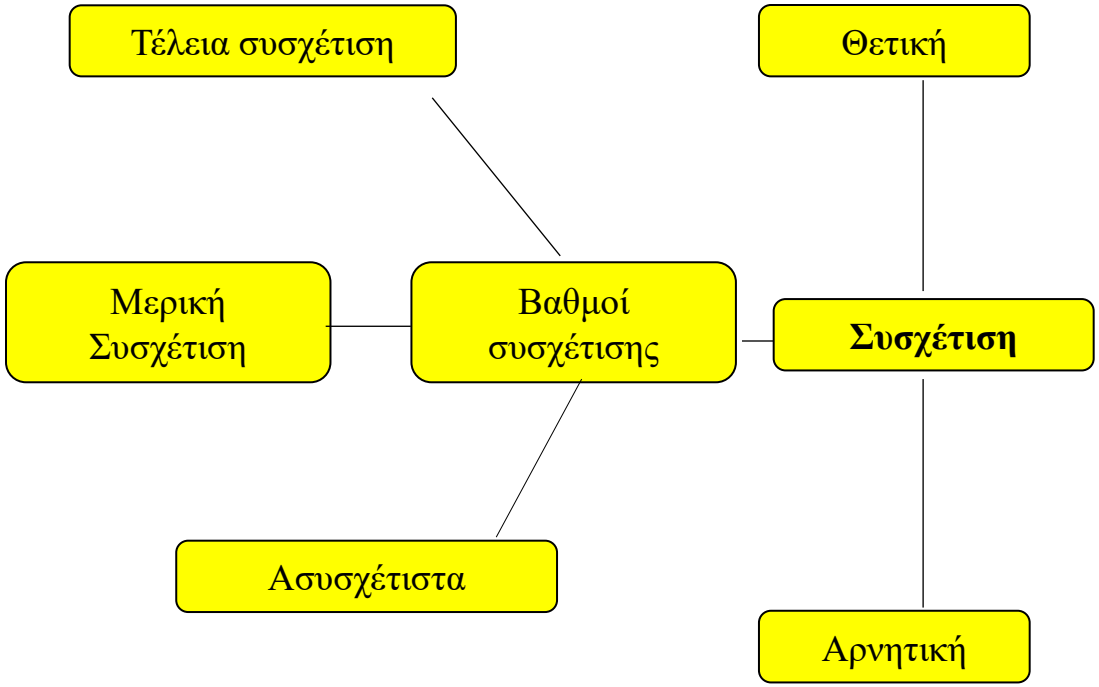
ΣΥΣΧΕΤΙΣΗ


Τέλεια συσχέτιση
Θετική
Μερική Συσχέτιση
Βαθμοί συσχέτισης
Συσχέτιση
Ασυσχέτιστα
Αρνητική

ΣΥΣΧΕΤΙΣΗ

Ορισμός:

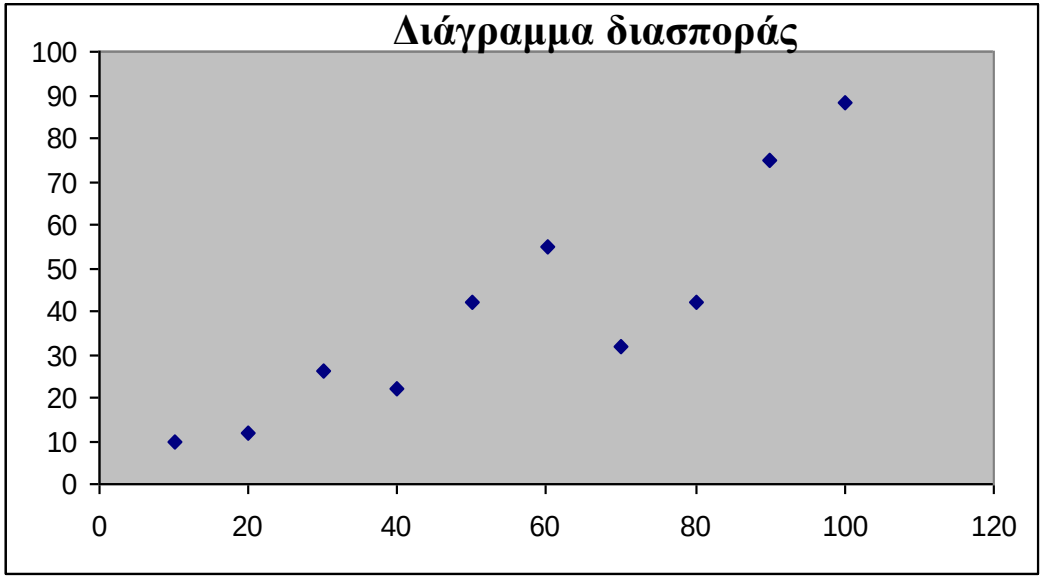
Συσχέτιση *σημαίνει μια σχέση ή «ανταπόκριση» ανάμεσα σε δύο μεταβλητές*

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)

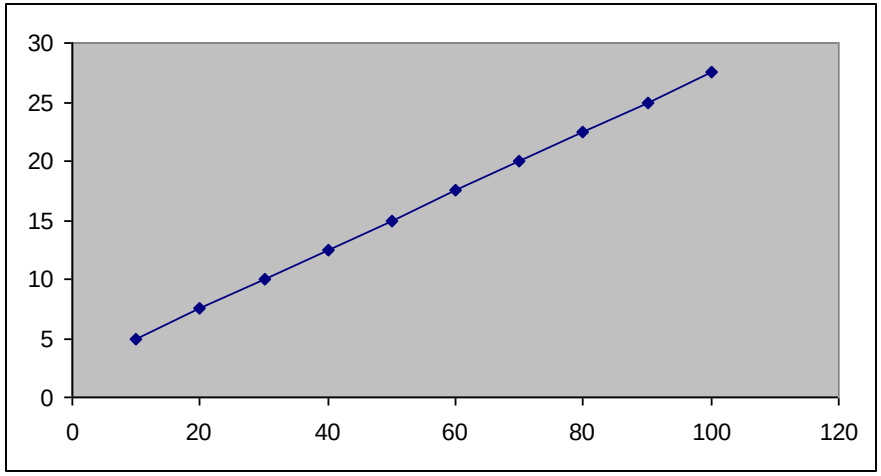
αυτές οι μεταβλητές λαμβάνονται σε ζευγάρια μετρήσεων (παρατηρήσεων) σε μια λογική σειρά στην πορεία του χρόνου



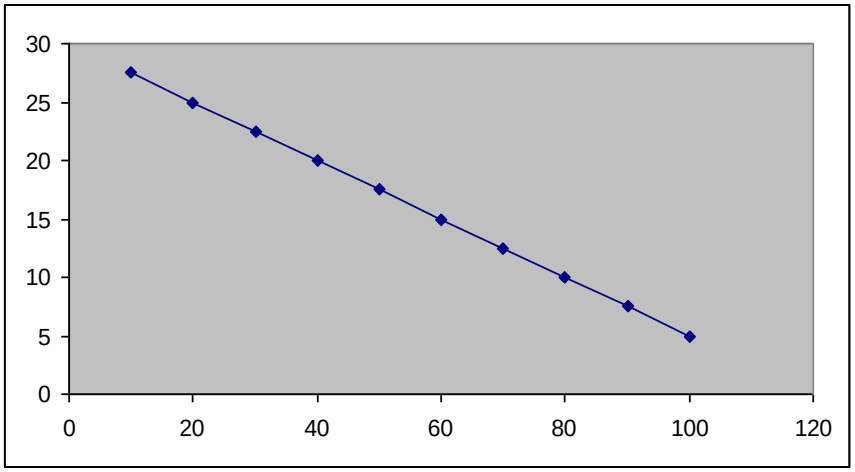
Μηνιαίες πωλήσεις ενός προϊόντος έναντι ***διαφημηστικής δαπάνης*** για αυτό το προϊόν

ΣΥΣΧΕΤΙΣΗ – Βαθμοί Συσχέτισης

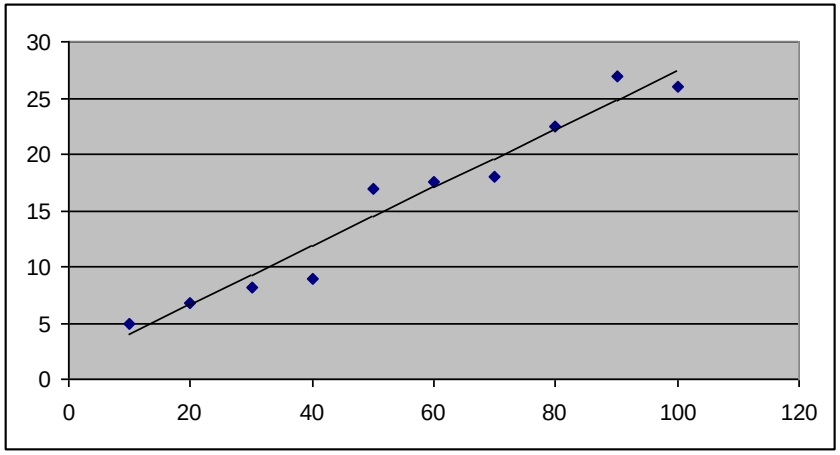
Τέλεια θετική συσχέτιση



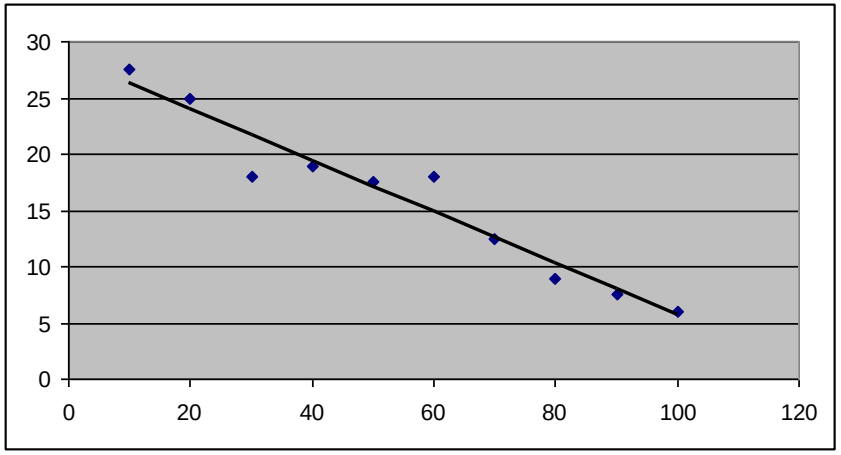
Τέλεια αρνητική συσχέτιση



Μερική θετική συσχέτιση



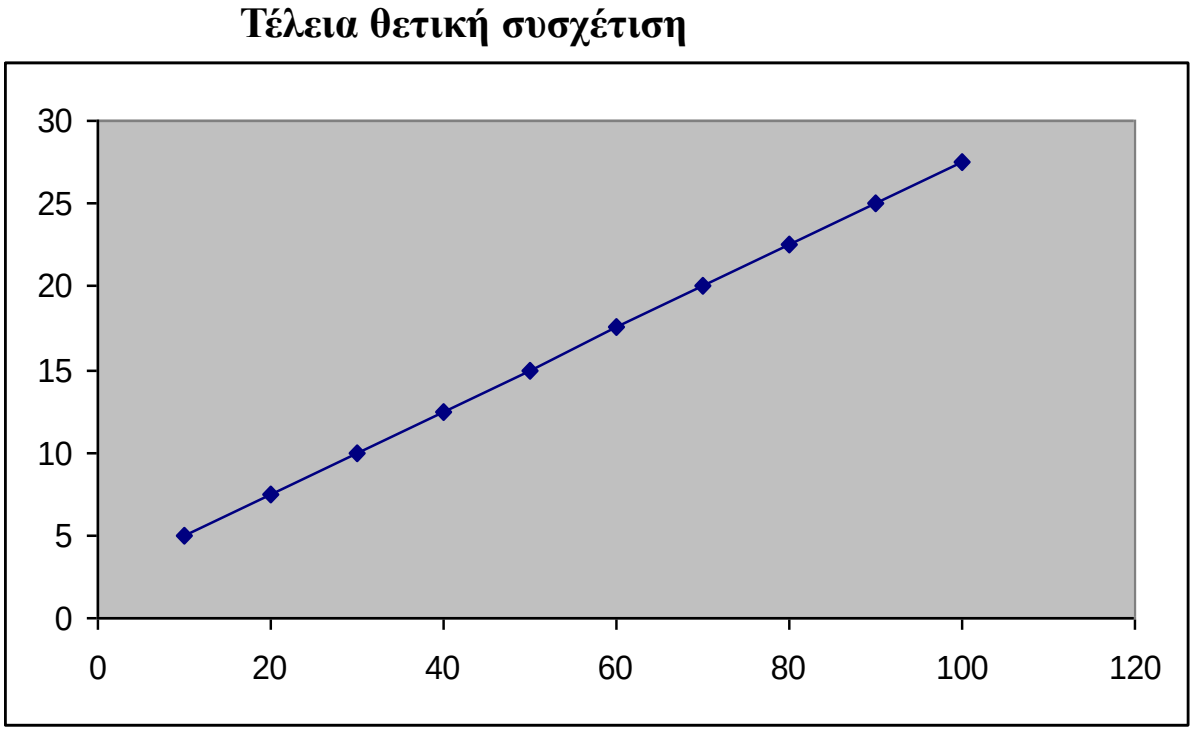
Μερική αρνητική συσχέτιση



ΣΥΣΧΕΤΙΣΗ – Βαθμοί Συσχέτισης

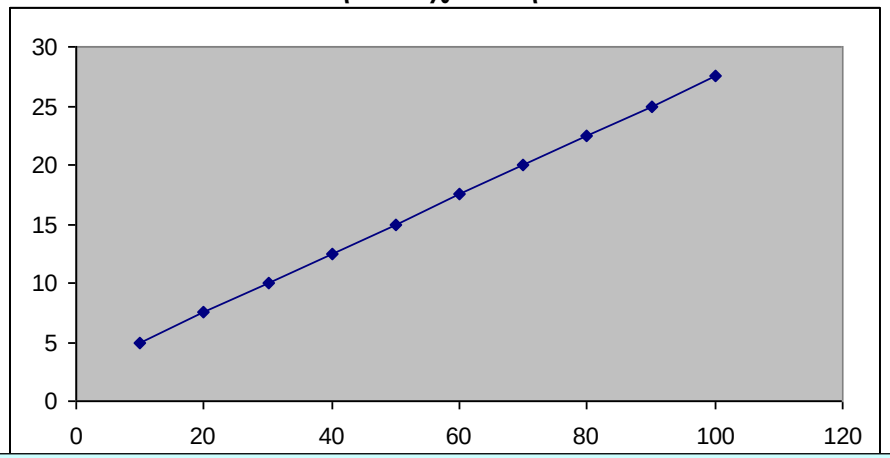
Τέλεια θετική συσχέτιση

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

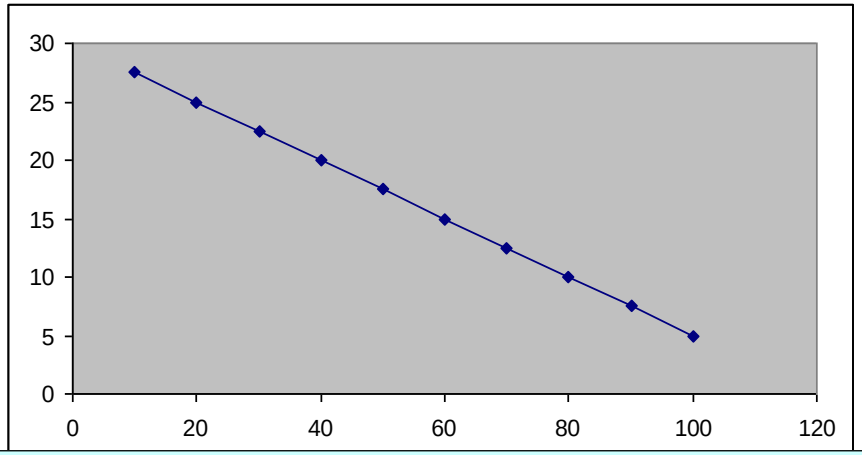


ΣΥΣΧΕΤΙΣΗ – Βαθμοί Συσχέτισης

Τέλεια θετική συσχέτιση



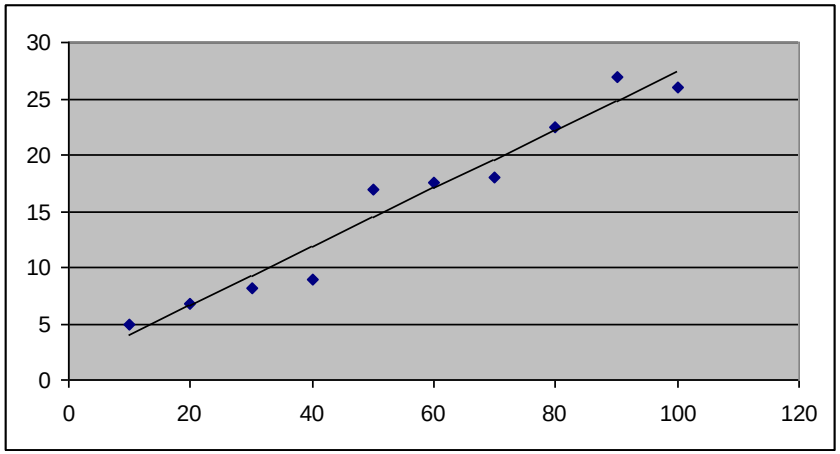
Τέλεια αρνητική συσχέτιση



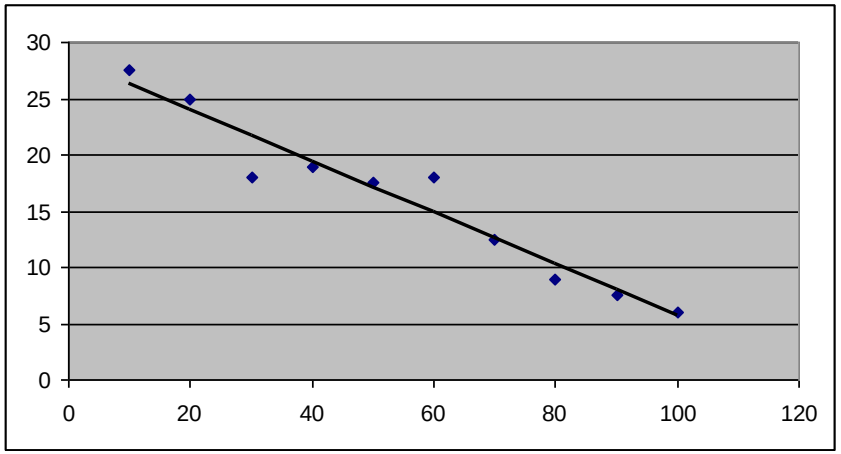
Θετική συσχέτιση σημαίνει οτι χαμηλές αυξήσεις (ή μειώσεις) στις τιμές της μιας μεταβλητής συνδέονται με χαμηλές αυξήσεις (ή μειώσεις) στις τιμές της άλλης και υψηλές αυξήσεις (ή μειώσεις) στις τιμές της μιάς μεταβλητής συνδέονται με υψηλές αυξήσεις (ή μειώσεις) στις τιμές της άλλης - οσο πιο παρόμοιες είναι οι παράλληλες αυξήσεις (ή μειώσεις) τόσο πιό 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 ειναι 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	nΣxy - Σx Σy =	-19.000
	20	25	500	400	625	παρονομαστής	$n\Sigma x^2 - (\Sigma x)^2$ =	82500
	30	18	540	900	324	παρονομαστής	$n\Sigma y^2 - (\Sigma y)^2$ =	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			
∑	**550**	**160**	**6900**	**38500**	**3026**			
$(\sum)^2$	302500	25600						

ΣΥΣΧΕΤΙΣΗ – Μέτρηση της συσχέτισης

Ορισμός:

Ο συντελεστης προσδιορισμού r^2 *or* R^2 μετράει το μέρος της συνολικης διακύμανσης της μιας μεταβλητής που μπορεί να ερμηνευθεί απο την διακύμανση της άλλης μεταβλητης.

What values can R^2 take ?

	x	y	xy	x^2	y^2			
	10	27,5	275	100	756,25	Αριθμητής	nΣxy - Σx Σy =	-19.000
	20	25	500	400	625	Παρονομαστής	$n\Sigma x^2 - (\Sigma x)^2$ =	82500
	30	18	540	900	324	Παρονομαστής	$n\Sigma y^2 - (\Sigma y)^2$ =	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		**R^2 =**	**0,939003772**
	80	9	720	6400	81			
	90	7,5	675	8100	56,25			
	100	6	600	10000	36			
∑	**550**	**160**	**6900**	**38500**	**3026**			
$(\sum)^2$	302500	25600						

Αυτό σημαίνει οτι το 94% της διακύμανσης της μεταβλητής y μπορεί να ερμηνευθεί απο τη διακύμανση της μεταβλητής της x

ΣΥΣΧΕΤΙΣΗ – Ανάλυση / Ερμηνεία

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) ειναι πολύ μεγάλο είναι πιο πιθανό αυτή η συσχέτιση να μην είναι «τυχαία» (δηλαδή υπάρχει κάποια αιτιώδης σχέση ανάμεσά τους)

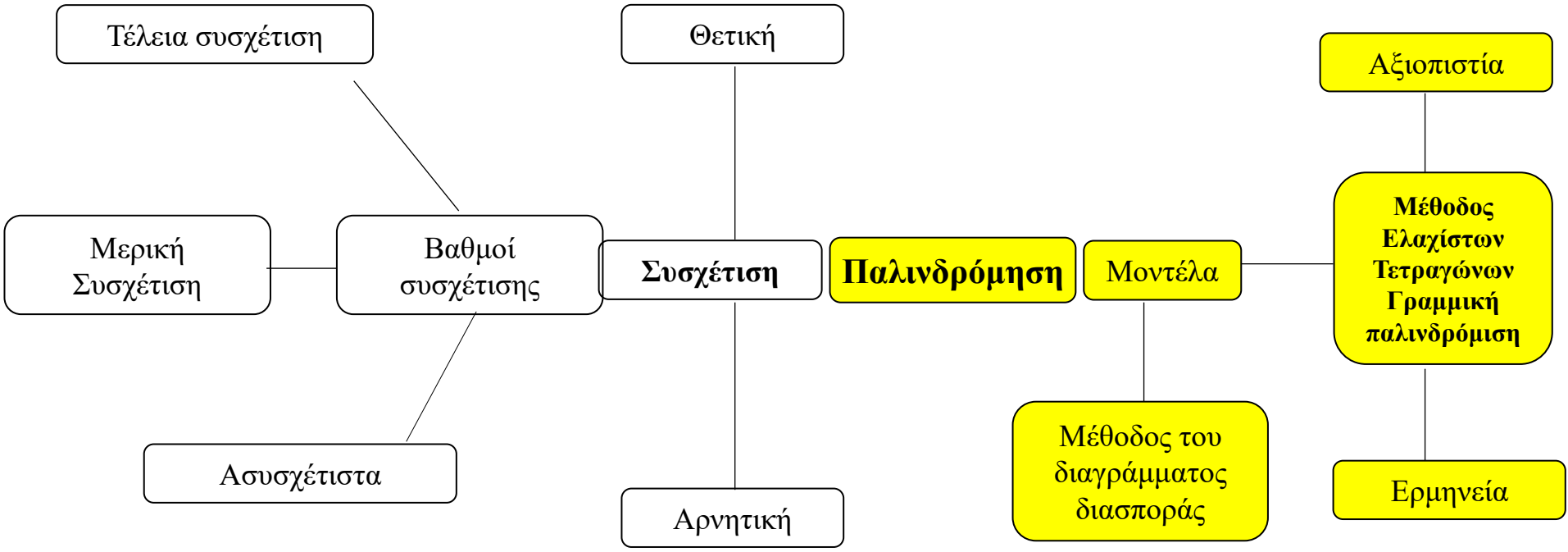
- Συνήθως οι ερευνητές ελέγχουν την συσχέτιση ανάμεσα σε μεταβλητές που συνδέονται λογικά δηλαδή πιστεύουν (στη βάση κάποιας θεωρίας) ότι *η μεταβλητή x είναι η **αιτία** των διακυμάνσεων και της συμπεριφοράς της άλλης μεταβλητής της y*. Μπορεί να υπάρχει κάποια "υποστηρικτική" θεωρία για την σχέση αυτών 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

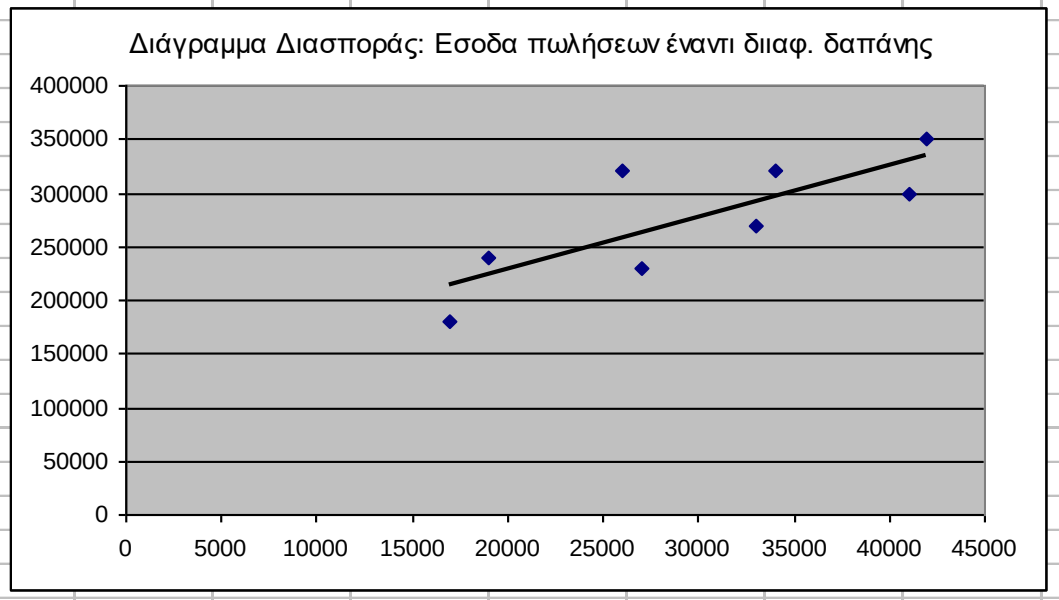
CORRELATION & REGRESION



- Η συσχέτιση δείχνει το επίπεδο σχέσης ανάμεσα σε δύο μεταβλητές αλλά δεν δείχνει το πως (τον τρόπο) που αυτή η σχέση συμβαίνει
- Αυτή τη σχέση την καλύπτει η παλινδρόμηση

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



Ενα πρώτο βήμα – προσέγγισης αυτής της σχέσης ανάμεσα στις δύο μεταβλητές είναι να τις απεικονίσουμε σε ενα 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 παλινδρομισης είναι μια πολύ συνηθισμένη τεχνική για την εκτίμηση της **εξίσωσης της γραμμής καλύτερης εφαρμογής**.*

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

y = a + bx

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

y ειναι η εξαρτημένη μεταβλητή

a είναι ο *σταθερός συντελεστής*: a = (Σy / n) – (b Σx / n)

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

b είναι ο *συντελεστής κλίσης*: b = [n ∑xy - ∑x ∑y] / [n ∑x² – (∑x)²]

$$b = \frac{\eta \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\eta \Sigma x_i^2 - (\Sigma x_i)^2}$$

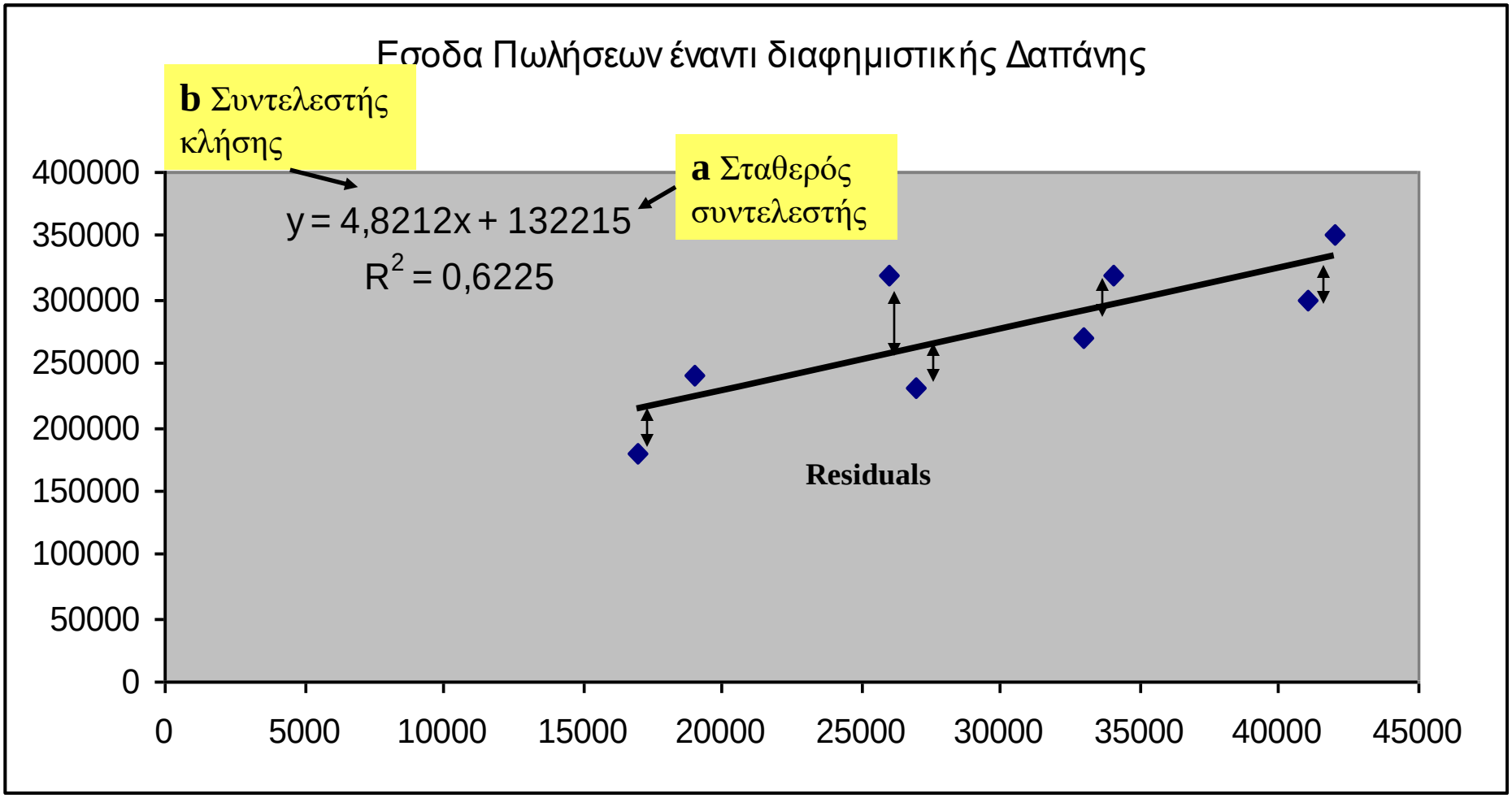
Η εξίσωση που προκύπτει με αυτόν τον τρόπο αντιπροσωπευει την παλινδρόμιση της y πάνω στην x.

	x	y	xy	x²		
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**		
(Σx)² =	57.121					
		n ∑xy - ∑x ∑y =	23330			
			b =	**4,82**	**a =**	**132**
		n ∑x2 – (∑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 υπολογίσθηκε με τέτοιο τρόπο ώστε το ***άθροισμα των τετραγώνων των αποκλίσεων ελαχιστοποιήθηκε*** (= ελάχιστα τετράγωνα)

ΠΑΛΙΝΔΡΟΜΗΣΗ – Γραμμική παλινδρόμηση (Μέθοδος ελαχίστων τετραγωνων) ενα παραδειγμα

Ο παρακάτω πίνακας δείχνει τα εξοδολόγια 10 πωλητών μιας εταιρίας
σχετίζει διανυούμενα χιλιομετρα 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 ενα γραμμικό μοντέλο
το οποίο θα μπορεί να προβλέπει το κόστος επι τη βάσει των χιλιομέτρων (που

y = a + bx

b) κάνε ενα διάγραμμα διασποράς (ενα γραφημα)

C) Επιπλεόν 3 πωλητές υποβάλλουν εξοδολόγια ως εξής

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

ειναι αυτά τα έξοδα λογικά?

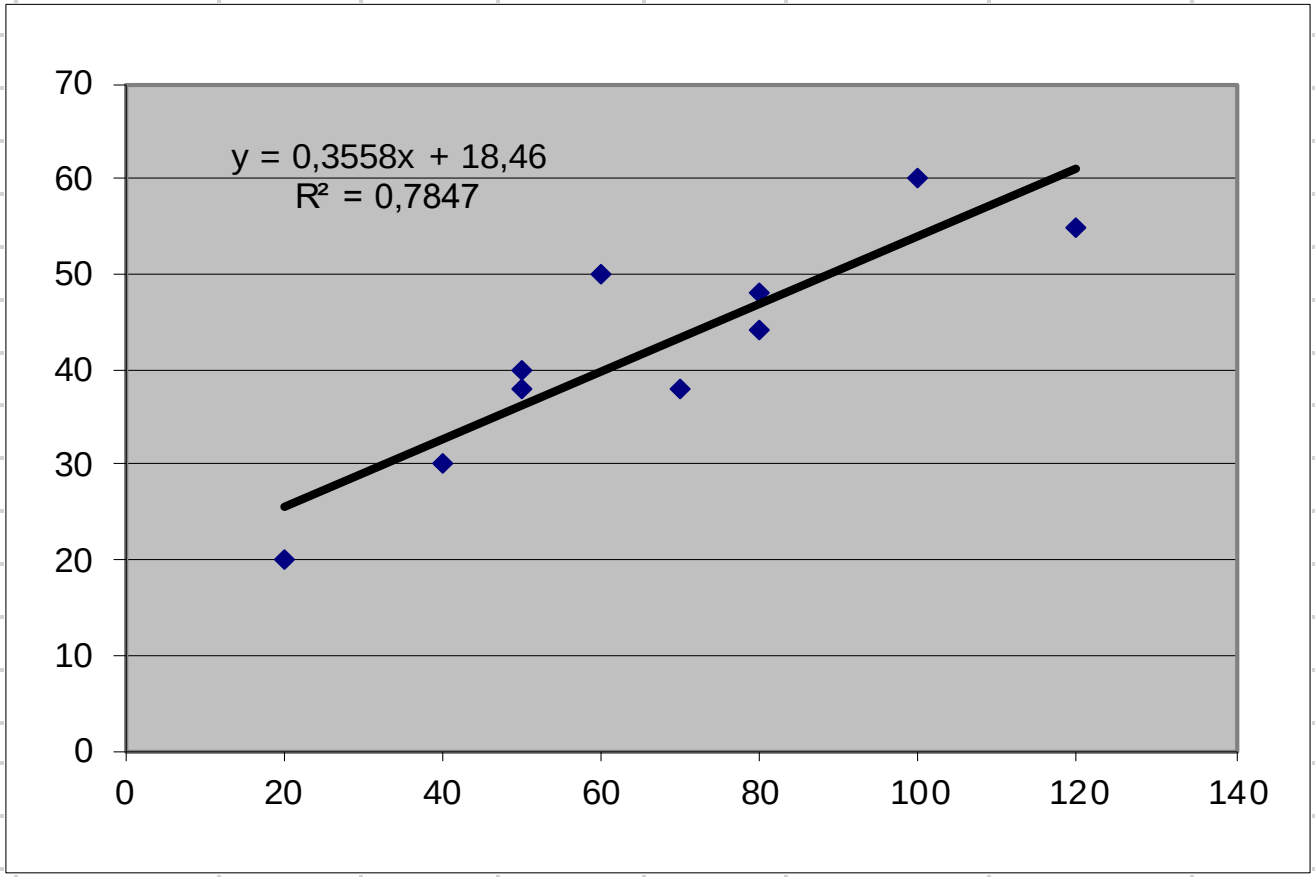
χιλιόμετρα x	εξοδα y	Χλμ * Εξοδα (xy)	(χιλιόμετρα) x^2	(εξοδα) y^2		
100	60	6.000	10.000	3.600		
80	48	3.840	6.400	2.304		
20	20	400	400	400		
120	55	6.600	14.400	3.025		
70	38	2.660	4.900	1.444	$\sum x \sum y =$	283.410
50	38	1.900	2.500	1.444		
80	44	3.520	6.400	1.936	`	
40	30	1.200	1.600	900		
50	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				

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

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

ΠΑΛΙΝΔΡΟΜΗΣΗ – Γραμμική παλινδρόμηση (Μέθοδος ελαχίστων τετραγωνων) ενα παραδειγμα

Το παραδειγμα με τα χιλιόμετρα και τα εξοδα των αυτοκινήτων



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

ΠΑΛΙΝΔΡΟΜΗΣΗ – Μέθοδος Ελαχίστων Τετραγώνων: Αξιοπιστία των συντελεστών του μοντέλου

Αξιοπιστία των εκτιμήσεων:

- Ο συντελεστής απλού προσδιορισμού (coefficient of determination) R^2 (ειναι 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 ειναι η αξία του R^2 τόσο μεγαλύτερη ειναι 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 ειναι πολύ κοντά στις πραγματικές τιμές.

- Εάν, αντιθέτως, το R^2 ειναι 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 ειναι πιο αξιόπιστη όταν το δείγμα ειναι 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

Παλινδρόμηση - Ανάλυση

- Στη στατιστική «μοντελοποίηση», η ανάλυση παλινδρόμησης είναι ένα σύνολο στατιστικών διεργασιών για την εκτίμηση των σχέσεων μεταξύ μιας ***εξαρτημένης*** μεταβλητής (συχνά αποκαλείται «μεταβλητή αποτελέσματος») και μιας ή περισσότερων ***ανεξάρτητων*** μεταβλητών (συχνά αποκαλούνται «παράγοντες προβλέψεων», «συν-μεταβλητές» ή «χαρακτηριστικά»).

- Η πιο κοινή μορφή ανάλυσης παλινδρόμησης είναι η ***γραμμική παλινδρόμηση***, στην οποία ο μελετής βρίσκει τη γραμμή που ταιριάζει περισσότερο στα δεδομένα του εκάστοτε δείγματος σύμφωνα με ένα συγκεκριμένο μαθηματικό κριτήριο.

Παλινδρόμηση - Ανάλυση

- Η μέθοδος των συνηθισμένων **«ελάχιστων τετραγώνων»** υπολογίζει τη μοναδική γραμμή ***που ελαχιστοποιεί το άθροισμα των τετραγώνων των διαφορών*** μεταξύ των πραγματικών δεδομένων και αυτής της γραμμής (που υπολογίστηκε)

- Η Γραμμική παλινδρόμηση επιτρέπει στον ερευνητή να «εκτιμήσει» την προσδοκόμενη τιμή (ή τη μέση προσδοκόμενη τιμή της *εξαρτημένης* μεταβλητής, όταν οι *ανεξάρτητες* μεταβλητές παίρνουν συγκεκριμένες τιμές.

Παλινδρόμηση - Ανάλυση

- Η ανάλυση παλινδρόμησης χρησιμοποιείται κυρίως για δύο εννοιολογικά διαφορετικούς σκοπούς.

1. Πρώτον, η ανάλυση παλινδρόμησης χρησιμοποιείται ευρέως για ***πρόβλεψη*** και προβολή (πέρα και έξω από το δείγμα που διαθέτουμε).

2. Δεύτερον, σε ορισμένες περιπτώσεις η ανάλυση παλινδρόμησης μπορεί να χρησιμοποιηθεί για να συναχθούν ***αιτιώδεις σχέσεις*** μεταξύ των ανεξάρτητων και εξαρτημένων μεταβλητών.

Παλινδρόμηση - Ανάλυση

- Είναι σημαντικό ότι οι παλινδρομήσεις αποκαλύπτουν μόνες τους σχέσεις που μπορεί να υπάρχουν μεταξύ μιας εξαρτημένης μεταβλητής και μιας «συλλογής» ανεξάρτητων μεταβλητών μέσα από ένα δείγμα (συλλογή δεδομένων).

- Ενας ερευνητής για να χρησιμοποιήσει παλινδρόμηση για ***πρόβλεψη*** ή για να συναγάγει ***αιτιώδεις σχέσεις***, αντίστοιχα, πρέπει να αιτιολογήσει προσεκτικά γιατί *οι υπάρχουσες σχέσεις έχουν προγνωστική ισχύ* ή γιατί μια σχέση μεταξύ δύο μεταβλητών έχει *αιτιώδη ερμηνεία*. Το τελευταίο είναι ιδιαίτερα σημαντικό όταν οι ερευνητές ελπίζουν να εκτιμήσουν τις αιτιώδεις σχέσεις χρησιμοποιώντας δείγματα (δεδομένα παρατήρησης).

Παλινδρόμηση - Ανάλυση

See simple linear regression for a derivation of these formulas and a numerical example

In linear regression, the model specification is that the dependent variable, y_i is a linear combination of the *parameters* (but need not be linear in the *independent variables*). For example, in simple linear regression for modeling n data points there is one independent variable: x_i, and two parameters, β_0 and β_1:

straight line: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \ldots, n.$ **απλή γραμμική παλινδρόμηση - ευθεία γραμμή**

In multiple linear regression, there are several independent variables or functions of independent variables.

Adding a term in x_i^2 to the preceding regression gives:

parabola: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \ i = 1, \ldots, n.$ **«Παραβολική» γραμμική παλινδρόμηση**

This is still linear regression; although the expression on the right hand side is quadratic in the independent variable x_i, it is linear in the parameters β_0, β_1 and β_2. **Είναι γραμμική παλινδρόμηση παρά το x^2 είναι γραμμική ως πρός τα β_o β_1 β_2**

In both cases, ε_i is an error term and the subscript i indexes a particular observation.

Returning our attention to the straight line case: Given a random sample from the population, we estimate the population parameters and obtain the sample linear regression model:

$$\hat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

The residual, $e_i = y_i - \hat{y}_i$, is the difference between the value of the dependent variable predicted by the model, $\hat{y}_i$, and the true value of the dependent variable, y_i. One method of estimation is ordinary least squares. This method obtains parameter estimates that minimize the sum of squared residuals, SSR:

$$SSR = \sum_{i=1}^{n} e_i^2.$$

Αυτή είναι **η μέθοδος ελαχίστων τετραγώνων** καθώς *ελαχιστοποιεί το* ***RSS*** = *άθροισμα των τετραγώνων των καταλοίπων (σφάλματα - errors)*

Minimization of this function results in a set of normal equations, a set of simultaneous linear equations in the parameters, which are solved to yield the parameter estimators, $\widehat{\beta}_0, \widehat{\beta}_1$.

Παλινδρόμηση – Ανάλυση

χρησιμα – κατατοπιστικά links

https://www.youtube.com/watch?v=m-k84cCves8

https://www.youtube.com/watch?v=dQNpSa-bq4M

https://www.youtube.com/watch?v=cXiZ_t2NK1k