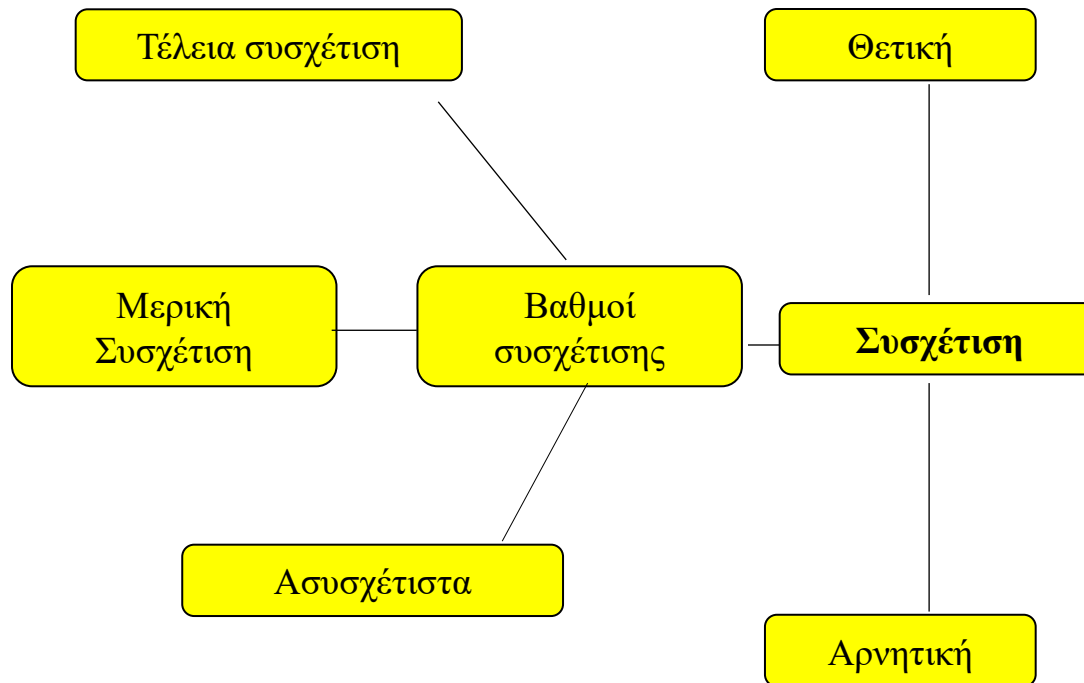

ΣΥΣΧΕΤΙΣΗ



ΣΥΣΧΕΤΙΣΗ

Ορισμός:

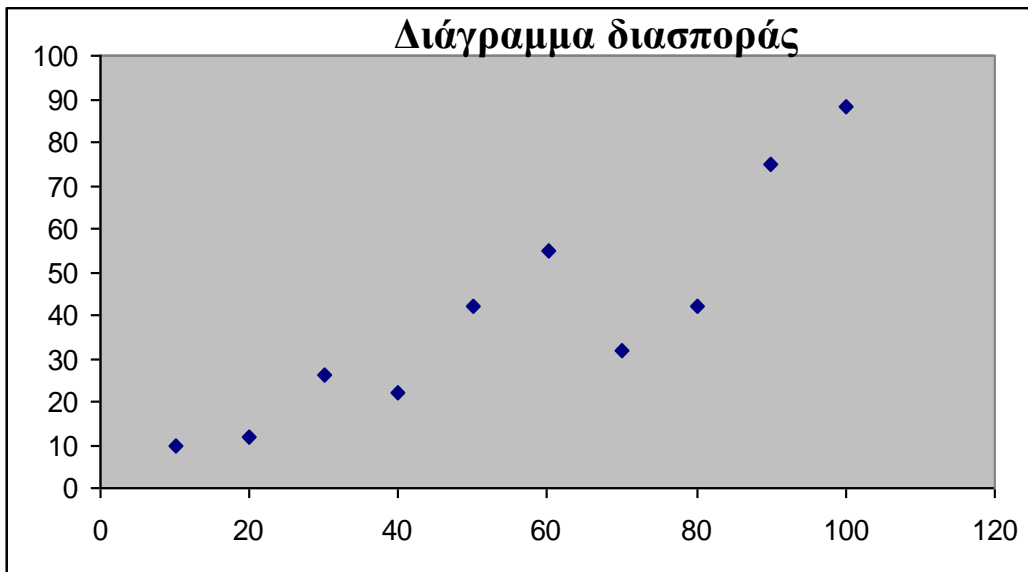
Συσχέτιση σημαίνει μια σχέση ή «ανταπόκριση» ανάμεσα σε δύο μεταβλητές

Συσχέτιση αναφέρεται στην συν-μεταβλητότητα δύο μεταβλητών που συμβολίζονται x και y ,

για παράδειγμα

- ❖ ανάμεσα στο υψος και στο βάρος μια ομάδας ανθρώπων (ενος πληθυσμού)
- ❖ ανάμεσα στις πωλήσεις και στην διαφημιστική δαπάνη ενός προϊόντος
- ❖ ανάμεσα στην θερμοκρασία που παρατηρείται σε μια πόλη και στην καταναλωση της ηλεκτρικής ενέργειας (για μια μακρά χρονική περίοδο)

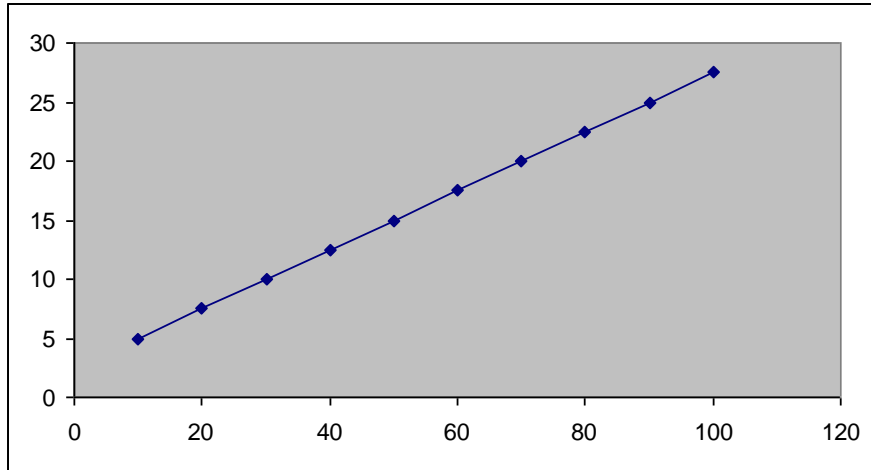
αυτές οι μεταβλητές λαμβάνονται σε ζευγάρια μετρήσεων (παρατηρήσεων) σε μια λογική σειρά στην πορεία του χρόνου



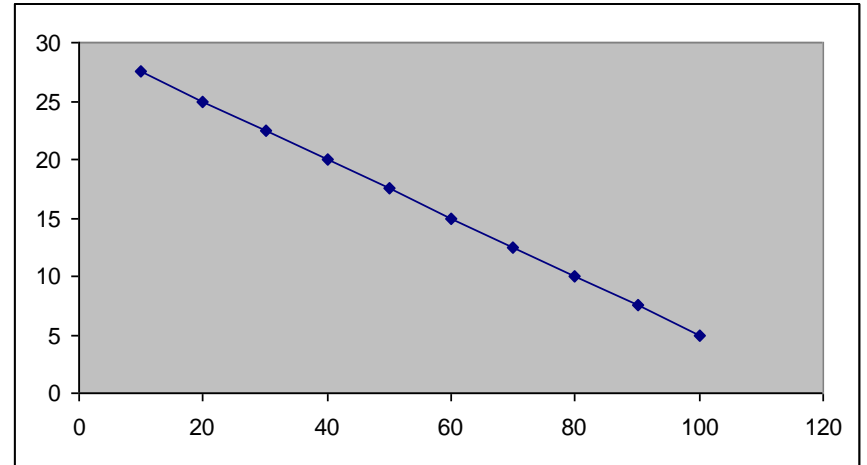
Μηνιαίες πωλήσεις ενός προϊόντος έναντι διαφημιστικής δαπάνης για αυτό το προϊόν

ΣΥΣΧΕΤΙΣΗ – Βαθμοί Συσχέτισης

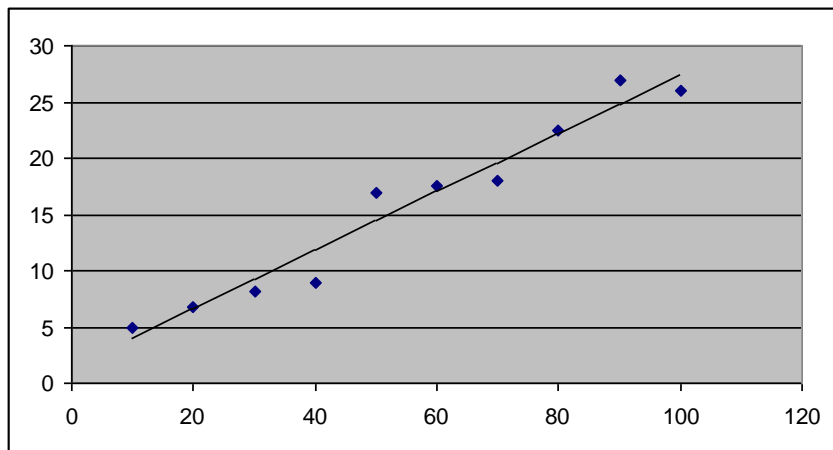
Τέλεια θετική συσχέτιση



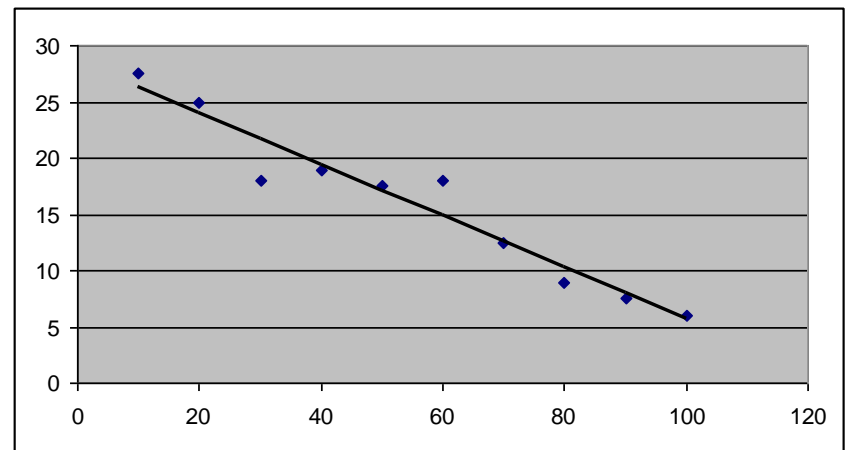
Τέλεια αρνητική συσχέτιση



Μερική θετική συσχέτιση



Μερική αρνητική συσχέτιση

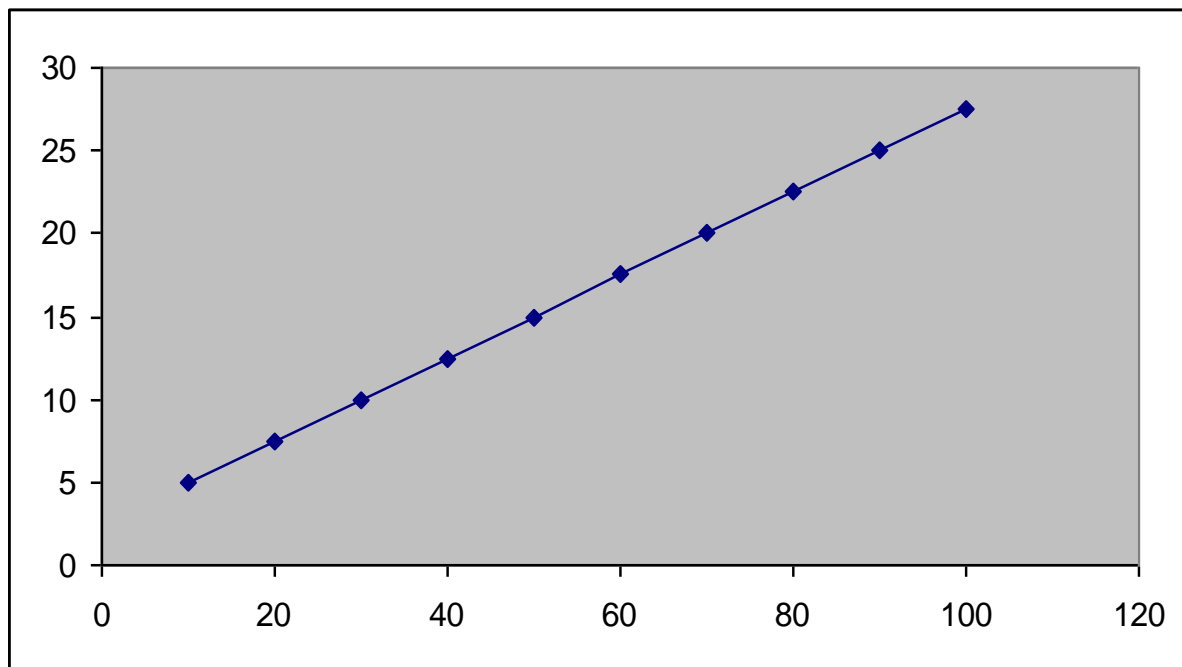


ΣΥΣΧΕΤΙΣΗ – Βαθμοί Συσχέτισης

Τέλεια θετική συσχέτιση

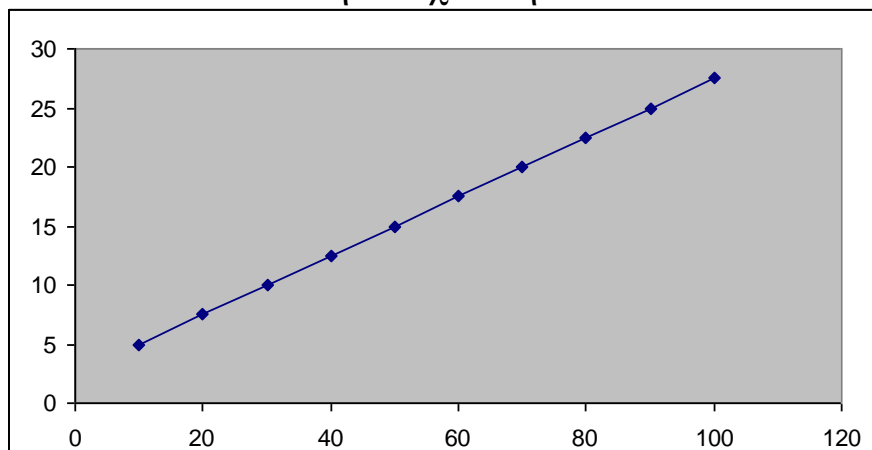
Τέλεια Θετική Συσχέτιση	
x	y
10	5
20	7,5
30	10
40	12,5
50	15
60	17,5
70	20
80	22,5
90	25
100	27,5

Τέλεια θετική συσχέτιση

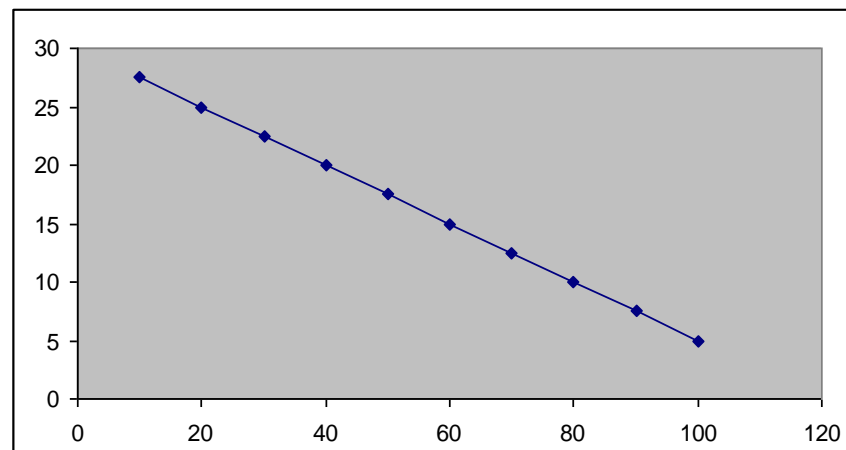


ΣΥΣΧΕΤΙΣΗ – Βαθμοί Συσχέτισης

Τέλεια θετική συσχέτιση



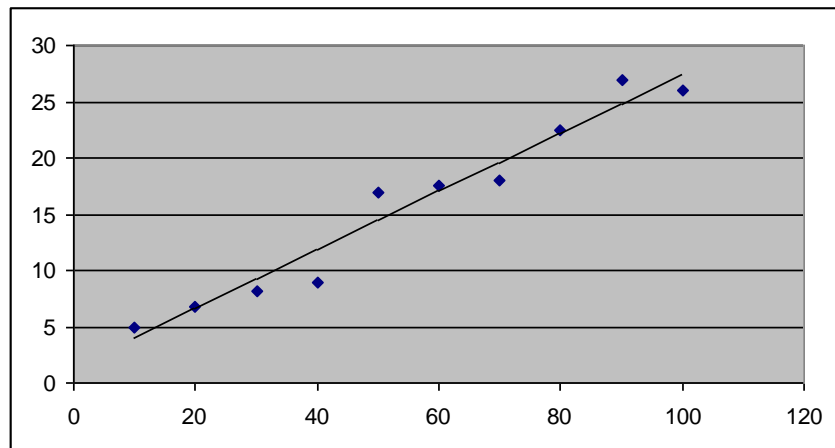
Τέλεια αρνητική συσχέτιση



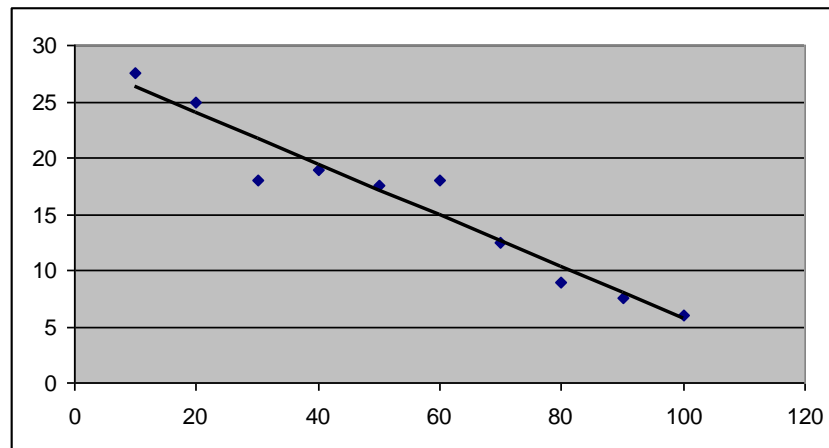
Θετική συσχέτιση σημαίνει ότι χαμηλές αυξήσεις (ή μειώσεις) στις τιμές της μιας μεταβλητής συνδέονται με χαμηλές αυξήσεις (ή μειώσεις) στις τιμές της άλλης και υψηλές αυξήσεις (ή μειώσεις) στις τιμές της μίας μεταβλητής συνδέονται με υψηλές αυξήσεις (ή μειώσεις) στις τιμές της άλλης - όσο πιο παρόμοιες είναι οι παράλληλες αυξήσεις (ή μειώσεις) τόσο πιο μεγάλη είναι η συσχέτιση

Στην **Αρνητική συσχέτιση** συμβαίνει το αντίστροφο ... οι αυξήσεις της μιας μεταβλητής συνδέονται με μειώσεις της άλλης

Μερική θετική συσχέτιση



Μερική αρνητική συσχέτιση



ΣΥΣΧΕΤΙΣΗ – Μέτρηση της συσχέτισης

Ορισμοί:

Ο *συντελεστής συσχέτισης* μετράει το βαθμό συσχέτισης ανάμεσα σε δυο μεταβλητές

$$r = \frac{n \sum xy - \sum x \sum y}{\sqrt{[n \sum x^2 - (\sum x)^2] [n \sum y^2 - (\sum y)^2]}}$$

Όπου x και y συμβολίζουν ζεύγη στοιχείων για δύο μεταβλητές x και y

και n είναι το πλήθος των ζευγαριών στοιχείων (το μέγεθος του δείγματος)

ο συντελεστής **r** παίρνει πάντα τιμές στο διάστημα από -1 έως 1

r = +1 σημαίνει ότι οι μεταβλητές συσχετίζονται θετικά τέλεια

r = -1 σημαίνει ότι οι μεταβλητές συσχετίζονται αρνητικά τέλεια

r = 0 σημαίνει ότι οι μεταβλητές είναι ασυσχέτιστες

	x	y	xy	x ²	y ²			
	10	27,5	275	100	756,25	Αριθμητής	$n \sum xy - \sum x \sum y =$	-19.000
	20	25	500	400	625	παρονομαστής	$n \sum x^2 - (\sum x)^2 =$	82500
	30	18	540	900	324	παρονομαστής	$n \sum y^2 - (\sum y)^2 =$	4660
	40	19	760	1600	361	παρονομαστής		384.450.000
	50	17,5	875	2500	306,25	παρονομαστής	sqrt	19607,39656
	60	18	1080	3600	324		r =	-0,96902207
	70	12,5	875	4900	156,25			
	80	9	720	6400	81			
	90	7,5	675	8100	56,25			
	100	6	600	10000	36			
\sum	550	160	6900	38500	3026			
$(\sum)^2$	302500	25600						

ΣΥΣΧΕΤΙΣΗ – Μέτρηση της συσχέτισης

Ορισμός:

Ο συντελεστής προσδιορισμού r^2 or R^2 μετράει το μέρος της συνολικής διακύμανσης της μιας μεταβλητής που μπορεί να ερμηνευθεί από την διακύμανση της άλλης μεταβλητής.

What values can R^2 take ?

	x	y	xy	x^2	y^2			
	10	27,5	275	100	756,25	Αριθμητής	$n\sum xy - \sum x \sum y =$	-19.000
	20	25	500	400	625	Παρονομαστής	$n\sum x^2 - (\sum x)^2 =$	82500
	30	18	540	900	324	Παρονομαστής	$n\sum y^2 - (\sum y)^2 =$	4660
	40	19	760	1600	361	Παρονομαστής		384.450.000
	50	17,5	875	2500	306,25	Παρονομαστής	Τετραγ. Ρίζα (sqrt)	19607,39656
	60	18	1080	3600	324		r =	-0,96902207
	70	12,5	875	4900	156,25		R² =	0,939003772
	80	9	720	6400	81			
	90	7,5	675	8100	56,25			
	100	6	600	10000	36			
\sum	550	160	6900	38500	3026			
$(\sum)^2$	302500	25600						

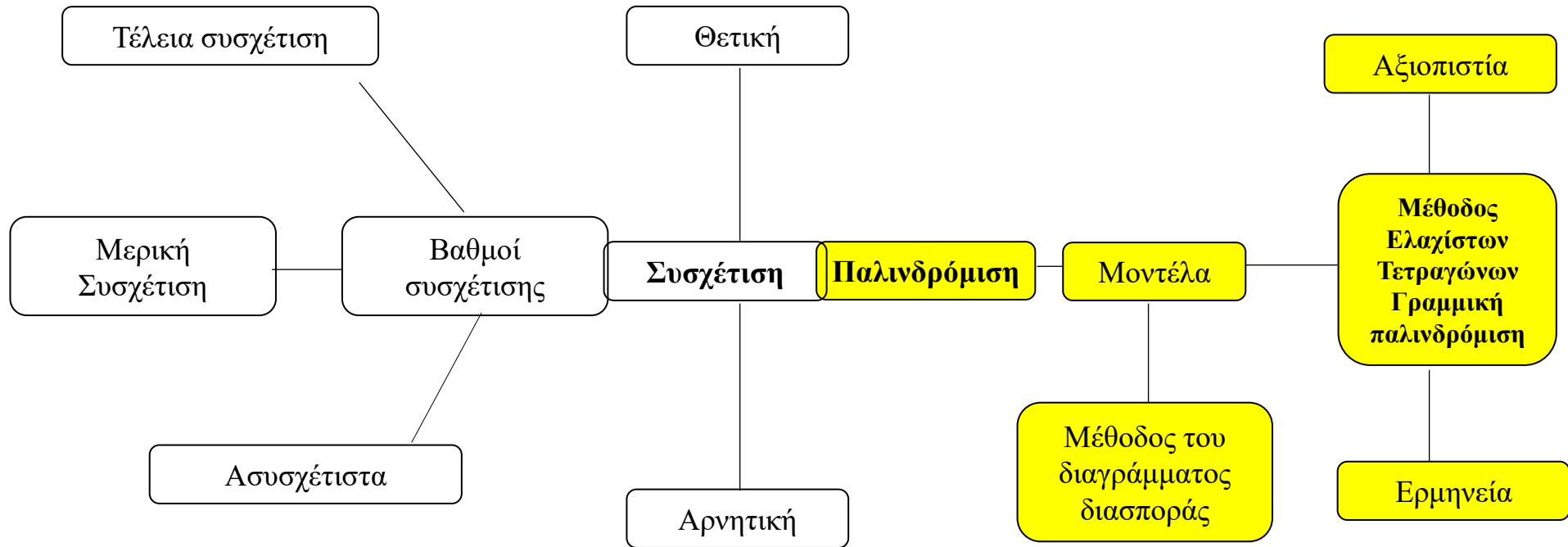
Αυτό σημαίνει ότι το 94% της διακύμανσης της μεταβλητής y μπορεί να ερμηνευθεί από τη διακύμανση της μεταβλητής της x

Ουσιώδη σημεία:

- Εάν δύο μεταβλητές συσχετίζονται έντονα (είτε θετικά είτε αρνητικά) αυτό μπορεί να είναι εντελώς τυχαίο δηλαδή να μην υπάρχει κάποια αιτιώδης ή λογική σχέση μεταξύ τους ή μπορεί και να υπάρχει Εάν το μέγεθος του δείγματος (το πλήθος των ζευγαριών τιμών των x και y) είναι πολύ μεγάλο είναι πιο πιθανό αυτή η συσχέτιση να μην είναι «τυχαία» (δηλαδή υπάρχει κάποια αιτιώδης σχέση ανάμεσά τους)
- Συνήθως οι ερευνητές ελέγχουν την συσχέτιση ανάμεσα σε μεταβλητές που συνδέονται λογικά δηλαδή πιστεύουν (στη βάση κάποιας θεωρίας) ότι **η μεταβλητή x είναι η αιτία των διακυμάνσεων και της συμπεριφοράς της άλλης μεταβλητής της y** . Μπορεί να υπάρχει κάποια “υποστηρικτική” θεωρία για την σχέση αυτών των δύο μεταβλητών.
- Ισχυρή συσχέτιση που μπορεί να υπάρχει ανάμεσα σε δύο μεταβλητές μπορεί να οφείλεται όχι σε άμεση μεταξύ τους σχέση αλλά «μεταβατικά» μέσω μιας *τρίτης μεταβλητής – παράγοντα* η οποία ερμηνεύει τις διακυμάνσεις και των δύο πρώτων
- Μετά από όλα αυτά βεβαίως μέσα από την διακύμανση κανείς μπορεί να «ανακαλύψει» ενδείξεις για την **σχέση – εξάρτηση** ανάμεσα σε δύο μεταβλητές

ΣΥΣΧΕΤΙΣΗ & ΠΑΛΙΝΔΡΟΜΗΣΗ

CORRELATION & REGRESION

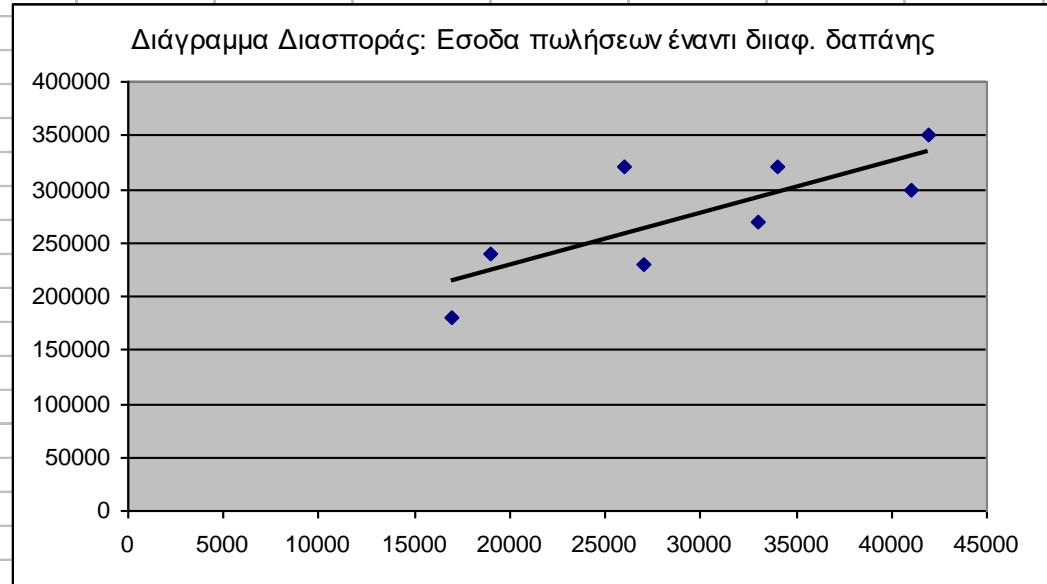


Η συσχέτιση δείχνει το επίπεδο σχέσης ανάμεσα σε δύο μεταβλητές αλλά δεν δείχνει το πώς (τον τρόπο) που αυτή η σχέση συμβαίνει

Αυτό το καλύπτει η παλινδρόμιση

ΣΥΣΧΕΤΙΣΗ – Γραμμές της «καλύτερης εφαρμογής» που αντιπροσωπεύει τη σχέση δύο μεταβλητών εικονιζόμενη στο Διάγραμμα Διασποράς τους

ΕΤΟΣ	Διαφημιστική Δαπάνη	Εσοδα πωλήσεων
1	17000	180000
2	33000	270000
3	34000	320000
4	42000	350000
5	19000	240000
6	41000	300000
7	26000	320000
8	27000	230000



Ένα πρώτο βήμα – προσέγγισης αυτής της σχέσης ανάμεσα στις δύο μεταβλητές είναι να τις απεικονίσουμε σε ένα διάγραμμα διασποράς

1. Οπτικά μπορούμε να αποφανθούμε για ένα *καλό επίπεδο συσχέτισης* και το *σχήμα* αυτής της σχέσης.
2. Με τη σχεδίαση μιας γραμμής (εμπειρικά) ανάμεσα στο νέφος (διασπορά) των σημείων μπορεί κανείς να το *προεκτείνει παραπέρα*.
3. Έτσι για ένα μελλοντικό επίπεδο εξόδων (πέρα από το δείγμα) με την χρήση της γραμμής μπορεί να γίνει μια πρόβλεψη - εκτίμηση των εσόδων πωλήσεων.

ΠΑΛΙΝΔΡΟΜΗΣΗ – Γραμμική Παλινδρόμηση

Ορισμός:

Η μέθοδος των ελαχίστων τετραγώνων της γραμμικής παλινδρόμησης είναι μια πολύ συνηθισμένη τεχνική για την εκτίμηση της εξίσωσης της γραμμής καλύτερης εφαρμογής.

Οι δύο μεταβλητές x και y σχετίζονται μέσω της γραμμικής εξίσωσης:

$$y = a + bx$$

σε αυτή τη περίπτωση η x μεταβλητή είναι η *ανεξάρτητη* ή *ερμηνευτική* μεταβλητή
 y είναι η εξαρτημένη μεταβλητή

a είναι ο σταθερός συντελεστής: $a = (\sum y / n) - (b \sum x / n)$

$$a = \frac{\sum y_i}{n} - b \frac{\sum x_i}{n}$$

b είναι ο συντελεστής κλίσης: $b = [n \sum xy - \sum x \sum y] / [n \sum x^2 - (\sum x)^2]$

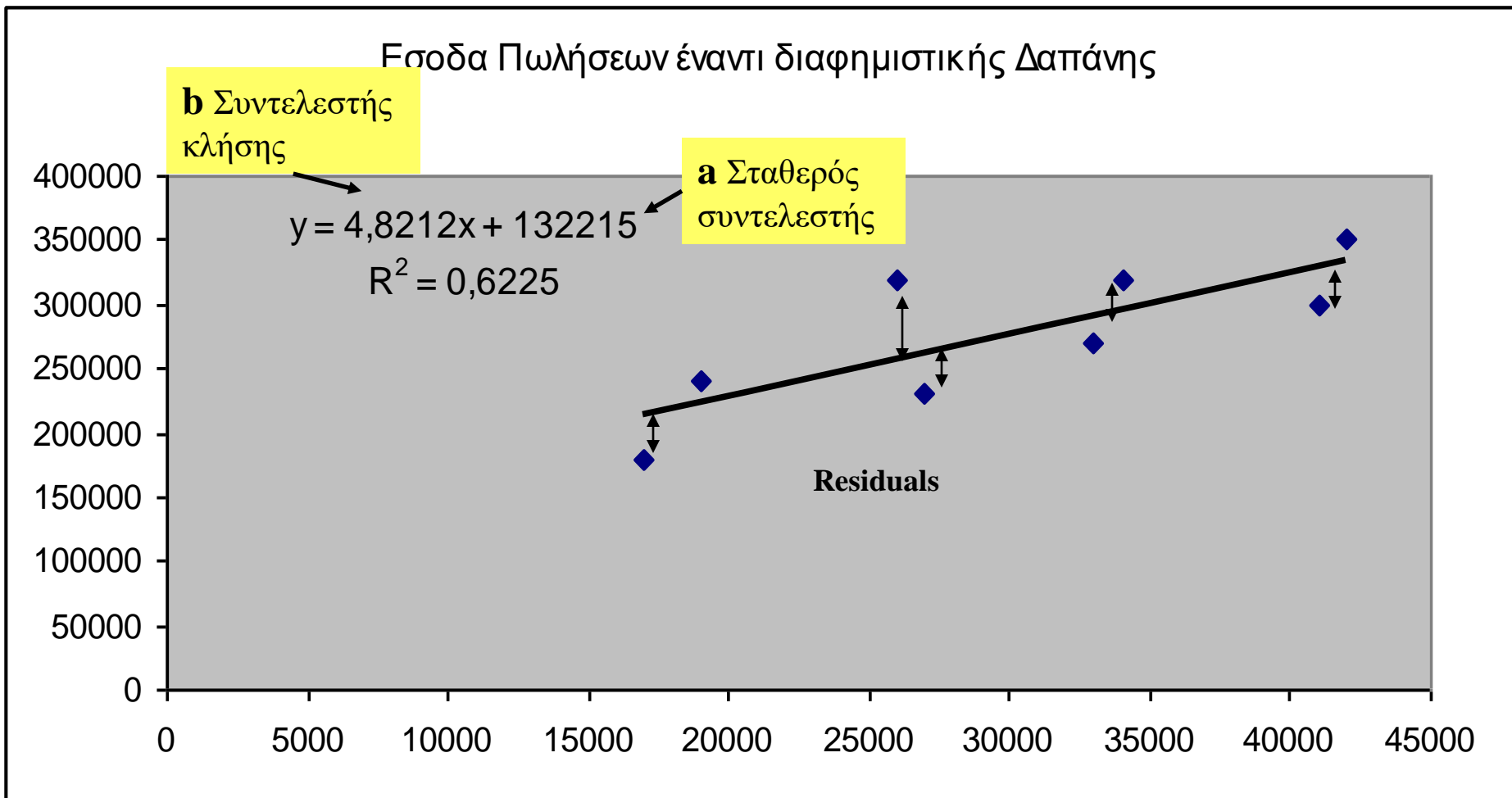
$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

Η εξίσωση που προκύπτει με αυτόν τον τρόπο αντιπροσωπεύει την παλινδρόμηση της y πάνω στην x .

	x	y	xy	x ²		
YEAR	Διαφημιστική Δαπάνη	Εσοδα Πωλήσεων				
1	17	180	3.060	289		
2	33	270	8.910	1.089		
3	34	320	10.880	1.156		
4	42	350	14.700	1.764		
5	19	240	4.560	361		
6	41	300	12.300	1.681		
7	26	320	8.320	676		
8	27	230	6.210	729		
	239	2.210	68.940	7.745		
($\sum x$) ² =	57.121					
		$n \sum xy - \sum x \sum y =$	23330			
		b =	4,82	a =	132	
		$n \sum x^2 - (\sum x)^2 =$	4839			

**Γραμμή Ελαχίστων
Τετραγώνων:**
 $y = 132 + 4,82 x$

ΠΑΛΙΝΔΡΟΜΗΣΗ – Γραμμική Παλινδρόμηση



Η σημασία του όρου των Ελαχίστων Τετραγώνων:

είναι ότι η εξίσωση της σχεδόν άριστης εφαρμογής υπολογίστηκε με τέτοιο τρόπο ώστε το *άθροισμα των τετραγώνων των αποκλίσεων ελαχιστοποιήθηκε* (= ελάχιστα τετράγωνα)

ΠΑΛΙΝΔΡΟΜΗΣΗ – Γραμμική παλινδρόμηση (Μέθοδος ελαχίστων τετραγώνων) ένα παραδειγμα

Ο παρακάτω πίνακας δείχνει τα εξοδολόγια 10 πωλητών μιας εταιρίας
σχετίζει διανυσόμενα χιλιόμετρα x και εξοδα βενζίνης y

Τι είναι τα y, και x ? >

αυξων αριθμό	πωλητής	χιλιόμετρα x	εξοδα y
1	A	100	60
2	B	80	48
3	C	20	20
4	D	120	55
5	E	70	38
6	F	50	38
7	G	80	44
8	H	40	30
9	I	50	40
10	J	60	50

a) εκτίμηση με την μέθοδο ελαχίστων τετραγώνων ένα γραμμικό μοντέλο
το οποίο θα μπορεί να προβλέπει το κόστος επι τη βάσει των χιλιομέτρων (πιο

$$y = a + bx$$

b) κάνε ένα διάγραμμα διασποράς (ένα γραφημα)

C) Επιπλέον 3 πωλητές υποβάλλουν εξοδολόγια ως εξής

πωλητής	χιλιόμετρα x	εξοδα y
K	110	64
L	30	48
M	160	80

είναι αυτά τα έξοδα λογικά?

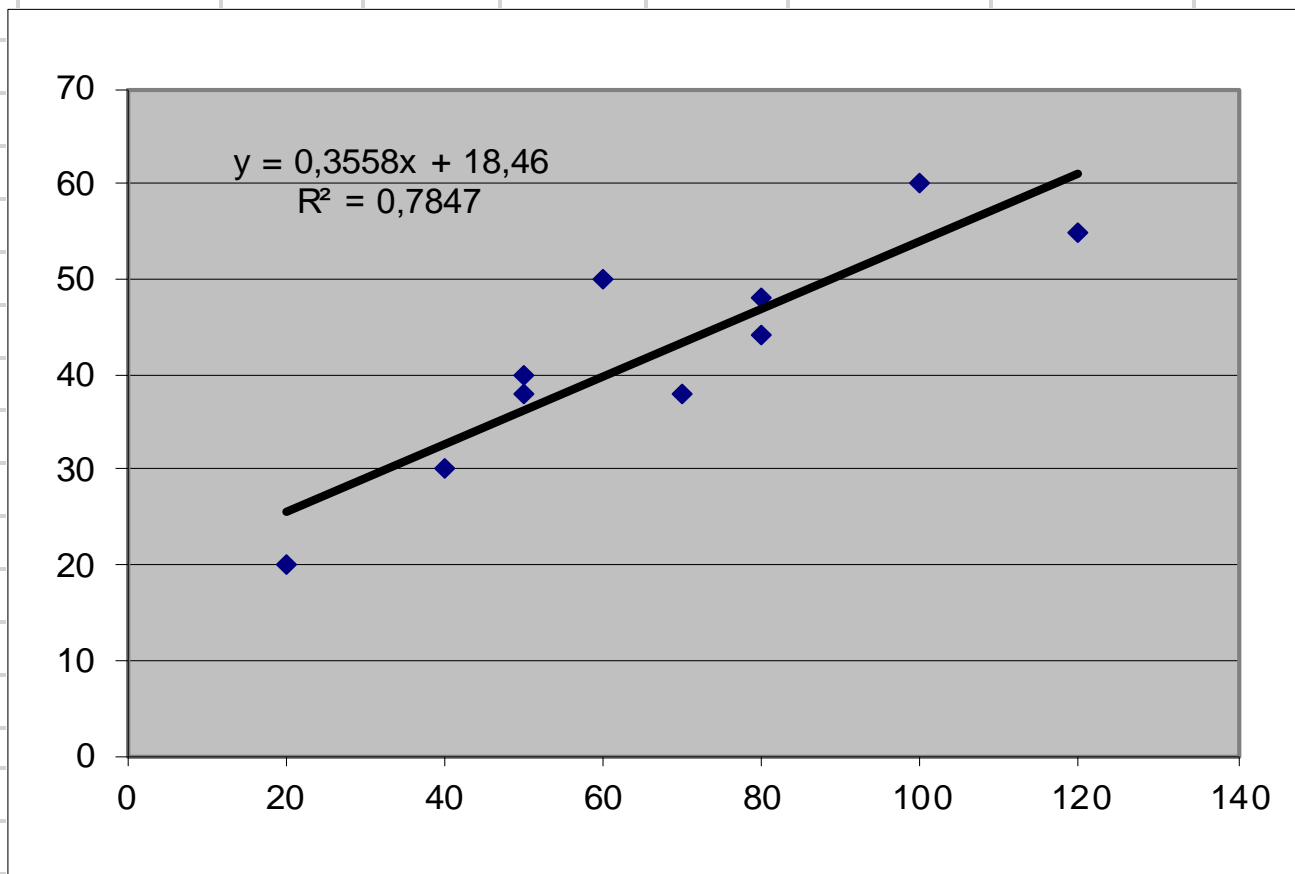
χιλιόμετρα x	εξοδα y	Χιλμ * Εξοδα (xy)	(χιλιόμετρα) x ²	(εξοδα) y ²	
100	60	6.000	10.000	3.600	
80	48	3.840	6.400	2.304	
20	20	400	400	400	
120	55	6.600	14.400	3.025	
70	38	2.660	4.900	1.444	ΣxΣy = 283.410
50	38	1.900	2.500	1.444	
80	44	3.520	6.400	1.936	
40	30	1.200	1.600	900	
50	40	2.000	2.500	1.600	
60	50	3.000	3.600	2.500	
Sums >	670	423	31.120	52.700	19.153
(Sums) ² >	448.900	178.929			

$$b = [n \sum xy - \sum x \sum y] / [n \sum x^2 - (\sum x)^2]$$

$$a = (\sum y / n) - (b \sum x / n)$$

ΠΑΛΙΝΔΡΟΜΗΣΗ – Γραμμική παλινδρόμηση (Μέθοδος ελαχίστων τετραγώνων) ένα παράδειγμα

Το παράδειγμα με τα χιλιόμετρα και τα έξοδα των αυτοκινήτων



$y = 18,46 + 0,3558x$		δηλωθέντα		
Η εξίσωση (μοντέλο) παλινδρόμησης προβλέπει		έξοδα	διαφορά	
x =	110	y = 58	64	6
x =	30	y = 29	48	19
x =	160	y = 75	80	5

Αξιοπιστία των εκτιμήσεων:

- Ο συντελεστής απλού προσδιορισμού (coefficient of determination) R^2 (είναι η 2^η δύναμη δηλ. το τετράγωνο του συντελεστή συσχέτισης) μετράει τη διακύμανση του y που οφείλεται στη διακύμανση της x (της ανεξάρτητης μεταβλητής).

Στο παράδειγμά μας $R^2 = 0,784$ που σημαίνει ότι το 78,4% της διακύμανσης των εξόδων των αυτοκινήτων *ερμηνεύεται* από την διακύμανση των χιλιομέτρων x μέσω της εκτιμηθείσας εξίσωσης μοντέλου $y = 18.46 + 0.3558 x$

- Όσο μεγαλύτερη είναι η αξία του R^2 τόσο μεγαλύτερη είναι η αξιοπιστία της εξίσωσης για την πρόβλεψη της εξαρτημένης μεταβλητής y . Στο παράδειγμα ($R^2 = 0,9$ και πάνω) οι εκτιμηθείσες τιμές της y θεωρείται ότι είναι πολύ κοντά στις πραγματικές τιμές.
- Εάν, αντιθέτως, το R^2 είναι χαμηλότερο ($R^2 = 0,7$ και κάτω) τότε οι προβλεφθείσες τιμές του y θεωρείται ότι είναι μόνο φτωχές εκτιμήσεις των πραγματικών τιμών.
- Όπως συμβαίνει σε κάθε στατιστική ανάλυση η εξίσωση πρόβλεψης – εκτίμησης είναι πιο αξιόπιστη όταν το δείγμα είναι μεγάλο (πάνω από 30 ζευγάρια στοιχείων)
- Interpolation – Εσωτερική «προβολή» σημαίνει η χρήση μιας εξίσωσης ελαχίστων τετραγώνων για εκτίμηση / «πρόβλεψη» της εξαρτημένης μεταβλητής μέσα στο εύρος του διαθέσιμου δείγματος στοιχείων.
Δηλαδή επιχειρούμε μια εσωτερική πρόβλεψη της μεταβλητής y με τα στοιχεία του δείγματος είτε για να ελέγξουμε την απόδοση του, είτε για να καλύψουμε κάποιο κενό στο δείγμα.
- Extrapolation - Υπολογισμός κατά «προέκταση» δηλαδή για εκτίμηση / «πρόβλεψη» της εξαρτημένης μεταβλητής *πέρα και εκτός από το διαθέσιμο δείγμα στοιχείων*. Εκεί υπάρχει μεγαλύτερη αβεβαιότητα. Αποτελεί και τη ουσία της πρόβλεψης