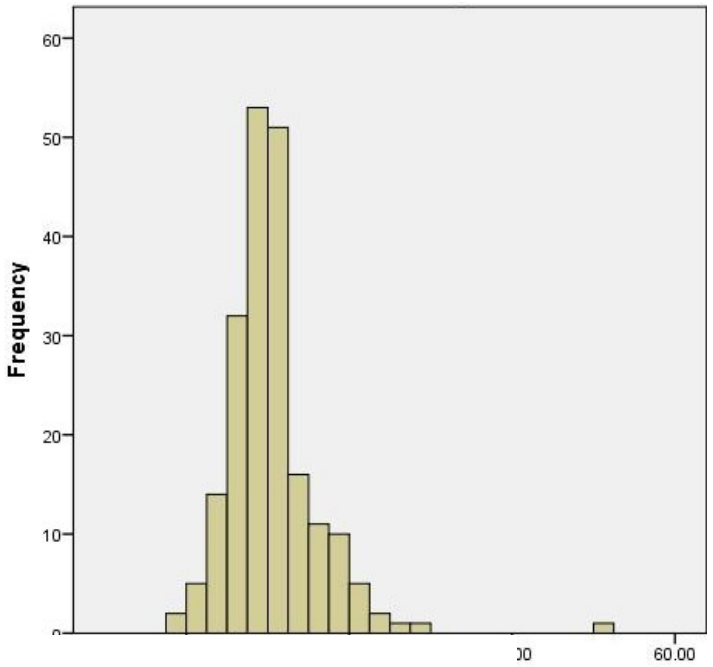
Στην διερεύνηση της εξάρτησης 2 ποσοτικών μεταβλητών:

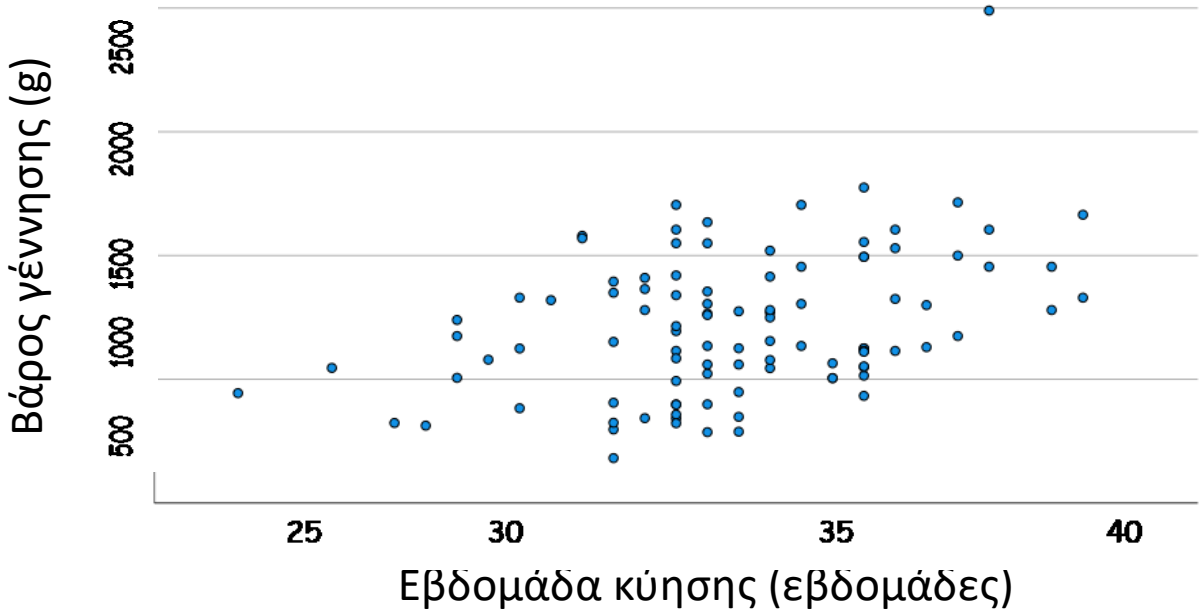
- Ορισμός ανεξάρτητης και εξαρτημένης μεταβλητής
- Έλεγχος κανονικότητας ΕΞΑΡΤΗΜΕΝΗΣ μεταβλητής
- Στικτόγραμμα για την ύπαρξη η μη γραμμικής σχέσης

Βάρος γέννησης



Εβδομάδα κύησης





Regression

[DataSet1] C:\WORK\ASKISEIS\Msc Εντατική\example data.sav

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	gestdur[b]	.	Enter

a. Dependent Variable: bweight

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.291[a]	.084	.080	470.09569

a. Predictors: (Constant), gestdur

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4321727.585	1	4321727.585	19.556	.000[b]
	Residual	46849870.78	212	220989.956		
	Total	51171598.36	213			

a. Dependent Variable: bweight

b. Predictors: (Constant), gestdur

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-1444.678	1120.906		-1.289	.199	-3654.227	764.871
	gestdur	123.907	28.019	.291	4.422	.000	68.675	179.138

a. Dependent Variable: bweight

Η εκτίμηση της σταθεράς a

Τυπικό σφάλμα της εκτίμησης του b

Στήλη με p-value

Η εκτίμηση του συντελεστή εξάρτησης b

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	-1444.678	1120.906		-1.289	.199	-3654.227	764.871
	gestdur	123.907	28.019	.291	4.422	.000	68.675	179.138

a. Dependent Variable: bweight

95% διάστημα εμπιστοσύνης

Ερμηνεία συντελεστή εξάρτησης και 95% διαστήματος εμπιστοσύνης:

Αν η διάρκεια κύησης (ανεξάρτητη) αυξηθεί κατά 1 εβδομάδα (μονάδα μέτρησης της ανεξάρτητης), αναμένουμε το βάρος του νεογνού (εξαρτημένη) να αυξηθεί (θετικός συντελεστής) κατά 123,9 (εκτίμηση συντελεστή εξάρτησης) γραμμάρια (μονάδα μέτρησης εξαρτημένης) **κατά μέσο όρο** (ενώ με 95% πιθανότητα η αύξηση αυτή μπορεί να κυμαίνεται από 68,7 μέχρι 179,1 γραμμάρια).

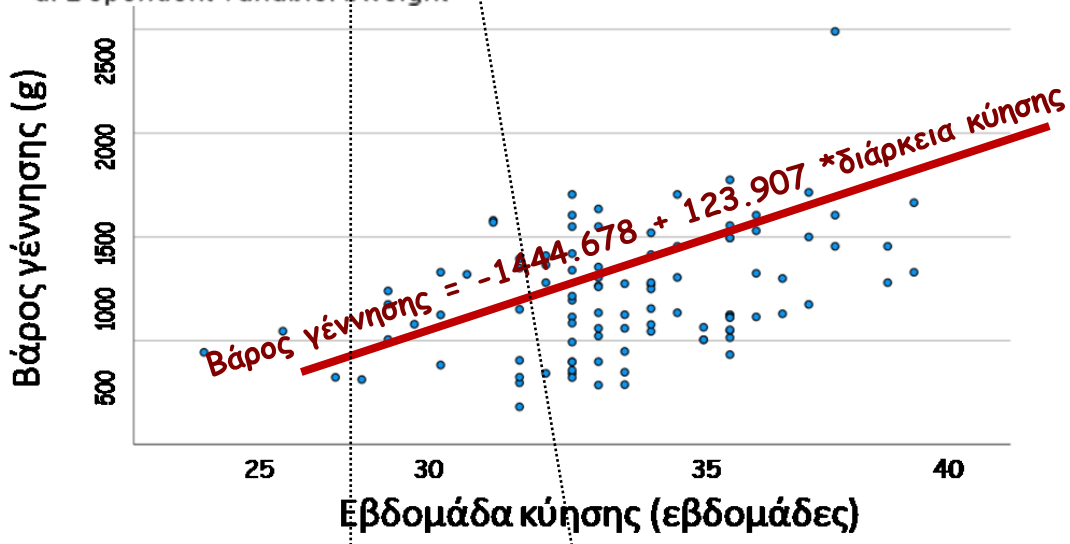
Αξιολόγηση στατιστικής σημαντικότητας:

Ο συντελεστής εξάρτησης διαφέρει στατιστικά σημαντικά από το 0 σε επίπεδο 5% εφόσον p-value< 0.05 και το 0 δεν περιλαμβάνεται στο 95% διάστημα αξιοπιστίας (και **οι 2 τρόποι αξιολόγησης** καταλήγουν στο ίδιο συμπέρασμα) -> στατιστικά σημαντική σχέση διάρκειας κύησης-βάρους γέννησης νεογνού

Εύρεση αναμενόμενης τιμής του βάρους γέννησης (Y) για συγκεκριμένη διάρκεια κύησης (X).

Model		Unstandardized Coefficients		95.0% Confidence Interval for B	
		B	Std. Error	Lower Bound	Upper Bound
1	(Constant)	-1444.678	1120.906	-3654.227	764.871
	gestdur	123.907	28.019	68.675	179.138

a. Dependent Variable: bweight



Η ευθεία που περιγράφει καλύτερη την σχέση βάρους γέννησης και διάρκειας κύησης είναι:
Y= a + bX =>
Βάρος γέννησης = -1444.678 + 123.907 *διάρκεια κύησης (η κόκκινη ευθεία στο στικτόγραμμα)

Τι βάρος γέννησης θα έχει κατά μέσο όρο ένα νεογνό αν η διάρκεια κύησης είναι 30 εβδομάδες;

Βάρος γέννησης = -1444.678 + 123.907*30
Βάρος γέννησης = 2272.532 γραμμάρια

Εύρεση 95% ΔΕ της αναμενόμενης τιμής του βάρους γέννησης (Y) για συγκεκριμένη διάρκεια κύησης (X).

Model		Unstandardized Coefficients		95.0% Confidence Interval for B	
		B	Std. Error	Lower Bound	Upper Bound
1	(Constant)	-1444.678	1120.906	-3654.227	764.871
	gestdur	123.907	28.019	68.675	179.138

a. Dependent Variable: bweight

Για να βρω το 95% ΔΕ της αναμενόμενης τιμής του βάρους γέννησης αν η διάρκεια κύησης είναι 30 εβδομάδες, στην ευθεία δε θα βάλω την ΕΚΤΙΜΗΣΗ του b αλλά **τα 95% όρια εμπιστοσύνης της εκτίμησης του Β.**

Κάτω όριο:

Βάρος γέννησης = -1444.678 + **68.675***30
Βάρος γέννησης = 615.572 γραμμάρια

Άνω όριο:

Βάρος γέννησης = -1444.678 + **179.871***30
Βάρος γέννησης = 3951.452 γραμμάρια

ΑΡΑ το αναμενόμενο βάρος γέννησης νεογνού για διάρκεια κύησης 30 εβδομάδων είναι 2272.532 γραμμάρια, ενώ με 95% πιθανότητα το βάρος γέννησης θα κυμαίνεται από 615.572 γραμ. έως 3951.452 γραμ.

Εύρεση αναμενόμενης διαφοράς βάρους γέννησης (Y) για συγκεκριμένη διαφορά στη διάρκειες κύησης (X).

Model		Unstandardized Coefficients		95.0% Confidence Interval for B	
		B	Std. Error	Lower Bound	Upper Bound
1	(Constant)	-1444.678	1120.906	-3654.227	764.871
	gestdur	123.907	28.019	68.675	179.138

a. Dependent Variable: bweight

Πόσο διαφέρει κατά μέσο όρο το βάρος γέννησης (Y) αν η διάρκεια κύησης (X) διαφέρει κατά 1 εβδομάδα?

Κατά b= 123.907 γραμμάρια με 95% ΔΕ της διαφοράς: (68.675, 179.138)

Πόσο διαφέρει κατά μέσο όρο το βάρος γέννησης (Y) αν η διάρκεια κύησης (X) διαφέρει κατά 2 εβδομάδες?

Κατά 2b= 2*123.907=247.814 γραμμάρια με 95% ΔΕ της διαφοράς: (2*68.675, 2*179.138) =(137.35, 358.276)

Πόσο διαφέρει κατά μέσο όρο το βάρος γέννησης (Y) αν η διάρκεια κύησης (X) διαφέρει κατά 6 εβδομάδες?

Κατά 6b= 6*123.907=743.442 γραμμάρια με 95% ΔΕ της διαφοράς: (6*68.675, 6*179.138)= (412.05, 1074.828)

Παράδειγμα SPSS: εξάρτηση βάρους γέννησης από βάρος που πήρε η γυναίκα κατά τη διάρκεια της εγκυμοσύνης

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3313.770	54.295		61.033	.000	3206.892	3420.648
	wtgain	19.630	4.526	.251	4.337	.000	10.721	28.540

a. Dependent Variable: bweight

Εξαρτημένη: βάρος γέννησης νεογνού (γραμμάρια),
Ανεξάρτητη: βάρος μητέρας (κιλά)

Εκτίμηση b: 19,6 γραμμάρια ανά κιλό (*μονάδες εξαρτημένης ανά μονάδα ανεξάρτητης*)

Ερμηνεία: Για αύξηση του βάρους της μητέρας κατά 1 κιλό, αναμένουμε το βάρος του νεογνού να αυξηθεί κατά 19,6 γραμμάρια **κατά μέσο όρο** (με 95% πιθανότητα η αύξηση αυτή να κυμαίνεται από 10,7 έως 28,5 γραμμάρια)

Στατιστικά σημαντική σχέση βάρους μητέρας-βάρους γέννησης νεογνού (p-value<0,05 και το 0 δεν περιλαμβάνεται στα 95% όρια αξιοπιστίας)