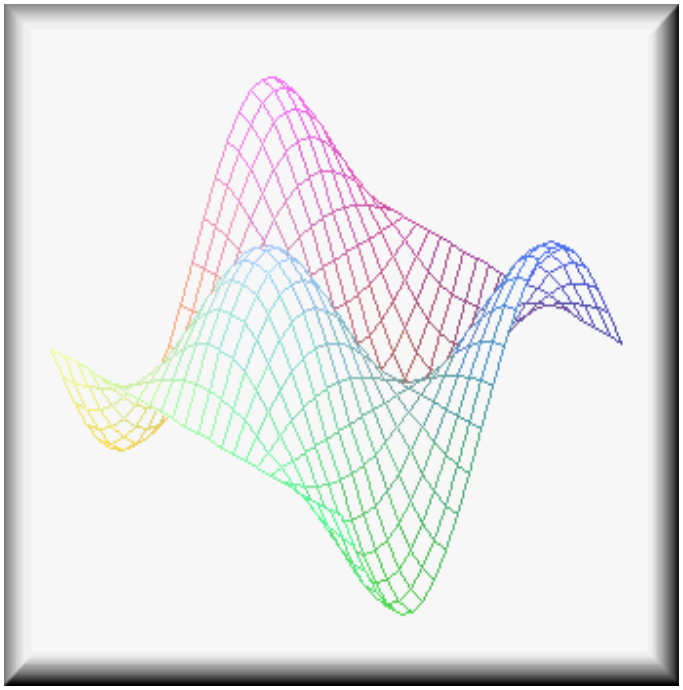
Τεχνολογικό Εκπαιδευτικό Ίδρυμα Πατρών

Σχολή Διοίκησης κ Οικονομίας
Τμήμα: Επιχειρηματικού Σχεδιασμού & Πληροφοριακών Συστημάτων

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Συσχέτιση Δύο Μεταβλητών και Παλινδρόμηση



Σπουδάστριες:
Γκερλέ Στεφανία
Καραγιώργου Ανδρομάχη
Τσακνάκη Μαρία

Εποπτεύουσα καθηγήτρια:
Μπουμπούλη Αθανασία

Πάτρα, 2010

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ

Οι έννοιες και οι μέθοδοι της επιστήμης της στατιστικής αποτελούν απαραίτητη προϋπόθεση για να αντιληφθούμε τον κόσμο γύρω μας. Στα πραγματικά φαινόμενα υπάρχει σχεδόν πάντα ο παράγοντας της αβεβαιότητας με αποτέλεσμα να είναι απαραίτητη η γνώση και κατά συνέπεια η χρήση των στατιστικών μεθόδων και μοντέλων για την επίλυση των προβλημάτων αυτών. Αντίθετα υπάρχουν και τα φαινόμενα αυτά, που ακριβώς επειδή απουσιάζει ο παράγοντας της αβεβαιότητας, περιγράφονται πλήρως από τα λεγόμενα καθοριστικά μοντέλα.

Ο Άγγλος γενετιστής και βιοστατικός Francis Galton (1822-1911) είναι «υπεύθυνος» για την ύπαρξη και τη σημερινή χρήση των όρων «**παλινδρόμηση**» και «**συσχέτιση**». Ο όρος **παλινδρόμηση** (regression) χρησιμοποιήθηκε από τον ίδιο για πρώτη φορά το 1990 ο οποίος παρατήρησε ότι «Τα πληθυσμιακά ακρότατα οπισθοχωρούν προς τη μέση τιμή τους». Πιο συγκεκριμένα μελετώντας τη σχέση μεταξύ του ύψους των γονέων και του ύψους των παιδιών τους, παρατήρησε ένα είδος επαναφοράς (παλινδρόμησης) του ύψους των παιδιών, στο ύψος των γονέων τους, δηλαδή τα παιδιά των ψηλών γονιών (αντίστροφα, κοντών) είναι κατά μέσο όρο κοντύτερα (αντίστροφα, ψηλότερα) από τους γονείς τους. Όσον αφορά τη συσχέτιση, δημιούργησε το 1885 το συντελεστή σχέσης (correlation) και η χρήση αυτού του μέτρου οδήγησε σε μία ευρεία αποδοχή του μέχρι και το 1892. Το διάστημα αυτό εμφανίζεται ο F.Y. Edgeworth και αλλάζει το όνομα σε συντελεστή συσχέτισης (coefficient of correlation), όρος που διατηρείτε μέχρι και σήμερα, ενώ τέσσερα χρόνια μετά έρχεται ο **Pearson** και μας παρέχει την αναλυτική μορφή του product-moment (συντελεστής συσχέτισης του **Pearson**).

Έχει αναφερθεί το πρόβλημα της παλινδρόμησης σε σχέση με τυχαίες μεταβλητές των οποίων γνωρίζαμε την από κοινού κατανομή.

Η **ανάλυση παλινδρόμησης** αποτελεί έναν από τους πιο σημαντικούς κλάδους της Στατιστικής με ευρείες εφαρμογές σε όλες τις σύγχρονες επιστήμες και μαζί με τη συσχέτιση έχουν εφαρμογές σε πάρα πολλούς τομείς της τεχνολογίας και είναι ιδιαίτερα σημαντικές σε περιπτώσεις όπου οι αναγκαίες σχέσεις μεταξύ διαφόρων ποσοτήτων πρέπει να προσδιοριστούν εμπειρικά.

Με τη φράση **ανάλυση παλινδρόμησης** εννοούμε διάφορες γραφικές και αναλυτικές μεθόδους που σκοπό έχουν την αναζήτηση σχέσεων μεταξύ της **εξαρτημένης μεταβλητής** (dependent variable) και των **ανεξάρτητων μεταβλητών** (independent variables). Όταν μια τέτοια σχέση βρεθεί, τότε μπορούμε να χρησιμοποιήσουμε το μοντέλο αυτό για να κάνουμε προβλέψεις, για να βρούμε ποια ή ποιες από τις ανεξάρτητες μεταβλητές επηρεάζουν περισσότερο την εξαρτημένη μεταβλητή, ή να ελέγξουμε διάφορες υποθέσεις.

Η **ανάλυση συσχετίσεως** χρησιμοποιείται όταν κάποιος θέλει να μελετήσει την ταυτόχρονη μεταβλητότητα ενός συνόλου μεταβλητών. Στην ανάλυση συσχετίσεως οι σχέσεις μεταξύ των διαφόρων μεταβλητών δεν έχουν γενικώς μία μόνο κατεύθυνση. Έχει αναφερθεί το πρόβλημα της παλινδρόμησης, σε σχέση με τυχαίες μεταβλητές των οποίων γνωρίζαμε την από κοινού κατανομή. Στην πράξη όμως υπάρχουν προβλήματα όπου η από κοινού κατανομή των τυχαίων μεταβλητών δεν είναι γνωστή και κατά συνέπεια η παλινδρόμηση της μίας εξ αυτών ως προς τις άλλες δεν μπορεί να βρεθεί. Στην περίπτωση αυτή, με βάση τα δεδομένα ενός τυχαίου δείγματος, πρέπει να εκτιμήσουμε την εν λόγω παλινδρόμηση.

Η **συσχέτιση** μετρά το βαθμό συνάφειας-αλληλεπίδρασης ανάμεσα σε δύο ή περισσότερες μεταβλητές. Πρακτικά σημαίνει ότι από την τιμή ενός δείκτη (συντελεστή συσχέτισης), η διαδικασία συσχέτισης παρουσιάζεται όχι μόνο σε ποσοτικές μεταβλητές (συντελεστής Pearson), αλλά και σε ποιοτικές ή κατηγορηματικές μεταβλητές. Θα πρέπει να διακρίνουμε μία διαφορά. Το γεγονός της ύπαρξης ή μη έντονης συνάφειας-συσχέτισης ανάμεσα σε δύο μεταβλητές δε συνεπάγεται απαραίτητα και την ύπαρξη μίας συναρτησιακής σχέσης αυτών. Το θέμα αυτό αναλύεται στη διαδικασία Regression.

Στο σημείο αυτό παρατηρούμε τη διαφορά μεταξύ της **ανάλυσης παλινδρόμησης** (regression analysis) και της **ανάλυσης συσχετίσεως** (correlation analysis). Στην ανάλυση παλινδρόμησης η σχέση είναι προς μία κατεύθυνση μόνο. Δηλαδή αγνοεί την πιθανή επίδραση της εξαρτημένης μεταβλητής στις ανεξάρτητες μεταβλητές. Σε μερικές περιπτώσεις, κυρίως σε εργαστηριακά πειράματα, ο ερευνητής μπορεί να προκαθορίσει τις τιμές των ανεξάρτητων μεταβλητών και στη συνέχεια να παρατηρήσει το μέγεθος της εξαρτημένης μεταβλητής. Στην περίπτωση αυτή είναι φανερό ότι η σχέση αυτή είναι προς μία μόνο κατεύθυνση. Σε πολλές περιπτώσεις όμως ο ερευνητής παρατηρεί συγχρόνως όλες τις μεταβλητές. Η ανάλυση παλινδρόμησης για τέτοια δεδομένα μπορεί και πάλι να χρησιμοποιηθεί όταν ο σκοπός της μελέτης μας είναι να εξετάσουμε τη μεταβολή μιας εξ αυτών σε σχέση με τις υπόλοιπες.

ΚΕΦΑΛΑΙΟ 1^{Ο}
ΣΥΣΧΕΤΙΣΗ ΜΕΤΑΒΛΗΤΩΝ

Karl Pearson

1.1 Εισαγωγή

Η **Συσχέτιση** έχει ως στόχο τη μέτρηση του βαθμού εξάρτησης, δηλαδή της φοράς και της έντασης της συμμεταβολής που υπάρχει μεταξύ των τυχαίων μεταβλητών **X** και **Y**. Πρακτικά σημαίνει, ότι από την τιμή ενός δείκτη (συντελεστή συσχέτισης) κατανοούμε πόσο έντονη ή χαλαρή είναι η συσχέτιση δύο μεταβλητών. Η διαδικασία συσχέτισης παρουσιάζεται όχι μόνο σε ποσοτικές μεταβλητές (συντελεστής Pearson) αλλά και σε ποιοτικές ή κατηγορικές μεταβλητές. Θα πρέπει να διακρίνουμε μία διαφορά. Το γεγονός της ύπαρξης ή μη έντονης συνάφειας-συσχέτισης ανάμεσα σε δύο μεταβλητές, δεν συνεπάγεται απαραίτητα και την ύπαρξη μίας συναρτησιακής σχέση.

Θεωρούμε δύο τυχαίες μεταβλητές **X** , **Y** και **ν** ζεύγη παρατηρήσεων από τυχαίο δείγμα μεγέθους **ν** .

Αναφερόμαστε, δηλαδή, σε **μη πειραματικά** δεδομένα (ο ερευνητής δεν προκαθορίζει-ελέγχει τις τιμές καμιάς από τις δύο μεταβλητές) όπως,

♦ **X** το ύψος των φοιτητών ενός πανεπιστημιακού τμήματος και **Y** το βάρος τους.

♦ **X** οι ώρες μελέτης των φοιτητών ενός πανεπιστημιακού τμήματος και **Y** η απόδοση τους σε ένα τεστ.

♦ **X** οι εβδομάδες εμπειρίας ενός εργάτη σε μια επιχείρηση και **Y** ο αριθμός των ελαττωματικών προϊόντων που παράγει.

♦ **X** η κατάταξη δέκα προϊόντων από ένα κριτή και **Y** η κατάταξη των ιδίων προϊόντων από έναν άλλο κριτή.

♦ **X** ο αριθμός των νέων στην ίδια περιοχή και **Y** ο αριθμός των πωλήσεων μουσικών CD σε μια περιοχή.

Δεν αναφερόμαστε όμως σε περιπτώσεις όπως,

♦ **X** ο αριθμός των ανοιχτών ταμείων ενός υποκαταστήματος τραπέζης (που καθορίζει ο διευθυντής) και **Y** ο χρόνος αναμονής των πελατών.

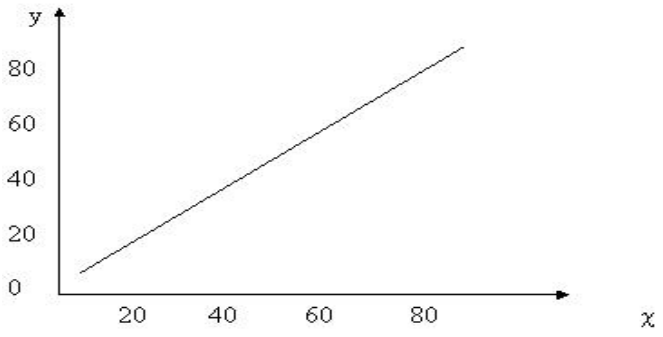
♦ **X** η ποσότητα λιπάσματος (που καθορίζει ο ερευνητής) και **Y** η απόδοση του αγρού.

♦ **X** το ύψος της διαφημιστικής δαπάνης ενός προϊόντος (που καθορίζει μια επιχείρηση) και **Y** το ύψος των πωλήσεων του προϊόντος.

Στις περιπτώσεις όπου από τον πληθυσμό επιλέγουμε ένα τυχαίο δείγμα και σε κάθε μονάδα του δείγματος μελετάμε δύο ή περισσότερα χαρακτηριστικά, είναι λογικό, να αναζητήσουμε μέτρα τα οποία να μπορούν να εκφράσουν και να ποσοτικοποιήσουν την πιθανή συμμεταβολή-συσχέτιση των χαρακτηριστικών. Για παράδειγμα, συσχετίζονται-συμμεταβάλλονται **ο μισθός** και **τα έτη σπουδών των εργαζομένων**.

Πώς συμμεταβάλλονται; Δηλαδή, όταν αυξάνονται τα έτη σπουδών, αυξάνεται ο μισθός του εργαζομένου; (μειώνεται μήπως;!). Πόσο ισχυρή είναι η συμμεταβολή των μεταβλητών **έτη σπουδών** και **μισθός**;

Στην γραφική παράσταση της συσχέτισης χρησιμοποιούμε ένα διάγραμμα που στον κάθετο και στον οριζόντιο άξονα βάζουμε τις τιμές των μεταβλητών **X** και **Y** όπως αυτό φαίνεται στο παρακάτω διάγραμμα.



Διάγραμμα 1. Γραφική παράσταση συσχέτισης των μεταβλητών x και y.

Ο βαθμός συσχετίσεως είναι ένας δείκτης που έχει μια αριθμητική έκφραση και συμβολίζεται με **r**. Η τιμή του **r** κυμαίνεται πάντοτε μεταξύ **-1** και **+1**.

1.2 ΔΙΑΓΡΑΜΜΑ ΔΙΑΣΠΟΡΑΣ

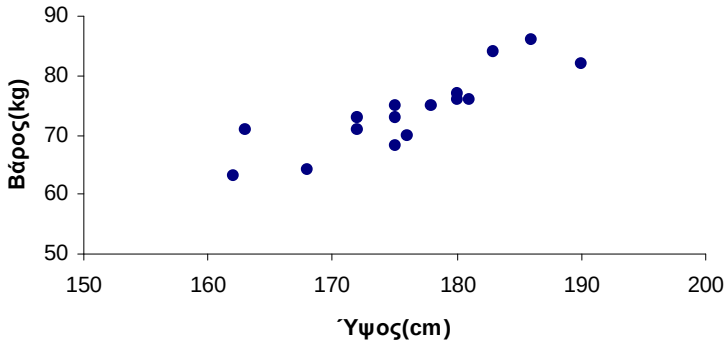
Ένας απλός τρόπος για να αποκτήσουμε μια πρώτη ιδέα για το αν και πώς δυο μεταβλητές συμμεταβάλλονται-συσχετίζονται, είναι να κατασκευάσουμε το **διάγραμμα διασποράς (Scatter Diagram)**. Να αναπαραστήσουμε δηλαδή τα ζεύγη των παρατηρήσεων σε ένα διάγραμμα. Ας δούμε ένα παράδειγμα.

Στον πίνακα που ακολουθεί φαίνονται οι παρατηρήσεις που πήραμε για το ύψος και το βάρος 16 εργατών μιας βιομηχανίας.

	Ύψος (cm)	**Βάρος (kg)**
1	183	84
2	162	63
3	172	71
4	181	76
5	180	77
6	168	64
7	176	70
8	180	76
9	190	82
10	175	68
11	178	75
12	175	73
13	186	86
14	172	73
15	175	75
16	163	71

Πίνακας 1

Από το **διάγραμμα διασποράς** φαίνεται ότι οι εργάτες στο δείγμα που έχουν μεγαλύτερο ύψος έχουν και μεγαλύτερο βάρος. Φαίνεται, δηλαδή, να υπάρχει μια ανάλογη σχέση μεταξύ του ύψους και του βάρους των εργατών.



Διάγραμμα 2:Διάγραμμα διασποράς του ύψους και του βάρους

1.2.1. Συντελεστής Γραμμικής Συσχέτισης του Pearson

Ο ***δειγματικός συντελεστής γραμμικής συσχέτισης του Pearson*** συμβολίζεται με ***r*** και ορίζεται από τον τύπο:
όπου,

$$s_{xy} = Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x}).(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i.y_i - n.\bar{x}.\bar{y}}{n-1}$$

$$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{και} \quad s_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Επομένως,

$$r=\frac{s_{xy}}{s_x.s_y}=\frac{\sum_{i=1}^{n}(x_i-\bar{x}).(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}.\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{\sum_{i=1}^{n}x_i.y_i-n.\bar{x}.\bar{y}}{\sqrt{\sum_{i=1}^{n}x_i^2-n.\bar{x}^2}.\sqrt{\sum_{i=1}^{n}y_i^2-n.\bar{y}^2}}$$

Ο **πληθυσμιακός συντελεστής γραμμικής συσχέτισης του Pearson** ορίζεται ανάλογα και συμβολίζεται με **ρ** .

Ας δούμε δύο αριθμητικά παραδείγματα υπολογισμού του συντελεστή γραμμικής συσχέτισης του **Pearson.**

Παράδειγμα 1

x_i	y_i	$x_i-\bar{x}$	$y_i-\bar{y}$	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x}).(y_i-\bar{y})$
1	4	-2	4	4	16	-8
2	2	-1	2	1	4	-2
3	0	0	0	0	0	0
4	-2	1	-2	1	4	-2
5	-4	2	-4	4	16	-8
$\sum x_i=15$	$\sum y_i=0$			$\sum(x_i-\bar{x})^2=10$	$\sum(y_i-\bar{y})^2=40$	$\sum=(x_i-\bar{x})^2.(y_i-\bar{y})^2=-20$

Πίνακας 2

$$\bar{x}=3,\bar{y}=0$$

$$r=\frac{\sum_{i=1}^{5}(x_i-\bar{x})\ .(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{5}(x_i-\bar{x})^2}.\sqrt{\sum_{i=1}^{5}(y_i-\bar{y})^2}}=\frac{-20}{\sqrt{10}.\sqrt{40}}=-1$$

Παράδειγμα *2*

X_i	Y_i	X_i^2	Y_i^2	$X_i.Y_i$
1	-2	1	4	-2
3	0	9	0	0
5	1	25	1	5
7	3	49	9	21
9	5	81	25	45
10	6	100	36	60
12	8	144	64	96
13	10	169	100	130
$\sum x_i = 60$	$\sum y_i = 31$	$\sum x_i^2 = 578$	$\sum y_i^2 = 239$	$\sum x_i.y_i = 355$

Πίνακας 3

$\overline{x} = 7,5$ και $\overline{y} = 3,9$

$$r = \frac{s_{xy}}{s_x.s_y} = \frac{\sum_{i=1}^{5} x_i.y_i - n.\overline{x}.\overline{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n.\overline{x}^2}.\sqrt{\sum_{i=1}^{n} y_i^2 - n.\overline{y}^2}} = \frac{355 - 8.7,5.3,9}{\sqrt{578 - 8.7,5^2}.\sqrt{239 - 8.3,9^2}} = 0,99$$

<u>Ερμηνεία και ιδιότητες του συντελεστή γραμμικής συσχέτισης *r*</u>

Ο συντελεστής **γραμμικής** συσχέτισης **r** δίνει ένα μέτρο του μεγέθους της **γραμμικής** συσχέτισης μεταξύ δύο μεταβλητών. Παίρνει τιμές στο κλειστό διάστημα **[-1 , 1]**

Για κάθε $\lambda \in R$ ισχύει: $\sum((x_i - \overline{x}) + \lambda(y_i - \overline{y}))^2 \geq 0.$

Άρα, $\sum((x_i - \overline{x})^2 + l^2(y_i - \overline{y})^2 + 2\,l\,(x_i - \overline{x})(y_i - \overline{y})) \geq 0$ ή

$\sum(x_i - \overline{x})^2 + l^2\Sigma(y_i - \overline{y})^2 + 2\,l\,\Sigma(x_i - \overline{x})(y_i - \overline{y})) \geq 0$ ή

$l^2.\sum(y_i - \overline{y})^2 + l.2\ \Sigma(x_i - \overline{x})(y_i - \overline{y}) + \Sigma(x_i - \overline{x})^2 \geq 0.$

Επειδή η τελευταία ανισότητα ισχύει για κάθε $\lambda \in R$, θα είναι $b^2 - 4ag \leq 0$ και άρα

$$4.(\sum (x_i - \bar{x})(y_i - \bar{y}))^2 \leq 4.\sum (y_i - \bar{y})^2 . \sum (x_i - \bar{x})^2 \Leftrightarrow$$

$$\left[\frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (y_i - \bar{y})^2} \sqrt{\Sigma (x_i - \bar{x})^2}} \right]^2 \leq 1 \Leftrightarrow r^2 \leq 1 \Leftrightarrow |r| \leq 1 \Leftrightarrow -1 \leq r \leq 1.$$

- Αν $r = \pm 1$ υπάρχει **τέλεια γραμμική** συσχέτιση.
- Αν $-0,3 \leq r < 0,3$ τότε **δεν υπάρχει γραμμική** συσχέτιση. Αυτό, όμως, δεν σημαίνει ότι δεν υπάρχει άλλου είδους συσχέτιση μεταξύ των δύο μεταβλητών.
- Αν $-0,5 < r \leq -0,3$ ή $0,3 \leq r < 0,5$ υπάρχει **ασθενής γραμμική** συσχέτιση.
- Αν $-0,7 < r \leq -0,5$ ή $0,5 \leq r < 0,7$ υπάρχει **μέση γραμμική** συσχέτιση.
- Αν $-0,8 < r \leq -0,7$ ή $0,7 \leq r < 0,8$ υπάρχει **ισχυρή γραμμική** συσχέτιση.
- Αν $-1 < r \leq -0,8$ ή $0,8 \leq r < 1$ υπάρχει **πολύ ισχυρή γραμμική** συσχέτιση.

Θετικές τιμές του **r** δεν υποδηλώνουν, κατ' ανάγκην μεγαλύτερο βαθμό γραμμικής συσχέτισης από το βαθμό γραμμικής συσχέτισης που υποδηλώνουν αρνητικές τιμές του **r** . Ο βαθμός γραμμικής συσχέτισης καθορίζεται από την απόλυτη τιμή του **r** και όχι από το πρόσημο του **r** . Το πρόσημο του **r** καθορίζει το είδος, μόνο, της συσχέτισης (θετική ή αρνητική). Μας πληροφορεί δηλαδή για το αν η αύξηση της μιας μεταβλητής αντιστοιχεί σε αύξηση ή σε μείωση της άλλη μεταβλητής.

Για παράδειγμα η τιμή **r = −0,9** δείχνει **ισχυρότερη γραμμική συσχέτιση** από την τιμή **r = 0,8** ενώ οι τιμές **r = −0,6** και **r = 0,6** δείχνουν **ίδιο βαθμό γραμμικής συσχέτισης αλλά αντίθετο είδος**.

Στην πράξη, υπολογίζουμε το συντελεστή γραμμικής συσχέτισης στις περιπτώσεις μόνο που το διάγραμμα διασποράς (στικτό διάγραμμα) έχει σχήμα **επιμήκους κεκλιμένης έλλειψης** ή **πλατυσμένου *J***. Αν, όμως, τον υπολογίσουμε και σε περιπτώσεις που το διάγραμμα διασποράς έχει άλλη μορφή, η τιμή του η οποία θα είναι μικρή, δεν συνεπάγεται μη συσχέτιση αλλά μη γραμμική συσχέτιση. Είναι, δηλαδή, δυνατόν να υπάρχει μεγάλη μη γραμμική συσχέτιση.

Ο **συντελεστής γραμμικής συσχέτισης r** χρησιμοποιείται ως μια εκτιμήτρια του πληθυσμιακού συντελεστή γραμμικής συσχέτισης **ρ** , μόνο όταν τα ζεύγη προέρχονται από τυχαία δειγματοληψία. Δεν έχει, επομένως, μεγάλη χρησιμότητα σε πειραματικές έρευνες, όπου οι τιμές της μιας μεταβλητής ελέγχονται-καθορίζονται από τον ερευνητή.

Συσχέτιση δε σημαίνει αιτιότητα. Όταν σε μια **μη** πειραματική έρευνα (δειγματοληψία) δύο μεταβλητές **X** και **Y** βρίσκονται συσχετισμένες αυτό σημαίνει μόνο ότι οι μεταβλητές αυτές συνδέονται με κάποια σχέση. Δε συνεπάγεται, κατ' ανάγκη, **αιτιότητα**. Οι δύο μεταβλητές μπορεί βέβαια να συνδέονται με σχέση αιτιότητας, μπορεί όμως, και όχι. Για παράδειγμα, μπορεί και οι δύο να επηρεάζονται από μια τρίτη μεταβλητή. Ας δούμε δύο παραδείγματα:

1) Παρατηρήθηκε ότι το **ύψος των μαθητών** ενός σχολείου, ηλικίας 6 έως 13 ετών, έχει ισχυρή θετική γραμμική συσχέτιση με την **αντιληπτική ικανότητα των μαθητών**. Προφανώς η αντιληπτική ικανότητα των μαθητών δεν επηρεάζεται από το ύψος τους. Απλώς τόσο η πνευματική όσο και η φυσική ανάπτυξη των μικρών μαθητών επηρεάζονται παράλληλα από άλλους παράγοντες.

2) Παρατηρήθηκε ότι οι **πωλήσεις ταχύπλοων στο Σίδνεϋ** είχαν, για μια μακρά περίοδο, ισχυρή θετική συσχέτιση με τις **πωλήσεις έγχρωμων τηλεοράσεων στη Μελβούρνη**. Προφανώς, τόσο οι πωλήσεις ταχύπλοων όσο και οι πωλήσεις έγχρωμων τηλεοράσεων ήταν συνάρτηση γενικότερων ευνοϊκών οικονομικών παραγόντων. Είναι, κατά συνέπεια, φανερό ότι η πρόχειρη ή επιπόλαιη ερμηνεία και χρήση του **r** οδηγεί πολλές φορές σε παρερμηνείες ή και σε λανθασμένα συμπεράσματα. Για αιτιολογικά συμπεράσματα, σχεδόν πάντοτε, απαιτείται πειραματισμός. Σε κάθε περίπτωση, αιτιώδη σχέση (αλληλεξάρτηση) μεταξύ δύο μεταβλητών δεχόμαστε μόνον όταν υπάρχει επιστημονική ή λογική βάση που την υπαγορεύει.

• Με **x** ***y s*** συμβολίζουμε τη ***δειγματική συνδιασπορά*** των μεταβλητών **X** και **Y**. Η ***πληθυσμιακή συνδιασπορά*** ορίζεται ανάλογα και συμβολίζεται με ***x y*** **σ** . Εκφράζει τη συμμεταβολή-συσχέτιση δύο μεταβλητών μέσω του αθροίσματος των γινομένων των αποκλίσεων των τιμών τους από τους αντίστοιχους μέσους. Μεγάλες τιμές της υποδηλώνουν ότι υπάρχει συμμεταβολή-συσχέτιση ενώ μικρές τιμές της υποδηλώνουν ότι δεν υπάρχει συμμεταβολή-συσχέτιση. Όμως, δε χρησιμοποιείται ως μέτρο συσχέτισης δύο μεταβλητών διότι επηρεάζεται από τις μονάδες στις οποίες εκφράζονται οι μεταβλητές.

1.2.2 Πίνακας Συσχετίσεων *R*

• Ο πίνακας συσχετίσεων είναι ο πίνακας που περιέχει σαν στοιχεία του τους συντελεστές συσχέτισης του **Pearson** για κάθε ζευγάρι μεταβλητών.

• Ο συντελεστής συσχέτισης του **Pearson** μετράει μόνο **γραμμική συσχέτιση** ανάμεσα στις μεταβλητές και επομένως δεν μπορεί να μας δώσει πληροφορία για άλλης μορφής συσχέτιση.

• Ο συντελεστής συσχέτισης του **Pearson** είναι κατάλληλος μόνο για ζεύγη **ποσοτικών** μεταβλητών.

$$R = \begin{pmatrix} 1 & r_{12} & \mathbf{K} & r_{1p} \\ r_{21} & 1 & \mathbf{K} & r_{2p} \\ \mathbf{K} & \mathbf{K} & \mathbf{K} & \\ r_{p1} & r_{p2} & \mathbf{K} & 1 \end{pmatrix} \textit{όπου}$$

$$r_{jk} = \frac{S_{jk}}{S_j . S_k} = \frac{S_{jk}}{\sqrt{S_{jj}} . \sqrt{S_{kk}}}, \; j,k = 1,2,\ldots, p$$

• Ο πίνακας έχει απαραίτητα τιμές ίσες με τη μονάδα στη διαγώνιο, είναι συμμετρικός και κανένα στοιχείο του **δεν μπορεί** να πάρει τιμή μεγαλύτερη σε απόλυτη τιμή από το 1.

• Τιμές **–1** και **1** σημαίνουν **απόλυτα γραμμική σχέση** των δύο μεταβλητών. Το πρόσημο υποδηλώνει την ύπαρξη **θετικής** ή **αρνητικής** σχέσης.

• Η θετική σχέση ερμηνεύεται ως εξής: Όσο αυξάνεται η τιμή της μιας μεταβλητής τόσο αυξάνεται και η τιμή της άλλης ενώ στην αρνητική σχέση όσο αυξάνεται η τιμή της μιας μεταβλητής μειώνεται η τιμής της άλλης.

Ο πίνακας διακυμάνσεων-συνδιακυμάνσεων **S** τυποποιημένων μεταβλητών ταυτίζεται με τον πίνακα συσχετίσεων ***R*** των αρχικών μεταβλητών πριν την τυποποίηση τους.

$$r_{jk} = \text{cov}(X*_j, X*_k)$$

Όπου $X^*_{j,} X^*_{k,}$, είναι οι τυποποιημένες μεταβλητές.

1.2.3 Στατιστική σημαντικότητα του *r*

Η ελεγχοσυνάρτηση που ελέγχει σε ένα δείγμα μεγέθους ***n*** αν η συσχέτιση του πληθυσμού είναι στατιστικά σημαντικά διαφορετική από το ***0*** είναι η εξής:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Αυτή η συνάρτηση ακολουθεί κάτω από τη μηδενική υπόθεση (και την υπόθεση πως τα δεδομένα προέρχονται από διμεταβλητή κανονική κατανομή) κατανομή ***t*** με ***n-2*** βαθμούς ελευθερίας.

Συνεπώς σε επίπεδο στατιστικής σημαντικότητας 5% απορρίπτω τη μηδενική υπόθεση αν

$$\left|\frac{r\sqrt{n-2}}{\sqrt{1-r^2}}\right| \geq t_{n-2,1-a/2} \quad \text{ή} \quad t^2 \geq \frac{t^2_{n-2,1-a/2}}{n-2+t^2_{n-2,1-a/2}}$$

όπου ***tn,1-a/2*** είναι το ***1-a/2*** ποσοστιαίο σημείο της κατανομής ***t*** με ***n*** βαθμούς ελευθερίας.

Μπορούμε συνεπώς να υπολογίσουμε για κάθε μέγεθος δείγματος ποια είναι η τιμή πάνω από την οποία απορρίπτουμε τη μηδενική υπόθεση περί μηδενικής συσχέτισης.

Μέγεθος δείγματος	5	10	15	20	30	50	100	200	500	1000
Τιμή συντελεστή	0,878	0,834	0,614	0,443	0,381	0,278	0,188	0,138	0,088	0,081

Πίνακας 4

Ακόμα και με σχετικά μικρά δείγματα μικρές συσχετίσεις είναι στατιστικά σημαντικές αν και ουσιαστικά αδιάφορες από στατιστικής απόψεως.

• Ο συντελεστής συσχέτισης σχετίζεται με τον συντελεστή προσδιορισμού της γραμμικής παλινδρόμησης (***r2 = R2)***. Έτσι μια συσχέτιση της τάξης του 0,20, (που για δείγμα μεγέθους 100 είναι στατιστικά σημαντική) σημαίνει πως ο συντελεστής προσδιορισμού σε μια παλινδρόμηση ανάμεσα στις δύο μεταβλητές θα είναι 4%, δηλαδή πάρα πολύ μικρός για οποιαδήποτε στατιστική χρήση.

• Συνεπώς μας ενδιαφέρουν μεγάλες σε απόλυτη τιμή συσχετίσεις και όχι απαραίτητα στατιστικά σημαντικές συσχετίσεις.

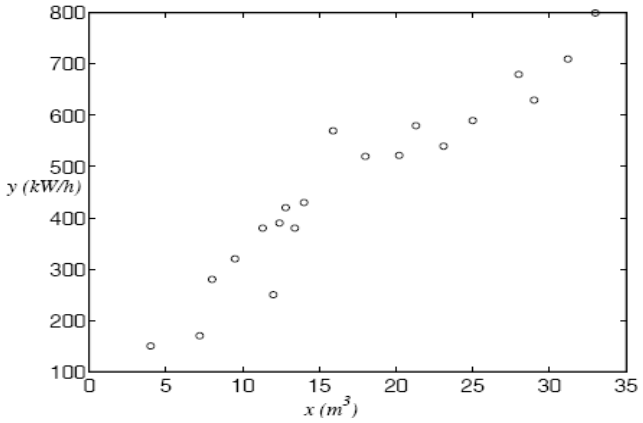
1.3 Παράδειγμα για τη συσχέτιση (κατανάλωση νερού και ρεύματος)

Θέλουμε να διερευνήσουμε τη συσχέτιση της κατανάλωσης νερού και κατανάλωσης ρεύματος νοικοκυριού και γι αυτό πήραμε τις μηνιαίες μετρήσεις της κατανάλωσης νερού και ρεύματος για κάποιο μήνα από 20 νοικοκυριά. Τα δεδομένα παρουσιάζονται στον Πίνακα 5.

Α/Α	**Νερό** *xi* **(**m^3**)**	**Ρεύμα** *yi* **(Kw/h)**
1	4.0	150
2	7.2	170
3	8.0	280
4	9.5	320
5	11.3	380
6	12.0	250
7	12.4	390
8	12.8	420
9	13.4	380
10	14.0	430
11	15.9	570
12	18.0	520
13	20.2	522
14	21.3	580
15	23.1	540
16	25.0	590
17	28.0	680
18	29.0	630
19	31.2	710
20	33.0	800

Πίνακας 5: Δεδομένα κατανάλωσης νερού (*xi*) και κατα νάλωσης ρεύματος (*yi*) από 20 νοικοκυριά.

Ποιοτικά μπορούμε να δούμε την συσχέτιση της κατανάλωσης νερού και ρεύματος από το διάγραμμα διασποράς στο διάγραμμα 3.



Διάγραμμα 3:Διάγραμμα διασποράς για την κατανάλωση νερού (*x*) και την κατανάλωση ρεύματος (*y*) από τα δεδομένα του Πίνακα 3.

Από το διάγραμμα διασποράς συμπεραίνουμε πως η συσχέτιση της κατανάλωσης νερού και ρεύματος είναι θετική και ισχυρή. Για να βρούμε τον συντελεστή συσχέτισης **r** υπολογίζουμε πρώτα τα παρακάτω:

$$\bar{x} = 17,465 \qquad \bar{y} = 465,6$$

$$\sum_{i=1}^{20} x_i^2 = 7471,53 \qquad \sum_{i=1}^{20} y_i^2 = 4944184 \qquad \sum_{i=1}^{20} x_i y_i = 190129,4$$

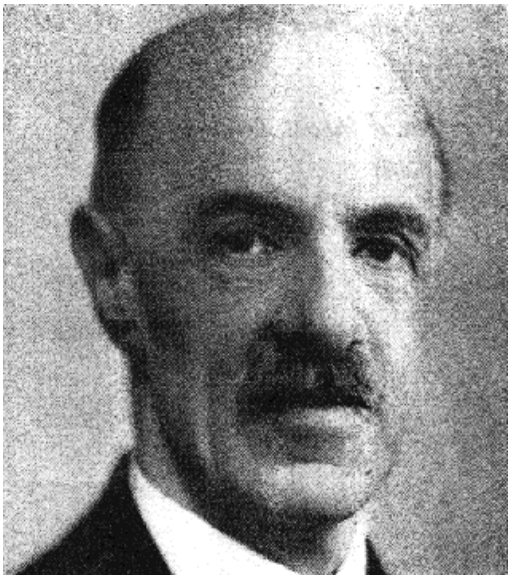
$$r = \frac{190129,4 - 20.17,465.465,6}{\sqrt{(7471,53 - 20.17,465^2).(4944184 - 20.465,6^2)}} = 0,952$$

Η τιμή του δειγματικού συντελεστή συσχέτισης δηλώνει επίσης την ισχυρή θετική συσχέτιση. Η μεταβλητότητα της μιας τυχαίας μεταβλητής. (κατανάλωση νερού ή ρεύματος) μπορεί να εξηγηθεί από τη συσχέτιση της με την άλλη κατά ποσοστό που δίνεται από το συντελεστή προσδιορισμού και είναι $r^2.100 = 100.0,952^2 = 90,6\%$.Συμπεραίνουμε λοιπόν πως η γνώση της μιας τυχαίας μεταβλητής. μας επιτρέπει να προσδιορίσουμε την άλλη με μεγάλη ακρίβεια.

Πρέπει επίσης να σημειωθεί ότι η εκτίμηση *r* του συντελεστή συσχέτισης μπορεί να αλλάξει σημαντικά με την πρόσθεση ή αφαίρεση λίγων παρατηρήσεων γιατί το μέγεθος του δείγματος είναι μικρό.

Για τον συντελεστή συσχέτισης **ρ** μπορούμε να υπολογίσουμε διαστήματα εμπιστοσύνης κάτω από την υπόθεση ότι η κοινή κατανομή των **X** και **Y** είναι κανονική.

ΚΕΦΑΛΑΙΟ 2^{Ο}
ΑΛΛΟΙ ΣΥΝΤΕΛΕΣΤΕΣ ΣΥΣΧΕΤΙΣΗΣ



Charles Edward Spearman

Maurice Kendall

2.1 Ο Συντελεστής Spearman rho

Ο συντελεστής συσχέτισης του **Spearman (rho)** Δίνει το μέγεθος της γραμμικής συσχέτισης **ποιοτικών μεταβλητών διάταξης**.

$$rho = 1 - \frac{6.\sum_{i=1}^{n} d_i^2}{n.(n^2 - 1)}$$

όπου,

Ν το μέγεθος του δείγματος και $d_i = x_i - y_{i,}\quad i = 1,2,...,n$

Παράδειγμα:

Φοιτητές i	**Σειρά κατάταξης στη Στατιστική (Χ)**	**Σειρά κατάταξης στα Μαθηματικά (Υ)**	d_i	d_i^2
Α	1ος	4ος	-3	9
Β	2ος	2ος	0	9
Γ	3ος	3ος	0	9
Δ	4ος	5ος	-1	1
Ε	5ος	1ος	4	16
ΣΤ	6ος	6ος	0	0
Ζ	7ος	8ος	-1	1
Η	8ος	7ος	1	1
				$\sum d_i^2 = 28$

Πίνακας 6

$$rho = 1 - \frac{6.28}{8.(8^2 - 1)} = 0,667$$

2.1.1 Ιδιότητες-χρήσεις του Συντελεστή Γραμμικής Συσχέτισης *rho*

• Παίρνει τιμές στο κλειστό διάστημα [-1 , 1]. Αν συμφωνούν πλήρως οι δύο κατατάξεις είναι ***rho* = 1**, ενώ όταν η μια διάταξη είναι ριζικά διαφορετική από την άλλη (για παράδειγμα, αν το μέγεθος δείγματος είναι 8,

τότε το **Χ είναι 1 όταν το Υ είναι 8**, το **Χ είναι 2 όταν το Υ είναι 7, κ.ο.κ*)* είναι ***rho* = −1**. Η τιμή 0 δείχνει το μικρότερο βαθμό συσχέτισης.

• Αν στην κατάταξη έχουμε ισοβαθμίες, δίνουμε, ως θέση, σε όλες τις θέσεις που ισοβαθμούν, τη μέση τιμή τους. Για παράδειγμα, αν η βαθμολογία οκτώ φοιτητών στα Μαθηματικά είναι: 10, 9, 9, 8, 7, 6, 6, 6 τότε η κατάταξη γίνεται ως εξής:

Φοιτητές i	**Βαθμός στα Μαθηματικά**	**Κατάταξη**
Α	10	1
Β	9	2,5
Γ	9	2,5
Δ	8	4
Ε	7	5
ΣΤ	6	7
Ζ	6	7
Η	6	7

Πίνακας 7

• Όταν υπάρχουν πολλές ισοβαθμίες ο συντελεστής **rho** δεν είναι αξιόπιστος. Σε αυτή την περίπτωση ενδείκνυται ο δείκτης **Kendall W.**

2.2 Ο Συντελεστής Συσχέτισης τ Του Kendall

Ο συντελεστής συσχέτισης **τ** του **Kendall** μοιάζει με τον συντελεστή ρ του **Spearman** ως προς το ότι υπολογίζεται με βάση την τάξη μεγέθους των παρατηρήσεων και όχι με βάση τις παρατηρήσεις αυτές καθαυτές και, επιπλέον, η κατανομή του δεν εξαρτάται από την κατανομή των μεταβλητών **Χ** και **Υ**, όταν αυτές είναι ανεξάρτητες και συνεχείς. Το κύριο πλεονέκτημα του μέτρου αυτού σε σχέση με το μέτρο ρ του **Spearman** είναι ότι τείνει στην κανονική κατανομή σχετικά γρήγορα.

(Πηγή www.stat-athens.aueb.gr/gr/prop/notes/np342.pdf)

Αποτέλεσμα αυτού είναι ότι η προσέγγιση της κατανομής του συντελεστή τ από την κανονική κατανομή είναι καλύτερη από την αντίστοιχη προσέγγιση της κατανομής του συντελεστή **ρ** του **Spearman**, όταν αληθεύει η μηδενική υπόθεση της ανεξαρτησίας μεταξύ των μεταβλητών **X** και **Y**. Ένα άλλο πλεονέκτημα του συντελεστή **τ** του **Kendall** βρίσκεται στο γεγονός ότι μπορεί άμεσα και απλά να ερμηνευθεί μέσω των πιθανοτήτων με τις οποίες παρατηρούμε **εναρμονισμένα ή συσχετισμένα (concordant)** ζεύγη τιμών και **μη εναρμονισμένα ή μη συσχετισμένα (discordant)** ζεύγη τιμών, όπως αυτά ορίζονται στην συνέχεια. Τα δεδομένα αποτελούνται από ένα διμεταβλητό τυχαίο δείγμα μεγέθους n παρατηρήσεων X_i, Y_i), i = 1, 2, ..., n, πάνω στο τυχαίο διάνυσμα. Δύο παρατηρήσεις, έστω (X_j, Y_j) και (X_K, Y_K), ονομάζονται **εναρμονισμένες ή συσχετισμένες (concordant)**, αν και τα δύο μέλη της μίας παρατήρησης είναι μεγαλύτερα (ή μικρότερα) από τα αντίστοιχα μέλη της άλλης παρατήρησης. Οι **παρατηρήσεις** (X_j, Y_j) και (X_K, Y_K) θα ονομάζονται **μη εναρμονισμένες ή μη συσχετισμένες (discordant),** αν η διάταξη των πρώτων μελών τους είναι αντίθετη από την διάταξη των δεύτερων μελών τους.

Ισοδύναμα, δύο ζεύγη παρατηρήσεων (X_j, Y_j) και (X_K, Y_K) θα ονομάζονται **εναρμονισμένα** αν οι διαφορές $X_j - X_K$ και $Y_j - Y_K$ έχουν το ίδιο πρόσημο (αν $(X_j - X_K), (Y_j - Y_K) > 0$). Τα ζεύγη (X_j, Y_j) και (X_K, Y_K) θα ονομάζονται **μη εναρμονισμένα** αν οι διαφορές $X_j - X_K$ και $Y_j - Y_K$ έχουν αντίθετο πρόσημο . (αν $(X_j - X_K)(Y_j - Y_K) < 0$).

Έστω Nc και Nd οι αριθμοί των εναρμονισμένων και μη εναρμονισμένων ζευγών παρατηρήσεων, αντίστοιχα. Τα ζεύγη των παρατηρή-

σεων (X_j, Y_j) και (X_K, Y_K), για τα οποία ισχύει ότι $X_j = X_K$ ή/και $Y_j = Y_K$,

δεν είναι ούτε εναρμονισμένα ούτε μη εναρμονισμένα. Τα ζεύγη αυτά ονομάζονται **ισοβαθμούντα (tied)**.

Έστω N_0 ο αριθμός των ισοβαθμούντων ζευγών παρατηρήσεων Επειδή οι n παρατηρήσεις μπορούν να συνδυασθούν ανά δύο με

$$\binom{n}{2} = n(n-1)/2$$

$$N_c + N_d + N_0 = \binom{n}{2}$$

Τα δεδομένα μπορούν, επίσης, να αποτελούνται από μη αριθμητικές παρατηρήσεις, οι οποίες εμφανίζονται κατά **n** ζεύγη, με την προϋπόθεση ότι οι παρατηρήσεις αυτές είναι τέτοιες, ώστε μπορούν να ορισθούν εναρμονισμένα και μη εναρμονισμένα ζεύγη παρατηρήσεων και να είναι δυνατός ο υπολογισμός των αριθμών **Nc** και **Nd**.

Το μέτρο συσχέτισης που προτάθηκε από τον **Kendall** το 1938 ορίζεται ως εξής Ο συντελεστής **τ**, δηλαδή, παριστάνει την διαφορά μεταξύ των ποσοστών των εναρμονισμένων και μη εναρμονισμένων ζευγών παρατηρήσεων.

Αν όλα τα ζεύγη παρατηρήσεων είναι εναρμονισμένα, τότε ο συντελεστής **τ** είναι ίσος με **1**. Αν όλα τα ζεύγη είναι μη εναρμονισμένα, τότε η τιμή του συντελεστή **τ** είναι **–1**. Είναι, δηλαδή, οι τιμές του συντελεστή **τ** μεταξύ **–1 και 1**. Επιπλέον, ο συντελεστής τα ικανοποιεί όλες τις προϋποθέσεις που προαναφέρθηκαν.

Ο υπολογισμός του συντελεστή **τ** γίνεται απλούστερος, αν οι παρατηρήσεις (X_i, Y_i), i = 1, 2, ..., n διαταχθούν σε μία στήλη κατά αύξουσα τάξη μεγέθους των τιμών των παρατηρήσεων πάνω στην τυχαία μεταβλητή **X**. Τότε, κάθε **Y** τιμή χρειάζεται να συγκριθεί μόνο με τις **Y** τιμές που είναι **"κάτω"** από αυτήν.

Έτσι, κάθε ζεύγος παρατηρήσεων εξετάζεται μόνο μία φορά και ο αριθμός των συσχετισμένων και μη συσχετισμένων ζευγών προσδιορίζεται γρηγορότερα.

Παράδειγμα:

Ας θεωρήσουμε τα δεδομένα πάνω στην επιθετικότητα των διδύμων. Διατάσσοντας τις παρατηρήσεις (Xi, Yi), i = 1, 2, ..., n κατά αύξουσα τάξη μεγέθους των τιμών των παρατηρήσεων Xi , i = 1, 2, ..., n, καταλήγουμε στον πίνακα 8.

	$(X^{(i)}, Y_i^*)$	Εναρμονισμένα ζεύγη κάτω από το $(X^{(i)}, Y_i^*)$	Μη εναρμονισμένα ζεύγη κάτω από το $(X^{(i)}, Y_i^*)$
	(68,64)	11	0
	(70,65)	9	0
ισοβαθμία {	(71,77)	4	4
	(71,80)	4	4
	(72,72)	5	1
ισοβαθμία {	(77,65)	5	0
	(77,76)	4	1
	(86,88)	2	2
	(87,72)	3	0
	(88,81)	2	0
ισοβαθμία {	(91,90)	0	0
	(91,96)	0	0
	Σύνολο	$N_c = 49$	$N_d = 12$

Πίνακας 8

Εδώ $X^{(i)}$, i = 1,…, n είναι η διατεταγμένη ακολουθία των παρατηρήσεων Xi, και Y_i^*, i = 1, …, n η προκύπτουσα αναδιάταξη των αντιστοιχών σ’ αυτές τιμών των Yi. Η δεύτερη στήλη του πίνακα δίνει, τον αριθμό των ζευγών ($X^{(i+1)}, Y_{i+1}^*$) για τα οποία $Y_i^* < Y_{i+1}^*$ όταν $X^{(i)} < X^{(i+1)}$,

(i = 1, …, n-1). Η τρίτη στήλη δίνει τον αριθμό των ζευγών ($X^{(i+1)}, Y^{*}_{i+1}$) για τα οποία $Y^{*}_{i} < Y^{*}_{i+1}$ όταν $X^{(i)} < X^{(i+1)}$ (i = 1, …, n-1).

Με βάση τα στοιχεία του πίνακα, προκύπτει ότι η τιμή του συντελεστή συσχέτισης **τ** του **Kendall** είναι:

$$t = \frac{N_C - N_d}{n(n-1)/2} = \frac{49-12}{(12)(11)/2} = 0,5606$$

Υπάρχει, επομένως, θετική συσχέτιση τάξης μεγέθους μεταξύ των μετρήσεων της επιθετικότητας των διδύμων, όπως προκύπτει από την μέτρηση του συντελεστή συσχέτισης τ του **Kendall**.

Ο συντελεστής **τ** μπορεί, επίσης, να χρησιμοποιηθεί ως ελεγχοσυνάρτηση για τον έλεγχο της μηδενικής υπόθεσης της ανεξαρτησίας μεταξύ των τυχαίων μεταβλητών **X** και **Y**, με αμφίπλευρες ή μονόπλευρες εναλλακτικές, όπως εξάλλου και στην περίπτωση του συντελεστή συσχέτισης ρ του Spearman. Περισσότερο συχνή, όμως, είναι η χρήση της διαφοράς **Nc – Nd** ως ελεγχοσυνάρτησης για τον έλεγχο των υποθέσεων αυτών. Χρησιμοποιούμε, δηλαδή, ως ελεγχοσυνάρτηση την στατιστική συνάρτηση **T = Nc – Nd**, την οποία ονομάζουμε ελεγχοσυνάρτηση του Kendall.

Είναι προφανές ότι μεγάλες τιμές της στατιστικής συνάρτησης **T** αποτελούν ένδειξη εναντίον της υπόθεσης H_0 και υπέρ της μονόπλευρης εναλλακτικής υπόθεσης θετικής συσχέτισης μεταξύ των μεταβλητών **X** και **Y**. Αντίστοιχα, μικρές τιμές της στατιστικής συνάρτησης **T** αποτελούν ένδειξη εναντίον της υπόθεσης H_0 και υπέρ της μονόπλευρης εναλλακτικής υπόθεσης αρνητικής συσχέτισης μεταξύ των μεταβλητών X και Y. Επομένως, θεωρώντας την ταξινόμηση των ζευγών υποθέσεων που θα μπορούσαν να μας ενδιαφέρουν στις κατηγορίες,

I. (Αμφίπλευρος έλεγχος συσχέτισης)

II. (Μονόπλευρος έλεγχος θετικής συσχέτισης)

III. (Μονόπλευρος έλεγχος αρνητικής συσχέτισης),

ο κανόνας απόφασης διαμορφώνεται ως εξής:

Α. Η υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α, αν $T > w_{1-a/2}$ ή αν $T < w_{a/2} = -w_{1-a/2}$.

Β. Η υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α, αν $T > w_{1-a}$

Γ. Η υπόθεση H_0 απορρίπτεται σε επίπεδο σημαντικότητας α, αν $T < w_a = -w_{1-a}$

Επιστρέφοντας στο παράδειγμά μας, ας υποθέσουμε ότι ενδιαφερόμαστε να ελέγξουμε την υπόθεση H_0: οι μεταβλητές **X** και **Y** είναι ασυσχέτιστες, εναντίον της εναλλακτικής H_1: οι μεταβλητές **X** και **Y** είναι συσχετισμένες (αμφίπλευρος έλεγχος).

Από τον πίνακα που κατασκευάσθηκε παραπάνω, έχουμε ότι η παρατηρούμενη τιμή της στατιστικής συνάρτησης **T** είναι:

$$Nc - Nd = 49 - 12 = 37.$$

Από τον σχετικό πίνακα του παραρτήματος, προκύπτει ότι τα ποσοστιαία σημεία για έναν αμφίπλευρο έλεγχο μεγέθους α=0.05, για n=12, είναι $w_{0,975} = 28$ και $w_{0,025} = -w_{0,975} = -28$. Επομένως, η τιμή της στατιστικής συνάρτησης T υπερβαίνει το 0.975-ποσοστιαίο σημείο της κατανομής της και, κατά συνέπεια, η μηδενική υπόθεση ότι δεν υπάρχει συσχέτιση μεταξύ των **X** και **Y** απορρίπτεται σε επίπεδο σημαντικότητας 0.05. Το κρίσιμο επίπεδο του ελέγχου αυτού είναι όπως φαίνεται από τον σχετικό πίνακα του παραρτήματος, περίπου ίσο με $\hat{a} \cong 2(0,005) = 0,01$.

Παρατήρηση: Η ακριβής κατανομή των στατιστικών συναρτήσεων ρ και τ είναι εύκολο να προσδιορισθεί, αν και, στην πράξη, η διαδικασία είναι πολύ χρονοβόρα ακόμα και για μέτριο μέγεθος δείγματος **n**. Και οι δύο κατανομές μπορούν να προσδιορισθούν κάτω από την υπόθεση ότι οι μεταβλητές **Xi** και **Yi** είναι ανεξάρτητες και ισόνομες. Τότε, οι n! διατάξεις των βαθμών (τάξεων μεγέθους) των μεταβλητών **Xi**, θεωρουμένων κατά ζεύγη με τους αντίστοιχους βαθμούς των μεταβλητών **Yi**, είναι ισοπίθανες. Οι συναρτήσεις κατανομής προκύπτουν με απλή απαρίθμηση του αριθμού των διατάξεων, οι οποίες οδηγούν σε μία συγκεκριμένη τιμή του ρ ή του τ και με διαίρεση του αριθμού αυτού των διατάξεων με n! για να προσδιορισθεί η πιθανότητα της συγκεκριμένης τιμής του **ρ** ή του **τ**. (πηγή *www.stat-athens.aueb.gr/gr/prop/notes/np342.pd)*

Επειδή, τόσο η στατιστική συνάρτηση **ρ** όσο και η **τ** αποτελούν αθροίσματα τυχαίων μεταβλητών, μπορεί κανείς να χρησιμοποιήσει μια μορφή του κεντρικού οριακού θεωρήματος για να προσεγγίσει κατανομές τους στην περίπτωση μεγάλων δειγμάτων. Και οι δύο συντελεστές έχουν συμμετρικές κατανομές γύρω από το μηδέν και, επομένως, έχουν και οι δύο μέση τιμή ίση με το μηδέν. Οι διασπορές των στατιστικών αυτών συναρτήσεων είναι δυσκολότερο να προσδιορισθούν. Διαίρεση, επομένως, των στατιστικών συναρτήσεων **ρ** και **τ** με τις αντίστοιχες διασπορές τους οδηγεί σε τυχαίες μεταβλητές, οι οποίες κατά προσέγγιση έχουν την κανονική κατανομή για μεγάλες τιμές του **n**. Όπως αναφέρθηκε και προηγουμένως, η προσέγγιση αυτή είναι καλύτερη στην περίπτωση της στατιστικής συνάρτησης **τ** για **n ≥ 8**. Δεν είναι, όμως, εξ ίσου ικανοποιητική όταν χρησιμοποιείται για τον κατά προσέγγιση προσδιορισμό των ποσοστιαίων σημείων της κατανομής της στατιστικής συνάρτησης **ρ**.

Στην περίπτωση που τα ζεύγη $(X_{i,}Y_i),\ i=1,2,...,n$ είναι ανεξάρτητες και ισόνομες δισδιάστατες κανονικές μεταβλητές, και οι δύο συντελεστές έχουν ασυμπτωτική σχετική αποτελεσματικότητα ίση με $9/p^2=0,912$, σε σχέση με τον παραμετρικό έλεγχο που χρησιμοποιεί τον συντελεστή **r** του Pearson ως στατιστική συνάρτηση (Stuart, 1954).

2.3 Ο συντελεστής biserial

Ο συντελεστής **biserial** είναι ένα μέτρο της ένωσης μεταξύ μιας συνεχούς μεταβλητής και δυαδικής μεταβλητής. Περιορίζεται για να είναι μεταξύ **-1 και + 1**.

Υποθέτουμε ότι το **X** είναι μια συνεχής μεταβλητή και το **Y** είναι κατηγορικό με τις τιμές 0 και 1. υπολογίζει το συντελεστή χρησιμοποιώντας τον τύπο

$$r=\frac{(\bar{x}_1-\bar{x}_0)\sqrt{p(1-p)}}{S_X}$$

Αυτό είναι από μαθηματική άποψη ισοδύναμο με τον παραδοσιακό τύπο συσχετισμού. Η ερμηνεία είναι παρόμοια. Ο συντελεστής είναι θετικός όταν συνδέονται οι μεγάλες τιμές του **X** με **Y=1** και οι μικρές τιμές του **X** συνδέονται με **Y=0**.

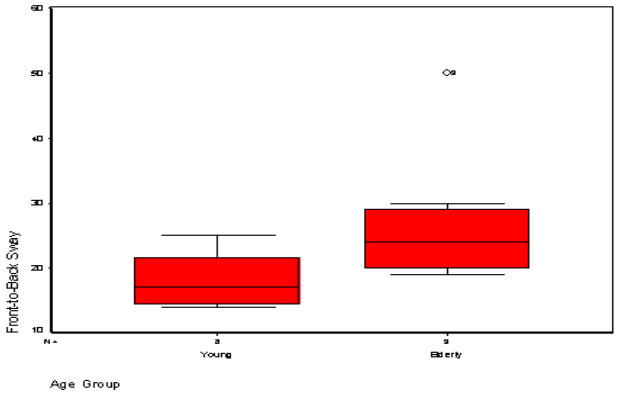
Παράδειγμα

Το FB αντιπροσωπεύει τη στάση ταλάντευση στην μπροστινός-οπίσθια κατεύθυνση και είναι συνεχές. Το SS αντιπροσωπεύει τη στάση ταλάντευση στη δευτερεύων-δευτερεύουσα κατεύθυνση και είναι επίσης συνεχές. AGE_GRP αντιπροσωπεύει την ομάδα ηλικίας (0=Young, 1=Elderly) και είναι δυαδικό.

	FB	SS	AGE
FB	1.00	0.70	0.49
SS	0.70	1.00	0.43
AGE	0.49	0.43	1.00

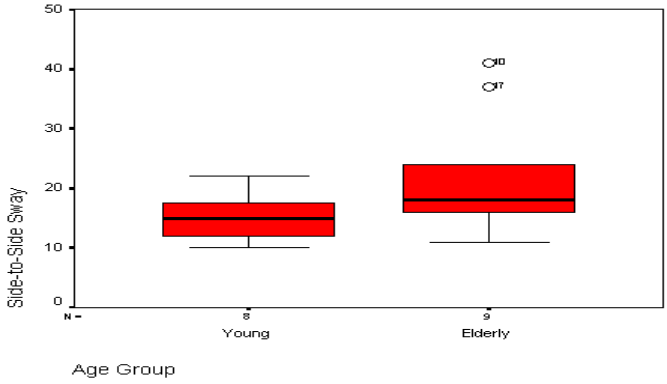
Πίνακας 9

Αυτό είναι ένα boxplot της ταλάντευσης FB για κάθε ομάδα ηλικίας.



Διάγραμμα 4

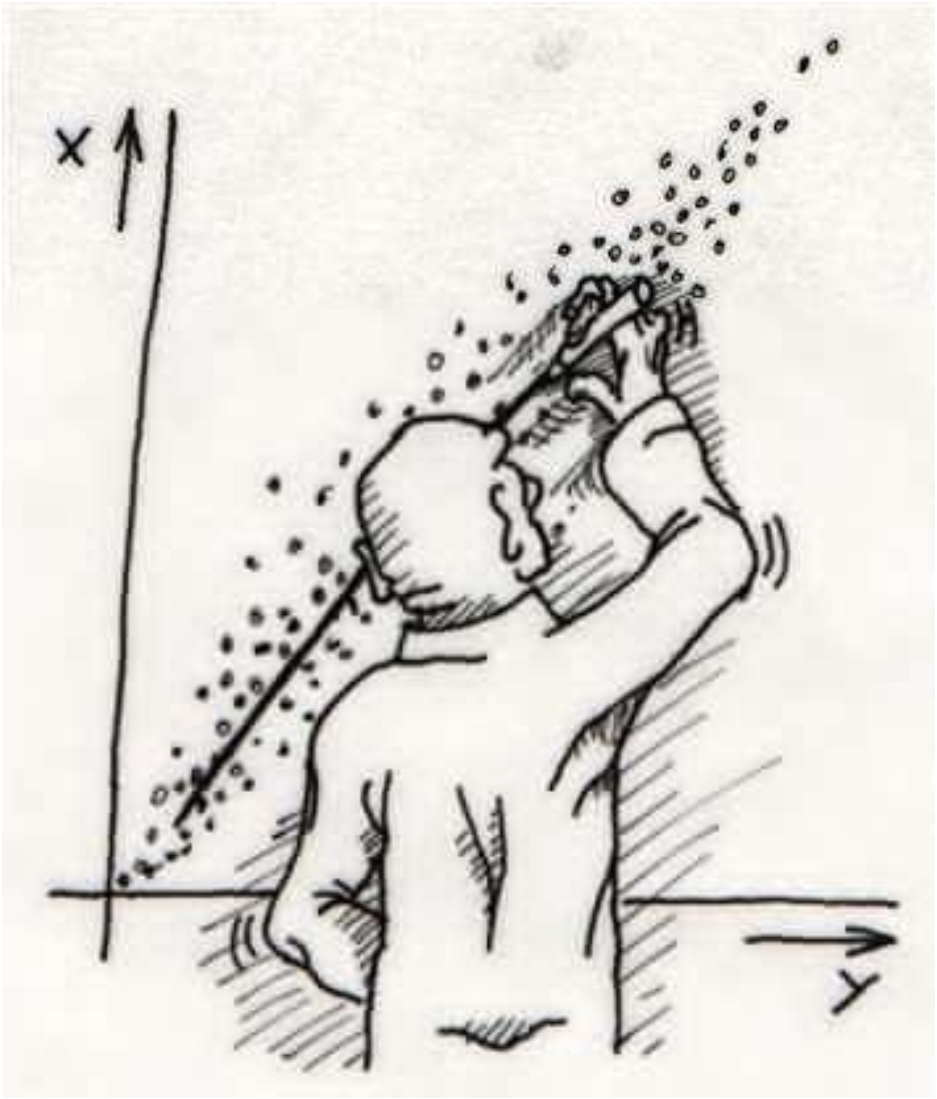
Το παρακάτω μας δείχνει την ταλάντευση SS για κάθε ομάδα ηλικίας. Ειδοποίηση και για αυτό και την προηγούμενη γραφική παράσταση ότι η ηλικιωμένη ομάδα ηλικίας τείνει να έχει τα υψηλότερα αποτελέσματα ταλάντευσης από τη νέα ομάδα Ακόμα κι έτσι, εξακολουθεί να υπάρχει ένα μεγάλο ποσό επικάλυψης μεταξύ αυτών των ομάδων κι αυτός είναι ο λόγος για τον οποίο υπάρχουν οι συντελεστές **biserial**.



Διάγραμμα 5

(Πηγή:www.cmh.edu)

ΚΕΦΑΛΑΙΟ 3^{Ο}
ΠΑΛΙΝΔΡΟΜΗΣΗ



3.1 Εισαγωγή

Με την **ανάλυση παλινδρόμησης (regression analysis)** εξετάζουμε τη σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών με σκοπό την πρόβλεψη των τιμών της μιας, μέσω των τιμών της άλλης (ή των άλλων). Σε κάθε πρόβλημα παλινδρόμησης διακρίνουμε δύο είδη μεταβλητών: τις **ανεξάρτητες ή ελεγχόμενες ή επεξηγηματικές** (independent, predictor, casual, input, explanatory variables) και τις **εξαρτημένες ή απόκρισης** (dependent, response variables). Σε πειραματικές έρευνες, **ανεξάρτητη μεταβλητή X** είναι εκείνη την οποία μπορούμε να ελέγξουμε, δηλαδή, να καθορίσουμε τις τιμές της (π.χ. το ύψος της διαφημιστικής δαπάνης ενός προϊόντος, ο αριθμός των λειτουργούντων ταμείων σε ένα υποκατάστημα τραπέζης, η ποσότητα λιπάσματος, η θερμοκρασία επεξεργασίας ενός προϊόντος). **Εξαρτημένη μεταβλητή Y** είναι εκείνη στην οποία αντανακλάται το αποτέλεσμα των μεταβολών στις ανεξάρτητες μεταβλητές (π.χ. η ζήτηση ενός προϊόντος, ο χρόνος αναμονής των πελατών ενός υποκαταστήματος τραπέζης, η απόδοση μιας καλλιέργειας, η αντοχή ενός υλικού).

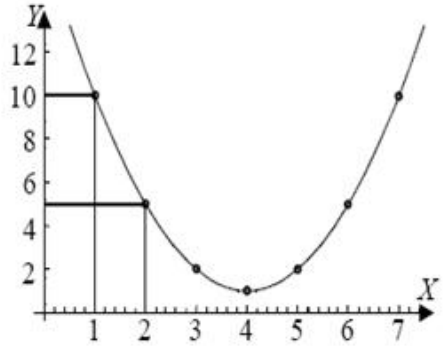
Σε μη πειραματικές έρευνες (δειγματοληψίες) η διάκριση μεταξύ **ανεξάρτητων** και **εξαρτημένων** μεταβλητών δεν είναι πάντοτε σαφής γιατί καμία μεταβλητή δεν είναι ελεγχόμενη αλλά όλες είναι τυχαίες (π.χ. το ύψος και το βάρος των φοιτητών, οι ώρες μελέτης των φοιτητών ενός πανεπιστημιακού τμήματος και η απόδοση τους σε ένα τεστ, οι εβδομάδες εμπειρίας ενός εργάτη σε μια επιχείρηση και ο αριθμός των ελαττωματικών προϊόντων που παράγει, η κατάταξη δέκα προϊόντων από έναν κριτή και η κατάταξη των ιδίων προϊόντων από έναν άλλο κριτή, ο αριθμός των πωλήσεων μουσικών CD σε μια περιοχή και ο αριθμός των νέων στην ίδια περιοχή).

Ας θεωρήσουμε δύο μεταβλητές **X , Y** . Αν οι μεταβλητές αυτές συνδέονται με μια σχέση της μορφής **Y = f (X)** μέσω της οποίας για κάθε τιμή της **X** μπορούμε να προβλέψουμε ακριβώς την τιμή της **Y** , δηλαδή, αν οι τιμές της **Y** δεν υπόκεινται σε σφάλματα, τότε λέμε ότι οι δύο μεταβλητές συνδέονται με τη ***συναρτησιακή-προσδιοριστική (deterministic) σχέση Y = f (X)*** . Για παράδειγμα, το ρεύμα που καταναλώνει μια οικογένεια σε ένα δίμηνο και το ποσό που πληρώνει για την κατανάλωση αυτή συνδέονται με συναρτησιακή-προσδιοριστική σχέση.

Επίσης, το ποσό που καταθέτει κάποιος στο Ταμιευτήριο και ο τόκος που παίρνει για το ποσό αυτό, συνδέονται με συναρτησιακή-προσδιοριστική σχέση. Σε αυτές τις περιπτώσεις τα σημεία του διαγράμματος διασποράς βρίσκονται όλα πάνω στην καμπύλη που έχει εξίσωση **Y = f (X)** και όσες φορές και αν επαναλάβουμε το πείραμα θέτοντας το **X** στο ίδιο επίπεδο $X = x_i$, θα παίρνουμε πάντα την ίδια τιμή για το **Y**. Για παράδειγμα, η εξίσωση $Y = (X-4)^2 + 1$ (που παριστάνει μια παραβολή) περιγράφει προσδιοριστικά τη σχέση μεταξύ των **X** και **Y** του παρακάτω πίνακα.

x_i	y_i
1	10
2	5
3	2
4	1
5	2
6	5
7	10

Πίνακας 10

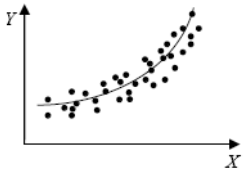


Διάγραμμα 6

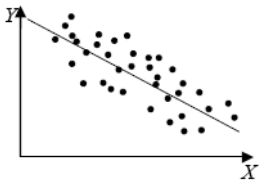
(πηγή www.aua.gr/gpapadopoulos/files/regression9.pdf)

Οι μη προσδιοριστικές σχέσεις μεταξύ μεταβλητών ονομάζονται **στοχαστικές στατιστικές** (stochastic, probabilistic) σχέσεις. Στην περίπτωση αυτή, αν επαναλάβουμε το πείραμα πολλές φορές θέτοντας το **X** στο ίδιο επίπεδο $X = x_i$ τότε στην τιμή x_i της **X** δεν αντιστοιχεί μια μόνο τιμή y_i της **Y** αλλά, γενικά, αντιστοιχεί ένα πλήθος διαφορετικών τιμών της **Y**. Για παράδειγμα, αν **X** είναι η τιμή ενός προϊόντος και **Y** είναι η ζήτησή του, η **Y** βρίσκεται σε στοχαστική σχέση-εξάρτηση από τη **X** , γιατί η ζήτηση ενός προϊόντος επηρεάζεται και από άλλους παράγοντες όπως είναι το ύψος του εισοδήματος των καταναλωτών, οι τιμές ομοειδών προϊόντων, οι καταναλωτικές συνήθειες, κ.ά

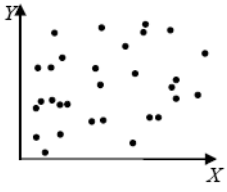
Σε μια στοχαστική σχέση το διάγραμμα διασποράς είναι, γενικά, ένα **νέφος σημείων** το οποίο πολλές φορές καθορίζει μια ιδεατή γραμμή η οποία δίνει μια πρώτη εικόνα της σχέσης που συνδέει τις δύο μεταβλητές. Η σχέση μάλιστα μεταξύ των δύο μεταβλητών είναι τόσο περισσότερο ισχυρή όσο πιο κοντά στην ιδεατή γραμμή βρίσκονται τα σημεία του διαγράμματος διασποράς. Στο πρώτο από τα παρακάτω σχήματα έχουμε το διάγραμμα διασποράς μιας ισχυρής σχέσης στην οποία όταν αυξάνουν οι τιμές της **X** αυξάνουν γενικά και οι τιμές της **Y**, ενώ στο δεύτερο σχήμα έχουμε μια λιγότερο ισχυρή σχέση στην οποία όταν αυξάνουν οι τιμές της **X** ελαττώνονται γενικά και οι τιμές της **Y**. Τέλος, στην περίπτωση του τρίτου σχήματος δε φαίνεται να υπάρχει κάποια σχέση μεταξύ των **X** και **Y**.



Διάγραμμα 7



Διάγραμμα 8



Διάγραμμα 9

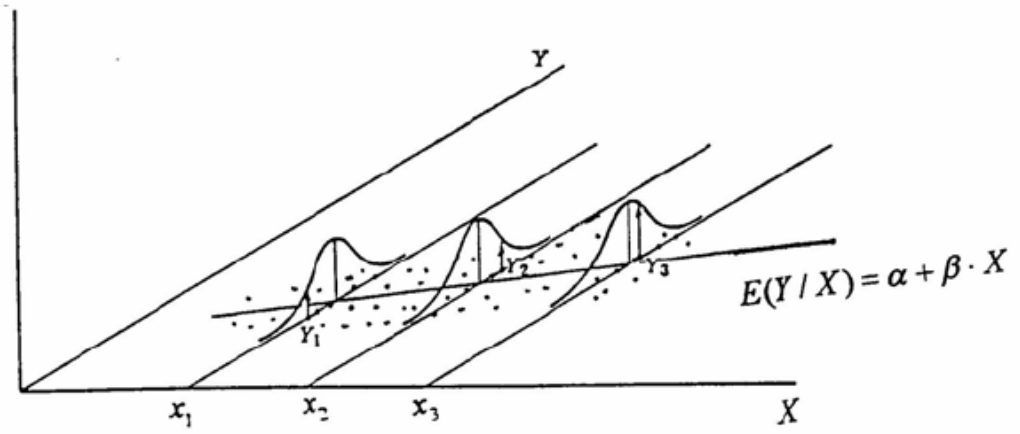
Γενικά, δύο μεταβλητές που συνδέονται είτε με συναρτησιακή-προσδιοριστική σχέση είτε με στοχαστική σχέση λέγονται **«εξαρτημένες».**

Αν υπάρχει εξάρτηση μεταξύ δύο μεταβλητών, τότε μπορούμε τη μια από αυτές να τη χαρακτηρίσουμε ως «αιτία» και την άλλη ως «αποτέλεσμα». Αυτό όμως, μόνο στην περίπτωση που η εξάρτηση οφείλεται σε σχέση αιτιότητας των δύο μεταβλητών και όχι σε μια απλή συμμεταβολή η οποία μπορεί να οφείλεται σε εξάρτηση των δύο μεταβλητών από μια τρίτη μεταβλητή. Αν, για παράδειγμα, **X** είναι το ετήσιο εισόδημα μιας οικογένειας και **Y, Z** είναι τα ποσά που ξοδεύει η οικογένεια αυτή σε ένα έτος για κρέας και για αγορά λογοτεχνικών βιβλίων, τότε: αν διαπιστώσουμε σε ένα σύνολο οικογενειών σχέση μεταξύ των **X** και **Y** (ή μεταξύ των **X** και **Z**) δεχόμαστε ότι υπάρχει εξάρτηση μεταξύ των δύο μεταβλητών και τότε μπορούμε να χαρακτηρίσουμε τη **X** ως «αιτία» και την **Y** (ή τη **Z**) ως «αποτέλεσμα». Αν όμως διαπιστωθεί σχέση μεταξύ των **Y** και **Z (**που είναι πολύ πιθανό, αφού και οι δύο μεταβάλλονται με το ετήσιο εισόδημα **X**) ασφαλώς θα πρόκειται για «νόθα» εξάρτηση.

Για να περιγράψουμε τη **στοχαστική εξάρτηση** δύο μεταβλητών **X** και **Y** προσπαθούμε να βρούμε, όπως και στην **προσδιοριστική εξάρτηση**, μια σχέση μεταξύ των **X** και **Y** η οποία όμως τώρα δε θα δίνει ακριβή αλλά προσεγγιστική μόνο εικόνα της εξάρτησης των **X** και **Y** και τα σημεία του διαγράμματος διασποράς των **X** και **Y** δε θα βρίσκονται πάνω, αλλά, γύρω από μια καμπύλη. Μια μέθοδος που χρησιμοποιείται για την περιγραφή της στοχαστικής εξάρτησης δύο μεταβλητών είναι η **μέθοδος των ελαχίστων τετραγώνων** και αυτή θα εφαρμόσουμε στη συνέχεια για να μελετήσουμε την πιο απλή μορφή στοχαστικής εξάρτησης, τη **γραμμική**.

3.2 Απλή γραμμική παλινδρόμηση

Αν το διάγραμμα διασποράς δύο μεταβλητών **X** και **Y** έχει μορφή επιμήκους κεκλιμένης έλλειψης ή πλατυσμένου **J**, η σχέση των **X** και **Y** είναι κατά προσέγγιση γραμμική. Στην περίπτωση αυτή έχουμε την απλούστερη μορφή παλινδρόμησης, την **απλή γραμμική παλινδρόμηση** όπου υπάρχει μόνο μια ανεξάρτητη μεταβλητή **X** και η εξαρτημένη μεταβλητή **Y** μπορεί να προσεγγισθεί ικανοποιητικά από μια γραμμική συνάρτηση του **X**. Η γραμμική σχέση $Y = a + b \cdot x$ δε μπορεί, ασφαλώς, να περιγράψει τη γραμμική στοχαστική εξάρτηση των μεταβλητών **X** και **Y** *α*φού αν, για παράδειγμα, **X** είναι η τιμή ενός προϊόντος και **Y** είναι η ζήτηση του προϊόντος αυτού, και διατηρήσουμε τη **X** στο ίδιο επίπεδο $X = x_1$ τότε οι αντίστοιχες τιμές του **Y** θα είναι φυσικά διαφορετικές στις διάφορες επαναλήψεις. Επίσης, αν **X** είναι η ποσότητα λιπάσματος και **Y** είναι η απόδοση μιας καλλιέργειας, και διατηρήσουμε τη **X** στο ίδιο επίπεδο $X = x_1$ τότε οι αντίστοιχες τιμές του **Y** θα είναι φυσικά διαφορετικές στις διάφορες επαναλήψεις αφού παράγοντες όπως, η θερμοκρασία, οι βροχοπτώσεις, η ποιότητα του εδάφους, θα επηρεάζουν, επίσης, την παραγωγή. Επιπλέον, συμβαίνει να παρατηρούνται και σφάλματα μέτρησης των τιμών της **Y** (λόγω οργάνων ή ελλιπούς πληροφόρησης). Έτσι, για $X = x_1$ το αντίστοιχο **Y** είναι μια τυχαία μεταβλητή Y_1 που ακολουθεί κάποια κατανομή. Ομοίως, για $X = x_2$ θα έχουμε κάποια άλλη κατανομή Y_2 κ.ό.κ..



Διάγραμμα 10

Επομένως, στην εξίσωση **Y =α + β.X** , πρέπει να προσθέσουμε έναν ακόμη όρο ε ο οποίος, για δεδομένη τιμή της X, να περιγράφει τη διαφορά της παρατηρούμενης από τη θεωρητική **(α + β.X)** τιμή της **Y**. Δηλαδή, $e = Y = -(a + b \cdot \mathrm{X})$. Προκύπτει, επομένως, το στοχαστικό μοντέλο

$$Y = a + b \cdot \mathrm{X} + e$$

Για λόγους απλούστευσης των υπολογισμών και εφικτότητας λύσης του προβλήματος, κάνουμε κάποιες υποθέσεις, όπως **E(ε) =0** και $\mathrm{E}(Y/\mathrm{X}) = a + b \cdot \mathrm{X}$. Δηλαδή, υποθέτουμε ότι τα σφάλματα έχουν μέση τιμή μηδέν και ότι για τις διάφορες τιμές της **X**, οι αντίστοιχες μέσες τιμές της **Y** βρίσκονται πάνω σε μια ευθεία. Η ευθεία αυτή $(\mathrm{E}(Y/\mathrm{X}) = a + b \cdot \mathrm{X})$, ονομάζεται ***πληθυσμιακή ευθεία παλινδρόμησης***.

3.3 Μέθοδος Ελαχίστων τετραγώνων

Με τη μέθοδο των ελαχίστων τετραγώνων θα προσδιορίσουμε στη συνέχεια μια εκτίμηση $\hat{Y} = \hat{a} + \hat{b} \cdot x$ της ευθείας των **α** και **β** αντίστοιχα.

Η εκτίμηση $\hat{Y} = \hat{a} + \hat{b} \cdot x$ της πληθυσμιακής ευθείας παλινδρόμησης $E(Y / X) = \alpha + \beta X$, ονομάζεται ***ευθεία ελαχίστων τετραγώνων*** από τη μέθοδο υπολογισμού των παραμέτρων της

<u>Μέθοδος ελαχίστων τετραγώνων</u>

Θεωρούμε ν ζεύγη παρατηρήσεων $(x, y), = (1, 2, 3, ..., n)$. Αναζητούμε μια προσέγγιση της μορφής:

$$y_i = a + b \cdot x_i + e_i$$

όπου τα **ε** παριστάνουν τις αποκλίσεις της πραγματικής τιμής yi από την προσαρμοσμένη (θεωρητική) $a + b x$. Δηλαδή, $(e) = y - a + b \cdot x$.

Είναι φανερό, ότι η εκλογή (εκτίμηση) των α και β θα πρέπει να γίνει έτσι ώστε να ελαχιστοποιηθούν οι ποσότητες ε. Για το σκοπό αυτό, θα αναζητήσουμε τις τιμές των α και β για τις οποίες ελαχιστοποιείται το άθροισμα των τετραγώνων των *ε*. Δηλαδή, η ποσότητα

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - a - b \cdot x_i)^2$$

(Η ελαχιστοποίηση του αθροίσματος Σe_i δεν αποτελεί ασφαλές κριτήριο επιλογής διότι κάποια αρνητικά e_i θα αναιρούν αντίστοιχες θετικές ποσότητες του αθροίσματος)

Παραγωγίζοντας την παραπάνω συνάρτηση ως προς **α** και **β** και εξισώνοντας με μηδέν παίρνουμε τις ακόλουθες δύο εξισώσεις που ονομάζονται **κανονικές εξισώσεις**

$$\sum_{i=1}^{n} y_i = n \cdot a + b \cdot \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = a \cdot \sum_{i=1}^{n} x_i + b \cdot \sum_{i=1}^{n} x_i^2$$

Λύνοντας το σύστημα των κανονικών εξισώσεων, παίρνουμε:

$$\hat{b} = \frac{n \cdot \sum_{i=1}^{n} X_i Y_i - \left(\sum_{i=1}^{n} X_i \right) \cdot \left(\sum_{i=1}^{n} y_i \right)}{n \cdot \sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i \right)^2} = \frac{\sum_{i=1}^{n} X_i Y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} X_i^2 - n \cdot \overline{X}^2}$$

$$\hat{a} = \overline{y} - \hat{b} \cdot \overline{x}$$

ή

$$\hat{b} = \frac{S_{xy}}{S_x^2} \text{ και } \hat{a} = \overline{y} - \hat{b} \cdot \overline{x}$$

Η **εκτίμηση ελαχίστων τετράγωνων** της **ευθείας παλινδρόμησης** από το δείγμα των ν ζευγών παρατηρήσεων είναι, επομένως, η

$$\hat{Y} = \hat{a} + \hat{b} \cdot \mathrm{X} = \overline{y} - \hat{b} \cdot \overline{x} + \hat{b} \cdot \mathrm{X} = \overline{y} + \hat{b} \cdot \left(\mathrm{X} - \overline{x} \right)$$

ή

$$\hat{Y} = \overline{y} + \frac{S_{xy}}{S_x^2} \cdot \left(\mathrm{X} - \overline{x} \right)$$

Προφανώς, η **ευθεία ελαχίστων τετραγώνων**, διέρχεται από το σημείο $(\overline{x}, \overline{y})$.Επισημαίνουμε ότι πρέπει να γίνεται διάκριση μεταξύ της

παρατηρούμενης τιμής του **Y** και της $\hat{Y}$ που εκτιμάμε. Η **παρατηρούμενη** τιμή y_i είναι η πραγματική τιμή της **Y**, ενώ η τιμή $\hat{y}_i$ της $\hat{Y}$, είναι εκτίμηση της μέσης τιμής $\mathrm{E}(Y/X) = x_i$.

Από την προφανή σχέση $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$, μπορεί εύκολα να δειχθεί (αλγεβρικά) ότι

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

Το άθροισμα $$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

λέγεται **ολικό άθροισμα τετραγώνων (total sum of squares)** ή **ολική μεταβλητότητα (total variation)** των y_i και όπως φαίνεται από τη αναλύεται σε δύο συνιστώσες: στο **άθροισμα τετραγώνων παλινδρόμησης (regression sum of squares)**

$$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

και στο άθροισμα τετραγώνων των σφαλμάτων (error sum of squares) ή υπόλοιπο μεταβλητότητας (residual variation)

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

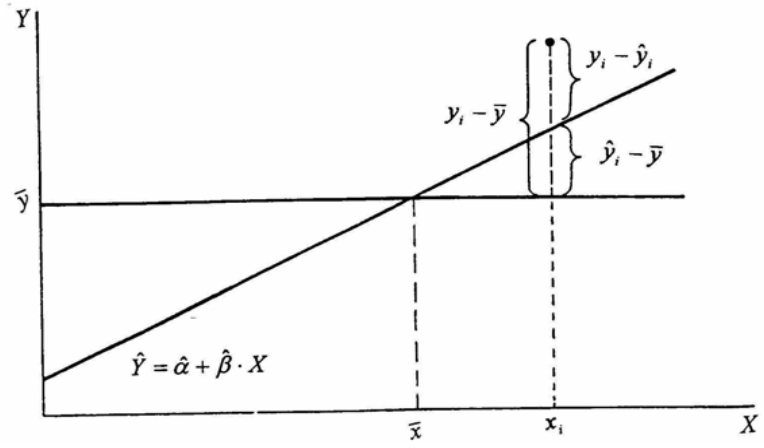
$$SSTO = SSR + SSE$$

Το $SSTO$ μετράει τη συνολική μεταβλητότητα των παρατηρήσεων y_i δηλαδή εκφράζει την αβεβαιότητα στην πρόβλεψη του **Y** όταν δε χρησιμοποιείται το **X**. Το SSR εκφράζει το μέρος της μεταβλητότητας που

μπορεί να οφείλεται στο **X** και το $SSTO = SSR + SSE$ εκφράζει την υπόλοιπη μεταβλητότητα που δεν εξηγείται από την παλινδρόμηση ενώ ο λόγος

$$r^2 = \frac{SSR}{SSTO} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

εκφράζει το ποσοστό της συνολικής μεταβλητότητας των y_i που εξηγείται (απορροφάται) από την παλινδρόμηση. Το r^2 λέγεται ***συντελεστής προσδιορισμού (coefficient of determination)*** και παίρνει τιμές στο κλειστό διάστημα **[0, 1]**. Όταν όλα τα σημεία $M_1(x_1, y_1), M_2(x_2, y_2), ..., M_n(x_n, y_n)$, βρίσκονται πάνω στην **ευθεία ελαχίστων τετραγώνων** θα έχουμε $y_i = \bar{y}_i$ και άρα $SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = 0$, οπότε $r^2 = 1$ ενώ όταν η κλίση της ευθείας ελαχίστων τετραγώνων είναι μηδέν δηλαδή $\hat{b} = 0$ θα είναι **r = 0** . Στις διάφορες πρακτικές εφαρμογές η τιμή του r^2 βρίσκεται μεταξύ 0 και 1 και όσο πλησιέστερα βρίσκεται προς το 1 τόσο καλύτερη είναι η ευθεία ελαχίστων τετραγώνων ως εκτίμηση της ευθείας παλινδρόμησης.



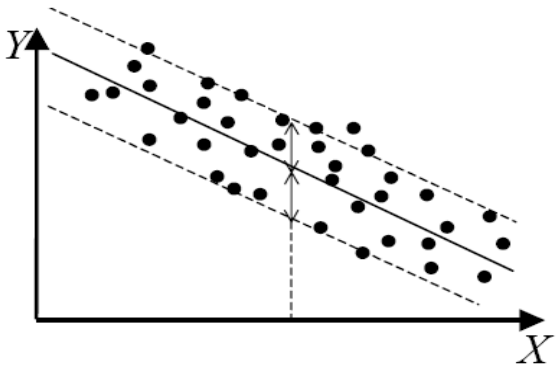
Διάγραμμα 11

Η μέση απόκλιση μεταξύ της πραγματικής και της εκτιμούμενης τιμής της μεταβλητής ονομάζεται ***τυπικό σφάλμα της εκτίμησης (standard error of the estimate)***, συμβολίζεται με **s** και δίνεται από τον τύπο

$$s = \sqrt{\frac{1}{n-2} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} = \sqrt{\frac{SSE}{n-2}}$$

Εάν το **τυπικό σφάλμα** της εκτίμησης είναι μικρό τότε οι παρατηρούμενες και οι εκτιμούμενες τιμές δε διαφέρουν πολύ και η ευθεία παλινδρόμησης μας δίνει μια καλή περιγραφή της σχέσης μεταξύ των **X** και **Y**. Αν το **τυπικό σφάλμα** της εκτίμησης είναι μεγάλο τότε δε μπορούμε να ισχυρισθούμε ότι έχουμε μια καλή περιγραφή της σχέσης.

Είναι φανερό, ότι το **τυπικό σφάλμα** της εκτίμησης, είναι ένα μέτρο της **διασποράς** των (x_i, y_i) γύρω από την **ευθεία ελαχίστων τετραγώνων** $\hat{y}_i = \hat{a} + \hat{b}.x_i$ (το s^2 είναι μια εκτίμηση της διασποράς των σφαλμάτων). Έχει, επομένως, ιδιότητες ανάλογες με περίπου το 68%, το 95% και το 99,7% των σημείων του διαγράμματος διασποράς αντίστοιχα. αυτές της **τυπικής απόκλισης**. Έτσι, αν φέρουμε δύο ευθείες παράλληλες προς την ευθεία ελαχίστων τετραγώνων και σε κατακόρυφες προς αυτήν αποστάσεις s,2s, 3s τότε, για μεγάλα ν (μεγαλύτερα του 30), μεταξύ των δύο αυτών ευθειών θα βρίσκεται



Διάγραμμα 12

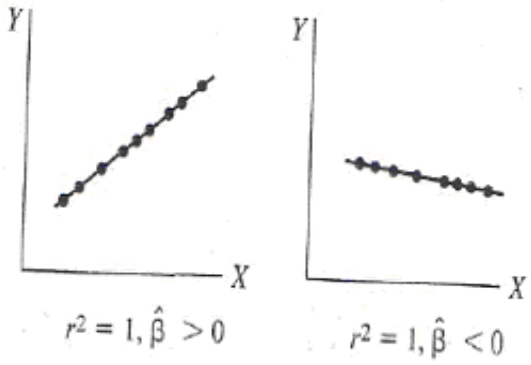
Σημείωση: Στο σχήμα οι παράλληλες έχουν σχεδιαστεί σε κατακόρυφη απόσταση από την ευθεία ελαχίστων τετραγώνων ίση με 2s.

Εύκολα μπορεί να αποδειχθεί ότι:

$$s=\sqrt{\frac{1}{n-2}\cdot\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}=\sqrt{\frac{n-1}{n-2}\cdot\left(s_x^2-\hat{b}^2\cdot s_x^2\right)}=\sqrt{\frac{n-1}{n-2}\cdot s_y^2\left(1-r^2\right)}$$

<u>Παρατηρήσεις για την ευθεία ελαχίστων τετραγώνων</u>

1. Είναι φανερό ότι το $\hat{b}$ της ευθείας των ελαχίστων τετραγώνων $\hat{Y}=\hat{a}+\hat{b}\cdot \mathrm{X}$ εκφράζει την αναμενόμενη μεταβολή της εξαρτημένης μεταβλητής Y (σε μονάδες μέτρησης της Y) όταν η ανεξάρτητη μεταβλητή X αυξηθεί κατά μια μονάδα (μέτρησής της).Πράγματι αν $\mathrm{X}=x_1$ έχουμε $\hat{y}_1=\hat{a}+\hat{b}\cdot x_1$ και αν $\mathrm{X}=x_1+1$ έχουμε $\hat{y}_2=\hat{a}+\hat{b}\cdot\left(x_1+1\right)=\hat{a}+\hat{b}\cdot x_1+\hat{b}=\hat{y}_1+\hat{b}$. Έτσι όταν το x_i αυξηθεί κατά μια μονάδα το $\hat{y}_i$ αυξάνεται κατά b μονάδες αν $\hat{b}>0$ ή ελαττώνεται κατά $\hat{b}$ μονάδες αν $\hat{b}<0$.
2. Το $\hat{a}$ της ευθείας ελαχίστων τετραγώνων $\hat{Y}=\hat{a}+\hat{b}\cdot \mathrm{X}$ εκφράζει την αναμενόμενη τιμή της εξαρτημένης μεταβλητής Y όταν η ανεξάρτητη μεταβλητή X πάρει την τιμή 0.
3. Η ποσότητα $1-r^2$ εκφράζει το ποσοστό της συνολικής μεταβλητότητας που οφείλεται στο τυχαίο σφάλμα.
4. Το r^2 **δεν μετρά** πόσο μεγάλη είναι η κλίση $\hat{b}$ της ευθείας παλινδρόμησης!



Διάγραμμα 13 Διάγραμμα 14

5. Όταν έχουμε πειραματικά δεδομένα όπου ο ερευνητής ελέγχει-καθορίζει τις τιμές της μιας μεταβλητής θεωρούμε τη μεταβλητή αυτή ανεξάρτητη (Χ) και την άλλη και την άλλη εξαρτημένη (Y). Σε αυτή την περίπτωση εκτιμάμε την ευθεία παλινδρόμησης **της** Y **πάνω στη** X, $\hat{Y} = \hat{a} + \hat{b} \cdot \mathrm{X}$. Όταν έχουμε μη πειραματικά δεδομένα όπου ο ερευνητής επιλέγει ένα τυχαίο δείγμα ατόμων και σε κάθε ένα από αυτά τα μέτρα τις τιμές των μεταβλητών, τότε μπορούμε να θεωρήσουμε ως ανεξάρτητη μεταβλητή οποιαδήποτε από τις δύο και να μελετήσουμε είτε την παλινδρόμηση **της** Y **πάνω στη** X είτε την παλινδρόμηση **της** X **πάνω στη** Y. Στην περίπτωση αυτή, και οι δύο μεταβλητές είναι τυχαίες, και ως μέτρο της γραμμικής συσχέτισης χρησιμοποιούμε το **συντελεστή γραμμικής συσχέτισης** $r = \frac{S_{xy}}{S_x \cdot S_y}$ και επειδή $\hat{b} = \frac{S_{xy}}{S_x^2}$ θα είναι, $r = \hat{b} \cdot \frac{S_x}{S_y}$ (Ι).Έτσι, αν το r πλησιάζει το 1 τότε τα σημεία του διαγράμματος διασποράς τείνουν να βρίσκονται σε μια ευθεία με συντελεστή διεύθυνσης $\hat{b} \rangle 0$ ενώ, αν το r πλησιάζει το -1 τότε τα σημεία του διαγράμματος διασποράς τείνουν να βρίσκονται σε

μια ευθεία με συντελεστή διεύθυνσης $\hat{b} \langle 0$. Αν $r \approx 0$ τότε $\hat{b} \approx 0$ και δεν υπάρχει γραμμική σχέση των μεταβλητών. Ο συντελεστής γραμμικής συσχέτισης έχει επομένως το ίδιο πρόσημο με το $\hat{b}$. Αν $\hat{X} = \hat{g} + \hat{d} \cdot Y$ είναι η εκτίμηση ελαχίστων τετραγώνων της ευθείας παλινδρόμησης **της** X **πάνω στη** Y θα ισχύει: $\hat{d} = \frac{S_{xy}}{S_Y^2}$ και $\hat{g} = \overline{x} - \hat{d} \cdot \overline{y}$, Συνεπώς, $r = \hat{d} \cdot \frac{S_y}{S_x}$ (II). Από τις (I) και (II) προκύπτει, επίσης ότι $r^2 = \hat{b} \cdot \hat{d}$.

6 Οι προβλέψεις που μπορούμε να κάνουμε για την εξαρτημένη μεταβλητή Y από τις τιμές της ανεξάρτητης μεταβλητής X μέσω της **ευθείας ελαχίστων τετραγώνων** $\hat{Y} = \hat{a} + \hat{b} \cdot X$ πρέπει να γίνονται μόνο για τις τιμές της ανεξάρτητης μεταβλητής, οι οποίες βρίσκονται στο διάστημα που έχει γίνει η μελέτη ή πολύ κοντά στα άκρα του διαστήματος αυτού.

7 Η εξίσωση της **ευθείας ελαχίστων τετραγώνων** $\hat{Y} = \hat{a} + \hat{b} \cdot X$, **δε μας επιτρέπει** να κάνουμε προβλέψεις για τις τιμές της X, όταν δίνονται οι τιμές της Y. Για να είναι δυνατόν, πρέπει να προσδιορίσουμε εξαρχής την **ευθεία ελαχίστων τετραγώνων** της **της** X **πάνω στη** Y**,** $\hat{X} = \hat{g} + \hat{d} \cdot Y$, η οποία γενικά είναι διαφορετική από την $\hat{Y} = \hat{a} + \hat{b} \cdot X$. Και στις δύο περιπτώσεις οι ευθείες διέρχονται από το σημείο $(\overline{x}, \overline{y})$.

8 Επισημαίνουμε ότι για δοσμένη τιμή x_i της X, η εκτίμηση $\hat{g}_i = \hat{a} + \hat{b} \cdot x_i$ αφορά **τη μέση τιμή** $E(Y / X = x_i)$ της Y και **όχι την πραγματική τιμή** του Y.

9 Αξίζει να σημειωθεί ότι πάντα ισχύει $\sum_{i=1}^{n} \hat{e}_i = 0$ αφού

$$\sum_{i=1}^{n} \hat{e}_i = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b} \cdot x_i)^2 = \sum_{i=1}^{n} y_i - n \cdot \hat{a} - \hat{b} \cdot \sum_{i=1}^{n} x_i = n \cdot \left(\overline{y} - \hat{a} - \hat{b} \cdot \overline{x} \right) = 0$$

Παράδειγμα 1:

Ο πίνακας που ακολουθεί δίνει τη **ζήτηση ενός προϊόντος (Y)**, για διάφορα επίπεδα **διαφημιστικής δαπάνης (X).**

y_i *(σε χιλιάδες τεμάχια)*	x_i *(σε χιλιάδες €)*	$x_i \cdot y_i$	x_i^2
12	*2*	*24*	*4*
13	*2*	*26*	*4*
13	*3*	*39*	*9*
14	*3*	*42*	*9*
15	*4*	*60*	*16*
15	*4*	*60*	*16*
14	*5*	*70*	*25*
16	*5*	*80*	*25*
17	*6*	*102*	*36*
18	*6*	*108*	*36*
$\sum y_i = 147$	$\sum x_i = 40$	$\sum x_i \cdot y_i = 611$	$\sum x_i^2 = 180$

Πίνακας 11

$$\overline{x} = \frac{40}{10} = 4$$

$$\overline{y} = \frac{147}{10} = 14,7$$

$$\hat{b} = \frac{\sum_{i=1}^{v} x_i y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum_{i=1}^{n} X_i^2 - n \cdot \overline{x}^2} = \frac{611 - 10 \cdot 4 \cdot 14,7}{180 - 10 \cdot 4^2} = 1,15$$

$$\hat{a} = \overline{y} - \hat{b} \cdot \overline{x} = 14,7 - 1,15 \cdot 4 = 10,1$$

και άρα η εξίσωση της **ευθείας ελαχίστων τετραγώνων** είναι η

Y = 10,1+1,15.X .

Ερμηνεία του $\hat{b}$

Επειδή $\hat{b} = 1{,}15 > 0$, αύξηση της διαφημιστικής δαπάνης συνεπάγεται αύξηση της ζήτησης του προϊόντος. Αν η διαφημιστική δαπάνη αυξηθεί κατά 1000 €, η μέση ζήτηση του προϊόντος εκτιμάται ότι θα αυξηθεί κατά 1,15 χιλιάδες τεμάχια.

Ερμηνεία του $\hat{a}$

Για μηδενική διαφημιστική δαπάνη, η μέση ζήτηση του προϊόντος εκτιμάται ότι θα είναι 10,1 χιλιάδες τεμάχια. Επειδή η τιμή 0 είναι μακριά από το διάστημα μελέτης, η ερμηνεία του $\hat{a}$ δεν έχει πρακτική αξία. Θα υπολογίσουμε το συντελεστή προσδιορισμού

x_i	y_i	$\hat{y}_i$	$\hat{y}_i - \overline{y}$	$(\hat{y}_i - \overline{y})^2$	$y_i - \overline{y}$	$(y_i - \overline{y})^2$
2	12	12,4	-2,3	5,29	-2,7	7,29
2	13	12,4	-2,3	5,29	-1,7	2,89
3	13	13,55	-1,15	1,32	-1,7	2,89
3	14	13,55	-1,15	1,32	-0,7	0,49
4	15	14,7	0	0	0,3	0,09
4	15	14,7	0	0	0,3	0,09
5	14	15,85	1,15	1,32	-0,7	0,49
5	16	15,85	1,15	1,32	1,3	1,69
6	17	17	2,3	5,29	2,3	5,29
6	18	17	2,3	5,29	3,3	10,89
$\sum x_i = 40$	$\sum y_i = 147$			$SSR = 26,44$		$SSTO = 32,1$

Πίνακας 12

$$r^2 = \frac{SSR}{SSTO} = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2} = \frac{26,44}{32,1} = 0,82$$

Ερμηνεία του r^2

Οι μεταβολές του ύψους της διαφημιστικής δαπάνης ερμηνεύουν το **82%** της μεταβλητότητας της ζήτησης του προϊόντος.

Ερμηνεία του $1-r^2$

Το **18%** της μεταβλητότητας της ζήτησης του προϊόντος, οφείλεται σε τυχαία σφάλματα

Το **τυπικό σφάλμα της εκτίμησης s** είναι

$$s=\sqrt{\frac{1}{n-2}\cdot\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}=\sqrt{\frac{SSE}{n-2}}=\sqrt{\frac{32,1-26,44}{8}}=\sqrt{0,7}=0,84$$

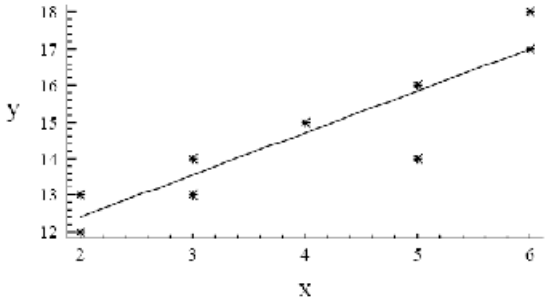
Πρόβλεψη με το μοντέλο $\hat{Y}=10,1+1,15\cdot X$ που εκτιμήσαμε.

Αν το ύψος της διαφημιστικής δαπάνης είναι π.χ. 3,5 χιλιάδες €, η μέση ζήτηση του προϊόντος, εκτιμάται ότι θα είναι $10,1+1,15\cdot 3,5=14,125$ χιλιάδες τεμάχια.

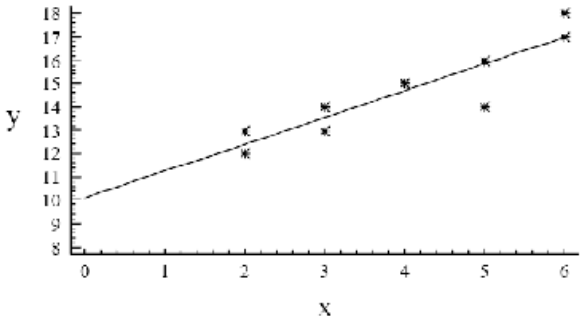
Αξιολόγηση του μοντέλου: Το μοντέλο $\hat{Y}=10,1+1,15\cdot X$ ερμηνεύει το 82% της μεταβλητότητας της ζήτησης του προϊόντος.

Παρατήρηση:

Για συγκεκριμένη ζήτηση του προϊόντος, δε μπορούμε από το μοντέλο αυτό, να προβλέψουμε το απαιτούμενο ύψος διαφημιστικής δαπάνης.



Διάγραμμα 15



Διάγραμμα 16

3.4 Πολλαπλή γραμμική παλινδρόμηση

Στην περίπτωση που θέλουμε να διερευνήσουμε τη μεταβολή των τιμών μίας μεταβλητής *(*εξαρτημένη*)* συναρτήσει των τιμών όχι μίας αλλά περισσότερων ανεξάρτητών μεταβλητών, εφαρμόζουμε πολλαπλή στατιστική εξάρτηση ή παλινδρόμηση.

Μοντέλο πολλαπλής γραμμικής παλινδρόμησης :

Το αντίστοιχο μοντέλο για τις μεταβολές της μέσης τιμής της εξαρτημένης μεταβλητής, συναρτήσει των τιμών των ανεξάρτητων είναι το:

$$E\left(\hat{Y}_i \middle| X_{1i}, X_{2i}, ..., X_{pi}\right) = b_0 + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + ... + b_p X_{pi}$$

Όσα αναφέρθηκαν όσων αφορά τις προϋποθέσεις εφαρμογής της παλινδρόμησης, ισχύουν και στην περίπτωση της πολλαπλής. Δεν υπάρχει ωστόσο ένας συντελεστής εξάρτησης αλλά τόσοι όσες οι ανεξάρτητες μεταβλητές. Δεδομένου ότι καθένας αντιπροσωπεύει την εξάρτηση της **Y** από την αντίστοιχη μεταβλητή **Xi** οι **bi** καλούνται συντελεστές μερικής εξάρτησης.

Συντελεστής μερικής εξάρτησης :

Οι ιδιότητες των συντελεστών μερικής εξάρτησης **(bi)** είναι ίδιες με αυτές που αναφέρθηκαν για τον **b1**. Υπάρχει όμως διαφορά στην ερμηνεία: Ο **bi** εκφράζει την αναμενόμενη μεταβολή της εξαρτημένης μεταβλητής, όταν η αντίστοιχη ανεξάρτητη **(Xi)** μεταβληθεί κατά μια μονάδα και όλες οι υπόλοιπες ανεξάρτητες μεταβλητές παραμείνουν σταθερές. Αυτό φαίνεται από τις αντίστοιχες εξισώσεις.

Ας θεωρήσουμε το μοντέλο

$$E\left(\hat{Y}_i \middle| X_{1i}, X_{2i}, ..., X_{pi}\right) = b_0 + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + ... + b_p X_{pi}$$

Αν υποθέσουμε ότι η X_{1i} αυξάνεται κατά μία μονάδα ενώ όλες οι υπόλοιπες παραμένουν σταθερές.

$$E\left(\hat{Y}_i \middle| (X_{1i}+1), X_{2i}, ..., X_{pi}\right) = b_0 + b_1 \cdot (X_{1i}+1) + b_2 \cdot X_{2i} + ... + b_p X_{pi}$$

Προκειμένου να υπολογίσουμε την μεταβολή στην μέση τιμή της **Y**, για μία μονάδα αύξησης της X1i, αφαιρούμε τις παραπάνω σχέσεις

$$E\left(\hat{Y}_i \middle| (X_{1i}+1), X_{2i}, ..., X_{pi}\right) - E\left(\hat{Y}_i \middle| X_{1i}, X_{2i}, ..., X_{pi}\right) =$$

$$b_0 + b_1 \cdot (X_{1i}+1) + b_2 \cdot X_{2i} + ... + b_p X_{pi} - b_0 - b_1 \cdot X_{1i} - b_2 \cdot X_{2i} - ... - b_p X_{pi} = b_1$$

3.5 Ο ΣΥΝΤΕΛΕΣΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ R^2

Ο Συντελεστής προσδιορισμού μας δείχνει πόση είναι η μεταβλητικότητα της μεταβλητής Υ που εξηγείται από την παλινδρόμηση και πόση μένει ανεξήγητη, δηλαδή οφείλεται στους τυχαίους παράγοντες που εκφράζουν τα κατάλοιπα.

$$R^2 = \frac{\sum \left(\hat{Y} - \bar{Y}\right)^2}{\sum \left(Y - \bar{Y}\right)^2}$$

ή

$$R^2 = \frac{1 - \sum \hat{u}^2}{\sum \left(Y - \bar{Y}\right)^2}$$

$$0 \le R^2 \le 1 \qquad (8)$$

Όσο μεγαλύτερος είναι ο συντελεστής προσδιορισμού τόσο καλύτερη είναι η προσαρμογή του υποδείγματος στα δεδομένα του δείγματος και αντίστροφα. Αξίζει να σημειώσουμε ότι η χαμηλή ή ακόμη και μηδενική τιμή του συντελεστή προσδιορισμού δεν σημαίνει αναγκαστικά έλλειψη εξαρτήσεως ανάμεσα στις μεταβλητές **X** και **Y**. Ο συντελεστής προσδιορισμού δεν είναι εκτιμητής μιας άγνωστης παραμέτρου του πληθυσμού. Αφορά τα δεδομένα του δείγματος και μόνον. Στην ουσία μας δείχνει το ποσοστό μεταβλητικότητας της εξαρτημένης μεταβλητής που ερμηνεύεται από την παλινδρόμηση.

Δηλαδή: Συνολική μεταβλητικότητα = μεταβλητικότητα που οφείλεται στην παλινδρόμηση + μεταβλητικότητα που οφείλεται σε τυχαίους παράγοντες.

Και μεταβλητικότητα που οφείλεται στην παλινδρόμηση / συνολική μεταβλητικότητα ή 1- μεταβλητικότητα που οφείλεται σε τυχαίους παράγοντες/συνολική μεταβλητικότητα μας δίνει το συντελεστή προσδιορισμού.

Ο συντελεστής προσδιορισμού μπορεί να αναφέρεται και σε περισσότερες από δύο μεταβλητές και αφορά, ως ήδη ανεφέρθη, τη προσαρμογή του υποδείγματος στα δεδομένα του δείγματος.

3.6 Ο ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ ρ

$r = s\ xy / s\ xy$.

Ο τύπος είναι αναλυτικά:

$$r = \frac{\sum \left(X - \bar{X}\right) \cdot \left(Y - \bar{Y}\right)}{\sqrt{\sum \left(X - \bar{X}\right)^2} \cdot \sqrt{\sum \left(Y - \bar{Y}\right)^2}}$$

Ο συντελεστής συσχέτισης του δείγματος **ρ** είναι ένας εκτιμητής του συντελεστή συσχέτισης στον πληθυσμό (ρ) και επομένως το (ρ) είναι μια άγνωστη παράμετρος του πληθυσμού της συνδυασμένης κατανομής δύο τυχαίων μεταβλητών.

Ο τύπος που μας δίνει τον συντελεστή ρ συσχέτισης του δείγματος είναι ένας εκτιμητής του συντελεστή συσχέτισης του πληθυσμού (ρ). Η συγκεκριμένη τιμή που παίρνει ο συντελεστής συσχέτισης του δείγματος ονομάζεται εκτίμηση.

Ο συντελεστής συσχέτισης αφορά το βαθμό γραμμικής συσχέτισης ανάμεσα στις μεταβλητές **X** και **Y**.

Επίσης, ένας τύπος που μας δίδει τον συντελεστή συσχέτισης είναι:

$$r = \pm\sqrt{R^2} \qquad (9)$$

3.7 ΙΔΙΟΤΗΤΕΣ ΤΩΝ ΕΚΤΙΜΗΤΩΝ

1. Οι εκτιμητές (ως τυχαίες μεταβλητές που ακολουθούν στατιστικές κανονικές κατανομές) που δίνονται από τις σχέσεις:

$$\hat{b}_0 = \frac{\sum X^2 \sum Y - \sum X \sum XY}{T\sum X^2 - \left(\sum X\right)^2}$$

$$\hat{b}_1 = \Sigma_{yx} \Big/ \Sigma_{x^2}$$

είναι αμερόληπτοι, δηλαδή οι προσδοκώμενες εκτιμημένες τιμές, που παίρνουμε από τους εκτιμητές του δείγματος είναι αυτές των τιμών του πληθυσμού.

$$E\left(\hat{b}_0\right) = b_0 \qquad (10)$$

$$E\left(\hat{b}_1\right) = b_1 \qquad (11)$$

2. Είναι γραμμικές συναρτήσεις των παρατηρήσεων της εξαρτημένης μεταβλητής **Y** και κατ' επέκταση των διαταρακτικών όρων **ui**.

3. Μεταξύ όλων των γραμμικών αμερόληπτων εκτιμητών έχουν τη μικρότερη διακύμανση, που δίνονται από τις ακόλουθες σχέσεις:

$$V\left(\hat{b}_0\right) = \frac{s^2 \sum X^2}{T\sum x^2} \qquad (12)$$

$$V\left(\hat{b}_1\right) = \frac{s^2}{\sum x^2} \qquad (13)$$

Επιπλέον η συνδιακύμανση των συντελεστών $\hat{b}_0, \hat{b}_1$ δίνεται από την ακόλουθη σχέση:

$$Cov\left(\hat{b}_0, \hat{b}_1\right) = -s^2 = \frac{\bar{X}}{\sum x^2} \qquad (14)$$

Με την τετραγωνική ρίζα των διακυμάνσεων παίρνουμε τα τυπικά σφάλματα των εκτιμητών:

$$s\ \hat{b}_0 = \sqrt{V\ \left(\hat{b}_0\right)} \qquad (15)$$

$$s\ \hat{b}_1 = \sqrt{V\ \left(\hat{b}_1\right)} \qquad (16)$$

Τα τυπικά σφάλματα των εκτιμητών αποτελούν το μέτρο ακρίβειας των εκτιμητών. Όσο πιο μικρό είναι το τυπικό σφάλμα τόσο πιο καλή είναι η εκτίμηση της άγνωστης παραμέτρου του πληθυσμού. Το τυπικό σφάλμα είναι τόσο πιο μικρό όσο πιο μικρή είναι η διακύμανση των εκτιμητών, καθώς η τετραγωνική ρίζα της διακύμανσης θα δώσει μικρότερο τυπικό σφάλμα. Από τους τύπους των διακυμάνσεων των εκτιμητών βλέπουμε πως όσο μικρότερη είναι η διακύμανση των διαταρακτικών όρων ui, στον αριθμητή τόσο μικρότερη θα είναι η συνολική διακύμανση των εκτιμητών, άρα τόσο μικρότερο θα είναι και το τυπικό σφάλμα.

Όπου $\boldsymbol{\sigma^2}$ είναι η διακύμανση του πληθυσμού, η τιμή της οποίας διαμορφώνεται με βάση τους διαταρακτικούς όρους **ui**. Επειδή η τιμή είναι άγνωστη χρησιμοποιούμε έναν εκτιμητή της **σ2**. Επομένως για να εκτιμήσουμε τις διακυμάνσεις των συντελεστών, που είναι απαραίτητες για την εφαρμογή των στατιστικών κριτηρίων, θα πρέπει να έχουμε μια εκτίμηση για τη **σ2** του πληθυσμού.

Η εκτίμηση της **σ2** βασίζεται στα κατάλοιπα $\hat{u}_i$ της γραμμής παλινδρόμησης του δείγματος. Έτσι, ένας αμερόληπτος εκτιμητής (ως τυχαία μεταβλητή) της σ^2 δίνεται από τη σχέση:

$$S^2 = \frac{\sum \hat{u}i^2}{T-2} \quad \text{ή} \quad S^2 = \frac{\sum \hat{y}i^2 - \hat{b}_1 \cdot \sum xy}{T-2} \qquad (17)$$

$$S = \sqrt{S^2} \qquad (18)$$

Η τετραγωνική ρίζα της διακύμανσης $\mathbf{S^2}$ ονομάζεται τυπικό σφάλμα **S** εκτιμήσεως της **Y**. Δηλαδή η **S** είναι μέτρο της διασποράς των τιμών της **Y** από τη γραμμή παλινδρομήσεως. Από τον τύπο βλέπουμε πως όσο λιγότερα είναι τα κατάλοιπα (δηλαδή όσο μικρότερη είναι η επίδραση των τυχαίων παραγόντων), τόσο μικρότερη θα είναι η διασπορά των τιμών της **Y** από τη γραμμή παλινδρόμησης.

3.8 ΣΤΑΤΙΣΤΙΚΗ ΕΠΑΓΩΓΗ. ΕΛΕΓΧΟΣ ΤΟΥ ΥΠΟΔΕΙΓΜΑΤΟΣ

Στο βαθμό που η τυχαία μεταβλητή **ui** (διαταρακτικός όρος) ακολουθεί την κανονική κατανομή με μέσο μηδέν και σταθερή διακύμανση **ui→ N (0, σ²)**, έπεται ότι και οι εκτιμητές $\hat{b}_0$ και $\hat{b}_1$, στο βαθμό που αποτελούν γραμμικές συναρτήσεις της **Y** και κατ' επέκταση των διαταρακτικών όρων ui, ακολουθούν την κανονική κατανομή. Είδαμε ότι οι εκτιμητές $\hat{b}_0$ και $\hat{b}_1$ είναι αμερόληπτοι των b_0 και b_1 με αντίστοιχες διακυμάνσεις:

$$V\left(\hat{b}_0\right) = s^2 \frac{\sum X^2}{T\sum x^2} \qquad V\left(\hat{b}_1\right) = \frac{s^2}{\sum x^2}$$

Επομένως:

$$\hat{b_0} \rightarrow N\left(b_0 , s^2 \frac{\sum X^2}{T\sum x^2} \right) \quad (19)$$

$$\hat{b_1} \rightarrow N\left(b_1 , \frac{s^2}{\sum x^2} \right) \quad (20)$$

3.9 ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ ΓΙΑ ΤΟΥΣ ΣΥΝΤΕΛΕΣΤΕΣ bo και b1

Όπως είδαμε όσο πιο μικρό είναι το τυπικό σφάλμα των εκτιμητών τόσο πιο καλή είναι η εκτίμηση της άγνωστης παραμέτρου του πληθυσμού. Βέβαια, η παράμετρος του πληθυσμού μας είναι άγνωστη. Μπορούμε όμως να φτιάξουμε διαστήματα εμπιστοσύνης για τους άγνωστους συντελεστές (παράμετρες) b_0 και b_1 του πληθυσμού.

Από τις σχέσεις 19 και 20 δημιουργούμε τις στατιστικές:

$$\frac{\hat{b}_0 - b_0}{s\ \hat{b}_0} \quad (21)$$

$$\frac{\hat{b}_1 - b_1}{s\ \hat{b}_1} \quad (22)$$

Όπου $s\,\hat{b}_0$ και $s\,\hat{b}_1$ τα τυπικά σφάλματα των εκτιμητών που είναι η τετραγωνική ρίζα των διακυμάνσεων των εκτιμητών (σχέσεις 12 και 13).

Οι στατιστικές ακολουθούν την τυποποιημένη κανονική κατανομή.

Όμως λόγω του γεγονότος ότι η διακύμανση **σ²** του πληθυσμού είναι άγνωστη (βλέπε τον αριθμητή των τύπων των διακυμάνσεων των εκτιμητών), αντικαθιστούμε την διακύμανση **σ2** του πληθυσμού με την αμερόληπτη εκτίμησή της **S2** που την υπολογίζουμε από το δείγμα.

Έτσι οι διακυμάνσεις των σχέσεων (12 και 13) γίνονται:

$$S^2\left(\hat{b}_0\right) = S^2 \frac{\sum X^2}{T\sum x^2} \qquad (23)$$

$$S^2\left(\hat{b}_1\right) = \frac{S^2}{\sum x^2} \qquad (24)$$

και τα τυπικά σφάλματα $S\left(\hat{b}_0\right) = \sqrt{S^2\left(\hat{b}_0\right)}$ (25)

$$S\left(\hat{b}_1\right) = \sqrt{S^2\left(\hat{b}_1\right)} \qquad (26)$$

Τα κάτωθι στατιστικά ακολουθούν την **t** κατανομή με **T-2** βαθμούς ελευθερίας:

$$\frac{\hat{b}_0 - b_0}{\sum\left(\hat{b}_0\right)} \qquad (27)$$

$$\frac{\hat{b}_1 - b_1}{\sum\left(\hat{b}_1\right)} \qquad (28)$$

Τα εν λόγω στατιστικά ακολουθούν τώρα την **t** κατανομή με **T-2** βαθμούς ελευθερίας.

Ορίζοντας ένα επίπεδο σημαντικότητας **α= 5% ή 0.05** κατασκευάζουμε ένα διάστημα εμπιστοσύνης **1-α** για τις άγνωστες παράμετρες του πληθυσμού b_0 και b_1.

$$\hat{b}_0 - t \cdot a/2 \cdot S\hat{b}_0 \le b_0 \le \hat{b}_0 + t \cdot a/2 \cdot S\hat{b}_0 \qquad (29)$$

$$\hat{b}_1 - t \cdot a/2 \cdot S\hat{b}_1 \le b_0 \le \hat{b}_1 + t \cdot a/2 \cdot S\hat{b}_1 \qquad (30)$$

ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ ΓΙΑ ΤΟΥΣ ΣΥΝΤΕΛΕΣΤΕΣ bO ΚΑΙ b1

Στην οικονομετρία ενδιαφερόμαστε να ελέγξουμε στατιστικά αν πραγματικά υπάρχει σχέση μεταξύ της εξαρτημένης μεταβλητής και της ερμηνευτικής μεταβλητής. Αν δεν υπάρχει σχέση ανάμεσα στις μεταβλητές **Y** και **X** τότε η γραμμή παλινδρόμησης είναι παράλληλη προς τον οριζόντιο άξονα, δηλαδή **b1=0** και φυσικά **R2 =0**.

Επομένως η μηδέν (H_0) και η εναλλακτική υπόθεση (H_1) που θέλουμε να ελέγξουμε είναι:

$$H_0 : b_1 = 0 \qquad (31)$$

$$H_1 : b_1 \neq 0 \qquad (32)$$

Έχουμε δηλαδή έλεγχο σημαντικότητας γιατί ελέγχουμε εάν ο συντελεστής **b1** είναι στατιστικά σημαντικά διαφορετικός από το μηδέν. Εάν δηλαδή ισχύει η εναλλακτική υπόθεση H_1 τότε πράγματι ο συντελεστής είναι στατιστικά σημαντικός και επομένως η ερμηνευτική μεταβλητή (**X**) είναι σημαντική στην ερμηνεία συμπεριφοράς της εξαρτημένης μεταβλητής (**Y**).

Για να ισχύει η εναλλακτική υπόθεση θα πρέπει:

$$\left|t\right| = \frac{\hat{b}_1}{S\left(\hat{b}_1\right)} \geq t_{T-2,\,a/2} \qquad (33)$$

ΚΕΦΑΛΑΙΟ 4^{Ο}
ΣΤΑΤΙΣΤΙΚΟΙ ΕΛΕΓΧΟΙ

4.1 Έλεγχοι υποθέσεων, τύποι σφαλμάτων και διαδικασία ενός ελέγχου.

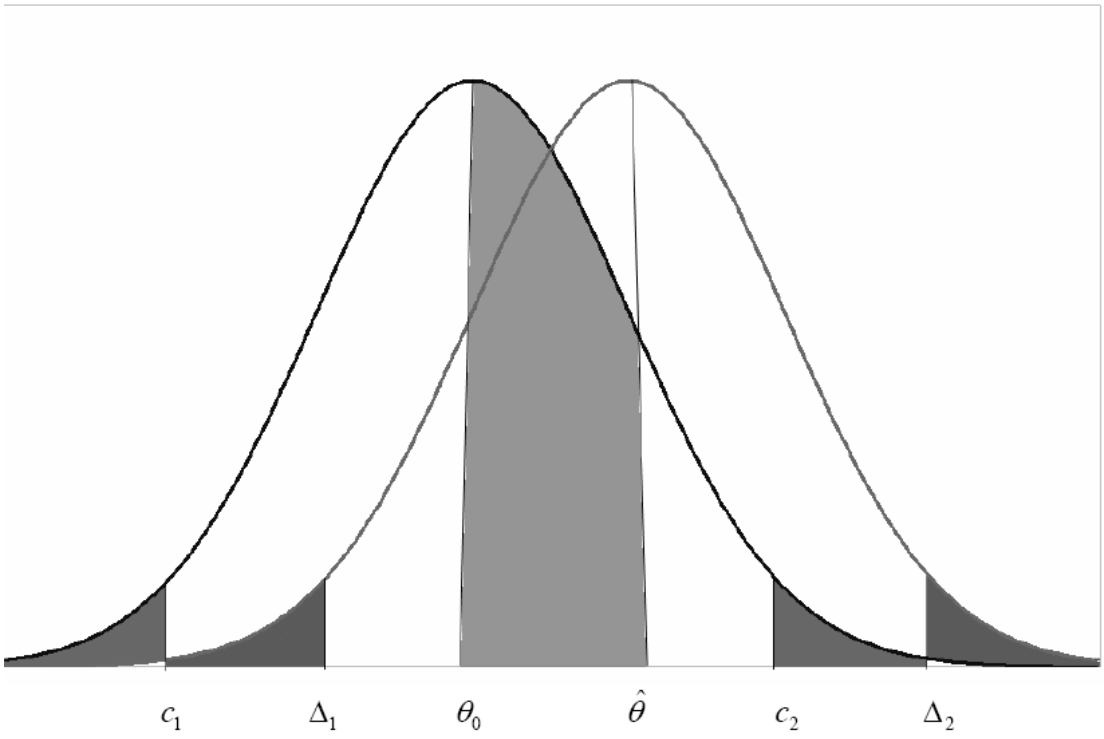
Στατιστική υπόθεση (statistical hypothesis) είναι μια υπόθεση που κάνουμε για την τιμή που μπορεί να πάρει μια άγνωστη παράμετρος της κατανομής ενός πληθυσμού ή για την κατανομή του πληθυσμού ή για την ανεξαρτησία μεταξύ δύο ή περισσότερων πληθυσμών**. Έλεγχος στατιστικής υπόθεσης** (test of a statistical hypothesis) είναι μια διαδικασία με την οποία ελέγχουμε την ισχύ μιας υπόθεσης, η οποία καλείται και **υπόθεση μηδέν** (null hypothesis) και συμβολίζεται με H_0, για την άγνωστη παράμετρο του πληθυσμού έναντι μιας **εναλλακτικής υπόθεσης** (alternative hypothesis), που συμβολίζεται με H_1, για αυτή τη παράμετρο. Το συμπέρασμα μιας διαδικασίας ελέγχου υπόθεσης θα είναι ότι είτε απορρίπτουμε την αρχική υπόθεση με πιθανότητα **1−α** ή όπως αλλιώς λέμε σε επίπεδο σημαντικότητας **α** , και δεχόμαστε την εναλλακτική υπόθεση, είτε δεχόμαστε την υπόθεση μηδέν (πιο σωστά δεν απορρίπτουμε την υπόθεση μηδέν) και απορρίπτουμε την εναλλακτική με πιθανότητα **1− α** ή σε **επίπεδο σημαντικότητας α**.

Οι υποθέσεις διακρίνονται σε απλές και σύνθετες υποθέσεις. Απλή υπόθεση είναι όταν προσδιορίζουμε μια συγκεκριμένη τιμή για την άγνωστη παράμετρο του πληθυσμού, ενώ σύνθετη υπόθεση είναι όταν δίνουμε ένα διάστημα τιμών.

Έστω πληθυσμός X ~ f(x |θ). Έστω ότι θέλουμε να ελέγξουμε την υπόθεση $H_0 : \theta_0 = 0$ έναντι της εναλλακτικής $H_1 : \theta_0 \neq 0$. Έστω $\hat{\Theta}$ είναι μια αμερόληπτη εκτιμήτρια της θ ($E(\hat{\Theta}) = \theta$) με κατανομή δειγματοληψίας $g(\hat{\Theta})$ και έστω $K \in \{x_1, x_2, ..., x_n\}$ ένα συγκεκριμένο τυχαίο δείγμα από τον πληθυσμό **X** και $\hat{q}$ μια εκτίμηση.

Είναι λογικό να πούμε ότι όσο πιο κοντά βρίσκεται η τιμή $\hat{q}$ στην τιμή q_0 τόσο πιο πειστικά μπορούμε να δεχθούμε ότι $q_0 = q$. Την πιθανότητα αυτή μπορούμε να την μετρήσουμε με την πιθανότητα :

P (0)=θ < Θ̂ < $\hat{q}$. Όσο πιο μικρή είναι αυτή η πιθανότητα τόσο πιο εύκολα μπορούμε να δεχθούμε ότι $q_0 = q$.



Διάγραμμα 19:Απόσταση της εκτίμηση της παραμέτρου από την πραγματική τιμή του πληθυσμού.

Πόσο μικρή όμως πρέπει να είναι αυτή η πιθανότητα; Η απάντηση είναι: ανάλογα με το επίπεδο εμπιστοσύνης που επιθυμούμε. Επομένως, θέτουμε κάποια όρια ανάλογα με το επίπεδο εμπιστοσύνης.

Περίπτωση 1η:

<u>1ος τρόπος</u>: Γνωρίζουμε ότι $P\left(-z_{a/2} < z < z_{a/2}\right) = 1-a$ όπου $z = \dfrac{\bar{x} - m}{\dfrac{s}{\sqrt{n}}}$.

Επομένως, τα όρια που καθορίζουν την περιοχή αποδοχής της αρχικής υπόθεσης του ελέγχου είναι, $\dfrac{c_1 - m_0}{\dfrac{s}{\sqrt{n}}} = -z_{a/2}$ και $\dfrac{c_2 - m_0}{\dfrac{s}{\sqrt{n}}} = z_{a/2}$. Λύνοντας

ως προς c_1, c_2 βρίσκουμε $c_1 = m_0 - z_{a/2} \cdot \dfrac{s}{\sqrt{n}}$ και $c_2 = m_0 + z_{a/2} \cdot \dfrac{s}{\sqrt{n}}$.

Οπότε, δεχόμαστε την H_0, σε επίπεδο σημαντικότητας a , τότε και μόνον τότε, όταν $m_0 - z_{a/2} \cdot \dfrac{s}{\sqrt{n}} < \bar{x} < m_0 + z_{a/2} \cdot \dfrac{s}{\sqrt{n}}$, ή διαφορετικά

$-z_{a/2} < \dfrac{\bar{x} - m_0}{\dfrac{s}{\sqrt{n}}} < z_{a/2}$, όπου $z = \dfrac{\bar{x} - m}{\dfrac{s}{\sqrt{n}}}$ είναι η τιμή της στατιστικής Z του

ελέγχου του δείγματος $\{x_1, x_2, ..., x_n\}$.

<u>2ος τρόπος</u>: Σύμφωνα με τα προηγούμενα, δεχόμαστε την H_0, σε επίπεδο σημαντικότητας a , τότε και μόνον τότε όταν $m_0 \in \Delta\text{Ε}(a)$ για τον m , δηλαδή όταν $m_0 \in \left(\bar{x} - z_{a/2} \cdot \dfrac{s}{\sqrt{n}}, \bar{x} + z_{a/2} \cdot \dfrac{s}{\sqrt{n}} \right) \Leftrightarrow \bar{x} - z_{a/2} \cdot \dfrac{s}{\sqrt{n}} < m_0 < \bar{x} + z_{a/2} \cdot \dfrac{s}{\sqrt{n}}$

$\Leftrightarrow -z_{a/2} < \dfrac{m_0 - \bar{x}}{\dfrac{s}{\sqrt{n}}} < z_{a/2} \Leftrightarrow -z_{a/2} < \dfrac{\bar{x} - m_0}{\dfrac{s}{\sqrt{n}}} < z_{a/2}$.

Η περιοχή $\left(-z_{a/2}, z_{a/2}\right)$ καλείται περιοχή αποδοχής του ελέγχου (acceptance range), ενώ η περιοχή $\left(-\infty,-z_{a/2}\right]\cup\left[z_{a/2},+\infty\right)$ είναι η περιοχή απόρριψης (rejection range) της H_0 και καλείται κρίσιμη περιοχή (critical range) του ελέγχου.

Περίπτωση 2η

Θέλουμε να ελέγξουμε την υπόθεση μηδέν $H_0: m = m_0$ έναντι της εναλλακτικής $H_1: m < m_0$ σε επίπεδο σημαντικότητας **α**. Ο έλεγχος αυτός, λόγω της μορφής της εναλλακτικής υπόθεσης, χαρακτηρίζεται ως μονόπλευρος έλεγχος (one-tail test).

<u>1ος τρόπος</u>: Δεχόμαστε σε επίπεδο σημαντικότητας α την $H_0 \Leftrightarrow \bar{x}\in\left(c,+\infty\right)$

$$\Leftrightarrow P\left(\bar{x} > c\right)_{m-m_0} = 1-a \Leftrightarrow P\left(\frac{\bar{x}-m_0}{\frac{s}{\sqrt{n}}} > \frac{c-m_0}{\frac{s}{\sqrt{n}}}\right)_{m-m_0} = 1-a$$

$\Leftrightarrow P(z > -z_a) = 1-a \Leftrightarrow \dfrac{c-m_0}{\frac{s}{\sqrt{n}}} = -z_a \Leftrightarrow c = m_0 - z_a \dfrac{s}{\sqrt{n}}$. Συνεπώς δεχόμαστε

την $H_0 \Leftrightarrow \bar{x} > m_0 - z_a \dfrac{s}{\sqrt{n}} \Leftrightarrow z = \dfrac{x-m_0}{\frac{s}{\sqrt{n}}} > -z_a$.

Η περιοχή αποδοχής του ελέγχου είναι τώρα η $\left(-z_a, +\infty\right)$, ενώ η κρίσιμη περιοχή του ελέγχου είναι η $\left(-\infty,-z_a\right]$. Να σημειωθεί ότι θέσαμε κάτω φράγμα για τον $\bar{x}$ στην πιθανότητα $P\left(\bar{x} > c\right)_{m-m_0} = 1-a$ για να ενισχύσουμε την H_1, κάνοντας έτσι αυστηρότερο τον έλεγχο για την H_0.

2ος τρόπος: Δεχόμαστε την H_0, σε επίπεδο σημαντικότητας α, τότε και μόνον τότε όταν $m_0 \in \Delta\Epsilon(a)$ για τον μ, δηλαδή όταν $m_0 \in \left(-\infty, \bar{x} + z_a \frac{s}{\sqrt{n}}\right)$

$$\Leftrightarrow m_0 < \bar{x} + z_a \frac{s}{\sqrt{n}} \Leftrightarrow \frac{m_0 - \bar{x}}{\frac{s}{\sqrt{n}}} < z_a \Leftrightarrow z = \frac{\bar{x} - m_0}{\frac{s}{\sqrt{n}}} > -z_a .$$

Περίπτωση 3η:

Θέλουμε να ελέγξουμε την υπόθεση μηδέν $H_0 : m = m_0$ έναντι της εναλλακτικής $H_1 : m < m_0$ σε επίπεδο σημαντικότητας α (μονόπλευρος έλεγχος).

1ος τρόπος: Δεχόμαστε σε επίπεδο σημαντικότητας **α** την $H_0 \Leftrightarrow \bar{x} \in (-\infty, c)$

$$\Leftrightarrow P(\bar{x} < c)_{m-m_0} = 1-a \Leftrightarrow P\left(\frac{\bar{x} - m_0}{\frac{s}{\sqrt{n}}} < \frac{c - m_0}{\frac{s}{\sqrt{n}}}\right)_{m-m_0} = 1-a$$

$$\Leftrightarrow P(z < -z_a) = 1-a \Leftrightarrow \frac{c - m_0}{\frac{s}{\sqrt{n}}} = z_a \Leftrightarrow c = m_0 + z_a \frac{s}{\sqrt{n}} .$$

Συνεπώς δεχόμαστε την $H_0 \Leftrightarrow \bar{x} < m_0 + z_a \frac{s}{\sqrt{n}} \Leftrightarrow z = \frac{\bar{x} - m}{\frac{s}{\sqrt{n}}} < z_a$.

Η περιοχή αποδοχής του ελέγχου είναι τώρα η $(-\infty, z_a)$, ενώ η κρίσιμη περιοχή του ελέγχου είναι η $[z_a, +\infty)$. Να σημειωθεί ότι θέσαμε κάτω φράγμα για τον $\bar{x}$ στην πιθανότητα $P(\bar{x} < c)_{m-m_0} = 1-a$ για να ενισχύσουμε την H_1, κάνοντας έτσι αυστηρότερο τον έλεγχο για την H_0.

<u>2ος τρόπος</u>: Δεχόμαστε την H_0, σε επίπεδο σημαντικότητας α, τότε και μόνον τότε όταν $m_0 \in \Delta\Epsilon(a)$ για τον μ, δηλαδή όταν $m_0 \in \left(\bar{x} - z_a \frac{s}{\sqrt{n}}, +\infty \right)$

$\Leftrightarrow m_0 > \bar{x} - z_a \frac{s}{\sqrt{n}} \Leftrightarrow \frac{m_0 - \bar{x}}{\frac{s}{\sqrt{n}}} > -z_a \Leftrightarrow z = \frac{\bar{x} - m_0}{\frac{s}{\sqrt{n}}} < z_a$.

Στους ελέγχους υποθέσεων υπάρχει η πιθανότητα να κάνουμε κάποιο σφάλμα, δηλαδή να απορρίψουμε μια υπόθεση H_0 ενώ αυτή είναι αληθής ή για να δεχθούμε μια υπόθεση H_0 ενώ αυτή δεν είναι αληθής.

Ειδικότερα, έχουμε δύο τύπους σφαλμάτων. Το σφάλμα τύπου I (type I error) είναι το γεγονός να δεχθώ την εναλλακτική υπόθεση, δηλαδή να απορρίψω την H_0 ενώ αυτή είναι αληθής, ενώ το σφάλμα τύπου II (type II error) είναι το γεγονός να δεχθώ την H_0 ενώ αυτή είναι ψευδής.

Η πιθανότητα αποφυγής του σφάλματος τύπου II καλείται **δύναμη του ελέγχου** (power of the test). Συγκεκριμένα για τον έλεγχο $H_0 : m = m_0$ έναντι της εναλλακτικής $H_1 : m \neq m_0$ ισχύει ο παρακάτω πίνακας (Πίνακας 13).

$H_0 : m = m_0$ $H_1 : m \neq m_0$	H_0: αληθής H_1: ψευδής	H_0: ψευδής H_1: αληθής
Δέχομαι H_0 Απορρίπτω H_1	Σωστή απόφαση $p = 1 - a$	Σφάλμα τύπου II $p(m) = b(m)$
Δέχομαι H_1 Απορρίπτω H_0	Σφάλμα τύπου I $p = a$	Σωστή απόφαση $p(m) = 1 - b(m)$

Πίνακας 13: Τύποι σφαλμάτων ενός ελέγχου υποθέσεων.

Η συνάρτηση που δίνει την πιθανότητα αποφυγής του σφάλματος τύπου ΙΙ και που εξαρτάται από την πραγματική τιμή της παραμέτρου του πληθυσμού, καλείται **δυναμοσυνάρτηση** (power function) του ελέγχου.

Είναι το μοντέλο στατιστικά σημαντικό ; .

Η μηδενική και η εναλλακτική υπόθεση, ελέγχονται σε επίπεδο σημαντικότητας α(=0.05 ή 0,01 ή…).

Ελέγχω την $H_0 : b_1 = b_2 = b_3 = ... = b_k = 0$

έναντι της H_1:Τουλάχιστον ένα b_i δεν είναι ίσο με 0 σε επίπεδο σημαντικότητας α.

Αν ισχύει η μηδενική υπόθεση και όλοι οι συντελεστές είναι 0, τότε το μοντέλο της παλινδρόμησης δεν είναι ικανό να προβλέψει ή να περιγράψει.

Ο έλεγχος της **F**, είναι μια μέθοδος με την οποία ελέγχουμε αν το μοντέλο παλινδρόμησης μπορεί να εξηγήσει ένα σημαντικό μέρος της μεταβλητότητας της εξαρτημένης μεταβλητής. Ο έλεγχος της στατιστικής **F** για την πολυδιάστατη παλινδρόμηση είναι:

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = \frac{MSR}{MSE}$$

Όπου: Πολυδιάστατος συντελεστής παλινδρόμησης $\left(R^2\right)$

$$R^2 = \frac{SSR}{TSS}$$

SSR: Άθροισμα τετραγώνων παλινδρόμησης
TSS: Συνολικό άθροισμα τετραγώνων
n: Αριθμός δεδομένων
k: Αριθμός ανεξάρτητων μεταβλητών
Βαθμοί ελευθερίας: $D_1 = k$ και $D_2 = n - k - 1$

Είναι οι μεταβλητές από μόνες τους σημαντικές;

$H_0 : b_i = 0$, δεδομένου ότι όλες οι άλλες μεταβλητές είναι ήδη στο μοντέλο

$H_1 : b_i \neq 0$, για κάθε i

Ο έλεγχος των υποθέσεων, μπορεί να γίνει χρησιμοποιώντας τον έλεγχο της t.

$$t = \frac{b_i - 0}{s_{b_i}}$$

όπου:

b_i : Συντελεστής κλίσης δείγματος για την ανεξάρτητη **i** μεταβλητή

s_{b_i} : Εκτίμηση τυπικού σφάλματος για τον **i** συντελεστή κλίσης του δείγματος

Βαθμοί ελευθερίας= **n-k-1**

4.2 Εκτίμηση διαστημάτων εμπιστοσύνης και έλεγχοι στατιστικής σημαντικότητας

Έστω το υπόδειγμα:

$$Y_i = b_0 + b_1 \cdot X_{1i} + b_2 \cdot X_{2i} + ... + b_k \cdot X_{ki} + e_i \qquad (1)$$

Ως γνωστόν: $\hat{b} = b + (X'X)^{-1} X'e$ και $e \in N(0, s^2 I)$ οπότε:

$$\hat{b} \in N\left(b_1, s^2 (X'X)^{-1}\right) \qquad (2)$$

αφού $E(\hat{b}) = b$ και $Cov(b) = s^2 (X'X)^{-1}$, $Cov(e) = s^2 I$

Από την (2) έχουμε ότι:

$$\hat{b_1} \in N\left(b_1, s^2 C_{ii}\right),\ i = 0,1,2,...,k \qquad (3)$$

όπου C_{ii} είναι το i διαγώνιο στοιχείο του πίνακα $(X'X)^{-1}$. Το γινόμενο $s^2 C_{ii}$ εκφράζει τη διακύμανση της εκτιμήτριας $\hat{b_1}$.

Για την απλή παλινδρόμηση $(k=1)$, σύμφωνα με τις σχέσεις

$V\left(\hat{b_0}\right)=\frac{s^2\sum X_{1i}^2}{n\sum x_{1i}^2}, V\left(\hat{b_1}\right)=\frac{s^2}{\sum x_{1i}^2}$ και $Cov\left(\hat{b_0},\hat{b_1}\right)=-\frac{s^2\bar{X}}{\sum x_{1i}^2}$ έχουμε:

$$\hat{b_0}\in \mathrm{N}\left(b_0, \frac{s^2\sum X_i^2}{n\sum x_i^2}\right) \qquad (4)$$

όπου $x_i = X_i - \bar{X}$ και

$$\hat{b_1}\in \mathrm{N}\left(b_1, \frac{s^2}{\sum x_i^2}\right) \qquad (5)$$

Για την πολλαπλή παλινδρόμηση με δύο ερμηνευτικές μεταβλητές $(k=2)$, σύμφωνα με τις σχέσεις

$$\begin{cases} V\left(\hat{b_1}\right)=s^2\dfrac{\sum x_{2i}^2}{\sum x_{1i}^2\sum x_{2i}^2-\left(\sum x_{1i}x_{2i}\right)^2} \\ \left\{V\left(\hat{b_2}\right)=s^2\dfrac{\sum x_{1i}^2}{\sum x_{1i}^2\sum x_{2i}^2-\left(\sum x_{1i}x_{2i}\right)^2}\right. \\ Cov\left(\hat{b_1},\hat{b_2}\right)=-s^2\dfrac{\sum x_{1i}x_{2i}}{\sum x_{1i}^2\sum x_{2i}^2-\left(\sum x_{1i}x_{2i}\right)^2} \end{cases}$$

έχουμε:

$$\hat{b_1}\in \mathrm{N}\left(b_{1,}\frac{s^2\sum x_{2i}^2}{\sum x_{1i}^2\sum x_{2i}^2-\left(\sum x_{1i}x_{2i}\right)^2}\right) \qquad (6)$$

και

$$\hat{b}_2 \in \mathrm{N}\left(b_{2,} \frac{s^2 \sum x_{1i}^2}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i}\right)^2} \right) \qquad (7)$$

όπου $x_{1i} = X_{1i} - \bar{X}_1$ και $x_{2i} = X_{2i} - \bar{X}_2$.

Προφανώς από (2):

$$\frac{\hat{b}_1 - b_1}{s\sqrt{C_{ii}}} \in \mathrm{N}(0,1), i = 0,1,2,...,k \qquad (8)$$

ενώ, όπως γνωρίζουμε $\frac{\hat{e}'\hat{e}}{s^2} = \frac{e'P'Pe}{s^2} = \frac{e'PPe}{s^2} = \frac{e'Pe}{s^2} \cdot x^2_{trace(P)} = x^2_{n-k-1}$, άρα

$$\frac{\sum_{i=1}^{n} \hat{e}_i^2}{s^2}' = \frac{\hat{e}'\hat{e}}{s^2} \cdot x^2_{n-k-1} \qquad (9)$$

Οπότε:

$$\frac{\hat{b}_1 - b_1}{\hat{s}_{\hat{b}_i}} = \frac{\hat{b}_1 - b_1}{\hat{s}\sqrt{C_{ii}}} = \frac{\frac{\hat{b}_1 - b_1}{s\sqrt{C_{ii}}}}{\sqrt{\frac{(n-k-1)\hat{s}^2}{s^2} \Big/ (n-k-1)}} \cdot t_{n-k-1} \qquad (10)$$

Άρα ένα διάστημα εμπιστοσύνης, σε επίπεδο σημαντικότητα **α** , για την παράμετρο b_1 του υποδείγματος (1) θα είναι

$$\left(\hat{b}_1 - t_{n-k-1,\frac{a}{2}} \cdot \hat{s}_{\hat{b}_1}, \hat{b}_1 + t_{n-k-1,\frac{a}{2}} \cdot \hat{s}_{\hat{b}_1} \right)$$

όπου $\hat{s}_{\hat{b}_i} = \hat{s}\sqrt{C_{ii}}$ είναι το τυπικό της εκτιμήτριας $\hat{b}_1$ της παραμέτρου b_1.

4.3 Έλεγχος της υπόθεσης

$$b_i = b_i^*$$

Με βάση την (9) μπορούμε να ελέγξουμε την υπόθεση $H_0 : b_i = b_i^*$, όπου b_i^* είναι η υποτιθέμενη τιμή για την άγνωστη παράμετρο **i** του υποδείγματος (1).

Έλεγχος της υπόθεσης $b_i = b_i^*$

Υπόθεση μηδέν Ho	Στατιστική ελέγχου	Εναλλακτική υπόθεση H1	Κρίσιμη περιοχή
$b_i = b_i^*$	$t = \frac{\hat{b_i} = b_i^*}{\hat{s}_{\hat{b_i}}}$	$\hat{b_i} \neq b_i^*$	$t \leq -t_{n-k-1,a/2}$ ή $t \geq -t_{n-k-1,a/2}$
		$\hat{b_i} < b_i^*$	$t \leq -t_{n-k-1,a}$
		$\hat{b_i} > b_i^*$	$t \geq -t_{n-k-1,a}$

Οι κρίσιμες τιμές λαμβάνονται από τους πίνακες της κατανομής **t** του student. Αν ***n−k−1>30*** (μεγάλο δείγμα), τότε οι κρίσιμες αυτές τιμές μπορούν να αντικατασταθούν από τις κρίσιμες τιμές a_z και $a_{z/2}$ της τυποποιημένης κανονική κατανομή στα αντίστοιχα επίπεδα σημαντικότητας.

Αν $b_i = 0$, τότε ο έλεγχος της υπόθεσης $H_0 : b_i = 0$ έναντι της εναλλακτικής $H_1 : b_i \neq 0$ (ή $H_0 : b_i < 0$ ή $H_1 : b_i > 0$) είναι μεγάλης σπουδαιότητας για την αξιολόγηση μιας εξίσωσης παλινδρόμησης και γίνεται πάντοτε για όλους τους συντελεστές , και καλείται έλεγχος στατιστικής σημαντικότητας των συντελεστών παλινδρόμησης.

Αν για κάποια παράμετρο b_i του υποδείγματος (1) απορριφθεί σε επίπεδο σημαντικότητας ***a*** η υπόθεση μηδέν $H_0 : b_i = 0$, οπότε γίνεται δεκτή η εναλλακτική υπόθεση $H_1 : b_i \neq 0$ (ή $H_0 : b_i < 0$ ή $H_1 : b_i > 0$), τότε η παράμετρος β***i*** αλλά και η αντίστοιχη εκτιμήτρια $\hat{b_i}$ καλείται στατιστικά σημαντική σε επίπεδο σημαντικότητας ***a*** , που σημαίνει στατιστικά σημαντική διάφορη του μηδενός (ή αλλιώς με πιθανότητα **1− *a*** η παράμετρος αυτή είναι διάφορη του μηδενός). Αυτό είναι καλό για το θεωρούμενο υπόδειγμα (1) γιατί σημαίνει ότι η αντίστοιχη ερμηνευτική μεταβλητή X_i επιδρά (με γραμμικό τρόπο) στην διαμόρφωση των τιμών της ερμηνευτικής μεταβλητής **Y** και καλώς συμπεριλήφθη στο β μέλος του υποδείγματος (1).

Αντιθέτως, αν η υπόθεση μηδέν $H_0 : b_i = 0$ γίνει δεκτή σε επίπεδο σημαντικότητας **α** , τότε η παράμετρος b_i (αλλά και η αντίστοιχη εκτιμήτρια b_i) καλείται μη στατιστικά σημαντική (διάφορη του μηδενός) σε επίπεδο σημαντικότητα **α** , δηλαδή με πιθανότητα **1− *a*** μπορεί να θεωρηθεί μηδέν. Αυτό είναι κακό για το θεωρούμενο υπόδειγμα (1) γιατί σημαίνει ότι η αντίστοιχη ερμηνευτική μεταβλητή X_i δεν ασκεί γραμμική επίδραση στην ερμηνευόμενη μεταβλητή **Y** και κακώς συμπεριλήφθη στο β μέλος του υποδείγματος (1).

<u>Ανάλυση της διακύμανσης της εξαρτημένης μεταβλητής</u>

Ας θεωρήσουμε το κλασσικό γραμμικό υπόδειγμα **y**=**Xβ** +**ε** με **k** ερμηνευτικές μεταβλητές. Ο έλεγχος της υπόθεση μηδέν:

- $H_0 : b_1 = b_2 = ... = b_K = 0$

ότι δηλαδή όλες οι παράμετροι του υποδείγματος, εξαιρουμένου του σταθερού όρου είναι μηδέν, έναντι της εναλλακτικής υπόθεσης:

- H_1: μία τουλάχιστον παράμετρος είναι διάφορος του μηδέν,

που είναι γνωστή ως ***F − test*** ή ως έλεγχος της ερμηνευτικότητας της εξίσωσης παλινδρόμησης. Ο έλεγχος αυτός στην περίπτωση της απλής παλινδρόμησης αφορά την υπόθεση: $H_0 : b = 0$, έναντι της εναλλακτικής $H_1 : b \neq 0$ και αποδεικνύεται ότι είναι ισοδύναμος με το αντίστοιχο ***t−test*** για την παράμετρο αυτή. Ο έλεγχος της υπόθεσης μπορεί να γίνει μέσω του πίνακα της ανάλυσης διακύμανσης της εξαρτημένης μεταβλητής.

Πηγή Διακύμανσης	**Άθροισμα Τετραγώνων**	**Βαθμοί Ελευθερίας**	**Μέσο Τετράγωνο**	**Στατιστική F**
Γραμμική επίδραση των $X_1, X_2, ..., X_k$	$\sum \hat{y}_i^2$	k	$\sum \hat{y}_i^2 \Big/ k$	$F = \dfrac{\sum \hat{y}_i^2 \Big/ k}{\sum \hat{e}_i^2 \Big/ n-k+1}$
Σφάλματα	$\sum \hat{e}_i^2$	n-k+1	$\sum \hat{e}_i^2 \Big/ n-k+1$	
Σύνολο	$\sum y^2{}_i$	n-1		

Πίνακας 14: Ανάλυση διακύμανσης της εξαρτημένης μεταβλητής

Αποδεικνύεται ότι η στατιστική **F** ακολουθεί την κατανομή *Fk,n- k*+1. Η κρίσιμη περιοχή του ελέγχου (περιοχή απόρριψης της H_0) σε επίπεδο σημαντικότητας *a* είναι:

$$F = \frac{\sum \hat{y}_i^2 \Big/ k}{\sum \hat{e}_i^2 \Big/ (n-k+1)} \geq F_{k,n-k+1,a} \; (= f^u_{k,n-k+1,a})$$

Είναι:

$$F = \frac{\sum \hat{y}_i^2 \Big/ k}{\sum \hat{e}_i^2 \Big/ (n-k+1)} = \frac{\dfrac{\sum \hat{y}_i^2}{\sum y_i^2}}{\dfrac{\sum \hat{e}_i^2}{\sum y_i^2}} \cdot \frac{(n-k+1)}{k} = \frac{R^2}{1-R^2} \cdot \frac{(n-k+1)}{k}$$

Άρα ο έλεγχος της υπόθεσης μπορεί να γίνει και μέσω της στατιστικής.

4.4 Έλεγχος της στατιστικής σημαντικότητας των συντελεστών μερικής συσχέτισης

Η κατανομή δειγματοληψίας των συντελεστών μερικής συσχέτισης ομοιάζει προς την κατανομή **t** , μόνον όμως όταν ο αντίστοιχος συντελεστής μερικής συσχέτισης στον πληθυσμό είναι μηδέν. Για τιμές του συντελεστή μερικής συσχέτισης στον πληθυσμό που αποκλίνουν από το μηδέν, η κατανομή δειγματοληψίας του αντίστοιχου δειγματικού συντελεστή συσχέτισης γίνεται ασύμμετρη και μάλιστα τόσο ασύμμετρη όσο η τιμή του συντελεστή μερικής συσχέτισης στον πληθυσμό ως προς τα άκρα του διαστήματος $[-1,1]$. Αν στο γραμμικό υπόδειγμα Y_j με *k* ερμηνευτικές μεταβλητές, *ρ* είναι ο συντελεστής μερικής συσχέτισης μεταξύ των Y_j και *X , j= 1, 2,...,k,* και *Y* είναι ο αντίστοιχος δειγματικός συντελεστής μερικής συσχέτισης τότε ο έλεγχος της υπόθεσης $H_0 : Y_j = 0$ έναντι της εναλλακτικής $H_1 : Y_j \neq 0$, δηλαδή ο έλεγχος της στατιστικής σημαντικότητας της Y_j, σε επίπεδο σημαντικότητας *a* γίνεται μέσω της στατιστικής:

$$t = \frac{r_{Yj}\sqrt{n-k+1}}{\sqrt{1-r_{Yj}^2}} t_{n-k+1}$$

Με κρίσιμη περιοχή την $t \leq -t_{n-k+1,\frac{a}{2}}$ ή $t \geq -t_{n-k+1,\frac{a}{2}}$

\

4.5 Εφαρμογή της Ανάλυσης της Παλινδρόμησης στον Επενδυτικό Κίνδυνο

Οι επενδύσεις στο χρηματιστήριο είναι ελκυστικές σε όλους. Παρόλα αυτά, οι χρηματιστηριακές επενδύσεις μεταφέρουν και το στοιχείο του κινδύνου. Ο κίνδυνος που αντιστοιχεί σε κάθε μετοχή, μπορεί να μετρηθεί με δύο τρόπους. Ο πρώτος είναι ο **συστηματικός κίνδυνος** (systematic risk), που εξηγεί την μεταβλητότητα που δημιουργείται στην αξία της μετοχής, καθώς η αγορά κινείται πάνω ή κάτω, η αξία τείνει να κινείται και αυτή προς την ίδια κατεύθυνση. Ο δείκτης Standar & Poor's (S&P) 500 είναι το πιο συνηθισμένο μέτρο που χρησιμοποιείται στην αγορά. Ο δεύτερος τύπος κινδύνου καλείται **ειδικός κίνδυνος** (specific risk), και δείχνει την μεταβλητότητα που οφείλεται σε άλλου παράγοντες, όπως είναι η δυνατότητα αποδοχών της εταιρείας , οι στρατηγικές αποκτήσεων και τα λοιπά. Ο ειδικός κίνδυνος υπολογίζεται από το τυπικό σφάλμα της εκτίμησης.

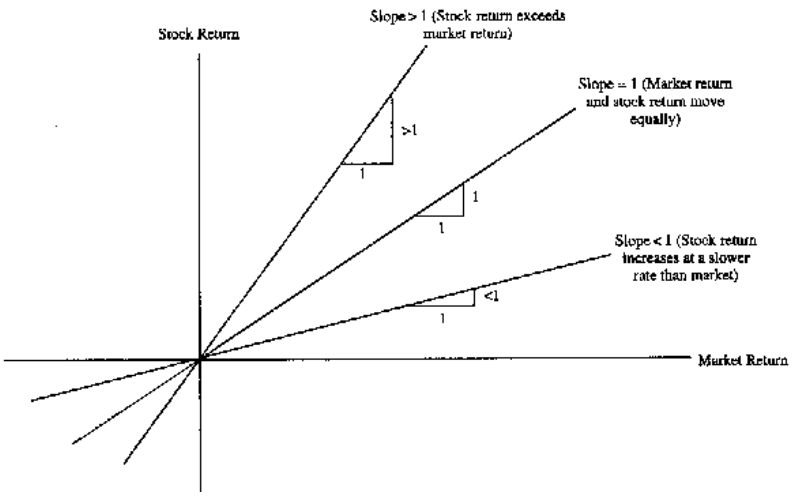
Ο συστηματικός κίνδυνος χαρακτηρίζεται από ένα μέτρο που καλείται **βήτα** (beta) . Οι τυποποιημένες τιμές, γνωστές και ως beta τιμές, προκύπτουν σαν εκτιμηθείσες τιμές των παραμέτρων του μοντέλου, αφού πρώτα

μετατρέψουμε τις τιμές της εξαρτημένης και ανεξάρτητης μεταβλητής σε τυποποιημένες.

Στην απλή γραμμική παλινδρόμηση, η beta τιμή της κλίσης ισούται με τον συντελεστή συσχέτισης των δύο μεταβλητών. Μια αξία beta που ισοδυναμεί με 1,0 δηλώνει ότι η συγκεκριμένη μετοχή θα ακολουθεί τις μετακινήσεις της αγοράς, ενώ ένα beta μικρότερο από 1,0 δείχνει ότι η μετοχή είναι λιγότερο ασταθής από τη αγορά . Ένα beta μεγαλύτερο από 1,0 δείχνει ότι η μετοχή έχει μεγαλύτερη διακύμανση από την αγορά. Ακόμα, μετοχές με μεγαλύτερες τιμές beta είναι περισσότερο επικίνδυνες από αυτές με χαμηλότερες τιμές beta.

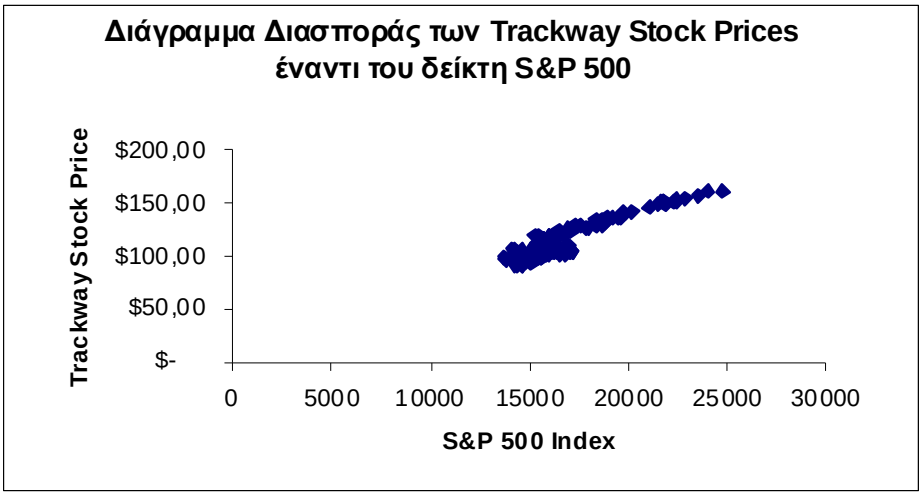
Οι τιμές beta μπορούν να υπολογισθούν αναπτύσσοντας ένα μοντέλο παλινδρόμησης με τις αποδόσεις των μετοχών (εξαρτημένη μεταβλητή) έναντι του μέσου όρου των αποδόσεων της αγοράς (ανεξάρτητη μεταβλητή). Η κλίση της γραμμής παλινδρόμησης ισοδυναμεί με τον κίνδυνο beta. $X_j, j = 1, 2, ..., k$. Αν σχεδιάσουμε τη μορφή των αποδόσεων της αγοράς έναντι της απόδοσης της κάθε μετοχής και βρούμε την γραμμή παλινδρόμησης, τότε θα παρατηρήσουμε ότι η κλίση ισούται με την μονάδα, δηλαδή οι μετοχές μεταβάλλονται κατά το ίδιο ποσοστό με την αγορά. Παρόλα αυτά, αν η τιμή της μετοχής μεταβάλλεται λιγότερο από ότι μεταβάλλεται η αγορά, τότε η κλίση της γραμμής παλινδρόμησης θα είναι μικρότερη από την μονάδα, ενώ η κλίση θα είναι μεγαλύτερη από την μονάδα, όταν η τιμή της μετοχής μεταβάλλεται περισσότερο από ότι μεταβάλλεται η αγορά. Η αρνητική κλίση δείχνει ότι η μετοχή μετακινείται προς την αντίθετη κατεύθυνση από εκείνη της αγοράς.

Για παράδειγμα αν η αγορά κινείται προς τα πάνω, η τιμή της μετοχής πέφτει προς τα κάτω.



Διάγραμμα 20

Το φύλλο εργασίας Stock, που παίρνουμε από την βάση δεδομένων της επιχείρησης Tracway , περιλαμβάνει τις ημερήσιες τιμές της μετοχής Tracway για την περίοδο από 30 Ιουνίου μέχρι της 31 Δεκεμβρίου του 2000. Το διάγραμμα (21) είναι ένα διάγραμμα διασποράς της απόδοσης του δείκτη S&P500 και της απόδοσης της μετοχής Tracway για μια περίοδο έξι μηνών, αντίστοιχα. Φαίνεται καθαρά η συσχέτιση που εμφανίζεται να υπάρχει.



Διάγραμμα 21

Η ποσοστιαία αύξηση (αρνητικές τιμές δείχνουν μείωση) και για τον S&P500 και για τις μετοχές της Tracway, υπάρχει στο φύλλο εργασίας Stock, από όπου παίρνουμε τα απαραίτητα στοιχεία για να δημιουργήσουμε το μοντέλο παλινδρόμησης.

Ημερήσια μεταβολή στην τιμή της μετοχής Tracway = β_0 + β_1 μεταβολή του S&P500

Ο πίνακας (15) δείχνει τα αποτελέσματα της εφαρμογής της παλινδρόμησης από το Excel. Το μοντέλο που προκύπτει είναι :

Ημερήσια μεταβολή στην τιμή της μετοχής Tracway = 0,00124 + 0,62124 μεταβολή του S&P500

Η τιμή του συντελεστή προσδιορισμού (R^2 = 0,90) δείχνει ότι ένα μεγάλο ποσοστό της μεταβλητότητας εξηγείται από το μοντέλο. Η κλίση της γραμμής παλινδρόμησης, β_1 , (ο κίνδυνος beta της μετοχής Tracway) είναι 0,62. Αυτό δείχνει ότι η Tracway έχει μικρότερο κίνδυνο από την μετοχή S&P500 .

Stock

Στατιστικά παλινδρόμησης	
Πολλαπλό R	*0.9506849*
R Τετράγωνο	*0.9038017*
Προσαρμοσμένο R Τετράγωνο	*0.9030196*
Τυπικό σφάλμα	*0.010344*
Μέγεθος δείγματος	*125*

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ

	βαθμοί ελευθερίας	*SS*	*MS*	*F*	*Σημαντικότητα F*
Παλινδρόμηση	*1*	*0.1236491*	*0.1236491*	*1155.609*	*2.2E-64*
Υπόλοιπο	*123*	*0.0131609*	*0.000107*		
Σύνολο	*124*	*0.1368099*			

	Συντελεστές	*Τυπικό σφάλμα*	*t*	*τιμή-P*	*Κατ. 95%*	*Υψηλ. 95%*	*Κατ. 95.0%*	*Υψηλ. 95.0%*
Τεταγμένη επί την αρχή	*0.0012396*	*0.0009261*	*1.338491*	*0.1832053*	*-0.0005936*	*0.0030728*	*-0.0005936*	*0.003073*
S&P % change	*0.62124*	*0.0182749*	*33.994249*	*2.2E-64*	*0.5850661*	*0.657414*	*0.5850661*	*0.657414*

Πίνακας 15

ΚΕΦΑΛΑΙΟ 5^{ο}
ΧΡΗΣΗ ΤΟΥ SPSS

Η πρώτη έκδοση του **SPSS** (Statistical Package for the Social Sciences) εμφανίστηκε για πρώτη φορά στην αγορά το 1968 από τους Norman H Nie και C. Hadlai Hull. Είναι το πιο διαδεδομένο πρόγραμμα στατιστικής ανάλυσης στην κοινωνία των επιστημών. Το 1970 χαρακτηρίστηκε ως «ένα από τα σημαντικότερα βιβλία που επηρεάζουν την Κοινωνιολογία».Μεταξύ του 2009 και 2010 το πρώτο λογοσμικό του SPSS ονομαζόταν PASW (Predictive Analytics SoftWare) Statistics. Η εταιρεία ανακοίνωσε τον Ιούλιο του 2009 ότι αποκτήθηκε από την IBM για 1,2 δισεκατομμύρια δολλάρια και από τον Ιανουάριο του 2010 μετονομάστηκε σε SPSS: An IBM Company.

Στον πίνακα που ακολουθεί έχουμε εισάγει τα ποσοστά της ανεργίας στην Ελλάδα σε χιλιάδες, από το 1ο τρίμηνο του 2005 μέχρι και το 2^{ο} τρίμηνο του 2009.Ως εξαρτημένη μεταβλητή θέτουμε την ανεργία, ενώ ως ανεξάρτητες, τους απόφοιτους μεταπτυχιακού τίτλου, ανώτατης εκπαίδευσης, ανώτερης τεχνικής σχολής, μέσης εκπαίδευσης, γυμνασίου, όσων έχουν απολυτήριο στοιχειώδους εκπαίδευσης, όσων δεν έχουν τελειώσει τη στοιχειώδη εκπαίδευση και όσων δεν έχουν πάει καθόλου σχολείο.

Με τη χρήση του SPSS δημιουργούμε γραφήματα που μας δείχνουν αν υπάρχει συσχέτιση μεταξύ της εξαρτημένης μεταβλητής (ανεργία) και κάθε ανεξάρτητης μεταβλητής ξεχωριστά, καθώς επίσης με τα δεδομένα που εξάγουμε από το SPSS κάνουμε έλεγχο υποθέσεων για να δούμε αν υφίσταται παλινδρόμηση, κι αν όντως υφίσταται , κάνουμε έλεγχο στατιστικής σημαντικότητας του εκτιμητή b. Ορίζουμε την εξίσωση παλινδρόμησης και βρίσκουμε αν έχουμε ασθενή, μέτρια ή ισχυρή γραμμική συσχέτιση στο μοντέλο μας.

5.1 Εισαγωγή δεδομένων στο spss

ΤΡΙΜΗΝΑ	ΑΝΕΡΓΟΙ	ΜΕΤΑΠΤΥΧΙΑ-ΚΟΣ ΤΙΤΛΟΣ	ΑΝΩΤΑΤΗ ΕΚΠΑΙΔΕΥΣΗ	ΑΝΩΤΕΡΗ ΤΕΧΝΙΚΗ ΣΧΟΛΗ	ΜΕΣΗ ΕΚΠΑΙ-ΔΕΥΣΗ	ΑΠΟΦΟΙΤΟΙ ΓΥΜΝΑΣΙΟΥ	ΑΠΟΛΥΤΗΡΙΟ ΣΤΟΙΧΕΙΩΔΟ-ΥΣ ΕΚΠΑΙΔΕΥ-ΣΗΣ	ΔΕΝ ΤΕΛΕΙΩ-ΣΑΝ ΤΗ ΣΤΟΙ-ΧΕΙΩΔΗ ΕΚΠΑΙΔΕΥΣΗ	ΔΕΝ ΠΗΓΑΝ ΚΑΘΟΛΟΥ ΣΧΟΛΕΙΟ
Α' ΤΡΙΜΗΝΟ 2005	502,40	3,4	55,3	94,3	182,5	64,2	97,6	2,6	2,6
Β' ΤΡΙΜΗΝΟ 2005	466,90	4,2	56,6	88,6	171,5	61,2	80,6	2,2	1,9
Γ' ΤΡΙΜΗΝΟ 2005	469,80	5,4	60,4	93,8	176,8	54,2	74,9	1,9	2,4
Δ' ΤΡΙΜΗΝΟ 2005	470,90	5,9	56,2	97,2	176,4	57,3	74,1	1,7	2,0
Α' ΤΡΙΜΗΝΟ 2006	473,10	4,0	51,8	94,1	180,0	59,0	81,1	1,1	2,0
Β' ΤΡΙΜΗΝΟ 2006	427,40	4,9	48,7	87,9	160,5	52,8	70,5	1,1	1,2
Γ' ΤΡΙΜΗΝΟ 2006	408,30	5,9	52,6	88,8	152,6	44,5	61,2	1,4	1,1
Δ' ΤΡΙΜΗΝΟ 2006	429,10	6,0	47,1	89,1	166,6	49,3	67,6	2,0	1,5
Α' ΤΡΙΜΗΝΟ 2007	445,60	5,8	47,7	92,2	166,2	51,6	77,1	1,9	3,0
Β' ΤΡΙΜΗΝΟ 2007	398,00	5,7	47,4	81,5	145,0	48,6	66,3	1,8	1,8
Γ' ΤΡΙΜΗΝΟ 2007	387,50	6,9	54,9	80,4	138,2	44,6	59,2	1,8	1,4
Δ' ΤΡΙΜΗΝΟ 2007	396,50	5,8	50,8	84,9	144,6	43,0	62,2	2,5	2,7
Α' ΤΡΙΜΗΝΟ 2008	406,50	5,2	44,2	86,7	149,1	51,0	65,9	2,1	2,2
Β' ΤΡΙΜΗΝΟ 2008	357,10	5,4	43,7	80,5	125,0	43,5	55,3	1,6	2,2
Γ' ΤΡΙΜΗΝΟ 2008	355,10	6,3	46,8	81,8	124,5	44,6	49,2	1,2	0,7
Δ' ΤΡΙΜΗΝΟ 2008	392,70	5,2	47,2	85,3	144,3	50,3	57,6	1,3	1,4
Α' ΤΡΙΜΗΝΟ 2009	462,30	6,0	49,5	93,0	171,4	60,7	77,5	2,0	2,2
Β' ΤΡΙΜΗΝΟ 2009	442,60	7,4	52,7	84,2	160,7	61,2	71,7	2,6	2,1

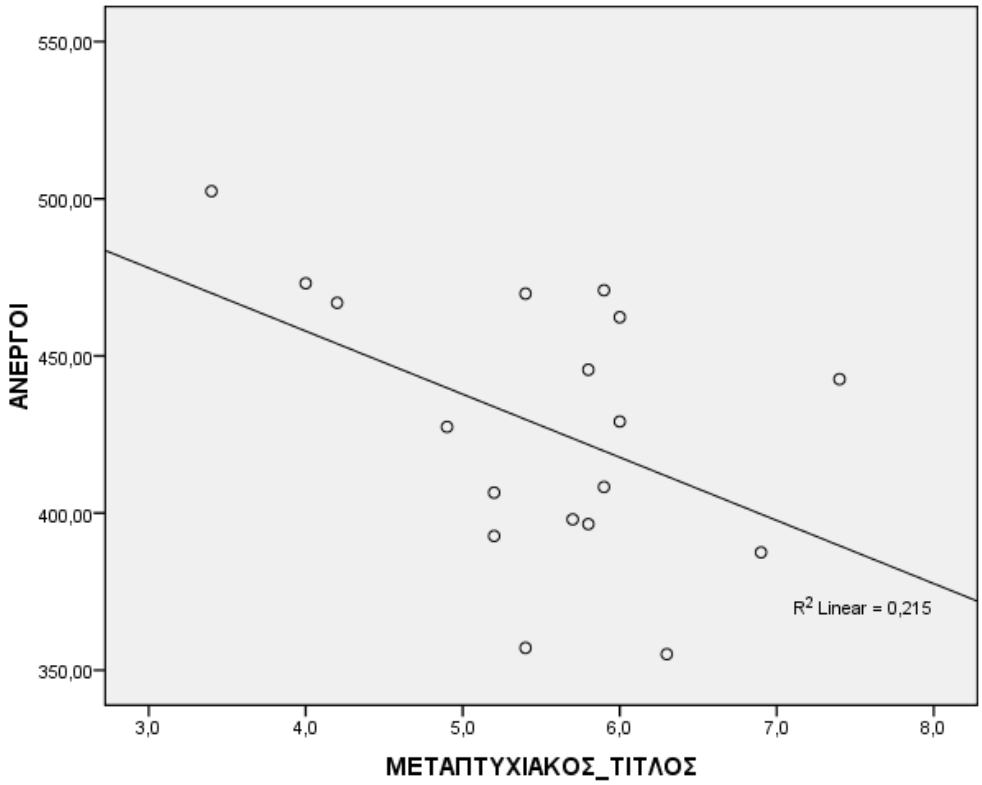
(πηγή:http://www.statistics.gr/portal/page/portal/ESYE/PAGE-customthemes?p_param=A0101&period=quarterly)

5.2 Αποτελέσματα από τη χρήση του spss

```
GET
FILE='F:\esye.sav'.
GRAPH
/SCATTERPLOT(BIVAR)=ΜΕΤΑΠΤΥΧΙΑΚΟΣ ΤΙΤΛΟΣ WITH
ΑΝΕΡΓΟΙ
/MISSING=LISTWISE.
```

Graph

[DataSet1] F:\esye.sav



Υπάρχει ασθενής γραμμική συσχέτιση.

```
REGRESSION
   /MISSING LISTWISE
   /STATISTICS COEFF OUTS R ANOVA
   /CRITERIA=PIN(.05) POUT(.10)
   /NOORIGIN
   /DEPENDENT ΑΝΕΡΓΟΙ
   /METHOD=ENTER ΜΕΤΑΠΤΥΧΙΑΚΟΣ_ΤΙΤΛΟΣ.
```

Regression

[DataSet1] F:\esye.sav

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	ΜΕΤΑΠΤΥΧΙΑΚΟΣ ΤΙΤΛΟΣ[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: ΑΝΕΡΓΟΙ

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,463[a]	,215	,166	38,58389

a.Predictors:(Constant),ΜΕΤΑΠΤΥΧΙΑΚΟΣ ΤΙΤΛΟΣ

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6513,620	1	6513,620	4,375	,053[a]
	Residual	23819,471	16	1488,717		
	Total	30333,091	17			

a. Predictors: (Constant), ΜΕΤΑΠΤΥΧΙΑΚΟΣ_ΤΙΤΛΟΣ

b. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients
	B	Std. Error	Beta
1 (Constant)	538,220	53,792	
ΜΕΤΑΠΤΥΧΙΑΚΟΣ_ΤΙΤΛΟΣ	-20,082	9,601	-,463

a. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	t	Sig.
1 (Constant)	10,006	,000
ΜΕΤΑΠΤΥΧΙΑΚΟΣ_ΤΙΤΛΟΣ	-2,092	,053

a. Dependent Variable: ΑΝΕΡΓΟΙ

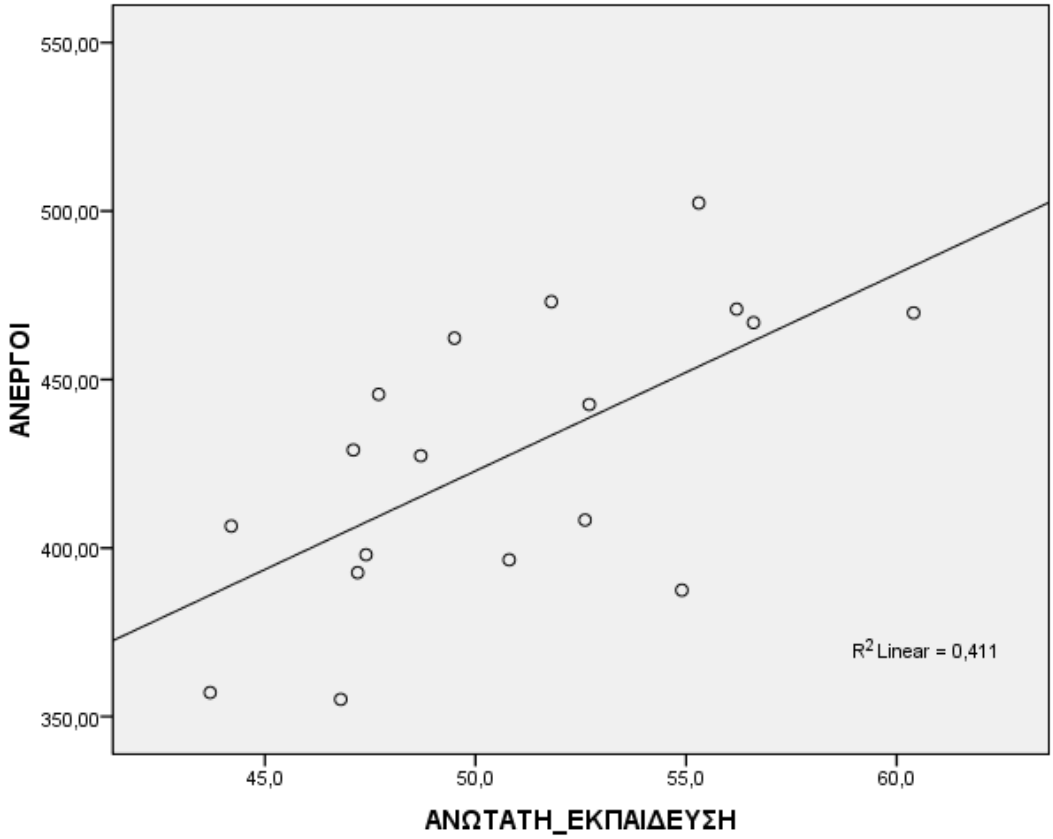
H_0: όλοι οι συντελεστές (εκτός του â) = 0 {δεν υφίσταται η έννοια της παλινδρόμησης}
H_1: όλοι οι συντελεστές ή έστω ένας $\neq 0$ {υφίσταται η έννοια της παλινδρόμησης}

$\rho=0,053 > \alpha=0,05$
Αποδέχομαι την $H_{0,}$ άρα δεν υφίσταται η έννοια της παλινδρόμησης.

```
GRAPH
  /SCATTERPLOT(BIVAR)=ΑΝΩΤΑΤΗ_ΕΚΠΑΙΔΕΥΣΗ WITH ΑΝΕΡΓΟΙ
  /MISSING=LISTWISE.
```

Graph

[DataSet1] F:\esye.sav



Υπάρχει μέση γραμμική συσχέτιση.

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
 /DEPENDENT ΑΝΕΡΓΟΙ
  /METHOD=ENTER ΑΝΩΤΑΤΗ_ΕΚΠΑΙΔΕΥΣΗ.
```

Regression

[DataSet1] F:\esye.sav

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	ΑΝΩΤΑΤΗ ΕΚΠΑΙΔΕΥΣΗ[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: ΑΝΕΡΓΟΙ

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,641[a]	,411	,374	33,43016

a.Predictors:(Constant),ΑΝΩΤΑΤΗ_ΕΚΠΑΙΔΕΥΣΗ

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	12451,887	1	12451,887	11,142	,004[a]
	Residual	17881,205	16	1117,575		
	Total	30333,091	17			

a. Predictors: (Constant), ΑΝΩΤΑΤΗ_ΕΚΠΑΙΔΕΥΣΗ

b. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients
	B	Std. Error	Beta
1 (Constant)	130,351	89,317	
ΑΝΩΤΑΤΗ ΕΚΠΑΙ-ΔΕΥΣΗ	5,851	1,753	,641

a. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	t	Sig.
1 (Constant)	1,459	,164
ΑΝΩΤΑΤΗ ΕΚΠΑΙΔΕΥΣΗ	3,338	,004

a. Dependent Variable: ΑΝΕΡΓΟΙ

H_0: όλοι οι συντελεστές (εκτός του â) = 0 {δεν υφίσταται η έννοια της παλινδρόμησης}
H_1: όλοι οι συντελεστές ή έστω ένας $\neq 0$ {υφίσταται η έννοια της παλινδρόμησης}

$\rho=0,04< \alpha=0,05$
Απορρίπτω την $H_{0,}$ άρα υφίσταται η έννοια της παλινδρόμησης.

H_0:b=0 {ο συντελεστής b δεν είναι στατιστικά σημαντικός}
H_1:b$\neq$ 0 {ο συντελεστής b είναι στατιστικά σημαντικός}

$\rho=0,04< \alpha=0,05$
Απορρίπτω την $H_{0,}$ άρα ο συντελεστής b είναι στατιστικά σημαντικός

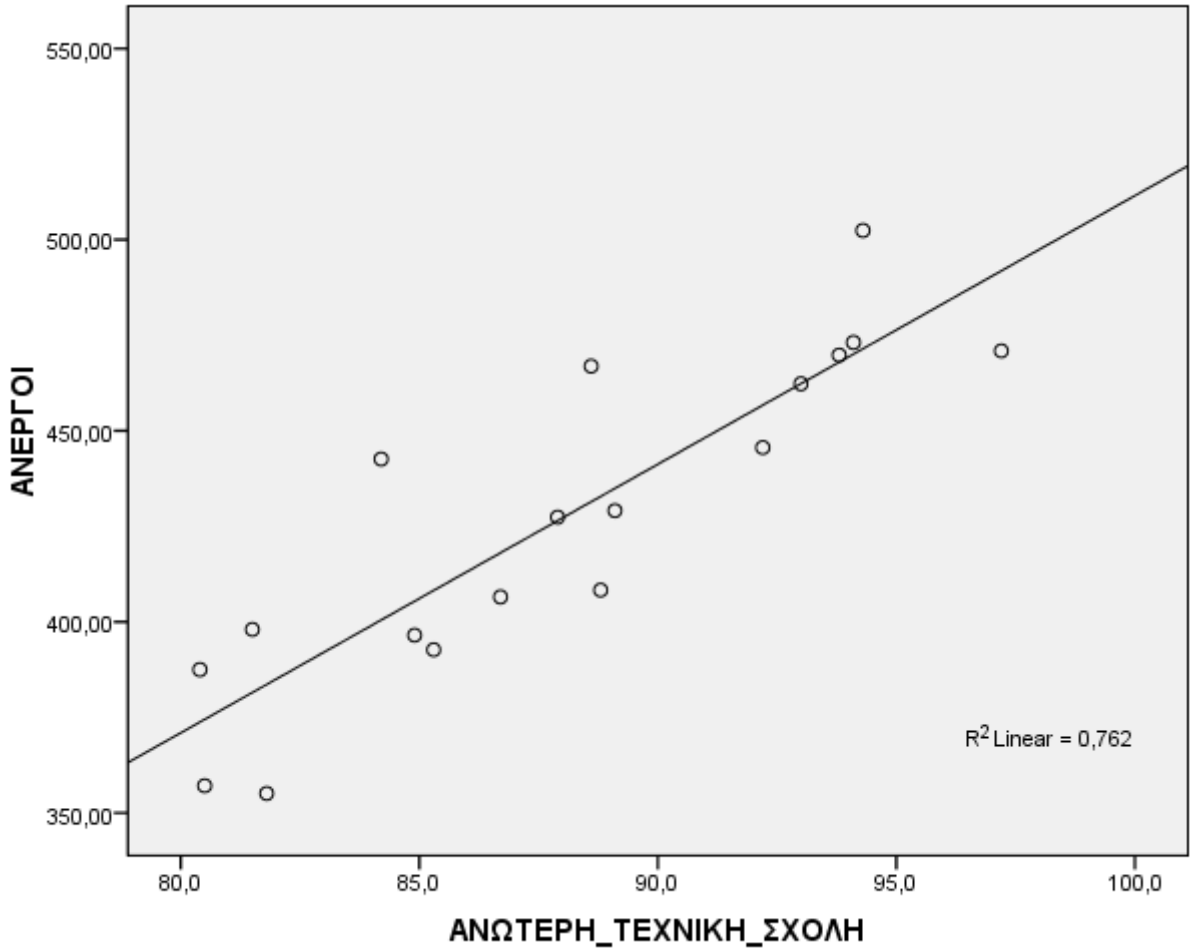
Εξίσωση παλινδρόμησης: y=130,351+5,851x

Το 41,1% της μεταβλητότητας των ανέργων ερμηνεύεται πλήρως από τη μεταβλητότητα των ατόμων με πτυχίο ανώτατης εκπαίδευσης.
R=0,641 Φανερώνει μια μέση γραμμική συσχέτιση.

```
GRAPH
  /SCATTERPLOT(BIVAR)=ΑΝΩΤΕΡΗ_ΤΕΧΝΙΚΗ_ΣΧΟΛΗ WITH
  ΑΝΕΡΓΟΙ
  /MISSING=LISTWISE.
```

Graph

[DataSet1] F:\esye.sav



Υπάρχει πολύ ισχυρή γραμμική συσχέτιση.

```
REGRESSION
    /MISSING LISTWISE
    /STATISTICS COEFF OUTS R ANOVA
    /CRITERIA=PIN(.05) POUT(.10)
    /NOORIGIN
    /DEPENDENT ΑΝΕΡΓΟΙ
    /METHOD=ENTER ΑΝΩΤΕΡΗ_ΤΕΧΝΙΚΗ_ΣΧΟΛΗ.
```

Regression

[DataSet1] F:\esye.sav

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	ΑΝΩΤΕΡΗ ΤΕΧΝΙΚΗ ΣΧΟΛΗ[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: ΑΝΕΡΓΟΙ

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,873[a]	,762	,747	21,25713

a.Predictors:(Constant),ΑΝΩΤΕΡΗ ΤΕΧΝΙΚΗ ΣΧΟΛΗ

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23103,245	1	23103,245	51,129	,000[a]
	Residual	7229,847	16	451,865		
	Total	30333,091	17			

a. Predictors: (Constant), ΑΝΩΤΕΡΗ_ΤΕΧΝΙΚΗ_ΣΧΟΛΗ

b. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients
	B	Std. Error	Beta
1 (Constant)	-191,246	86,653	
ΑΝΩΤΕΡΗ ΤΕΧΝΙΚΗ ΣΧΟΛΗ	7,028	,983	,873

a. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	t	Sig.
1 (Constant)	-2,207	,042
ΑΝΩΤΕΡΗ_ΤΕΧΝΙΚΗ_ΣΧΟΛΗ	7,150	,000

a. Dependent Variable: ΑΝΕΡΓΟΙ

H_0: όλοι οι συντελεστές (εκτός του â) = 0 {δεν υφίσταται η έννοια της παλινδρόμησης}
H_1: όλοι οι συντελεστές ή έστω ένας $\neq 0$ {υφίσταται η έννοια της παλινδρόμησης}

$\rho=0< \alpha=0,05$
Απορρίπτω την H_0, άρα υφίσταται η έννοια της παλινδρόμησης.

H_0:b=0 {ο συντελεστής b δεν είναι στατιστικά σημαντικός}
H_1:b$\neq$ 0 {ο συντελεστής b είναι στατιστικά σημαντικός}

$\rho=0< \alpha=0,05$
Απορρίπτω την H_0, άρα ο συντελεστής b είναι στατιστικά σημαντικός

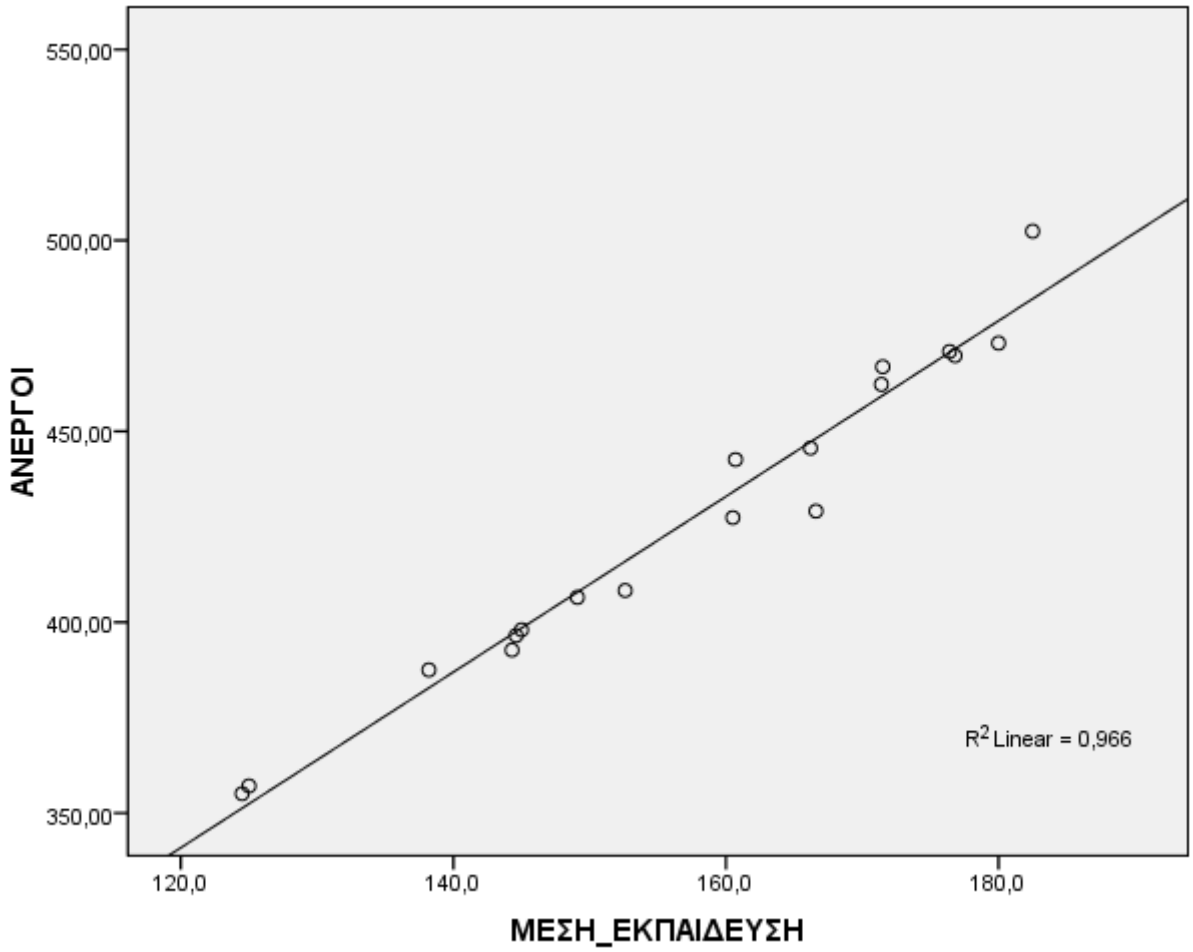
Εξίσωση παλινδρόμησης: y=-191,246+7,028x

Το 76,2% της μεταβλητότητας των ανέργων ερμηνεύεται πλήρως από τη μεταβλητότητα των ατόμων με πτυχίο ανώτερης τεχνικής σχολής.
R=0,873 Φανερώνει μια πολύ ισχυρή γραμμική συσχέτιση.

```
GRAPH
  /SCATTERPLOT(BIVAR)=ΜΕΣΗ_ΕΚΠΑΙΔΕΥΣΗ WITH ΑΝΕΡΓΟΙ
  /MISSING=LISTWISE.
```

Graph

[DataSet1] F:\esye.sav



Υπάρχει πολύ ισχυρή γραμμική συσχέτιση.

```
REGRESSION
    /MISSING LISTWISE
    /STATISTICS COEFF OUTS R ANOVA
    /CRITERIA=PIN(.05) POUT(.10)
    /NOORIGIN
    /DEPENDENT ΑΝΕΡΓΟΙ
    /METHOD=ENTER ΜΕΣΗ_ΕΚΠΑΙΔΕΥΣΗ
```

Regression

[DataSet1] F:\esye.sav

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	ΜΕΣΗ ΕΚΠΑΙΔΕΥΣΗ[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: ΑΝΕΡΓΟΙ

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,983[a]	,966	,964	8,01429

a. Predictors: (Constant), ΜΕΣΗ ΕΚΠΑΙΔΕΥΣΗ

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	29305,430	1	29305,430	456,266	,000[a]
	Residual	1027,661	16	64,229		
	Total	30333,091	17			

a. Predictors: (Constant), ΜΕΣΗ ΕΚΠΑΙΔΕΥΣΗ

b. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
1 (Constant)	64,868	17,073		3,799	,002
ΜΕΣΗ ΕΚΠΑΙ-ΔΕΥΣΗ	2,301	,108	,983	21,360	,000

a. Dependent Variable: ΑΝΕΡΓΟΙ

H_0: όλοι οι συντελεστές (εκτός του â) = 0 {δεν υφίσταται η έννοια της παλινδρόμησης}
H_1: όλοι οι συντελεστές ή έστω ένας $\neq 0$ {υφίσταται η έννοια της παλινδρόμησης}

ρ=0< α=0,05
Απορρίπτω την H_0, άρα υφίσταται η έννοια της παλινδρόμησης.

H_0:b=0 {ο συντελεστής b δεν είναι στατιστικά σημαντικός}
H_1:b≠ 0 {ο συντελεστής b είναι στατιστικά σημαντικός}
ρ=0< α=0,05
Απορρίπτω την H_0, άρα ο συντελεστής b είναι στατιστικά σημαντικός

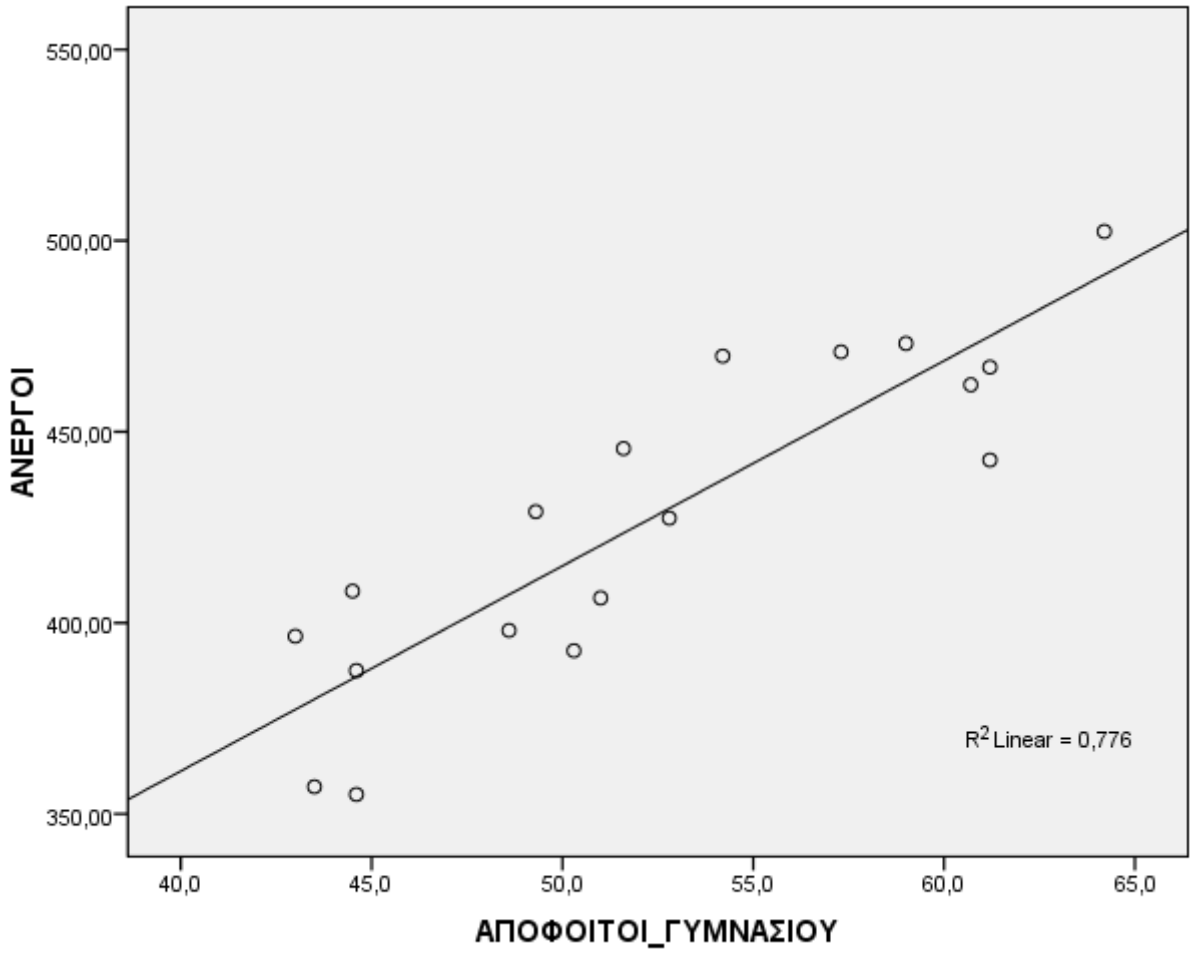
Εξίσωση παλινδρόμησης: y=64,868+2,301x

Το 96,6% της μεταβλητότητας των ανέργων ερμηνεύεται πλήρως από τη μεταβλητότητα των ατόμων με απολυτήριο μέσης εκπαίδευσης.
R=0,983 Φανερώνει μια πολύ ισχυρή γραμμική συσχέτιση.

```
GRAPH
    /SCATTERPLOT(BIVAR)=ΑΠΟΦΟΙΤΟΙ ΓΥΜΝΑΣΙΟΥ WITH
   ΑΝΕΡΓΟΙ
    /MISSING=LISTWISE.
```

Graph

[DataSet1] F:\esye.sav



Υπάρχει πολύ ισχυρή γραμμική συσχέτιση.

```
REGRESSION
    /MISSING LISTWISE
    /STATISTICS COEFF OUTS R ANOVA
    /CRITERIA=PIN(.05) POUT(.10)
    /NOORIGIN
    /DEPENDENT ΑΝΕΡΓΟΙ
    /METHOD=ENTER ΑΠΟΦΟΙΤΟΙ_ΓΥΜΝΑΣΙΟΥ.
```

Regression

[DataSet1] F:\esye.sav

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	ΑΠΟΦΟΙΤΟΙ ΓΥΜΝΑΣΙΟΥ[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: ΑΝΕΡΓΟΙ

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,881[a]	,776	,762	20,62244

a.Predictors:(Constant),ΑΠΟΦΟΙΤΟΙ ΓΥΜΝΑΣΙΟΥ

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	23528,529	1	23528,529	55,324	,000[a]
	Residual	6804,562	16	425,285		
	Total	30333,091	17			

a. Predictors: (Constant), ΑΠΟΦΟΙΤΟΙ ΓΥΜΝΑΣΙΟΥ

b. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients
	B	Std. Error	Beta
1 (Constant)	146,522	38,064	
ΑΠΟΦΟΙΤΟΙ ΓΥΜ-ΝΑΣΙΟΥ	5,368	,722	,881

a. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	t	Sig.
1 (Constant)	3,849	,001
ΑΠΟΦΟΙΤΟΙ ΓΥΜΝΑΣΙΟΥ	7,438	,000

a. Dependent Variable: ΑΝΕΡΓΟΙ

H_0: όλοι οι συντελεστές (εκτός του â) = 0 {δεν υφίσταται η έννοια της παλινδρόμησης}
H_1: όλοι οι συντελεστές ή έστω ένας $\neq 0$ {υφίσταται η έννοια της παλινδρόμησης}
$\rho=0 < \alpha=0,05$
Απορρίπτω την H_0, άρα υφίσταται η έννοια της παλινδρόμησης.

H_0:b=0 {ο συντελεστής b δεν είναι στατιστικά σημαντικός}
H_1:b$\neq$ 0 {ο συντελεστής b είναι στατιστικά σημαντικός}
$\rho=0 < \alpha=0,05$
Απορρίπτω την H_0, άρα ο συντελεστής b είναι στατιστικά σημαντικός

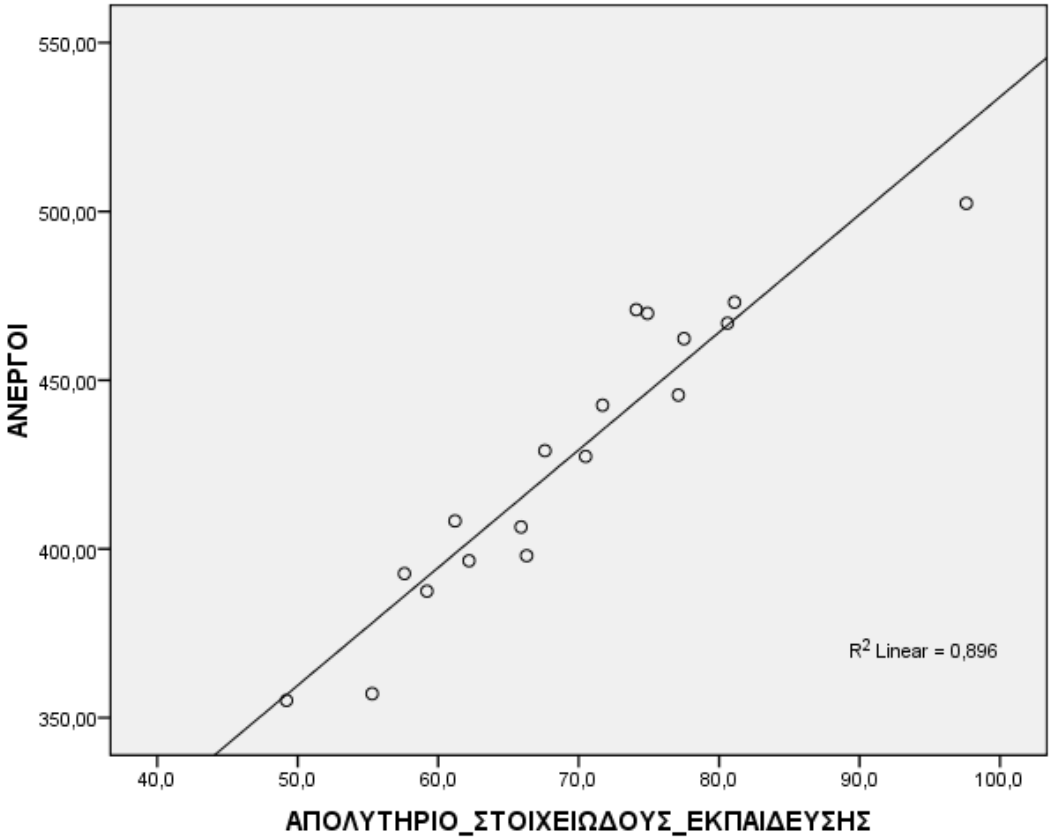
Εξίσωση παλινδρόμησης: y=146,522+5,368x

Το 77,6% της μεταβλητότητας των ανέργων ερμηνεύεται πλήρως από τη μεταβλητότητα των ατόμων με απολυτήριο γυμνασίου.
R=0,881 Φανερώνει μια πολύ ισχυρή γραμμική συσχέτιση.

GRAPH
/SCATTERPLOT(BIVAR)=**ΑΠΟΛΥΤΗΡΙΟ ΣΤΟΙΧΕΙΩΔΟΥΣ ΕΚΠΑΙΔΕΥΣΗΣ** WITH **ΑΝΕΡΓΟΙ**
/MISSING=LISTWISE.

Graph

[DataSet1] F:\esye.sav



Υπάρχει πολύ ισχυρή γραμμική συσχέτιση.

REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT ΑΝΕΡΓΟΙ
/METHOD=ENTER ΑΠΟΛΥΤΗΡΙ-
Ο_ΣΤΟΙΧΕΙΩΔΟΥΣ_ΕΚΠΑΙΔΕΥΣΗΣ

Regression

[DataSet1] F:\esye.sav

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	ΑΠΟΛΥΤΗΡΙΟ ΣΤΟΙΧΕΙΩΔΟΥΣ ΕΚΠΑΙΔΕΥΣΗΣ[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: ΑΝΕΡΓΟΙ

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,947[a]	,896	,890	14,02412

a.Predictors:(Constant),ΑΠΟΛΥΤΗΡΙΟ ΣΤΟΙΧΕΙΩΔΟΥΣ ΕΚΠΑΙΔΕΥ-ΣΗΣ

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	27186,278	1	27186,278	138,229	,000[a]
	Residual	3146,813	16	196,676		
	Total	30333,091	17			

a. Predictors: (Constant), ΑΠΟΛΥΤΗΡΙΟ ΣΤΟΙΧΕΙΩΔΟΥΣ ΕΚΠΑΙ-ΔΕΥΣΗΣ

b. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients
	B	Std. Error	Beta
1 (Constant)	185,233	20,855	
ΑΠΟΛΥΤΗΡΙΟ ΣΤΟΙΧΕΙΩ-ΔΟΥΣ ΕΚΠΑΙΔΕΥΣΗΣ	3,487	,297	,947

a. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	t	Sig.
1 (Constant)	8,882	,000
ΑΠΟΛΥΤΗΡΙΟ ΣΤΟΙΧΕΙΩΔΟΥΣ ΕΚΠΑΙΔΕΥΣΗΣ	11,757	,000

a. Dependent Variable: ΑΝΕΡΓΟΙ

H_0: όλοι οι συντελεστές (εκτός του â) = 0 {δεν υφίσταται η έννοια της παλινδρόμησης}
H_1: όλοι οι συντελεστές ή έστω ένας $\neq 0$ {υφίσταται η έννοια της παλινδρόμησης}

ρ=0< α=0,05
Απορρίπτω την $H_{0,}$ άρα υφίσταται η έννοια της παλινδρόμησης.

H_0:b=0 {ο συντελεστής b δεν είναι στατιστικά σημαντικός}
H_1:b≠ 0 {ο συντελεστής b είναι στατιστικά σημαντικός}

ρ=0< α=0,05
Απορρίπτω την $H_{0,}$ άρα ο συντελεστής b είναι στατιστικά σημαντικός

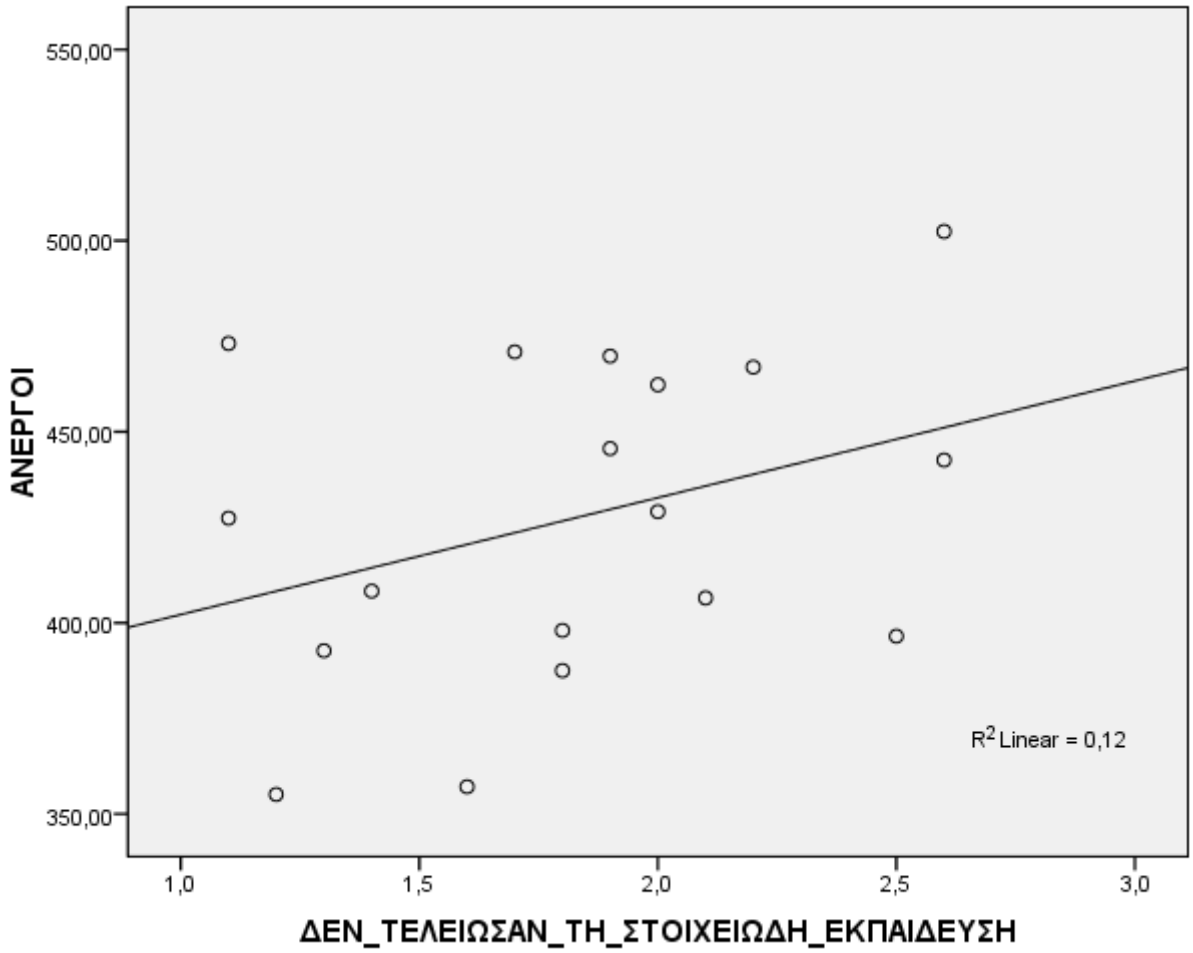
Εξίσωση παλινδρόμησης: y=158,233+3,487x

Το 89,6% της μεταβλητότητας των ανέργων ερμηνεύεται πλήρως από τη μεταβλητότητα των ατόμων με απολυτήριο στοιχειώδους εκπαίδευσης.
R=0,947 Φανερώνει μια πολύ ισχυρή γραμμική συσχέτιση.

GRAPH
/SCATTERPLOT(BIVAR)=**ΔΕΝ ΤΕΛΕΙΩΣΑΝ ΤΗ ΣΤΟΙΧΕΙΩΔΗ ΕΚΠΑΙΔΕΥΣΗ** WITH **ΑΝΕΡΓΟΙ**
/MISSING=LISTWISE.

Graph

[DataSet1] F:\esye.sav



Υπάρχει ασθενής γραμμική συσχέτιση.

REGRESSION
/MISSING LISTWISE
/STATISTICS COEFF OUTS R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT ΑΝΕΡΓΟΙ
/METHOD=ENTER
ΔΕΝ_ΤΕΛΕΙΩΣΑΝ_ΤΗ_ΣΤΟΙΧΕΙΩΔΗ_ΕΚΠΑΙΔΕΥΣΗ

Regression

[DataSet1] F:\esye.sav

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	ΔΕΝ ΤΕΛΕΙΩΣΑΝ ΤΗ ΣΤΟΙΧΕΙΩΔΗ ΕΚΠΑΙΔΕΥΣΗ[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: ΑΝΕΡΓΟΙ

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,347[a]	,120	,066	40,83362

a. Predictors: (Constant), ΔΕΝ ΤΕΛΕΙΩΣΑΝ ΤΗ ΣΤΟΙΧΕΙΩΔΗ ΕΚΠΑ-ΙΔΕΥΣΗ

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3654,936	1	3654,936	2,192	,158[a]
	Residual	26678,155	16	1667,385		
	Total	30333,091	17			

a. Predictors: (Constant), ΔΕΝ ΤΕΛΕΙΩΣΑΝ ΤΗ ΣΤΟΙΧΕΙΩΔΗ ΕΚΠΑ-ΙΔΕΥΣΗ

b. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients
	B	Std. Error	Beta
1 (Constant)	371,618	38,836	
ΔΕΝ ΤΕΛΕΙΩΣΑΝ ΤΗ ΣΤΟΙΧΕΙΩΔΗ ΕΚΠΑΙΔΕΥΣΗ	30,570	20,648	,347

a. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	t	Sig.
1 (Constant)	9,569	,000
ΔΕΝ ΤΕΛΕΙΩΣΑΝ ΤΗ ΣΤΟΙΧΕΙΩΔΗ ΕΚΠΑΙΔΕΥΣΗ	1,481	,158

a. Dependent Variable: ΑΝΕΡΓΟΙ

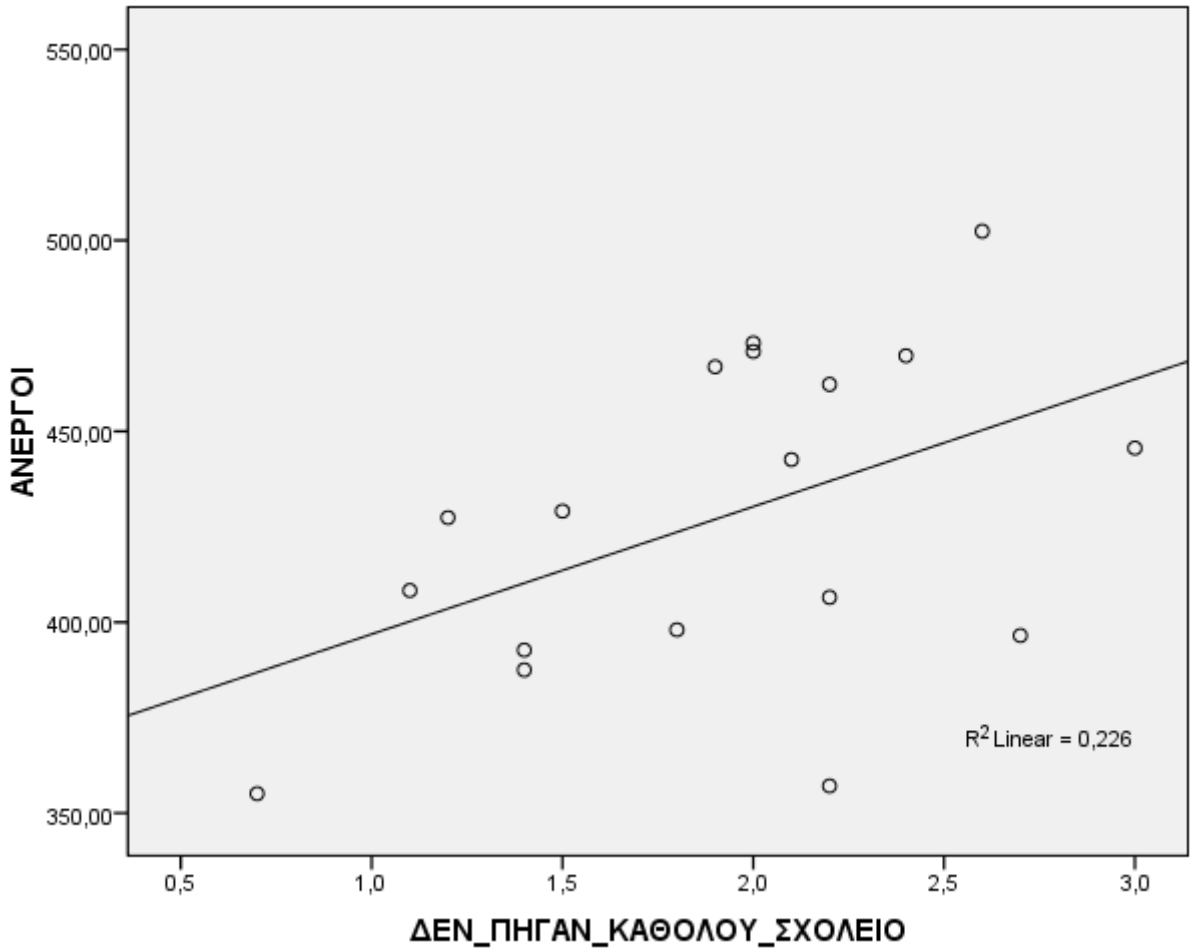
H_0: όλοι οι συντελεστές (εκτός του â) = 0 {δεν υφίσταται η έννοια της παλινδρόμησης}
H_1: όλοι οι συντελεστές ή έστω ένας $\neq 0$ {υφίσταται η έννοια της παλινδρόμησης}

$\rho=0,158 > \alpha=0,05$
Αποδέχομαι την H_0, άρα δεν υφίσταται η έννοια της παλινδρόμησης.

```
GRAPH
    /SCATTERPLOT(BIVAR)=ΔΕΝ ΠΗΓΑΝ ΚΑΘΟΛΟΥ ΣΧΟΛΕΙΟ
   WITH ΑΝΕΡΓΟΙ
    /MISSING=LISTWISE.
```

Graph

[DataSet1] F:\esye.sav



Υπάρχει ασθενής γραμμική συσχέτιση.

```
REGRESSION
    /MISSING LISTWISE
    /STATISTICS COEFF OUTS R ANOVA
    /CRITERIA=PIN(.05) POUT(.10)
    /NOORIGIN
    /DEPENDENT ΑΝΕΡΓΟΙ
    /METHOD=ENTER ΔΕΝ_ΠΗΓΑΝ_ΚΑΘΟΛΟΥ_ΣΧΟΛΕΙΟ.
```

Regression

[DataSet1] F:\esye.sav

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	ΔΕΝ_ΠΗΓΑΝ_ΚΑΘΟΛΟΥ_ΣΧΟΛΕΙΟ[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: ΑΝΕΡΓΟΙ

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,476[a]	,226	,178	38,29382

a. Predictors: (Constant), ΔΕΝ_ΠΗΓΑΝ_ΚΑΘΟΛΟΥ_ΣΧΟΛΕΙΟ

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6870,425	1	6870,425	4,685	,046[a]
	Residual	23462,667	16	1466,417		
	Total	30333,091	17			

a. Predictors: (Constant), ΔΕΝ_ΠΗΓΑΝ_ΚΑΘΟΛΟΥ_ΣΧΟΛΕΙΟ

b. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients
	B	Std. Error	Beta
1 (Constant)	363,486	30,842	
ΔΕΝ_ΠΗΓΑΝ_ΚΑΘΟΛΟΥ_ΣΧΟΛΕΙΟ	33,403	15,432	,476

a. Dependent Variable: ΑΝΕΡΓΟΙ

Coefficients[a]

Model	t	Sig.
1 (Constant)	11,785	,000
ΔΕΝ_ΠΗΓΑΝ_ΚΑΘΟΛΟΥ_ΣΧΟΛΕΙΟ	2,165	,046

a. Dependent Variable: ΑΝΕΡΓΟΙ

H_0: όλοι οι συντελεστές (εκτός του â) = 0 {δεν υφίσταται η έννοια της παλινδρόμησης}
H_1: όλοι οι συντελεστές ή έστω ένας $\neq 0$ {υφίσταται η έννοια της παλινδρόμησης}

$\rho=0,046 < \alpha=0,05$
Απορρίπτω την $H_{0,}$ άρα υφίσταται η έννοια της παλινδρόμησης.

H_0:b=0 {ο συντελεστής b δεν είναι στατιστικά σημαντικός}
H_1:b≠ 0 {ο συντελεστής b είναι στατιστικά σημαντικός}
$\rho=0,046 < \alpha=0,05$
Απορρίπτω την $H_{0,}$ άρα ο συντελεστής b είναι στατιστικά σημαντικός

Εξίσωση παλινδρόμησης: y=363,486+33,403x

Το 22,6% της μεταβλητότητας των ανέργων ερμηνεύεται πλήρως από τη μεταβλητότητα των ατόμων που δεν πήγαν καθόλου σχολείο.
R=0,476 Φανερώνει μια ασθενής γραμμική συσχέτιση.

ΠΗΓΕΣ - ΒΙΒΛΙΟΓΡΑΦΙΑ

- Χαλικιάς, Ιωάννης Γ, ΣΤΑΤΙΣΤΙΚΗ, Rosil, ΑΘΗΝΑ, 2003
- Ιωαννίδης, Δημήτρης, Στατιστικές μέθοδοι, Ζήτη, Θεσσαλονίκη, 2005
- Χρήστου, Γεώργιος Κ, Εισαγωγή στην οικονομετρία Α και Β ΤΟΜΟΣ, Gutenberg, Αθήνα, 2003
- Ζαχαροπούλου, Χρυσούλα, Στατιστική, Σοφία, 2005
- ΛΑΖΑΡΙΔΗΣ ΑΛΕΞΑΝΔΡΟΣ, Στατιστική, Δίαυλος, 2008
- Γναρδέλλης, Χαράλαμπος, Εφαρμοσμένη στατιστική, Παπαζήσης, 2003

ΔΙΑΔΙΚΤΥΟ

- http://www.aueb.gr
- http://www.uoc.gr
- http://www.uom.gr
- http://www.teipat.gr
- http://www.aegean.gr
- http://www.aua.gr
- http://www.statistics.gr/portal/page/portal/ESYE
- http://en.wikipedia.org/wiki/Spss
- http://www.itia.ntua.gr/en/docinfo/122/