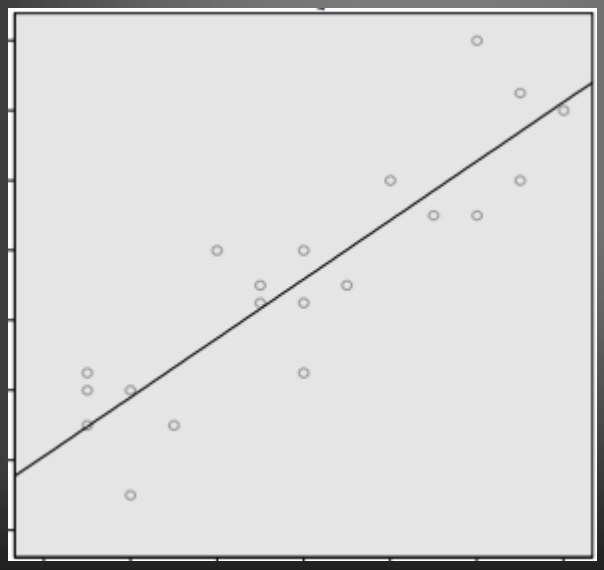
ΤΕΙ ΔΥΤΙΚΗΣ ΕΛΛΑΔΑΣ

Σχολή Διοίκησης και Οικονομίας
Τμήμα:Διοίκησης Επιχειρήσεων

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

Απλή Γραμμική Παλινδρόμηση καί Απλοί Συντελεστές Συσχέτισης



Σπουδάστριες:

Κωστοπούλου Θεοδώρα

Παναγοπούλου Θεοδώρα

Εποπτεύουσα Καθηγήτρια:

Βάσιου Γεωργία

Πάτρα,14/01/1014

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΡΟΛΟΓΟΣ

Στην εργασία αυτή παρουσιάζεται ένα μέρος της στατιστικής μεθοδολογίας που χρησιμοποιείται στο πλαίσιο της ανάλυσης οικονομικών μεγεθών.

Σε αυτό το πλαίσιο της ανάλυσης στατιστικών δεδομένων μας ενδιαφέρει να ορίσουμε τη σχέση μεταξύ δύο παραγόντων (μεταβλητών).Για παράδειγμα μας ενδιαφέρει να διερευνήσουμε και να ορίσουμε την σχέση που έχουν σε κάποια τράπεζα οι επενδύσεις με την αύξηση των επιτοκίων ή σε κάποια επιχείρηση, πως οι πωλήσεις επηρεάζονται απο τις τιμές των προιόντων. Η πιο διαδεδομένη στατιστική μεθοδολογία για την διερεύνηση αυτού του τύπου είναι η Ανάλυση Παλινδρόμησης.Τέτοιες σχέσεις είναι κατα κανόνα περίπλοκες.Τα διάφορα φαινόμενα και διαδικασίες επηρεάζονται συχνά απο μία σειρά παραγόντων.

Στο πλαίσιο της διοίκησης επιχειρήσεων και της ανάλυσης οικονομικών διαδικασιών γίνεται εκτεταμένη χρήση των σχέσεων μεταξύ δυο μεταβλήτων.Αυτές οι σχέσεις μπορεί να εκφραστούν μαθηματικά ως

$$Y = f(x)$$

Όπου Y και X οι δύο υπό μελέτη μεταβλητές.Η συνάρτηση $f(x)$ μπορεί να έχει κάποια γραμμική ή μη γραμμική μορφή πράγμα που θα αναλύσουμε μέσα σε αυτήν την εργασία

Για την περάτωση αυτής της εργασίας θα θέλαμε να ευχαριστήσουμε την εισηγήτρια μας κυριά Βάσιου καθώς και την κυρία Μπουμπουλή με την οποία ξεκινήσαμε αυτήν την εργασία αρχικά.Ενα μεγάλο ευχαριστώ στον κύριο Κακαρελίδη για την βοηθεία του η οποία ήταν καταλυτική σε ότι αφορά την εργασία αυτή.

Τέλος να πούμε ένα μεγάλο ευχαριστώ στις οικογενειές μας για την ηθική, οικονομική συμπαράσταση τους σε όλη την διάρκεια των φοιτητικών μας σπουδών καθώς και στην εκπόνηση της πτυχιακής μας εργασίας.

ΚΕΦΑΛΑΙΟ 1

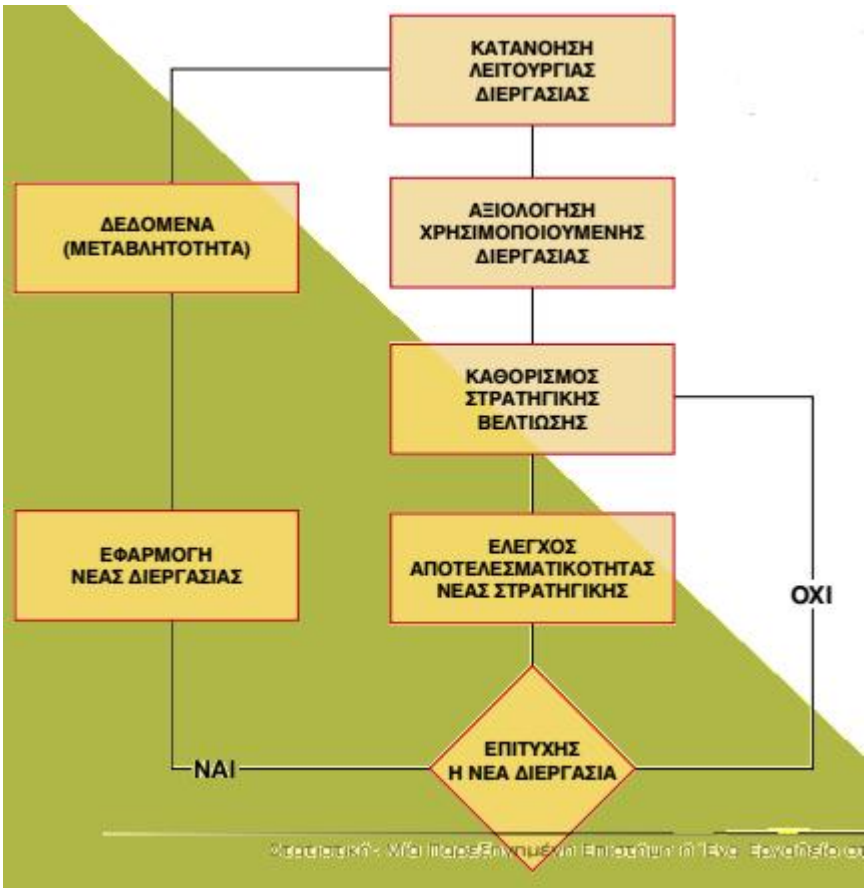
1.1.ΕΙΣΑΓΩΓΗ

Η μέθοδος της Στατιστικής μας βοηθαεί να αντιληφθούμε καλύτερα το κόσμο γύρω μας.Η στατιστική είναι μια μεθοδική μαθηματική παλαιότερα τεχνική και σήμερα επιστήμη που επιχειρεί να εξαγάγει έγκυρη γνώση.Η στατιστική ως έννοια εμφανιζεται απο τους μυθικούς χρόνους απο την αρχή της δημιουργίας οργανομένων κοινωνιών.Μια πρώτη γραφή στατιστικής μορφής με αριθμητικά δεδομένα είναι ο νέων κατάλογος (κατάλογος των πλοίων) των Αχαιών στον Τρωικό πόλεμο από τον Όμηρο.Στοιχειώδεις τέτοιες απογραφές είχαν πραγματοποιήσει και άλλοι αρχαίοι λαοί όπως οι αρχαίοι Αιγύπτιοι , οι Βαβυλώνιοι, οι Πέρσες ,οι αρχαίοι 'Ελληνες και οι Ρωμαίοι με χαρακτηριστικότερη την απογραφή του Οκταβιανου του Αυγούστου.Η συστηματική συλλογή δεδομένων για πληθυσμό και οικονομία άρχισε στη διάρκεια τηα Αναγγένησης ειδικότερα στην Βενετία και την Φλωρεντία όπου γρήγορα επεκτάθηκε σ'όλα τα τότε Βασιλέια της Ευρώπης.Το 1620 ο Άγγλος έμπορος Τζον Κραούντ ξεκίνησε πρώτος τη δειγματοληπτική έρευνα σε οικογένειες του Λονδίνου για τους θανάτους την εποχή της πανώλης.Έτσι πολλοί επιστήμονες θέτουν αφετηρία της στατιστικής το έτος 1663 με την έκδοση του βιβλίου Φυσικές και Πολιτικές Παρατηρήσεις της Θνησιμότητας.Σήμερα η στατιστική έρευνα απο μαθηματική τεχνική έχει αναχθεί σε σπουδαία αυτοτελή επιστήμη ακολουθώντας ιδιαίτερες μεθόδους ανάλυσης.Η λέξη στατιστική προέρχεται απο τη λατινική λέξη ΄΄status΄΄ που σημαίνει κράτος και δήλωνε αρχικά τη συλλογή στοιχείων για κρατικές ανάγκες (έκταση,παραγωγή,πληθυσμό κ.α).

Στην αρχή η στατιστική είχε περιγραφικό χαρακτήρα και ασχολιόταν κυρίως με θέματα Δημογραφίας όπως οι απογραφές οληθυσμού που ξεκίνησαν απο την Κίνα απο τόν αυτοκράτορα Υ-αο το έτος 2238 π.χ ενώ στους Ρωμαίους η πρώτη απογραφή πήθυσμού έγινε επί Ρωμύλλου 1753-717 π.χ) και ουτέ καθ΄΄εξης εως το έτος 1741 απο την Αγγλία εως και την Γερμανία.Το 1853 γράφεται απο το Fr.sansovino το πρώτο βιβλίο στατιστικού περιεχομένου και λίγο αργότερα εισάγεται απο το Korning η στατιστική στην ανώτερη παιδεία.Η στατιστική θα ξεφύγει απο τον περιγραφικό χαρακτήρα με την ανάπτυξη ενός νέου κλάδου ,του Λογισμικού των Πιθανοτήτων ο οποιός προήθλε απο τη μελέτη των τυχερών παιχνιδίων.Θεμελιωτές του ΄΄Λογισμικού των Πιθανοτήτων΄΄ αναφέρουμε τον Bernoulli, ο οποίος στο βιβλίο του ΄΄η τέχνη των προβλέψεων΄΄ διατυπώνει τον περίφημο νόμο των μεγάλων αριθμών και

το Γάλλο μαθηματικό Laplace στον οποίο οφείλεται η εφαρμογή του Λογισμικού των Πιθανοτήτων στη σπονδή των φυσικών φαινομένων με πολυσύνθετες αιτίες.
Η στατιστική όπως την γνωρίζουμε στην σύγχρονη εποχή βασίζεται στην συγκέντρωση ,ανάλυση και παράθεση δεδομένων.Η ανάλυση είναι το πιο βασικό απο τα τρία βήματα για την εφαρμογή της επιστήμης της στατιστικής.Καθοριστικό πόλο στην ανάλυση δεδομένων παίζει η αντίληψη της έννοιας μεταβλητότητας.Η μεταβλητότητα είναι αναπόφευκτη σε όλες τις πλευρές της ανθρώπινης δραστηριότητας.Κατανόηση της μεταβλητότητας και των λόγων που την προκαλούν είναι απαραίτητη για την ερμηνεία των δεδομένων.Θα μπορούσε κανείς να ισχυριστεί ότι η κατανόηση και ερμηνεία της μεταβλητότητας σε ένα σύνολο δεδομένων είναι ακριβώς αυτό με το οποίο ασχολείται η Στατιστική.Η έννοια της μεταβλητότητας είναι ίσως αυτή ακριβώς που οδήγησε στην σημαντική ανάπτυξη και αξιοποίηση των μεθόδων των πιθανοτήτων και της Στατιστικής τα τελευταία χρόνια σε κατεύθυνση διαφορερική από εκείνη των μαθηματικών.Τα Μαθηματικά ασχολούνται με συγκεκριμένες και σαφώς καθορισμένες διαδικασίες όπου ένα σύνολο συγκεκριμένων υποθέσεων μπορεί να οδηγήσει σε ενα μονοσήμαντο αποτέλεσμα.Αντίθετα η στατιστική δημιουργήθηκε απο την ανάγκη μελέτης φαινομένων που υπό συνθήκες είναι αδύνατον να καταλήξουν σε διαφορετικά αποτελέσματα λόγω της ύπαρξης μεταβλητότητας.
Στην καθομιλουμένη ,Στατιστική σημαίνει συστηματική απαρίθμηση και παρουσίαση αριθμητικών δεδομένων ή στοιχείων τα οποία προέρχονται απο πολλές παρατηρήσεις ή μετρήσεις.Στην επιστημονική γλώσσα η λέξη στατιστική έχει ευρήτερη σημασία και κάποιος μπορεί να ισχυρισθεί ότι είναι η διαδικασία συλλογισμών που αναγνωρίζει ότι υπάρχει μεταβλητότητα σε όλα τα φαινόμενα και ότι η μελέτη της μεταβλητότητας οδηγεί σε νεές γνώσεις και καλύτερες αποφάσεις.Μια από τις κυρίαρχες εφαρμογές της Στατιστικής είναι η χρήση μεθόδων για την υποβοήθηση της λήψης αποφάσεων .Στο πλαίσιο αυτό στατιστική σκέψη είναι ο τρόπος σκέψης που μας επιτρέπει να καταλάβουμε και τελικά να βελτιώσουμε κάποιες διεγρασίες μέσω της συνεχόμενης μελέτης της μεταβλητότητας των δεδομένων.

Το παρακάτω σχήμα εξηγεί τη χρησιμοποίηση της Στατιστικής Σκέψης σε συνδυασμό με τη γνώση του συντελεστή και την κρίση αυτού που λαμβάνει τις αποφάσεις ,όπως αυτά συνδέονται με το σχεδιασμό , τη λήψη αποφάσεων και τη βελτίωση συστημάτων



Το σχήμα έχει πολλά κοινά στοιχεία με τον κύκλο αξιολόγησης των πληροφοριών

Μια απλή απαρίθμηση των εφαρμογών της δείχνει ότι η Στατιστική η οποία είναι βασικά εφαρμοσμένη επιστήμη ,χρησιμοποιείται σε όλους σχεδόν τους τομείς της ανθρώπινης δραστηριότητας.Η στατιστική είναι απαραίτητη στη Διοίκηση γενικά όπου η λήψη ορθών αποφάσεων έχει μεγάλη σημασία για την πρόοδο ενός κράτους ,ενός οργανισμού ,μιας βιομηχανίας ή μιας επιχείρησης.Γι'αυτό και δεν υπάρχει σήμερα τομέας σε καμμία σύγχρονη επιχείρηση που να μην χρησιμοποιεί στατιστικές μεθόδους.

Η κακή χρήση της μεθόδου αυτής μπορεί να οδηγήσει σε λανθασμένα συμπεράσματα και συνεπώς σε λάθος επιλογές.Στη παράθεση πληροφορίων που γίνεται παρακάτω θα ασχολήθουμε με συγκεκριμένους τομείς και κλάδους της στατιστικής επιστήμης καθώς και με συγκεκριμένες διαδικασίες.

1.2 ΣΚΟΠΟΣ ΤΗΣ ΕΡΓΑΣΙΑΣ ΚΑΙ ΓΕΝΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ ΣΤΑΤΙΣΤΙΚΗΣ.

Σε αυτήν την εργασία αποζητούμε την ανάλυση και κατανόηση των εννοιών της απλής γραμμικής παλινδρόμησης και των απλών συντελεστών συσχέτισης μέσω της θεωρητικής επεξήγησης καθώς και με την παράθεση διάφορων παραδειγμάτων τα οποία θα παρουσιαστούν παρακάτω στη διάρκεια της εκπόνησης αυτής της εργασίας.

Στις παρακάτω σελιδές θα παρουσιαστούν οι έννοιες της απλής γραμμίκης παλινδρόμησης και απο ποίες μεταβλητές αποτελούνται κάθως και τον ρόλο που απαρτίζουν στην εξίσωση γραμμικής παλινδρόμησης.Σε δεύτερο επίπεδο θα δούμε την σχέση που έχουν οι συντελεστές συσχέτισης και πως βοηθούν στην πραγματοποίηση και ολοκλήρωση της έρευνας που θέλουμε να διεξάγουμε.

Για να γίνουν πιο κατανοητές οι έννοιες που θέλουμε να παρουσιάσουμε παρακάτω θα σας δώσουμε κάποιες γενικές έννοιες για τη στατιστική. Η στατιστική είναι η επιστήμη που επιχειρεί να εξαγάγει γνώση χρησιμοποιώντας εμπειρικά δεδομένα .Βασίζεται στην χρήση της στατιστικής θεωρίας των εφαρμοσμένων μαθηματικών. Στη στατιστική η τυχαία και η απροσδιοριστία ορίζονται στα πλαίσια της θεωρίας πιθανοτήτων. Η πρακτική της στατιστικής περιλαμβάνει τη σχεδίαση , τη συλλογή και ερμηνεία δεδομένων που προκύπτουν από αβέβαιες παρατηρήσεις . Επειδή η στατιστική αποσκοπεί στην εξαγωγή των ΄΄καλύτερων΄΄ πληροφοριών από τα διαθέσιμα δεδομένα κατατάσσεται από μερικούς σαν κλάδος της θεωρίας των αποφάσεων.

Μερικές επιστήμες χρησιμοποιούν την Εφαρμοσμένη στατιστική τόσο εκτεταμένα ώστε έχουν ειδική ορολογία. Τέτοιοι επιστημονικοί τομείς είναι οι εξής: Επιχειρηματική Στατιστική , Οικονομική Στατιστική , Μηχανική Στατιστική, Ψυχολογική Στατιστική , Ανάλυση Διαδικασιών στη Χημειομετρία. Υπάρχουν και οι τομείς της Βιοστατικής ,της Δημοσιογραφίας ,της Επαγωγικής, της Περιγραφικής καθώς και της Οικονομετρίας.

Η στατιστική αποτελείται από διάφορα μέτρα υπολογισμών σε κάθε τομέα στον οποίο εφαρμόζεται η επιστήμη της στατιστικής . Στους τομείς που θα αναλύσουμε παρακάτω χρησιμοποιούνται τα μέτρα τα οποία χαρακτηρίζονται αριθμητικά περιληπτικά μέτρα και διακρίνονται

1.Στα μέτρα θέσης ή κεντρικής τάσης που περιλαμβάνουν:

A. Τη μέση τιμή ή μέσο(mean)

B. Τη διάμεσο (median)

C. Την επικρατούσα τιμή (mode)

2. Στα μέτρα μεταβλητότητας ή διασποράς που περιλαμβάνουν :

A. Το εύρος (range)

B. Το ενδοτεταρτημοριακό εύρος (Interquartile range)

C. Τη διασπορά (variance)

D. Τη τυπική απόκλιση (standard deviation)

E. Το συντελεστή μεταβλητότητας (coefficient of variance)

Ο μέσος χρησιμοποιείται ως περιληπτικό μέτρο για συνεχή ή ασυνεχή δεδομένα, ενώ επηρεάζεται σημαντικά από την ύπαρξη ακραίων τιμών. Η διάμεσος επηρεάζεται σε πολύ μικρότερο βαθμό από τις ακραίες τιμές σε σχέση με το μέσο και χρησιμοποιείται τόσο για συνεχή ή ασυνεχή δεδομένα.

Αφού γίνει η επεξεργασία , η οργάνωση και η ταξινόμηση των στατιστικών στοιχείων ακολουθεί το στάδιο της συνοπτικής παρουσίασης των συγκεντρωθέντων στοιχείων. Η παρουσίαση γίνεται με τρείς τρόπους:

Α. Με τη μορφή πινάκων .Υπάρχουν δυο είδη πινάκων. Οι πίνακες απλής εισόδου και οι πίνακες διπλής εισόδου.

Β. Με τη μορφή γραφικών παραστάσεων. Αποτελεί το καλύτερο μέσο στατιστικής παρουσίασης και δεν περιέχει πολλές λεπτομέρειες.

Γ. Με τη μορφή εκθέσεων ή αναφορών.

Η παρουσίαση των στατιστικών στοιχείων με μορφή πινάκων ή γραφικών παραστάσεων γίνεται με βάση διάφορα κριτήρια που εξαρτώνται από τη φύση του πληθυσμού που ερευνάται.Υπάρχει μια ακόμη μορφή παρουσίασης στατιστικών στοιχείων και είναι τα **Διαγράμματα**. Υπάρχουν πολλά είδη διαγραμμάτων αλλά στην πράξη χρησιμοποιούνται ορισμένα όπως:

Α. Τα ακιδωτά Διαγράμματα

Β. Τα χρονολογικά Διαγράμματα

Γ. Τα Κυκλικά Διαγράμματα

Δ. Τα Σπειροειδή Διαγράμματα

Ε. Τα Ημιλογαριθμικά και Λογαριθμικά Διαγράμματα.

1.3 ΔΙΑΓΡΑΜΜΑ ΔΙΑΣΠΟΡΑΣ

Για να μπορέσουμε να διεξάγουμε μια διαγραμματική απεικόνιση της γραμμικής παλινδρόμησης η οποία θα μας βοηθήσει στην εξήγηση του φαινομένου αυτού θα πρέπει απαραιτήτως να χρησιμοποιήσουμε το διάγραμμα διασποράς. Διάγραμμα διασποράς καλείται το σύστημα των ορθογώνιων αξόνων του επιπέδου στο οποίο σημειώνουμε τα σημεία τα οποία έχουν συντεταγμένες τα ζεύγη που παριστάνουν τις παρατηρήσεις μας με σκοπό το σχηματισμό ενός πλήθους σημείων που ονομάζεται νέφος σημείων ή και διάγραμμα διασποράς. Ο λόγος δημιουργίας του διαγράμματος διασποράς είναι η εύρεση μέτρων τα οποία μπορούν να εκφράσουν και να ποσοτικοποιήσουν τη πιθανή συμμεταβολή και συσχέτιση των χαρακτηριστικών ενός τυχαίου δείγματος από το πλήθος των στοιχειών που έχουμε και θέλουμε να το μελετήσουμε με βάση δύο ή περισσότερα χαρακτηριστικά.

Σε περίπτωση που το δείγμα των μεταβλητών δεν ακολουθεί τη νοητή γραμμή που δημιουργεί το διάγραμμα διασποράς συμπεραίνουμε ότι δεν υπάρχει αλληλεξάρτηση οπότε οι μεταβλητές είναι ανεξάρτητες μεταξύ τους.

Υπάρχουν δυο τρόποι αλληλεξάρτησης ή συμμεταβολής δυο μεταβλητών:

- *Η συναρτησιακή εξάρτηση:* Δυο μεταβλητές Χ και Υ έχουν συναρτησιακή εξάρτηση όταν για κάθε τιμή Xi της μεταβλητής Χ αντιστοιχεί μια μόνο τιμή Yi της μεταβλητής Υ. Ο τύπος Υ = f(X) επαληθεύεται από όλα τα ζεύγη τιμών (Xi ,Yi) που εμφανίζονται. Η σχέση αυτή επιτρέπει τον υπολογισμό των τιμών της Υ από τις αντίστοιχες τιμές της Χ με απόλυτη ακρίβεια. Σε μια τέτοια συναρτησιακή εξάρτηση η Χ χαρακτηρίζεται ***ανεξάρτητη μεταβλητή*** και η Υ ***εξαρτημένη ή ερμηνευτική μεταβλητή***. Η συναρτησιακή εξάρτηση είναι γενικά αντικείμενο των Μαθηματικών και γι' αυτό δεν ασχολείται ιδιαίτερα η στατιστική.
- *Η στοχαστική ή στατιστική εξάρτηση***:** Δυο μεταβλητές Χ και Υ έχουν στοχαστική εξάρτηση όταν κάθε τιμή της μεταβλητής Χ δεν αντιστοιχεί σε μια ορισμένη τιμή της μεταβλητής Υ. Αντιθέτως αντιστοιχεί μια τιμή της μεταβλητής Y η οποία προκύπτει από ένα πλήθος δυνατών τιμών της η οποία δεν μπορεί να προβλεφθεί ακριβώς όπως για παράδειγμα από το εισόδημα μιας οικογένειας δεν μπορούμε να προβλέψουμε με ακρίβεια τις δαπάνες διατροφής.

Για τη μελέτη μιας στοχαστικής εξάρτησης αφού συγκεντρώσουμε όλες τις παρατηρήσεις μας $(X_1,Y_1)(X_2,Y_2),.....,(X_9,Y_9)$ και σχηματίσουμε το νέφος των σημείων τους. Σε γενικές γραμμές εργαζόμαστε ως εξής:

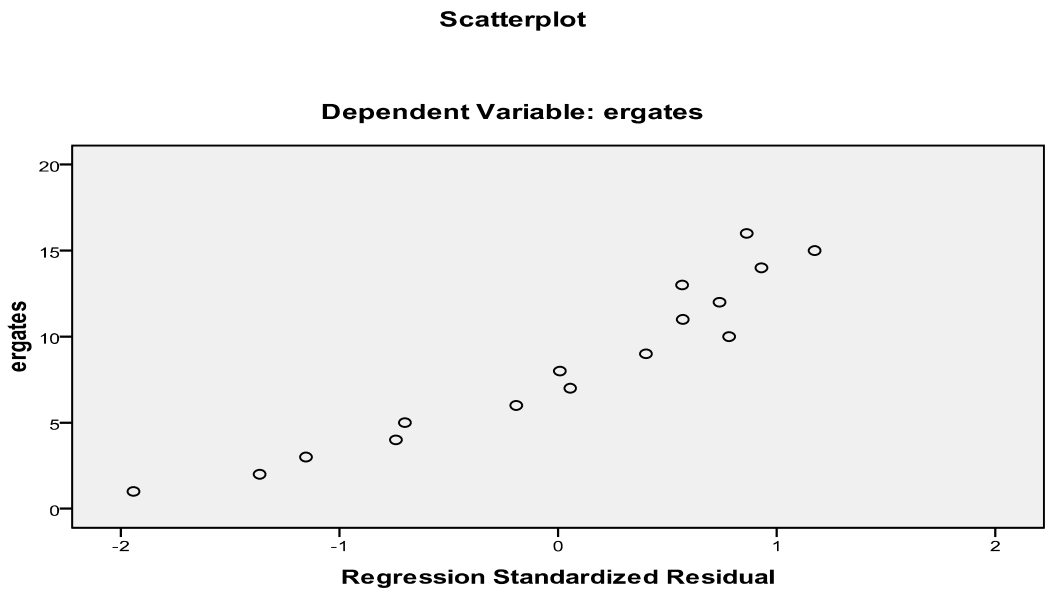
1. Παίρνουμε όλα τα σημεία του νέφους που έχουν την ίδια τετμημένη π.χ. την $x = aj$ και βρίσκουμε τον αριθμητικό μέσο Y_i των τεταγμένων τους. Ο Y_i λέγεται δεσμευμένος μέσος της μεταβλητής Y για $x = aj$.Αν έχουμε δυο διαφορετικές τιμές για το aj θα έχουμε και διαφορετικούς δεσμευμένους μέσους της Y .
2. Σημειώνουμε στο σύστημα αξόνων μας τα σημεία $K_j(aj, yj)$ που έχουν τετμημένες όλες τις διαφορετικές τιμές της X και τεταγμένες τις αντίστοιχες δεσμευτικές μέσες τιμές της Y .
3. Φέρνουμε μια καμπύλη (γ) που να διέρχεται από τα σημεία $K_1, K_2, K_3....K_P$ (ή που να διέρχεται πολύ κοντά σε αυτά) η οποία ονομάζεται *καμπύλη παλινδρόμησης*.

   Για παράδειγμα το διάγραμμα διασποράς μπορεί να μας δείξει πως σχετίζονται και συμμεταβάλλονται το ύψος και το βάρος 16 εργατών . Όταν αυξάνει το ύψος , αυξάνει και το βάρος των εργατών. Μας δείχνει επίσης πόσο ισχυρή είναι η συμμεταβολή του ύψους και του βάρους.

   Έστω ότι έχουμε 16 εργάτες και τα ακόλουθα στοιχεία:

Ύψος	183	162	172	181	180	168	176	180	190
Βάρος	84	63	71	76	77	64	70	76	82

175	178	175	186	172
68	75	73	86	73



Στο παραπάνω διάγραμμα διασποράς βλέπουμε ότι το νέφος των σημείων ακολουθεί μια νοητή γραμμή η οποία μας δείχνει μια ισχυρή σχέση αφού τα σημεία βρίσκονται κοντά το ένα στο άλλο και δεν έχουν μεγάλες αποστάσεις μεταξύ τους και επίσης από τον τρόπο που χαράσσουν τη διαδρομή τους βλέπουμε ότι όταν αυξάνουν οι τιμές της μεταβλητής X γενικά έτσι αυξάνουν και οι τιμές της μεταβλητής Y γενικά.

Το διάγραμμα διασποράς χρησιμοποιείται τόσο στην απλή γραμμική παλινδρόμηση σαν βασικό εργαλείο όσο και στους συντελεστές συσχέτισης.

ΚΕΦΑΛΑΙΟ 2

2.1 ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Σε αυτό το σημείο θα πρέπει να ξεκινήσουμε την ανάλυση της γραμμικής παλινδρόμησης ετσι ώστε να μπορέσουμε να κατανοήσουμε το σκοπό αυτής της εργασίας.

Η απλή γραμμική παλινδρόμηση αναφέρεται στην σχέση μεταξύ δυο μεταβλητών από τις οποίες η μια μεταβλητή είναι εξαρτημένη και η άλλη ανεξάρτητη. Η γραμμική παλινδρόμηση έχει τη μορφή ενός γραμμικού υποδείγματος το οποίο αποτελείται από τιμές και παρατηρήσεις των δύο μεταβλητών που εξετάζονται. Στη γενικότερη περίπτωση θέλοντας να ερευνήσουμε τη σχέση μεταξύ των δύο μεταβλητών (X,Y) παίρνουμε το

απλό γραμμικό μοντέλο και με βάση αυτό το μοντέλο θεωρούμε ότι τα X_i, Y_i συνδέονται με τη σχέση $Y_i = a + b\,x_i + E_i, i = 1, 2, ..., n$ όπου a, b είναι δυο άγνωστες σταθερές (καλούνται και τεταγμένη ή intercept) και κλίση ή slope αντίστοιχα). Οι $E_1, E_2, ...E_n$ είναι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν κανονική κατανομή

$N(0, s^2)$ (s^2 άγνωστο) και συνήθως καλούνται ΄΄σφάλματα΄΄ των μετρήσεων. Μπορεί να θεωρηθεί ότι τα σφάλματα $E_1, E_2, ...E_n$ εμπεριέχουν όλους τους άλλους παράγοντες (εκτός της X) επηρεάζουν την τιμή της μεταβλητής Y . Υπογραμμίζεται και πάλι ότι οι τιμές $X_1, X_2, ..., X_n$ δεν είναι τυχαίες, αντίθετα με τις $Y_1, Y_2, ..., Y_n$ οι οποίες προφανώς είναι τυχαίες και μάλιστα ακολουθούν κανονική κατανομή (αφού είναι γραμμικές συναρτήσεις των κανονικών τιμών E_i) με παραμέτρους: $E(Y_i) = E(a + b\,X_i + E_i) = a + b\,X_i + E(e_i) = a + b\,X_i$

$V(Y_i) = V(a + b\,X_1 + E_i) = V(E_i) = s^2$ για $i = 1, 2, ..., n$ δηλαδή $Y_i \square\ N(a + b\,X_i, s^2)$. Επίσης οι τιμές $\mathrm{Y}_1, \mathrm{Y}_2, ..., \mathrm{Y}_n$ είναι ανεξάρτητες αφού τα σφάλματα $E_1, E_2, ...E_n$ είναι ανεξάρτητα (το ΄΄τυχαίο΄΄ ενός Y_i οφείλεται αποκλειστικά από σφάλμα E_i). Αρχικά θα πρέπει με βάση τα (X_i, Y_i) , $i = 1, 2, ..., n$ να εκτιμήσουμε τις παραμέτρους a, b και s^2 ενώ φυσικά είναι απαραίτητο να διερευνήσουμε πόσο ικανοποιητικά προσαρμόζονται τα δεδομένα μας στο μοντέλο αυτό.

Γενικά στη στατιστική η απλούστερη μορφή παλινδρόμησης είναι η απλή γραμμική παλινδρόμηση (simple linear Regression), κατά την οποία υπάρχει μόνο μια ***ανεξάρτητη μεταβλητή*** *X* (independent or input variable), και η **εξαρτημένη μεταβλητή** *Y* (dependent or response variable), η οποία μπορεί να προσεγγιστεί ικανοποιητικά από μία γραμμική συνάρτηση του Χ. Η περίπτωση αυτή εμφανίζεται τόσο σε πειραματικές όσο και σε μη πειραματικές μελέτες. Στις πειραματικές μελέτες ο ερευνητής καθορίζει, για παράδειγμα, από πριν τις δόσεις ενός φαρμάκου (ανεξάρτητη μεταβλητή) που δίνει στα πειραματόζωα και μετρά τις αντιδράσεις τους (εξαρτημένη μεταβλητή). Με την παλινδρόμηση ενδιαφέρεται να προσδιορίσει μία σχέση δόσης-αντίδρασης για το συγκεκριμένο φάρμακο. Στις μη πειραματικές μελέτες ή δειγματοληψίες, γίνονται μετρήσεις σε δύο χαρακτηριστικά (μεταβλητές) για κάθε άτομο (μονάδα) του δείγματος. Σε ένα δείγμα 10 μαθητών μετράμε, για παράδειγμα, το βάρος και το ύψος τους. Η διάκριση εδώ μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής είναι

δύσκολη. Αν αυτό που μας ενδιαφέρει είναι το "τι συμβαίνει με το βάρος των παιδιών όταν αλλάζει το ύψος τους", τότε θεωρούμε ως ανεξάρτητη μεταβλητή Χ το ύψος και ως εξαρτημένη μεταβλητή Υ το βάρος. Οπότε, ενδιαφερόμαστε για την ***παλινδρόμηση του βάρους (Υ) πάνω στο ύψος (Χ).*** Αντίθετα, αν μας ενδιαφέρει το "τι συμβαίνει με το ύψος των παιδιών όταν αλλάζει το βάρος τους", τότε θεωρούμε ως ανεξάρτητη μεταβλητή Χ το βάρος και ως εξαρτημένη μεταβλητή Υ το ύψος. Τότε έχουμε *παλινδρόμηση του ύψους (Υ) πάνω στο βάρος (Χ).*

Αφού έχουμε ορίσει σε γενικό επίπεδο τον ορισμό της γραμμικής παλινδρόμησης θα πρέπει να την αναλύσουμε και σε κάποιους τομείς της στατιστικής οι οποίοι έχουν ασχοληθεί με το φαινόμενο αυτό. Θα εστιάσουμε στους τομείς της Εφαρμοσμένης Στατιστικής και της Επαγωγικής Στατιστικής και Περιγραφικής Στατιστικής. Αρχικά όμως θα πρέπει να αναφερθούμε στους συντελεστές συσχέτισης καθώς και να αναλύσουμε τον ορισμό της ίδιας της συσχέτισης.

2.2 ΓΕΝΙΚΕΣ ΠΛΗΡΟΦΟΡΙΕΣ ΓΙΑ ΤΗ ΣΥΣΧΕΤΙΣΗ-CORRELATION

Η συσχέτιση μετρά το βαθμό συνάφειας- αλληλεπίδρασης ανάμεσα σε δύο ή περισσότερες μεταβλητές. 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 σχέσης αυτών.
Το θέμα αυτό αναλύεται στη ΔΙΑΔΙΚΑΣΙΑ Regression.

Οι συντελεστές συσχέτισης που θα δούμε χωρίζονται σε δύο κατηγορίες. Η πρώτη αφορά το συντελεστή γραμμικής συσχέτισης του Pearson και αναφέρεται σε ποσοτικές μεταβλητές και η δεύτερη κατηγορία αφορά τους συντελεστές Spearman και Kendall, οι οποίοι χρησιμοποιούνται σε ποιοτικές μεταβλητές και κατηγορικές μεταβλητές (δηλαδή μεταβλητές των οποίων οι τιμές δεν επιδέχονται ιεράρχηση)

Η ποσοτική μέτρηση της έντασης (γραμμικής) σχέσης μεταξύ δύο μεταβλητών ονομάζεται συντελεστής συσχέτισης (r) (correlation coefficient)

-Το εύρος τιμών του συντελεστή συσχέτισης είναι από -1,00 εώς +1,00.

-Τιμές κοντά στο -1,00 και 1,00 υποδεικνύουν τέλεια (ισχυρή) συσχέτιση.

-Τιμές του δείκτη κοντά στο 0 υποδηλώνουν ότι οι δύο μεταβλητές δεν σχετίζονται γραμμικά.

-Αρνητικές τιμές υποδεικνύουν αρνητική συσχέτιση, ενώ θετικές τιμές υποδεικνύουν θετική συσχέτιση.

-Η συσχέτιση μεταξύ δυο μεταβλητών μπορεί να είναι: Τέλεια θετική(αρνητική), έντονη θετική (αρνητική), ασθενής θετική (αρνητική)

Συσχέτιση δε σημαίνει αιτιότητα

Όταν σε μια μη πειραματική έρευνα (δειγματοληψία) δύο μεταβλητές Χ και Υ βρίσκονται συσχετισμένες αυτό σημαίνει μόνο ότι οι μεταβλητές αυτές συνδέονται με κάποια σχέση. Δε συνεπάγεται, κατ' ανάγκη, αιτιότητα. Οι δύο μεταβλητές μπορεί βεβαία να συνδέονται με σχέση αιτιότητας, μπορεί όμως, 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 Sidney είχαν, για μια μακρά περίοδο, ισχυρή θετική συσχέτιση με τις πωλήσεις έγχρωμων τηλεοράσεων στη Melbourne. Προφανώς, τόσο οι πωλήσεις ταχύπλοων όσο και οι πωλήσεις έγχρωμων τηλεοράσεων ήταν συνάρτηση γενικότερων ευνοϊκών οικονομικών παραγόντων.

Η σχέσητου συντελεστή συσχέτισης και παλινδρόμησης είναι η εξής.

Η παλινδρόμηση ορίζεται θεωρώντας την ανέξαρτητη μεταβλητή X καθορισμένη και την εξαρτημένη μεταβλητή Y τυχαία, ενώ για τη συσχέτιση θεωρούμε και τις δύο μεταβλητές X και Y τυχαίες. Για τις μεταβλητές X και Y της παλινδρόμησης, μπορούμε να αγνοήσουμε ότι η X δεν είναι τ.μ. και να ορίσουμε το συντελεστή συσχέτισης ρ όπως και πριν. Η σχέση μεταξύ του r (της εκτιμήτριας του ρ από το δείγμα) και του συντελεστή της παλινδρόμησης b (της εκτιμήτριας του β από το δείγμα) δίνεται ως εξής (συνδιάζοντας τις σχέσεις $r = \frac{s_{XY}}{s_X s_Y}$ Και $b = \frac{s_{XY}}{s_X^2}$)

$$r = b\frac{s_X}{s_Y} \quad \text{ή} \quad b = r\frac{s_Y}{s_X}.$$

Και τα δύο μεγέθη, r και b, εκφράζουν ποιοτικά τη γραμμική συσχέτιση των μεταβλητών X και Y , αλλά το b εξαρτάται από τη μονάδα μέτρησης των X και Y ενώ το r παίρνει τιμές στο διάστημα [−1, 1]. Έτσι αν η συσχέτιση είναι θετική (r > 0) τότε η κλίση της ευθείας παλινδρόμησης b είναι επίσης θετική, αν η συσχέτιση είναι αρνητική (r < 0) τότε είναι b < 0 και αν οι μεταβλητές X και Y δε συσχετίζονται (r = 0) τότε η ευθεία παλινδρόμησης είναι οριζόντια (b = 0).

Επίσης μπορούμε να εκφράσουμε το συντελεστή προσδιορισμού r^2 ως προς τη δειγματική

διασπορά του σφάλματος s^2 και αντίστροφα

$$s^2 = \frac{n-1}{n-2}s_Y^2(1-r^2) \quad \text{ή} \quad r^2 = 1 - \frac{n-2}{n-1}\frac{s^2}{s_Y^2}.$$

Η παραπάνω σχέση δηλώνει πως όσο μεγαλύτερο είναι το r^2 (ή το |r|) τόσο μικρότερη είναι η διασπορά του σφάλματος της παλινδρόμησης, δηλαδή τόσο καλύτερη είναι η πρόβλεψη που βασίζεται στην ευθεία παλινδρόμησης.

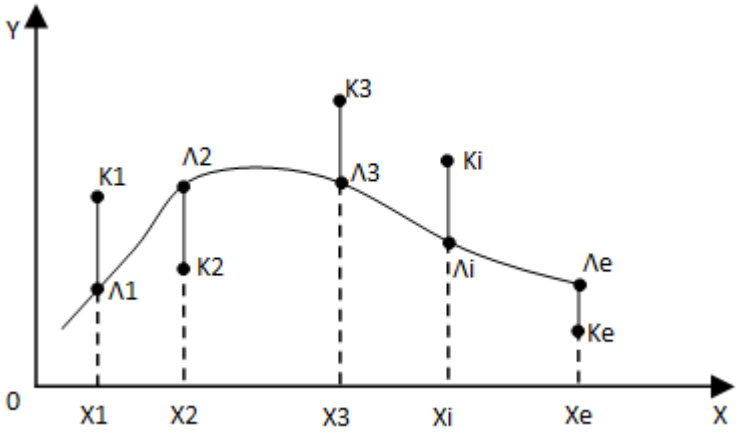
ΚΕΦΑΛΑΙΟ 3

Το θέμα της γραμμικής παλινδρόμησης αντιμετωπίζεται από πολλές πλευρές της στατιστικής . Για να ξεκινήσει η ανάλυση της γραμμικής παλινδρόμησης θα πρέπει να αναλυθεί η μέθοδος των ελάχιστων τετραγώνων η οποία είναι το πρώτο βήμα για να κατανοήσουμε την απλή γραμμική παλινδρόμηση. Στο παρακάτω κείμενο που ακολουθεί αναλύεται η έννοια της μεθόδου ελάχιστων τετραγώνων καθώς και κάποιες παράμετροι που την αφορούν.

3.1 ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Όπως αναφέρεται στη βιβλιογραφία το πρόβλημα της στοχαστικής εξάρτησης εντοπίζεται στην εύρεση μιας καμπύλης η οποία να διέρχεται πολύ κοντά από ορισμένα σημεία .Να βρεθεί η εξίσωση $y = j\,(x)$ μιας καμπύλης η οποία να διέρχεται πολύ κοντά από p ορισμένα σημεία $K_1(x_1, y_1), K_2(x_2, y_2), ..., K_p(x_p, y_p)$. Για να έχει νόημα αυτή η διατύπωση και κυρίως η έκφραση ΄΄ πολύ κοντά ΄΄ θα πρέπει να βρούμε κάποιο μέτρο το οποίο εκφράζει την απόσταση των p σημείων από οποιοδήποτε καμπύλη του επιπέδου. Αν λοιπόν ονομάσουμε $\Lambda_1, \Lambda_2, ..., \Lambda_p$ τα σημεία μιας οποιασδήποτε καμπύλης (g) που έχουν τις ίδιες τετμημένες $c_1, c_2, ..., c_p$ ως τέτοιο μέτρο παίρνουμε το άθροισμα $A = (K_1\Lambda_1)^2 + (K_2\Lambda_2)^2 + ... + (K_P\Lambda_P)^2$. Έτσι μια καμπύλη θα θεωρείται τόσο πιο κοντά στα σημεία $K_1, K_2, ..., K_P$ όσο πιο μικρό είναι το άθροισμα A δηλαδή $(K_1\Lambda_1)^2 + (K_2\Lambda_2)^2 + ... + (K_P\Lambda_P)^2 =$ ελάχιστο. Επειδή υπάρχουν πολλές καμπύλες με διαφορετικά σχήματα προσδιορίζουμε από την αρχή ανάλογα με τη θέση που έχουν τα p σημεία , το είδος της καμπύλης που θα τοποθετήσουμε ανάμεσα τους. Αυτό σημαίνει ότι παίρνουμε

αυθαίρετα την μορφή της εξίσωσης $y = j\,(x)$ και μετά προσδιορίζουμε τα διάφορα σημεία της με τη βοήθεια συντεταγμένων των δεδομένων σημείων. Με τις παραπάνω προϋποθέσεις το παραπάνω γενικό πρόβλημα λύνεται με μια μέθοδο που ονομάζεται *Μέθοδος Ελάχιστων Τετραγώνων* και η καμπύλη που βρίσκουμε *Καμπύλη Ελάχιστων Τετραγώνων*.



Με τη μέθοδο των ελάχιστων τετραγώνων θα προσδιορίσουμε στη συνέχεια μια εκτίμηση $\hat{Y} = \hat{a} + \hat{b}\,x$ της πληθυσμιακής ευθείας παλινδρόμησης $E(Y / X) = a + b\,x$ ονομάζεται ευθεία ελάχιστων τετραγώνων από τον τρόπο υπολογισμού των παραμέτρων της. Θεωρούμε v ζεύγη παρατηρήσεων $(x, y) = (1, 2, 3...., v)$ προσεγγίζουμε ως προς τη μορφή $Y_i = a + b\,x_i + E_i$ όπου E παριστάνουν τις αποκλίσεις της πραγματικής τιμής Y_i από την θεωρητική $a + b\,x$.Δηλώνει $E = Y - a + b\,x$.Η εκτίμηση των a και b θα πρέπει να γίνει έτσι ώστε να ελαχιστοποιηθούν οι ποσότητες E .Οπότε θα αναζητήσουμε τις τιμές των a και b για τις οποίες ελαχιστοποιείται το άθροισμα των τετραγώνων των E .Δηλαδή η ποσότητα $\sum_{i=1}^{v} E_i^{\,2} = \sum_{i=1}^{v} (y_i - a - b\,x_i)^2$.Παραγωγίζοντας την παραπάνω εξίσωση ως προς a και b και εξισώνονται με μηδέν έχουμε τις ακόλουθες εξισώσεις οι οποίες ονομάζονται κανονικές εξισώσεις. $\sum_{i=1}^{v} y_i = va + b\sum_{i=1}^{v} x_i$, $\sum_{i=1}^{v} x_i y_i = a\sum_{i=1}^{v} x_i + b\sum_{i=1}^{v} x_i^{\,2}$.Λύνοντας το

σύστημα παίρνουμε

$$\hat{b} = \frac{n\sum_{i=1}^{n} x_i y_i - [\sum_{i=1}^{v} x_i][\sum_{i=1}^{v} y_i]}{v\sum_{i=1}^{v} x_i^{\,2} - [\sum_{i=1}^{v} x_i]^2} = \frac{\sum_{i=1}^{v} x_i y_i - v\overline{x}\overline{y}}{\sum_{i=1}^{v} x_i^{\,2} - v - \overline{x}^2}, \hat{a} = \overline{y} - \hat{b}\,\overline{x} \text{ ή } \hat{b} = \frac{s_{xy}}{s_x^2} \text{ και}$$

$\hat{a} = \overline{y} - \hat{b}\,\overline{x}$. Η εκτίμηση ελάχιστων τετραγώνων της ευθείας παλινδρόμησης από το δείγμα των v ζευγών παρατηρήσεων είναι

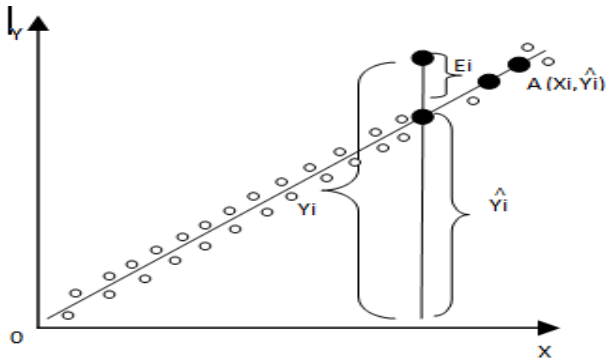
$$\hat{Y} = \hat{a} + \hat{b}\,x = \overline{y} - b\,\overline{x} + \hat{b}\,X = \overline{y} + \hat{b}\,(X - \overline{x}) \text{ ή } \hat{Y} = \overline{y} + \frac{s_{xy}}{s^2 x}(X - \overline{x}).$$

Αφού έχουμε αναλύσει τη μέθοδο προχωρούμε στο αποτέλεσμα που δίνει δηλαδή την ευθεία παλινδρόμησης καθώς και τις παραμέτρους που αφορούν την ευθεία αυτή καθώς και δυο παραδείγματα για την πρακτική παρουσίαση της.

3.2 ΕΥΘΕΙΑ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ (ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ)

Ας υποθέσουμε τώρα ότι η θέση που έχουν τα p σημεία $K_1(x_1, y_1), K_2(x_2, y_2), ..., K_p(x_p, y_p)$ μας επιτρέπει να ζητήσουμε μια ευθεία η οποία να περνάει πολύ κοντά από τα σημεία αυτά. Στην περίπτωση αυτή θεωρούμε μια εξίσωση πρώτου βαθμού ως προς X και Y: $y_i = a + \hat{b}\,x_i$ η οποία αντιπροσωπεύει όλες τις ευθείες του επιπέδου. Από όλες τις ευθείες της μορφής $y_i = a + \hat{b}\,x_i$ εμείς θα διαλέξουμε εκείνη την ευθεία που θα δώσει τις μικρότερες διαφορές (αποκλίσεις) μεταξύ εμπειρικών (y_i) και θεωρητικών $(\hat{y}_i)$ τιμών δηλαδή των τιμών:

$\hat{e} : y_i - \hat{y}_i = y_i - (\hat{a} + \hat{b}\,x_i)$. Η διαφορά $\hat{e} = y_i - \hat{y}_i$ ονομάζεται σφάλμα ή απόκλιση της παρατήρησης y_i από τη θεωρητική τιμή $\hat{y}_i$ του σημείου $A(x_i, y_i)$ της ευθείας $\hat{y}_i = \hat{a} + \hat{b}\,x_i$.



Από όλες τις ευθείες η ευθεία με την ιδιότητα

$\sum_{i=1}^{N} \hat{e}_i^{\,2} = \hat{e}_1^{\,2} + ... + \hat{e}_N^{\,2} = \sum [y_i - (\hat{a} + \hat{b}\, x_i)]^2$ =ελάχιστο είναι η ευθεία με την καλύτερη προσαρμογή. Οι ελάχιστες αποκλίσεις $\sum [y_i - \hat{a} - \hat{b}\, x_i)^2]$ μεταξύ εμπειρικών και θεωρητικών τιμών υπάρχουν όταν οι μερικές παραγωγοί ως προς $\hat{a}$ και $\hat{b}$ εξισωθούν με μηδέν δηλαδή όταν: $\frac{\partial}{\partial a}(\sum \hat{e}_i^{\,2}) = 0, \frac{\partial}{\partial \hat{b}}(\sum \hat{e}_i^{\,2}) = 0$.

Η διαδικασία η οποία ακολουθούμε για τον υπολογισμό των παραμέτρων $\hat{a}$ και $\hat{b}$ είναι γνώστη από το διαφορικό λογισμό ως ***Μέθοδος των Ελαχίστων Τετραγώνων.***

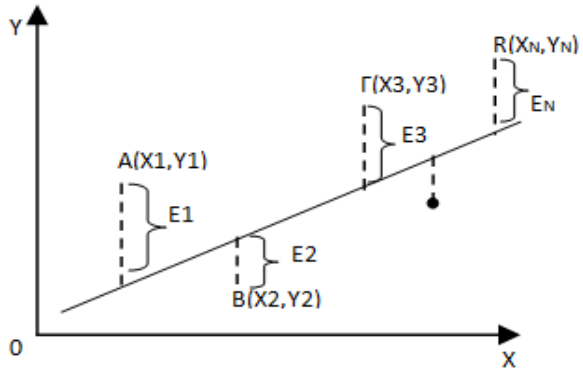
Η μέθοδος αυτή συνιστάται στον προσδιορισμό των τιμών $\hat{a}, \hat{b}$ των παραμέτρων a, b που ελαχιστοποιούν το άθροισμα των τετραγώνων όλων των αποκλίσεων $\hat{e} : y_i - \hat{y}_i = y_i - (\hat{a} + \hat{b}\, x_i)$. Την ευθεία ελάχιστων τετραγώνων θα μελετήσουμε :

Ø *Οταν τα δεδομένα της παρατήρησης είναι απλά*

Στα απλά δεδομένα τα ζεύγη των παρατηρήσεων μας (x_i, y_i) εμφανίζονται χωρίς συχνότητες όπως δείχνει ο πίνακας.

Η γραφική τους παράσταση εμφανίζεται όπως δείχνει

x_i	$x_1 \quad x_2 \quad x_3$ □□□□□ x_i □□□□□ x_N	$\sum x_i$
y_i	$y_1 \quad y_2 \qquad\qquad y_3$	$\sum y_i$



Στην περίπτωση των απλών δεδομένων το σύστημα των κανονικών εξισώσεων προκύπτει με τη μερική παραγώγιση ως προς $\hat{a}$ και $\hat{b}$ ως εξής:

$$\frac{\partial}{\partial \hat{a}} \sum \hat{e}_i^{\,2} = 0 \Rightarrow 2\sum (y_i - \hat{a} - \hat{b}\, x_i) = 0 \Rightarrow$$
$$\sum y_i = N\hat{a} + \hat{b} \sum x_i$$

Επίσης: $$\frac{\partial}{\partial \hat{b}} \sum \hat{e}_i^{\,2} = 0 \Rightarrow 2\sum (y_i - \hat{a} - \hat{b}\, x_i)(-x_i) = 0 \Rightarrow$$
$$\sum x_i y_i = \hat{a} \sum x_i + \hat{b} \sum x_i^{\,2}$$

Παρακάτω δίνεται παράδειγμα για την εφαρμογή των τύπων.

ΠΑΡΑΔΕΙΓΜΑ 1

Δίνεται η βαθμολόγια πέντε σπουδαστών μιας ανώτατης σχολής στα μαθήματα ΄΄ Φυσικής ΄΄ (x) και ΄΄Μαθηματικά΄΄ (y)

	Βαθμολογία *Στη Φυσική* x_i	*Βαθμολογία* *Στα Μαθηματικά* y_i
	2 3 5 6 8	1 2 5 6 7
Σύνολο	24	21

Να βρεθεί η ευθεία ελάχιστων τετραγώνων των πέντε αυτών ζευγών.

Με τη βοήθεια της μεθόδου των ελάχιστων τετραγώνων υπολογίζουμε:

$$\sum y_i = N\hat{a} + \hat{b}\sum x_i$$
$$\sum x_i y_i = \hat{a}\sum x_i + \hat{b}\sum x_i^2$$

Αν λύσουμε το σύστημα ως προς $\hat{a}$ και $\hat{b}$ έχουμε:

$$\hat{b} = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2} \quad \text{όπου } my = \frac{\Sigma y_i}{N}, mx = \frac{\sum x_i}{N}$$
$$\hat{a} = my - \hat{b}mx$$

	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
	2	1	2	4	1
	3	2	6	9	4
	5	5	25	25	25
	6	6	36	36	36
	8	7	56	64	49
ΣΥΝΟΛΟ	24	21	125	138	115

Με βάση τα στοιχεία αυτού του πίνακα έχουμε:

$$\hat{b} = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{2*125 - 24*21}{5*138 - 24^2} = 1,06$$

$$mx = \frac{24}{5} = 4,8, my = \frac{21}{5} = 4,2$$

$$\hat{a} = my - \hat{b}mx = 4,2 - 1,06*4,8 = -0,888$$

Άρα η ζητούμενη ευθεία παλινδρόμησης θα είναι: $y_i = 0,888 + 1,06x_i$.

Αν τώρα υποθέσουμε ότι η μεταβλητή X είναι εξαρτημένη και η Y ανεξάρτητη μεταβλητή, η εξίσωση παλινδρόμησης θα είναι $x_i = a' + b' y_i$.Με τη μέθοδο ελάχιστων τετραγώνων σχηματίζουμε το παρακάτω σύστημα κάνοντας εξίσωση που θα μας επιτρέψει τον υπολογισμό των παραμέτρων a' και b'.Λύνουμε ως προς a' και b'.

$$\sum x_i = Na' + \hat{b}' \sum y_i$$
$$\sum x_i y_i = \hat{a}' \sum y_i + \hat{b}' \sum y_i^2$$
$$\hat{b}' = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{n\sum y_i^2 - (\sum y_i)^2} (\hat{a}' - mx - \hat{b}' - my)$$

- *Όταν έχουμε ταξινομημένα δεδομένα.*

Όταν τα δεδομένα είναι ταξινομημένα σε ένα πίνακα διπλής εισόδου , που περιέχει για παράδειγμα αριθμητικά ζεύγη (x_i, y_i) κάθε ένα από τα οποία επαναλαμβάνεται με συχνότητα fij . Στην περίπτωση των ταξινομημένων δεδομένων το σύστημα των κανονικών εξισώσεων της ευθείας $\hat{y}_i = \hat{a} + \hat{b} x_i$ που προέκυψε από τη μέθοδο των ελάχιστων τετραγώνων .

Οι εξισώσεις είναι: $\sum f_{ij} y_i = N\hat{a} + \hat{b} \sum f_i x_i$
$\sum\sum f_{ij} x_i y_j = \hat{a}_i \sum f_i x_i + \hat{b} \sum f_i x_i^2$

x_i	y_j	f_i
	y_1 y_2 y_j y_l	
x_1	f_{11} f_{12} f_{1j} f_{1l}	f_1
x_2	f_{21} f_{22} f_{2j} f_{2l}	f_2
x_i	f_{i1} f_{i2} f_{ij} f_{il}	f_i
x_k	f_{k1} f_{k2} f_{kj} f_{kl}	f_k
$f \bullet j$	$f \bullet 1$ $f \bullet 2$ $f \bullet j$ $f \bullet l$	N

Αν λύσουμε ως προς $\hat{a}$ και $\hat{b}$ θα έχουμε:

$$\hat{b} = \frac{N\sum\sum f_{ij}x_i y_j - \sum f_i x_i \sum f_j y_j}{N\sum f_i x_i^2 - (\sum f_i x_i)^2}$$

$\hat{a} = m y - \hat{b} m x$.Μέσο Τετραγωνικό Σφάλμα σε πίνακα διπλής εισόδου:

$s^2 = \frac{\sum f_j y_j^2 - \hat{a}\sum f_j y_j - \hat{b}\sum\sum f_{ij}x_i y_j}{N}$.Ενώ ο δείκτης προσδιορισμού :

$p^2 = 1 - \frac{s^2}{s_y^2}, s^2 = \frac{\sum f_j y_j^2}{\sum f_i} - (\frac{\sum f_i y_j}{\sum f_i})^2$.

ΠΑΡΑΔΕΙΓΜΑ 2: Κατά τη διάρκεια ενός δεκαπενθήμερου η θερμοκρασία σε δυο σταθμούς παρατήρησης:

x_i	y_i				
2	2 3 4 5	f_i	$f_i x_i$	$f_i x_i^2$	$\sum f_{ij}x_i y_j$
4 5 6 7	1 1 – – 1 2 1 – – 1 3 2 – – 2 1	2 4 6 3	8 20 36 21	32 100 216 174	20 60 150 91
f_j	2 4 6 3	15	85	495	321
$f_j y_j$	4 12 24 15	55			
$f_j y_j^2$	8 36 96 75	215			
$\sum f_{ij} y_j x_i$	18 60 148 95	321			

Ευθεία παλινδρόμησης με τη μέθοδο των ελάχιστων τετραγώνων:

$$\sum f_j y_j = \hat{a}N + \hat{b}\sum f_i x_i$$

$$\sum\sum f_{ij} x_i y_j = \hat{a}\sum f_i x_i + \hat{b}\sum f_i x_i^2$$

$$35 = 15\hat{a} + 85\hat{b}$$

$$321 = 85\hat{a} + 495\hat{b}$$

$$\hat{b} = \frac{N\sum\sum f_{ij} x_i y_j - \sum f_i y_i}{N\sum f_i x_i^{\,2} - (\sum f_i x_i)^2}$$

$$\Rightarrow \hat{b} = \frac{15*321 - 85*55}{15*495 - 85^2} = 0,7$$

$$\hat{a} = mx - \hat{b}mx = \frac{55}{15} - 0,7\frac{85}{15} = -0,32$$

Άρα η ευθεία παλινδρόμησης: $\hat{y}_i = -0,32 + 0,7x_i$

Μέσο Τετραγωνικό Σφάλμα:

$$(215 + 0,3s^{\,2} = \frac{1}{\mathrm{N}}(\sum f_j y_j^{\,2} - \hat{a}\sum f_i y_j - \hat{b}\sum\sum(f_{ij} x_i y_j))$$

$$\Rightarrow s^{\,2} = \frac{1}{15}2*55 - 0,7*321) = \frac{1}{15}(215 + 17,60 - 224,7)$$

$$= 0,50$$

Δείκτης προσδιορισμού : $p^2 = 1 - \frac{s^{\,2}}{s\,y^2} = 1 - \frac{0,50}{1,4} = 1 - 0,36 = 0,64$.

Αντίστοιχα παρουσιάζεται και η συσχέτιση η οποία παρουσιάζεται στη βιβλιογραφία ως ποσοτικό μέγεθος και αντικατοπτρίζει το πόσο έντονα μια αλλαγή στο ένα μέγεθος επηρεάζει το άλλο μέγεθος. Παρακάτω παρουσιάζεται ο τύπος υπολογισμού του συντελεστή συσχέτισης καθώς και ένα παράδειγμα για την πρακτική του παρουσίαση.

-Υπολογίζουμε τον συντελεστή συσχέτισης με τον παρακάτω τύπο

$$r = \frac{\sum(x - \overline{x})(y - \overline{y})}{(n-1)S_x S_Y}$$

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

ΣΥΝΤΕΛΕΣΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ (r^2): Ο συντελεστής προσδιορισμού παριστάνει το ποσοστό της συνολικής μεταβλητότητας της εξαρτημένης μεταβλητής το οποίο εξηγείται από την ανεξάρτητη μεταβλητή.
Ο συντελεστής προσδιορισμού υπολογίζεται υψώνοντας στο τετράγωνο τον συντελεστή συσχέτισης.
- Το εύρος του συντελεστή προσδιορισμού είναι από 0 έως 1.

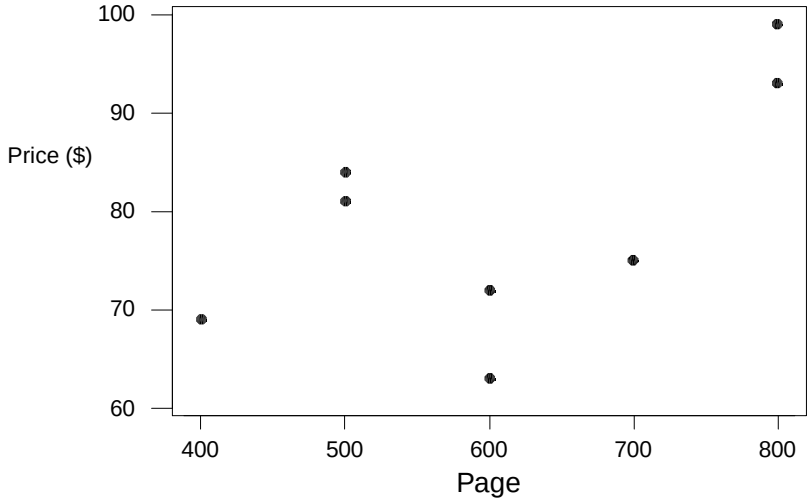
- Ο συντελεστής προσδιορισμού δεν μας παρέχει καμία πληροφορία για την κατεύθυνση ή την σχέση μεταξύ των μεταβλητών.

ΠΑΡΑΔΕΙΓΜΑ 1

Ο Πρόεδρος του σωματίου φοιτητών στο κρατικό πανεπιστήμιο του Τολέδο, ανησυχεί για το κόστος των πανεπιστημιακών εγχειριδίων. Θεωρεί ότι υπάρχει σχέση μεταξύ του αριθμού των σελίδων και της τιμής πώλησης του βιβλίου. Για να αποδείξει την προαναφερθείσα σχέση, ο πρόεδρος επιλέγει ένα δείγμα οκτώ εγχειριδίων που πωλούνται στο βιβλιοπωλείο του πανεπιστημίου. Α) Κατασκευάστε το διάγραμμα διασποράς. Β) Υπολογίστε το συντελεστή συσχέτισης.

ΒΙΒΛΙΟ	ΣΕΛΙΔΕΣ	ΤΙΜΗ
ΙΣΤΟΡΙΑ	500	84
ΑΛΓΕΒΡΑ	700	75
ΨΥΧΟΛΟΓΙΑ	800	99
ΚΟΙΝΩΝΙΟΛΟΓΙΑ	600	72
ΜΙΚΡΟΟΙΚΟΝΟΜΙΚΗ	400	69
ΜΑΚΡΟΟΙΚΟΝΟΜΙΚΗ	500	81
ΟΙΚΟΝΟΜΕΤΡΙΑ	600	63
ΣΤΑΤΙΣΤΙΚΗ	800	93

Scatter Diagram of Number of Pages and Selling Price of Text



ΒΙΒΛΙΟ	ΣΕΛΙΔΕΣ X	ΤΙΜΗ Y	XY	X^2	Y^2
ΙΣΤΟΡΙΑ	500	84	42,000	250,000	7,056
ΑΛΓΕΒΡΑ	700	75	52,500	490,000	5,625
ΨΥΧΟΛΟΓΙΑ	800	99	79,200	640,000	9,801
ΚΟΙΝΩΝΙΟΛΟΓΙΑ	600	72	43,200	360,000	5,184
ΜΙΚΡΟΟΙΚΟΝΟΜΙΚΗ	400	69	27,600	160,000	4,761
ΜΑΚΡΟΟΙΚΟΝΟΜΙΚΗ	500	81	40,500	250,000	6,561
ΟΙΚΟΝΟΜΕΤΡΙΑ	600	63	37,800	360,000	3,969
ΣΤΑΤΙΣΤΙΚΗ	800	93	74,400	640,000	8,649
ΣΥΝΟΛΟ	4900	636	397,200	3,150,000	51,606

Ο τύπος είναι:

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

$$= \frac{8(397,200) - (4,900)(636)}{\sqrt{[8(3,150,000 - (4,900)^2][8(51,606) - (636)^2]}}$$

$$= 0,614$$

Η συσχέτιση μεταξύ του αριθμού σελίδων και της τιμής πώλησης του βιβλίου είναι 0,614. Αυτό δείχνει μια ασθενή θετική συσχέτιση μεταξύ των δυο μεταβλητών.

Η απλή γραμμική παλινδρόμηση μπορεί να αναλυθεί από αρκετές πλευρές της στατιστικής και μας δίνει αρκετές πληροφορίες. Όπως αναφέραμε παραπάνω οι πλευρές που θα αναλύσουμε την γραμμική παλινδρόμηση είναι η επαγωγική, η εφαρμοσμένη και η περιγραφική που αναφέραμε παραπάνω.

ΚΕΦΑΛΑΙΟ 4

4.1 ΤΟΜΕΑΣ ΕΦΑΡΜΟΣΜΕΝΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΠΑΝΩ ΣΤΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ.

Στα πλαίσια της εφαρμοσμένης στατιστικής η απλή γραμμική παλινδρόμηση ορίζεται ως η ποσοτικοποίηση της σχέσης δυο συνεχών τυχαίων μεταβλητών Χ και Υ υπό τη μορφή ενός γραμμικού υποδείγματος στο οποίο οι τιμές της μιας μεταβλητής *Y* εκτιμώνται από τις τιμές της μεταβλητής Χ. Όποτε η μεταβλητή Υ ονομάζεται εξαρτημένη μεταβλητή και η μεταβλητή Χ ονομάζεται ανεξάρτητη.

Για να γίνει η εκτίμηση των τιμών της μεταβλητής Υ από τις τιμές της μεταβλητής Χ με βάση το υπόδειγμα της απλής γραμμικής παλινδρόμησης θα πρέπει αρχικά να διασφαλιστούν κάποιες προϋποθέσεις οι οποίες είναι οι εξής:

- Ο προσδιορισμός των τιμών της μεταβλητής Χ γίνεται χωρίς σφάλμα. Επειδή στην πραγματικότητα καμία συνεχής μέτρηση δεν είναι απαλλαγμένη από σφάλματα το μέγεθος του σφάλματος κατά τη μέτρηση της Χ είναι αμελητέο.
- Σε κάθε τιμή της μεταβλητής Χ αντιστοιχεί ένας υπό-πληθυσμός τιμών της μεταβλητής Υ ο οποίος ακολουθεί την κανονική κατανομή.

- Οι διακυμάνσεις των υπό-πληθυσμών της μεταβλητής Y που ορίζονται για τις διάφορες τιμές της μεταβλητής X , είναι ίσες. Η κοινή διακύμανση των υπό-πληθυσμών της μεταβλητής Y συμβολίζεται με $s^2_{Y/X}$. Η αποδοχή της ισότητας των διακυμάνσεων των τιμών της μεταβλητής Y ονομάζεται ομοσκεδαστικότητα (homoscedasticity) και είναι ανάλογη με την αποδοχή της ισότητας των διακυμάνσεων που απαιτείται σε ένα t-test για ανεξάρτητα δείγματα ή στην ανάλυση διακύμανσης με έναν παράγοντα.
- Οι μέσες τιμές των υπό-πληθύσμων της μεταβλητής Y συνδέονται με τις αντίστοιχες τιμές της μεταβλητής X δια μέσου μιας γραμμικής σχέσης της μορφής $m_{Y/X} = a + b_c$. $m_{Y/X}$ (μέση τιμή του υπό-πληθυσμού της μεταβλητής Y που αντιστοιχεί σε μια συγκεκριμένη τιμή χ της μεταβλητής X.)

  α ,β = πληθυσμιακοί συντελεστές της παλινδρόμησης .

  Το παραπάνω υπόδειγμα ορίζει μια ευθεία γραμμή επί της οποίας είναι τοποθετημένες οι μέσες τιμές $m_{Y/X}$ των διάφορων υπό-πληθυσμών της Y. Η ευθεία αυτή γραμμή ονομάζεται πληθυσμιακή ευθεία της παλινδρόμησης . Γεωμετρικά οι συντελεστές α και β αναπαριστούν αντίστοιχα την τεταγμένη στο σημείο 0 και την κλίση της ευθείας της παλινδρόμησης.

- Οι τιμές της μεταβλητής Y είναι ανεξάρτητες η μια της άλλης. Όλες οι προηγούμενες προϋποθέσεις συνοψίζονται στην παρακάτω εξίσωση η οποία ονομάζεται ***υπόδειγμα απλής γραμμικής παλινδρόμησης***.

  $$E = y - (a + b_c) = y - m_{Y/c}$$

Ø Η ποσότητα E ονομάζεται σφάλμα και εκφράζει τη διαφοροποίηση της μεταβλητής Y από τη μέση τιμή του υπό-πληθυσμού της μεταβλητής Y ο οποίος εκφράζει την απόκλιση της μεταβλητής Y από την ευθεία της παλινδρόμησης.

Το αποτέλεσμα της αποδοχής ότι οι διάφοροι υπό-πληθυσμοί της μεταβλητής Y ακολουθούν κανονική κατανομή με κοινή διακύμανση ίση με την κοινή διακύμανση τότε οι ποσότητες E για κάθε τιμή της μεταβλητής X ακολουθούν επίσης κανονική κατανομή με διακύμανση ίση με τη κοινή διακύμανση των $s^2 m_{Y/X}$ αντίστοιχων υπό-πληθυσμών της μεταβλητής Y. Επιπλέον από τον ορισμό των σφαλμάτων προκύπτει ότι η μέση τιμή τους είναι ίση με 0.Αυτά όσον αφόρα τη διακύμανση στο γραμμικό πρόβλημα της παλινδρόμησης.

Στη δειγματική εξίσωση πάνω σε ένα τυπικό πρόβλημα γραμμικής παλινδρόμησης το ενδιαφέρον εστιάζεται στον προσδιορισμό της *πληθυσμιακής ευθείας της παλινδρόμησης* δηλαδή της ευθείας που περιγράφει την πραγματική σχέση που υπάρχει μεταξύ των μεταβλητών X και Y.

Ο προσδιορισμός αυτής της ευθείας ισοδυναμεί με την εκτίμηση των συντελεστών της παλινδρόμησης a και b .

- Οι πληθυσμιακοί συντελεστές μπορεί να εκτιμηθούν με τη βοήθεια ενός τυχαίου δείγματος το οποίο λαμβάνεται από τον πληθυσμό και για το οποίο υπολογίζεται η αντίστοιχη δειγματική ευθεία της παλινδρόμησης. Ο προσδιορισμός των συντελεστών του δειγματικού υποδείγματος της παλινδρόμησης αποτελεί τη βάση για την εκτίμηση των αντίστοιχων πληθυσμιακών συντελεστών. Πριν όμως προσδιοριστεί η δειγματική ευθεία της παλινδρόμησης είναι απαραίτητο να επιβεβαιωθεί η γραμμική σχέση που υπάρχει μεταξύ των δύο μεταβλητών στα δειγματικά δεδομένα . Η διαδικασία αυτή μπορεί να γίνει με τη βοήθεια ενός διαγράμματος διασποράς.

  Ένα απλό παράδειγμα είναι το παρακάτω:

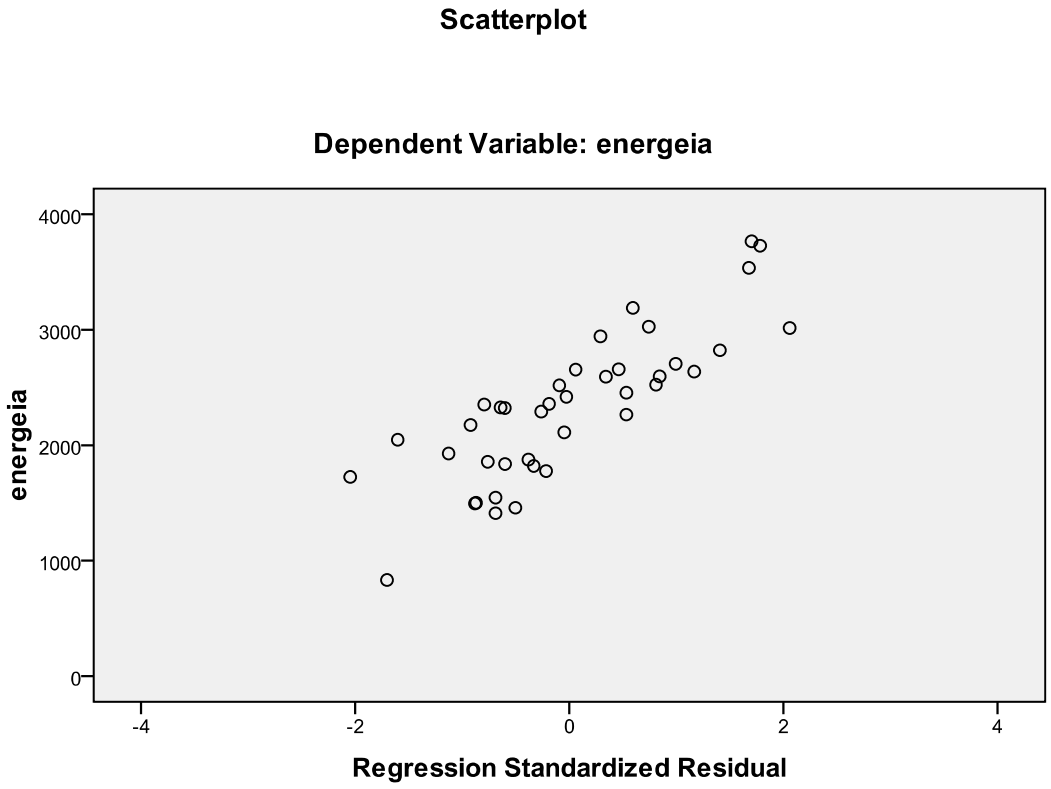
  Έστω οι τιμές της ημερήσιας ενεργειακής πρόσληψης 40 ενήλικων ατόμων μαζί με την ηλικία τους. Ανάμεσα στην ημερήσια ενεργειακή πρόσληψη και της ηλικίας υπάρχει γραμμική σχέση σύμφωνα με την οποία οποιαδήποτε αύξηση της ηλικίας σημαίνει ταυτόχρονη και αντίστοιχη μείωση της ενεργειακής πρόληψης. Η γραμμική σχέση επιβεβαιώνεται από τη μορφή του διαγράμματος διασποράς.

  Στο συγκεκριμένο διάγραμμα η ηλικία θεωρούμενη ως ανεξάρτητη μεταβλητή κατά τον προσδιορισμό της ενεργειακής πρόσληψης τοποθετείται στον οριζόντιο άξονα ενώ στο κατακόρυφο άξονα τοποθετείται η ενεργειακή πρόσληψη .

Πίνακας παραδείγματος

ΑΤΟΜΑ	1	2	3	4	5	6	7	8	9	10
ΗΛΙΚΙΑ	63	49	44	69	55	62	41	42	36	50
ΕΝΕΡΓΕΙΑ	2822	2419	2518	3015	1857	1857	2322	3536	2443	2558

ΑΤΟΜΑ	11	12	13	14	15	16	17	18	19	20
ΗΛΙΚΙΑ	60	50	42	37	34	40	65	62	45	42
ΕΝΕΡΓΕΙΑ	2596	2543	2655	2728	2766	2327	2637	2524	1908	3027



Βλέπουμε στο παραπάνω Διάγραμμα Διασποράς ότι το νέφος των σημείων ακολουθεί μια νοητή γραμμή του επιπέδου οπότε συμπεραίνουμε ότι υπάρχει αλληλεξάρτηση. Ο προσδιορισμός της ευθείας είναι απαραίτητο να γίνει με τρόπο αντικειμενικό ώστε να διασφαλίζεται η βέλτιστη προσέγγιση των σημείων από αυτήν. Η μέθοδος η οποία συνήθως χρησιμοποιείται για τον σκοπό αυτόν είναι γνωστή ως μέθοδο των ελάχιστων τετραγώνων ,ενώ η ευθεία που ορίζεται ονομάζεται ευθεία των ελάχιστων τετραγώνων .

Ο λόγος για τον οποίο χρησιμοποιείται η συγκεκριμένη ονομασία για τη μέθοδο αυτή προκύπτει από την γεωμετρική διαδικασία προσδιορισμού της ευθείας.

Έστω ένα οποιοδήποτε σημείο του διαγράμματος διασποράς με συντεταγμένες (Xi, Yi) και έστω Ei η κατακόρυφη απόσταση του σημείου από μια οποιαδήποτε ευθεία που προσεγγίζει τα σημεία του διαγράμματος. Η απόσταση Ei ονομάζεται υπόλοιπο (Residual) ή σφάλμα (Error). Αν όλα τα υπόλοιπα είναι ίσα με 0 θα έχουμε πλήρη προσαρμογή της ευθείας επί των σημείων του διαγράμματος. Η πλήρης προσαρμογή της ευθείας είναι απίθανο να προκύψει (έκτος αν οι δυο μεταβλητές είναι απολύτως γραμμικά εξαρτημένες η μια από την άλλη) μπορούν όμως να ελαχιστοποιηθούν οι κατακόρυφες αποστάσεις (τα υπόλοιπα δηλαδή) των σημείων από την ευθεία.

Η ελαχιστοποίηση αυτή ισοδυναμεί με την ελαχιστοποίηση της ποσότητας $\sum_{i=1}^{n} ei^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ η οποία ονομάζεται άθροισμα τετραγώνων των υπολοίπων ή άθροισμα τετραγώνων των σφαλμάτων. Η ευθεία δηλαδή των ελάχιστων τετραγώνων κατασκευάζεται με την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων. Η διαδικασία προσδιορισμού της ευθείας των ελάχιστων τετραγώνων η οποία συμβολικά ορίζεται από την εξίσωση $\hat{y} = \hat{a} + \hat{b}c$ απαιτεί τον προσδιορισμό των ποσοτήτων $\hat{a}$ και $\hat{b}$ οι οποίες είναι εκτιμήσεις των πληθυσμιακών συντελεστών της παλινδρόμησης a και b . Από την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων

$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_{i-} \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{a} - \hat{b}c_i)^2$ από αυτήν την εξίσωση και με τη βοήθεια του διαφορικού λογισμού προκύπτει ότι

$$\hat{b} = \frac{\sum_{i=1}^{n} (x_{i-} \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_{i-} \bar{x})^2} \quad \text{και} \quad \hat{a} = \bar{y} - \hat{b}\bar{c} \quad (\text{διότι ισχύει } \bar{y} = \hat{a} + \hat{b}\bar{c})$$

Οι παραπάνω εξισώσεις δίνουν την κλίση και την τεταγμένη στο σημείο 0 της ευθείας των ελάχιστων τετραγώνων.

4.2 ΤΟΜΕΑΣ ΕΠΑΓΩΓΙΚΗΣ ΣΤΑΤΙΣΤΙΚΗΣ ΠΑΝΩ ΣΤΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ.

Στο τομέα της Επαγωγικής Στατιστικής θα αναλύσουμε την γραμμική παλινδρόμηση από την πλευρά της εξίσωσης της γραμμικής παλινδρόμησης, από την πλευρά των υποθέσεων που μπορούν να γίνουν και να εφαρμοστούν στην γραμμική παλινδρόμηση. Επίσης την εκτίμηση των συντελεστών α και β του υποδείγματος και την μέθοδο ελάχιστων τετραγώνων. Η ανάλυση μας θα ξεκινήσει με το *Γενικό Γραμμικό Υπόδειγμα* . Πρόκειται να μελετήσουμε τη σχέση που υπάρχει μεταξύ μιας ποσοτικής μεταβλητής Y και ενός συνόλου άλλων μεταβλητών $x_1, x_2 ..... x_n$. Αυτή η σχέση μεταξύ των $n+1$ μεταβλητών ερμηνεύει την επίδραση που έχουν αυτές σε ένα πραγματικό φαινόμενο είτε αυτό είναι οικονομικό είτε κοινωνικό είτε δημογραφικό κλπ. και συγχρόνως δημιουργεί ένα υπόδειγμα του πραγματικού φαινομένου που μελετούμε.

Η βασική επιδίωξη μας σε αυτό τον τομέα της στατιστικής είναι να παρουσιάσουμε με όσο το δυνατόν μεγαλύτερη ακρίβεια:

- Μία μέθοδο που επιτρέπει την εκτίμηση των παραμέτρων του υποδείγματος.
- Τους ελέγχους πάνω στη δομή του υποδείγματος που επιτρέπουν να εκτιμήσουμε την ποιότητα του
- Θα περιοριστούμε στις περιπτώσεις εκείνες όπου η σχέση που συνδέει την μεταβλητή Y με τις $x_1, x_2 ..... x_n$ είναι γραμμική . Μ αυτήν την βασική προϋπόθεση το μαθηματικό υπόδειγμα που θα προσδιορίσουμε δεν θα ανταποκρίνεται πάντοτε στην πραγματικότητα . Θα υπάρχει διαφορά μεταξύ της τιμής της Y που παρατηρήσαμε και αυτής που παρέχει το γραμμικό υπόδειγμα. Αν υποθέσουμε ότι αυτή η διαφορά που στην πραγματικότητα πρόκειται για μια άλλη μεταβλητή E που ονομάζουμε σφάλμα ή υπόλοιπο ή κατάλοιπο οφείλεται σε μια σειρά από παραμέτρους και μεταβλητές που δεν λαμβάνουμε υπόψιν το υπόδειγμα μπορεί να πάρει τη μορφή $y_t = f(x_{t1}, x_{t2}, ..., x_{t4}) + E_t$ για κάθε παρατήρηση t.

Παρατηρούμε λοιπόν ότι σ' αυτήν την μορφή εμφανίζεται και ο όρος E_t που συχνά ονομάζεται και τυχαίος παράγοντας ή σφάλμα. Με αυτόν τον τρόπο μεταφερόμαστε από το ορισμένο υπόδειγμα σ 'ένα άλλο τυχαίο. Αυτή η μεταφορά είναι ουσιώδης γιατί μας επιτρέπει ξεκινώντας από μία υπόθεση για την κατανομή πιθανότητας του τυχαίου παράγοντα να αντιστοιχίσουμε διαστήματα εμπιστοσύνης στις άγνωστες παραμέτρους του υποδείγματος να προβούμε σε έλεγχους της συμβολής της κάθε παραμέτρου καθώς και να ελέγξουμε αν υπάρχει πραγματικά

αυτό το υπόδειγμα . Επίσης ο τυχαίος παράγοντας μας επιτρέπει να προβούμε σε προβλέψεις. Όταν οι μεταβλητές είναι οποιασδήποτε μορφής τότε το υπόδειγμα ονομάζεται *γενικό γραμμικό υπόδειγμα.*

Σε αυτό τον τομέα της στατιστικής υποθέτουμε ότι σε κάθε στοιχείο i αντιστοιχεί και μια τυχαία μεταβλητή E_i ,καθώς και ότι διαφορετικές τυχαίες μεταβλητές E_i αντιστοιχούν σε διαφορετικά στοιχεία παρόλα αυτά οι τυχαίες μεταβλητές E_i ακολουθούν την ίδια κατανομή. Παρατηρούμε λοιπόν ότι για κάθε στοιχείο i για το οποίο η ανεξάρτητη μεταβλητή παίρνει την τιμή X_i, η τιμή της εξαρτημένης μεταβλητής X_i είναι άθροισμα δυο όρων οι οποίοι είναι ο $(a + b\,X_i)$ και ο άλλος είναι ο τυχαίος παράγοντας E_i. Ο πρώτος παράγοντας $(a + b\,X_i)$ οφείλεται στην μεταβλητή X και ο δεύτερος (E_i) στους διάφορους άλλους παράγοντες που επιδρούν στην μεταβλητή Y και τους οποίους δεν μετράμε. Η μαθηματική έκφραση αυτής της γραμμικής σχέσης είναι : $Y = a + b\,X + E$.

Για να εκτιμήσουμε τους συντελεστές a και b καθώς και τα χαρακτηριστικά του τυχαίου παράγοντα E πρέπει να έχουμε στη διάθεσή μας έναν αριθμό T παρατηρήσεων των τιμών των X και Y. Τα δεδομένα μας σε αυτή τη περίπτωση παρουσιάζονται σαν διατεταγμένα ζεύγη της μορφής $(X_i, Y_i) i = 1, 2, ..., T$.

Θα υποθέσουμε στη συνέχεια ότι για ένα δείγμα μεγέθους T έχουμε τις παρακάτω T εξισώσεις :

$$y_1 = a + b\,x_1 + e_1$$

$$y_2 = a + b\,x_2 + e_2$$

$$y_t = a + b\,x_t + e_t \quad \longrightarrow t = 1, 2, ...T$$

Σημαίνει ότι στην τιμή x_t της ανεξάρτητης μεταβλητής X αντιστοιχεί όχι μόνο στη τιμή της εξαρτημένης Y αλλά ένα σύνολο τιμών της με μια ορισμένη πιθανότητα πραγματοποίησης διότι ο όρος e_t είναι όπως αναφέραμε μια τυχαία παρέκκλιση.

$$y_T = a + b\,x_T + e_t$$

Μπορούμε ακόμη να πούμε ότι η σχέση : $y_t = a + b\,x_t + e_t$ σημαίνει ότι στην t-τάξης παρατήρηση αντιστοιχεί η τυχαία μεταβλητή $y_t = a + b\,x_t + e_t$ της οποίας μετρούμε την πραγματική τιμή y_t. Στην πράξη οι συντελεστές a και b είναι άγνωστοι και γι 'αυτό το λόγο δεν είμαστε σε θέση να ξεχωρίσουμε ή να διακρίνουμε την επίδραση επί της y_t της $(a + b\,x)$ από αυτών της e_t.

Ας θεωρήσουμε σαν παράδειγμα ένα ταχυδρομικό υποκατάστημα που δέχεται 25 μέρες τον μήνα επιστολές. Το σύνολο των επιστολών χαρακτηρίζεται από τον αριθμό τους (πλήθος) και από το βάρος τους. Αν θελήσουμε να προσδιορίσουμε τη σχέση μεταξύ αυτών των δυο μεταβλητών θα χρησιμοποιήσουμε το υπόδειγμα της μορφής $y = a + b\,x + e$.Αριθμός επιστολών $y = a + b\,(x \text{ επιστολές}) + e$.Η μελέτη μιας τέτοιας σχέσης έχει αρχικό σκοπό την εκτίμηση των συντελεστών a και b που θα προκύψουν από τις διαθέσιμες πληροφορίες (c_t, y_t) των παρατηρήσεων για την ακρίβεια που θέλουμε γι' αυτές τις εκτιμήσεις. Στην συνέχεια με τη βοήθεια στατιστικών ελέγχων επαληθεύουμε το υπόδειγμα πράγμα που μας επιτρέπει να μελετήσουμε την επίδραση της X και Y.

Για να πραγματοποιηθούν αυτοί οι δυο βασικοί στόχοι θα πρέπει να κάνουμε υποθέσεις για τη δομή της κατανομής πιθανότητας της κάθε τυχαίας απόκλισης οι οποίες να δικαιολογούνται από τις γνώσεις που έχουμε για το φαινόμενο που μελετούμε.

Όταν εφαρμόζουμε τη μέθοδο της παλινδρόμησης ακολουθούμε δύο τελείως διαφορετικές μεταξύ τους διαδικασίες: την ανάλυση των σχέσεων σύνδεσης και την πρόβλεψη. Με την ανάλυση σχέσεων σύνδεσης προσπαθούμε να αντιληφθούμε τον μηχανισμό της αιτίας που επιτρέπει στις μεταβλητές x_i να επιδρούν στην Y. Με τη δεύτερη διαδικασία (της πρόβλεψης) προσπαθούμε να προβλέψουμε την τιμή της εξαρτημένης μεταβλητής Y ενός στοιχείου από την τιμή που έχουν οι ανεξάρτητες μεταβλητές x_i γι' αυτό.

Αποδεχόμαστε ότι οι τυχαίες αποκλίσεις ακολουθούν την ίδια κατανομή για την οποία υπάρχει μια ορισμένη μέση τιμή και μια ορισμένη διακύμανση. Επίσης αποδεχόμαστε ότι είναι ασυσχέτιστες μεταξύ τους και συσχετισμένες με την ανεξάρτητη μεταβλητή X .Οι υποθέσεις για τη κατανομή των τυχαίων ανωμαλιών πρέπει να προσδιορίζονται με

ακρίβεια για το υπόδειγμα της απλής γραμμικής παλινδρόμησης και είναι οι εξής :

- H_1 : Υποθέτουμε ότι οι παρεκκλίσεις δεν εκφράζονται με σφάλματα μετρήσεων υποθέτουμε δηλαδή ότι οι μετρήσεις των τιμών των μεταβλητών X και Y έγιναν χωρίς λάθη.
- H_2 : Κάθε παρέκκλιση E_t έχει μέση τιμή μηδέν $E(e_t) = 0 \ \forall_t$
- H_3 : Η διακύμανση της E_t δεν εξαρτάται από την παρατήρηση t.
  $Var(e_t) = E(e^2_t) = s^2 \ \forall_t$.
- H_4 : Οι αποκλίσεις e_t δεν είναι αυτοσυσχετισμένες

  $Cov(e_i, e_j) = 0 \ \forall_{i,j}$ με $i \neq j$

- H_5 : Η κατανομή πιθανότητας των αποκλίσεων e_t δεν εξαρτάται από την ανεξάρτητη μεταβλητή X .

  $E(e_t / x_1, x_2, ..., x_t) = 0 \ \forall_t$

  $E(e^2_t / x_1, x_2, ..., x_t) = 0 \ \forall_t$

  $Cov((e_i, e_j) / x_1, x_2, ..., x_t) = 0 \ \forall_t, j$ με $i \neq j$

- H_6 : Κάθε απόκλιση e_t ακολουθεί μια κανονική κατανομή $N(0, s^2)$.

Η υπόθεση H_3 λέγεται υπόθεση ομοσκεδαστικότητας .

Η υπόθεση H_4 είναι πολύ ενδιαφέρουσα στις διαχρονικές παρατηρήσεις και φανερώνει ότι η απόκλιση μιας χρονικής στιγμής t δεν εξαρτάται από τις αποκλίσεις προηγούμενων χρονικών στιγμών.

Επίσης η υπόθεση H_5 είναι πολύ χρήσιμη γιατί μας επιτρέπει να μεταφερόμαστε από την κατανομή πιθανότητας των αποκλίσεων στην κατανομή Y .

4.3 ΔΕΙΚΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ (ΜΕΣΟ ΤΕΤΡΑΓΩΝΙΚΟ ΣΦΑΛΜΑ)

Μετά τον προσδιορισμό της ευθείας των ελάχιστων τετραγώνων $\hat{y}_i = \hat{a} + \hat{b}\, x_i$ τίθεται το ερώτημα: πόσο καλά η ευθεία αυτή περιγραφεί το βαθμό εξάρτησης ανάμεσα στις μεταβλητές X και Y ?

Ένα μέτρο αξιολόγησης της καλής προσαρμογής της εξίσωσης $\hat{y}_i = \hat{a} + \hat{b}\, x_i$ στο διάγραμμα διασποράς είναι το μέσο τετραγωνικό σφάλμα το οποίο παριστάνουμε με s^2 και δίνεται από τον τύπο : $s^2 = \dfrac{\sum (y_i - \hat{y}_i)^2}{N}$ ή $s^2 = \dfrac{\sum y_i^2 - \hat{a}\sum y_i - \hat{b}\sum x_i y_i}{N}$.

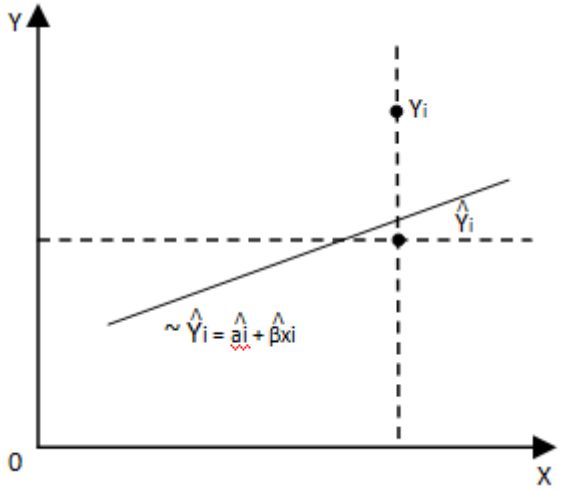
Το μέσο τετραγωνικό σφάλμα (s^2) είναι τόσο μεγαλύτερο όσο περισσότερο διεσπαρμένα είναι τα N σημεία γύρω από την ευθεία $\hat{y}_i = \hat{a} + \hat{b}\, x_i$ ενώ η τιμή s^2 είναι μικρή αν η ευθεία περνάει κοντά από το νέφος των σημείων. Ο δείκτης αυτός παίρνει τιμές: $0 \le s^2 \le \infty$ και επομένως ο χαρακτηρισμός μιας τιμής του s^2 ως μεγάλης ή μικρής είναι πολλές φορές υποκειμενικός και δεν προσφέρεται για συγκρίσεις και εκφράζεται σε τετραγωνικές μονάδες μέτρησης της y_i.Εκείνος ο δείκτης χρησιμοποιείται για τον έλεγχο της καλής προσαρμογής της ευθείας $\hat{y}_i = \hat{a} + \hat{b}\, x_i$ στα ζεύγη των δεδομένων μας είναι αυτό που ονομάζεται δείκτης προσδιορισμού (ή προσαρμογής). Ο δείκτης αυτός συμβολίζεται με p^2: $p^2 - 1 - \dfrac{s^2}{s^2 y}, s^2 y = \dfrac{\sum y_i^2}{N} - m^2 y$.

Ο δείκτης p^2 είναι καθαρός αριθμός (χωρίς μονάδες μέτρησης) και επομένως πάντοτε συγκρίσιμος. Παίρνει τιμές στο κλειστό διάστημα $[0,1]$ δηλαδή $0 \le p^2 \le 1$. Όσο η τιμή p^2 τείνει προς τη μονάδα τόσο τέλεια είναι η προσαρμογή της ευθείας , δηλαδή $\hat{y}_i = \hat{a} + b\, x_i$ περιγράφει πολύ καλά τα δεδομένα μας. Ειδικότερα αν $p^2 = 1$ η ευθεία περνάει από όλα τα σημεία (x_i, y_i) του διαγράμματος διασποράς.

Ο δείκτης προσαρμογής ή προσδιορισμού (p^2) δείχνει το ποσοστό της εξαρτημένης μεταβλητής που ερμηνεύεται από τις μεταβολές της ανεξάρτητης μεταβλητής όπως για παράδειγμα : αν $p^2 = 0,95$ σημαίνει ότι το 95% της συνολικής μεταβλητής της εξαρτημένης μεταβλητής Y οφείλεται στη σχέση που υπάρχει ανάμεσα στις μεταβλητές X και Y και μόνο το υπόλοιπο 5% της διακύμανσης της μεταβλητής οφείλεται

σε άλλες άγνωστες αιτίες δηλαδή με ελεγχόμενους παράγοντες. Ο δείκτης προσδιορισμού έχει τη μορφή: $p^2 = \frac{\sum (\hat{y}_i - my)^2}{\sum (y_i - my)^2}$.

Ο αριθμητής μας δίνει το άθροισμα των αποκλίσεων εξ ΄ αιτίας της παλινδρόμησης ενώ ο παρονομαστής μας δίνει το άθροισμα των τετραγώνων των αποκλίσεων από τη γραμμή παλινδρόμησης .



Οι αποκλίσεις $(y_i - \hat{y}_i)$ οφείλονται στην επίδραση ανερμήνευτων παραγόντων. Οι αποκλίσεις $(y_i - my)$ οφείλονται κατά ένα μέρος στην επίδραση της μεταβλητής X επί της μεταβλητής Y (δηλαδή στην παλινδρόμηση) και κατά ένα άλλο μέρος σε άλλους ανερμήνευτους από την παλινδρόμηση δηλαδή τυχαίους παράγοντες. Με βάση το προηγούμενο παράδειγμα: Οι παράμετροι $\hat{a}$ και $\hat{b}$ δίνονται από τις εξής σχέσεις:

$$\hat{b} = \frac{N\sum x_i y_i - \sum x_i \sum y_i}{N\sum x_i^2 - (\sum x_i)^2} = \frac{5*121 - 10*40}{5*30 - 10^2} = 4,1$$

$$\hat{a} = my - \hat{b}mx = 8*4,12 = 0,2$$

Άρα $\hat{y}_i = -0,2 + 4,1x_i$.

Το μέσο τετραγωνικό σφάλμα δίνεται από τη σχέση:

$$s^2 = \frac{[y_i^2 - \hat{a}\sum y_i - \hat{b}\sum x_i y_i]}{N} = \frac{492 - (-0,2)*40 - 4,1*121}{5} = 0,78$$

Ο δείκτης προσδιορισμού θα είναι:

$$p^2 = 1 - \frac{s^2}{s^2 y} = 1 - \frac{0,78}{34,4} = 1 - 0,02 = 0,98.$$

4.4 Ο ΣΥΝΤΕΛΕΣΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ.

Μετά τον προσδιορισμό της ευθείας ελάχιστων τετραγώνων δια μέσου της εξίσωσης $\hat{y} = \hat{a} + \hat{b}c$ εναπομένει η αξιολόγηση της προσαρμογής της ευθείας αυτής επί των δειγματικών τιμών. Ένας τρόπος για να αξιολογήσουμε την προσαρμογή της ευθείας των ελάχιστων τετραγώνων είναι να υπολογίσουμε το συντελεστή προσδιορισμού (***coefficient of determination).*** Ο συντελεστής προσδιορισμού της δειγματικής ευθείας της παλινδρόμησης συμβολισμένος με R^2 ,ορίζεται ως το τετράγωνο του δειγματικού συντελεστή συσχέτισης , δηλαδή $R^2 = r^2$.

Επειδή ο δειγματικός συντελεστής συσχέτισης παίρνει τιμές στο διάστημα $[-1,1]$ ο συντελεστής προσδιορισμού παίρνει τις τιμές στο διάστημα $[0,1]$. Όταν $R^2 = 1$,όλα τα σημεία που αναπαριστούν τις δειγματικές τιμές των X και Y βρίσκονται τοποθετημένα επί της ευθείας των ελάχιστων τετραγώνων . Όταν $R^2 = 0$ δεν υπάρχει γραμμική σχέση μεταξύ των δειγματικών τιμών X και Y .

Ο συντελεστής προσδιορισμού , ως μέτρο της προσαρμογής της ευθείας των ελάχιστων τετραγώνων επί των δειγματικών τιμών, ορίζεται πρωτογενώς από την ανάλυση της συνολικής διασποράς της εξαρτημένης μεταβλητής Y σε επιμέρους συνιστώσες. Χρησιμοποιώντας την ταυτότητα $(y_i - \hat{y}_i) = (y_i - \overline{y}) - (\hat{y}_i - \overline{y})$, $i = 1, 2, ...n$, η οποία ισχύει για τις δειγματικές τιμές της μεταβλητής Y , υψώνοντας και τα δύο μέλη της στο τετράγωνο και αθροίζοντας για $i = 1, 2, ..., n$ παίρνουμε

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left[(y_i - \overline{y}) - (\hat{y}_i - \overline{y}) \right]^2 =$$

$$\sum_{i=1}^{n} \left[(y_i - \overline{y})^2 + (\hat{y}_i - \overline{y})^2 - 2(y_i - \overline{y})(\hat{y}_i - \overline{y}) \right] =$$

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

Επειδή

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

$$-2\sum_{i=1}^{n}(y_i-\bar{y})(\hat{a}+\hat{b}c_i-\hat{a}-\hat{b}\bar{c})=$$

$$-2\hat{b}c\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$$

Θέτοντας όπου

$\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})=\hat{b}\sum_{i=1}^{n}(x_i-\bar{x})^2$, (από τον υπολογισμό $\hat{b}$)

Και

$x_i-\bar{x}=\frac{\hat{y}_i-\bar{y}}{\hat{b}}$ (από την αντικατάσταση του $\hat{a}=\bar{y}-\hat{b}c$ στην εξίσωση της παλινδρόμησης), προκύπτει

$$-2\hat{b}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})=$$

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

Και τελικά

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

Στη παραπάνω εξίσωση η ποσότητα $\sum_{i=1}^{n}(y_i-\bar{y})^2$ ονομάζεται *συνολικό άθροισμα τετραγώνων* ***(total sum of squares)*** και αποτελεί μέτρο της διασποράς των δειγματικών τιμών της Y γύρω από τη μέση τιμή τους $\bar{y}$.

Η ποσότητα $\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$ ονομάζεται *άθροισμα τετραγώνων επεξηγούμενου από τη γραμμική παλινδρόμηση****(regression sum of squares)*** και εκφράζει τη διασπορά των εκτιμώμενων τιμών της Y γύρω από τη δειγματική μέση τιμή $\bar{y}$.Η ποσότητα αυτή αποτελεί μέτρο της διασποράς των δειγματικών τιμών της Y , που ερμηνεύεται από το υπόδειγμα της γραμμικής παλινδρόμησης (της διασποράς δηλαδή που οφείλεται στη γραμμική επίδραση της X επί της Y).Τέλος η ποσότητα $\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$ είναι το γνωστό *άθροισμα τετραγώνων των σφαλμάτων* ***(error sum of squares)*** και

εκφράζει τη διασπορά των δειγματικών τιμών της Y γύρω από την εκτιμώμενη ευθεία της παλινδρόμησης. Όσο μικρότερο είναι το άθροισμα τετραγώνων των σφαλμάτων τόσο πλησιέστερα βρίσκονται οι δειγματικές τιμές της εξαρτημένης μεταβλητής Y στην ευθεία των ελάχιστων τετραγώνων.

Ισχύει επομένως ότι:

Συνολικό άθροισμα τετραγώνων= άθροισμα τετραγώνων επεξηγούμενο από τη γραμμική παλινδρόμηση + άθροισμα τετραγώνων σφαλμάτων.

Για να είναι η προσαρμογή της ευθείας των ελάχιστων τετραγώνων επί των δειγματικών δεδομένων όσο το δυνατόν καλύτερη, θα πρέπει το άθροισμα των τετραγώνων των σφαλμάτων να είναι όσο το δυνατόν μικρότερο και, επομένως , σύμφωνα με την προηγούμενη εξίσωση , το άθροισμα τετραγώνων ,το επεξηγούμενο από τη γραμμική παλινδρόμηση ,να είναι όσο το δυνατόν μεγαλύτερο. Το ποσοστό επομένως του συνολικού αθροίσματος τετραγώνων που επεξηγείται από τη γραμμική παλινδρόμηση υπολογισμένο από το λόγο

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

Αποτελεί μέτρο της προσαρμογής της ευθείας των ελάχιστων τετραγώνων επί των δειγματικών τιμών και ορίζει το συντελεστή προσδιορισμού. Ο συντελεστής προσδιορισμού επομένως μπορεί να ερμηνευθεί ως το *ποσοστό της μεταβλητότητας των τιμών της Y που επεξηγείται από το υπόδειγμα της γραμμικής παλινδρόμησης.*

Με απλούς αλγεβρικούς μετασχηματισμούς αποδεικνύεται ότι

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \frac{\sum_{i=1}^{n}(\hat{a} + \hat{b}c_i - \hat{a} - \hat{b}\overline{c})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} =$$

$$= b^2 \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = \left[\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right]^2 \left[\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}\right] =$$

$$= \frac{\left[\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})\right]^2}{\left[\sum_{i=1}^{n}(x_i - \overline{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \overline{y})^2\right]} = r^2$$

Δηλαδή ότι $R^2 = r^2$.

Ο δειγματικός συντελεστής προσδιορισμού αποτελεί σημειακή εκτίμηση (όχι όμως αμερόληπτη) του πληθυσμιακού συντελεστή προσδιορισμού , ο οποίος ισούται με r^2, όπου r ο συντελεστής συσχέτισης των μεταβλητών X και Y .Ο δειγματικός συντελεστής προσδιορισμού τείνει να υπερεκτιμά τον αντίστοιχο πληθυσμιακό συντελεστή (η δειγματοληπτική κατανομή του είναι θετικά ασύμμετρη), ιδιαίτερα όταν το μέγεθος του δείγματος είναι σχετικά μικρό. Μια αμερόληπτη εκτιμήτρια του πληθυσμιακού συντελεστή προσδιορισμού είναι η ποσότητα

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n-2)}{\sum_{i=1}^{n} (y_i - \bar{y})^2 / (n-2)} = 1 - \frac{s^2_{y/x}}{s^2_y}$$

η οποία ονομάζεται *προσαρμοσμένος (ή διορθωμένος) συντελεστής προσδιορισμού* ***(adjusted coefficient of determination).***

Στο παράδειγμα της παλινδρόμησης της ενεργειακής πρόληψης επί της ηλικίας, ο δειγματικός συντελεστής προσδιορισμού ισούται με

$R^2 = r^2 = (0,554)^2 = 0,307$ ενώ ο προσαρμοσμένος συντελεστής προσδιορισμού ισούται με

$$R^2 = 1 - \frac{s^2_{y/x}}{s^2_y} = 1 - \frac{(544,32)^2}{(645,25)^2} = 0,288 .$$

Ο έλεγχος της προσαρμογής της ευθείας της παλινδρόμησης επί των πληθυσμιακών τιμών των μεταβλητών X και Y (δηλαδή ο έλεγχος της ύπαρξης γραμμικής σχέσης μεταξύ των μεταβλητών X και Y) μπορεί να γίνει με τη βοήθεια της ανάλυσης διακύμανσης της μεταβλητής Y .

Υποθέτοντας ότι ισχύουν οι προϋποθέσεις του υποδείγματος της απλής γραμμικής παλινδρόμησης, ο έλεγχος της ισότητας

$H_0 : b = 0$

Έναντι της εναλλακτικής

$H_A : b \neq 0$

Γίνεται με τη βοήθεια του λόγου

$$F = \frac{MSR}{MSE}.$$

Η ποσότητα *MSR* ονομάζεται *μέσο τετράγωνο της παλινδρόμησης **(regression mean square)*** και ισούται με το άθροισμα των τετραγώνων , το επεξηγούμενο από την παλινδρόμηση διαιρούμενο με τους αντίστοιχους βαθμούς ελευθερίας. Οι βαθμοί ελευθερίας που αντιστοιχούν στο συγκεκριμένο άθροισμα τετραγώνων ορίζονται από τον αριθμό των συντελεστών του υποδείγματος ελαττωμένο κατά 1. Στην προκειμένη περίπτωση οι συντελεστές του υποδείγματος είναι 2, άρα οι βαθμοί ελευθερίας είναι 2-1=1.επομένως

$$MSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 .$$

Η ποσότητα MSE ονομάζεται *μέσο τετράγωνο σφαλμάτων **(error mean square ή residual mean square)*** και ισούται με το άθροισμα των τετραγώνων των σφαλμάτων , διαιρούμενο επίσης με τους αντίστοιχους βαθμούς ελευθερίας , οι οποίοι είναι ίσοι με $n-2$.Δηλαδή

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

Όταν ισχύει η μηδενική υπόθεση $H_{0:} b = 0$, ο λόγος

$$F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n-2)}$$

Ακολουθεί την κατανομή F με 1 και $n-2$ βαθμούς ελευθερίας. Χρησιμοποιώντας επομένως τις κρίσιμες τιμές της αντίστοιχης κατανομής F , μπορούμε να απορρίψουμε ή να μην απορρίψουμε την H_0.

Στην περίπτωση της απλής γραμμικής παλινδρόμησης , η παραπάνω διαδικασία είναι απολύτως ισοδύναμη με τον έλεγχο της υπόθεσης $H_{0:} b = 0$ μέσω του t-test.

Εφαρμόζοντας τον έλεγχο της προσαρμογής της ευθείας των ελάχιστων τετραγώνων επί των δειγματικών τιμών για τα δεδομένα της ενεργειακής πρόσληψης προκύπτει ότι

$$F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / (n-2)} = 16{,}81$$

Παρατηρούμε ότι η πιθανότητα p να προκύψει μια τιμή τόσο μεγάλη ή και μεγαλύτερη από την τιμή 16,81 για μια κατανομή F με 1 και 38 βαθμούς ελευθερίας είναι $p\langle 0,001$. Επομένως η μηδενική υπόθεση $H_{0:} b = 0$ απορρίπτεται εκ νέου .

Για να συνεχίσουμε την εργασία μας παρακάτω θα αναλύσουμε με τον ίδιο τρόπο τον συντελεστή συσχέτισης και τα τρία είδη του στα οποία χωρίζεται και τα οποία μας ενδιαφέρει να αναλύσουμε.

ΚΕΦΑΛΑΙΟ 5

5.1 Ο ΣΥΝΤΕΛΕΣΤΗΣ Ρ ΤΟΥ SPEARMAN

Έστω (X1, Y1), (X2, Y2), ..., (Xn, Yn) ένα δείγμα n παρατηρήσεων πάνω στο τυχαίο διάνυσμα (X, Y). Έστω R(Xi) ο βαθμός ή η τάξη μεγέθους της μεταβλητής X όταν αυτή συγκρίνεται με τις άλλες X τιμές, για i = 1, 2, ..., n. Δηλαδή, R(Xi) = 1, αν Xi είναι η μικρότερη από τις τιμές X1, X2, ..., Xn, R(Xi) = 2, αν η μεταβλητή Xi είναι η επόμενη μικρότερη τιμή, κ.ο.κ, με τον βαθμό n να αντιστοιχεί στην μεγαλύτερη τιμή από τις X1, X2, ..., Xn. Με όμοιο τρόπο, έστω ότι R(Yi) έχει την τιμή 1, 2, ..., n ανάλογα με το σχετικό μέγεθος της μεταβλητής Yi, όταν αυτή συγκρίνεται με τις υπόλοιπες Y τιμές.

Τα δεδομένα μπορούν να αποτελούνται και από μη αριθμητικές παρατηρήσεις, οι οποίες εμφανίζονται σε ζεύγη, αν οι παρατηρήσεις είναι τέτοιες που να μπορούν να διαταχθούν κατά αύξουσα σειρά μεγέθους με τον τρόπο που μόλις περιγράψαμε.

Στην περίπτωση αυτή, η διάταξη μπορεί να βασίζεται στην ποιότητα των παρατηρήσεων (απ την χειρότερη παρατήρηση στην καλύτερη παρατήρηση) ή στον βαθμό προτίμησης που μπορεί να αντιστοιχηθεί στις παρατηρήσεις κ.ο.κ.

Στις περιπτώσεις όπου δύο ή περισσότερες από τις τιμές ταυτίζονται (tie), αντιστοιχίζουμε σε κάθε μία από τις ίσες αυτές τιμές τον μέσο των βαθμών που θα είχαν αν δεν ταυτίζονταν.

Το μέτρο συσχέτισης που προτάθηκε από τον Spearman το 1904 δεν είναι άλλο από τον συντελεστή r του Pearson υπολογιζόμενο, όμως, με βάση τις τάξεις μεγέθους των παρατηρήσεων και όχι αυτές κάθε αυτές τις παρατηρήσεις. Δηλαδή,

$$\rho = \frac{\sum_{i=1}^{n}\left[R(X_i)-\overline{R(X)}\right]\left[R(Y_i)-\overline{R(Y)}\right]}{\left(\sum_{i=1}^{n}\left[R(X_i)-\overline{R(X)}\right]^2\right)^{1/2}\left(\sum_{i=1}^{n}\left[R(Y_i)-\overline{R(Y)}\right]^2\right)^{1/2}},$$

όπου $\overline{R(X)}=\sum_{i=1}^{n}R(X_i)/n$ και $\overline{R(Y)}=\sum_{i=1}^{n}R(Y_i)/n.$

Είναι προφανές, ότι εάν δεν υπάρχουν περιπτώσεις ίσων X τιμών (αντίστοιχα Y τιμών), τότε

$$\overline{R(X)}=\frac{1}{n}\sum_{i=1}^{n}R(X_i)=\frac{1}{n}\sum_{i=1}^{n}i=\frac{1}{n}\frac{n(n+1)}{2}=\frac{n+1}{2},$$

με αντίστοιχη έκφραση για τον μέσο βαθμό των Y τιμών. Επιπλέον,

$$\sum_{i=1}^{n}\left[R(X_i)-\overline{R(X)}\right]^2=\sum_{i=1}^{n}\left[i-\frac{n+1}{2}\right]^2$$

$$=\sum_{i=1}^{n}\left[i^2+\left[\frac{n+1}{2}\right]^2-2\frac{i\,(n+1)}{2}\right]$$

$$=\sum_{i=1}^{n}i^2+\frac{n\,(n+1)^2}{4}-(n+1)\sum_{i=1}^{n}i$$

$$=\frac{n\,(n+1)\,(2n+1)}{6}+\frac{n\,(n+1)^2}{4}-\frac{n\,(n+1)^2}{2}$$

$$=\frac{n\,(n+1)\,(2n+1)}{6}-\frac{n\,(n+1)^2}{4}$$

$$=\frac{n\,(n+1)}{2}\left[\frac{2n+1}{3}-\frac{n+1}{2}\right]$$

$$=\frac{n\,(n+1)}{12}(n-1)$$

$$=\frac{n\,(n^2-1)}{12},$$

με αντίστοιχη έκφραση για τις Y τιμές. Επομένως, αν όλες οι παρατηρήσεις είναι διακεκριμένες, ο συντελεστής ρ του Spearman μπορεί να γραφεί με την ισοδύναμη μορφή

$$\rho=\frac{\sum_{i=1}^{n}\left(R(X_i)-\frac{n+1}{2}\right)\left[R(Y_i)-\frac{n+1}{2}\right]}{n\left(n^2-1\right)/12}$$

Στην περίπτωση αυτή, συχνά, χρησιμοποιείται μία ισοδύναμη μορφή του συντελεστή ρ, η οποία προσφέρεται περισσότερο για ταχύτερους υπολογισμούς:

$$\rho=1-\frac{6T}{n\left(n^2-1\right)},$$

όπου,

$$T = \sum_{i=1}^{n} \left[R(X_i) - R(Y_i) \right]^2 .$$

Αν οι X τιμές (αντίστοιχα οι Y τιμές) δεν είναι όλες διακεκριμένες, δηλαδή υπάρχουν περιπτώσεις ίσων τιμών, τότε χρησιμοποιείται η εξής μορφή του συντελεστή ρ:

$$\rho = \frac{\sum_{i=1}^{n} \left[R(X_i) - \frac{n+1}{2} \right] \left[R(Y_i) - \frac{n+1}{2} \right]}{\sqrt{\sum_{i=1}^{n} \left[R(X_i) - \frac{n+1}{2} \right]^2 \sum_{i=1}^{n} \left[R(Y_i) - \frac{n+1}{2} \right]^2}}$$

$$= \frac{\sum_{i=1}^{n} R(X_i) R(X_i) - n \left[\frac{n+1}{2} \right]^2}{\sqrt{\left[\sum_{i=1}^{n} R(X_i)^2 - n \left(\frac{n+1}{2} \right)^2 \right] \left[\sum_{i=}^{n} R(Y_i)^2 - n \left(\frac{n+1}{2} \right)^2 \right]}} .$$

5.2 Ο ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ 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 καθεαυτές και, επιπλέον, η κατανομή του δεν εξαρτάται από την κατανομή των μεταβλητών X και Y, 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. 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. Ενα 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 (Xi, Yi), i = 1, 2, ..., n, πάνω στο τυχαίο διάνυσμα (X, Y).

Ορισμός: Δύο παρατηρήσεις, έστω (Xj, Yj) και (Xk, Yk), ονομάζονται εναρμονισμένες ή συσχετισμένες (concordant), αν και τα δύο μέλη της μίας παρατήρησης είναι μεγαλύτερα (ή μικρότερα) από τα αντίστοιχα μέλη της άλλης παρατήρησης. Δηλαδή, αν Xj> Xk (αντίστοιχα, Xj< Xk), τότε Yj> Yk (αντίστοιχα, Yj< Yk).

Οι παρατηρήσεις (Xj, Yj) και (Xk, Yk) θα ονομάζονται μη εναρμονισμένες ή μη συσχετισμένες (discordant), αν η διάταξη των πρώτων μελών τους είναι αντίθετη από την διάταξη των δεύτερων μελών

τους, δηλαδή, αν Xj > Xk (αντίστοιχα, Xj < Xk), τότε Yj < Yk (αντίστοιχα, Yj > Yk). Ισοδύναμα, δύο ζεύγη παρατηρήσεων (Xj, Yj) και (Xk, Yk) θαονομάζονται εναρμονισμένα αν οι διαφορές Xj- Xk και Yj - Yk έχουν το ίδιο πρόσημο (αν (Xj - Xk) (Yj - Yk) > 0). Τα ζεύγη (Xj, Yj) και (Xk,Yk) θα ονομάζονται μη εναρμονισμένα αν οι διαφορές Xj- Xk και Yj - Yk έχουν αντίθετο πρόσημο (αν (Xj - Xk)(Yj- Yk) < 0).

Έστω Nc και Nd οι αριθμοί των εναρμονισμένων και μη εναρμονισμένων ζευγών παρατηρήσεων, αντίστοιχα. Τα ζεύγη των παρατηρήσεων (Xj, Yj) και (Xk, Yk), για τα οποία ισχύει ότι Xj= Xk ή/και Yj = Yk , 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 $\frac{n}{2}$ = n (n-1) /2 διαφορετικούς τρόπους έπεται ότι Nc=Nd+N0= $\frac{n}{2}$

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:

$$\tau = \frac{N_c - N_d}{\binom{n}{2}} = \frac{N_c - N_d}{n(n-1)/2}.$$

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 τ ικανοποιεί όλες τις προϋποθέσεις που προαναφέρθηκαν. Ο υπολογισμός του συντελεστή τ γίνεται απλούστερος, αν οι παρατηρήσεις (Xi, Yi), i = 1, 2, ..., n διαταχθούν σε μία στήλη κατά αύξουσα τάξη μεγέθους των τιμών των παρατηρήσεων πάνω στην τυχαία μεταβλητή Χ. Τότε, κάθε Υ τιμή χρειάζεται να συγκριθεί μόνο με τις Υ 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

5.3 ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ PEARSON

Ο συντελεστής συσχέτισης (correlation coefficient) είναι ένας αριθμητικός δείκτης που **δείχνει την ισχύ και την κατεύθυνση της σχέσης μεταξύ δύο μεταβλητών**. Ο πιο συνηθισμένος και πιο χρήσιμος από αυτούς είναι ο συντελεστής συσχέτισης Pearson που κυμαίνεται σε μέγεθος από -0 έως +1. Το πρόσημο "+" υποδηλώνει θετική συσχέτιση, δηλ. ότι οι τιμές μιας μεταβλητής αυξάνονται όταν αυξάνονται και οι τιμές της άλλης. Το πρόσημο "-" σημαίνει αρνητική συσχέτιση, δηλ. ότι οι τιμές μιας μεταβλητής αυξάνονται καθώς μειώνονται οι τιμές της άλλης μεταβλητής. Όταν ο συντελεστής συσχέτισης είναι με +1 τότε λέμε ότι υπάρχει η τέλεια θετική γραμμική συσχέτιση και όταν είναι -1 τότε λέμε ότι υπάρχει η τέλεια αρνητική γραμμική συσχέτιση. Ένας συντελεστής συσχέτισης -0,5 δηλώνει ότι υπάρχει μία μέτρια αρνητική συσχέτιση.

Αναλυτικότερα, ο συντελεστής συσχέτισης Pearson r καταδεικνύει την ύπαρξη ή όχι σχέσης μεταξύ δύο μεταβλητών και υπολογίζει την μορφή αυτής της σχέσης (θετική ή αρνητική συσχέτιση) αλλά και την ένταση της (επίπεδο στατιστικής σημαντικότητας). Το στατιστικό αυτό κριτήριο ελέγχει τη μηδενική υπόθεση ότι δεν υπάρχει συσχέτιση μεταξύ δύο μεταβλητών. Ο συντελεστής συσχέτισης Pearson r είναι παραμετρικό κριτήριο και έτσι τα δεδομένα και στις δύο υπό μελέτη μεταβλητές θα πρέπει να είναι καταχωρημένα σε τουλάχιστον ισοδιαστημική κλίμακα, να ακολουθούν κανονική κατανομή και να έχουν όμοιες διασπορές. Οι τιμές που μπορεί να πάρει ο συντελεστής συσχέτισης r είναι από –1 μέχρι +1. Όταν το πρόσημο του συντελεστή είναι θετικό (θετική συσχέτιση) η μία μεταβλητή αυξάνεται καθώς αυξάνεται και η άλλη. Όταν το πρόσημο του συντελεστή είναι αρνητικό (αρνητική συσχέτιση) η μία μεταβλητή αυξάνεται καθώς η άλλη μειώνεται. Αν ο συντελεστής έχει τιμή 1 (μέγιστη τιμή) έχουμε απόλυτη συσχέτιση ενώ όταν είναι 0 δεν έχουμε καθόλου συσχέτιση μεταξύ των δύο μεταβλητών.

ΠΑΡΑΔΕΙΓΜΑ: Οι πιο κάτω πίνακες παρουσιάζουν τον συντελεστή συσχέτισης Pearson r και το αντίστοιχο επίπεδο στατιστικής σημαντικότητας μεταξύ των μεταβλητών «*Χρόνια υπηρεσίας στην εκπαίδευση*» - «*Πιστεύω ότι θα μάθω αρκετά καλά το SPSS*» και «*Χρόνια υπηρεσίας στην εκπαίδευση*» - «*Έχω αρκετές γνώσεις στατιστικής*» με βάση τα δεδομένα που προέκυψαν από το ερωτηματολόγιο που δόθηκε σε σεμινάριο του τμήματος μαθηματικών του πανεπιστημίου Πατρών.

Correlations

Variables	Statistics	Πιστεύω ότι θα μάθω αρκετά καλά το SPSS
Χρόνια υπηρεσίας στην εκπαίδευση	Pearson Correlation	,238**
	Sig. (2-tailed)	,003
	N	154

**. Correlation is significant at the 0.01 level (2-tailed).

Correlations

Variables	Statistics	Έχω αρκετές γνώσεις στατιστικής
Χρόνια υπηρεσίας στην εκπαίδευση	Pearson Correlation	-,185*
	Sig. (2-tailed)	,022
	N	154

*. Correlation is significant at the 0.05 level (2-tailed).

Με βάση τα παραπάνω ο **ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ (r)** ονομάζεται ο δείκτης που είναι αποτέλεσμα της ποσοτικής μέτρησης της έντασης (γραμμικής) σχέσης μεταξύ δύο μεταβλητών ονομάζεται συντελεστής συσχέτισης (correlation coefficient)

-Το εύρος τιμών του συντελεστή συσχέτισης είναι από -1,00 εώς +1,00.

-Τιμές κοντά στο -1,00 και 1,00 υποδεικνύουν τέλεια (ισχυρή) συσχέτιση.

-Τιμές του δείκτη κοντά στο 0 υποδηλώνουν ότι οι δύο μεταβλητές δεν σχετίζονται γραμμικά.

-Αρνητικές τιμές υποδεικνύουν αρνητική συσχέτιση, ενώ θετικές τιμές υποδεικνύουν θετική συσχέτιση.

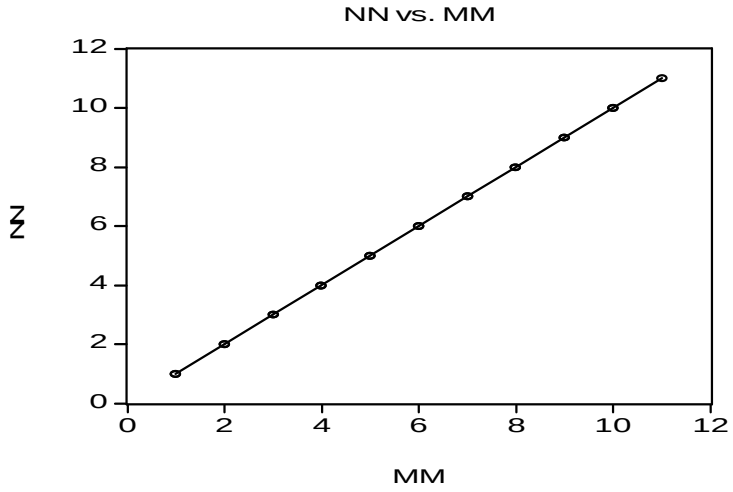
-Η συσχέτιση μεταξύ δυο μεταβλητών μπορεί να είναι: Τέλεια θετική(αρνητική), έντονη θετική (αρνητική), ασθενής θετική (αρνητική)

Είναι φανερό ότι η πρόχειρη ή επιπόλαιη ερμηνεία και χρήση του r οδηγεί πολλές φορές σε παρερμηνείες ή και σε λανθασμένα συμπεράσματα. Για αιτιολογικά συμπεράσματα, σχέδον πάντοτε, απαιτείται πειραματισμός. Σε κάθε περίπτωση, αιτιώδη σχέση (αλληλεξάρτηση) μεταξύ δύο μεταβλητών δεχόμαστε μόνον όταν υπάρχει επιστημονική ή λογική βάση που την υπαγορεύει.

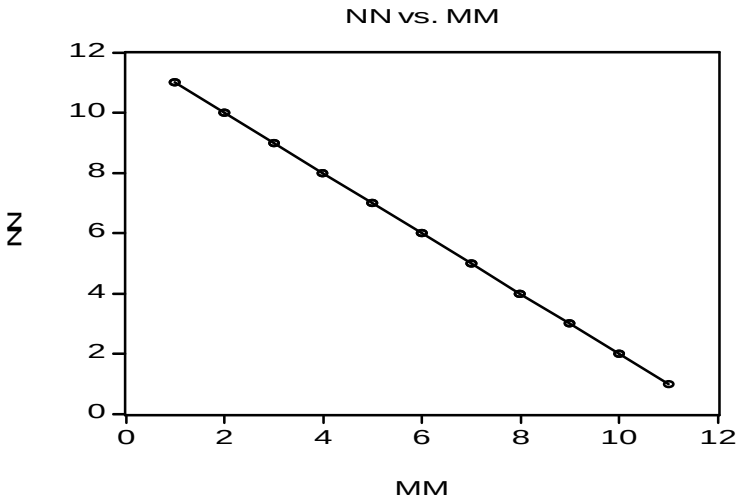
5.4 ΓΡΑΦΙΚΗ ΑΠΕΙΚΟΝΙΣΗ ΣΥΣΧΕΤΙΣΕΩΝ

Η γραφική απεικόνιση των συντελεστών συσχέτισης πραγματοποιείται με βάση το διάγραμμα διασποράς.Κινείται πάνω σε αυτά τα πλαίσια απεικόνισης.Όπως αναφέραμε παραπάνω το διάγραμμα διασποράς είναι ένα γράφημα που αναδεικνύει την σχέση μεταξύ δύο μεταβλητών. Στον οριζόντιο άξονα μετράται η μία μεταβλητή (εξαρτημένη) και στον κάθετο η άλλη μεταβλητή (ανεξάρτητη).

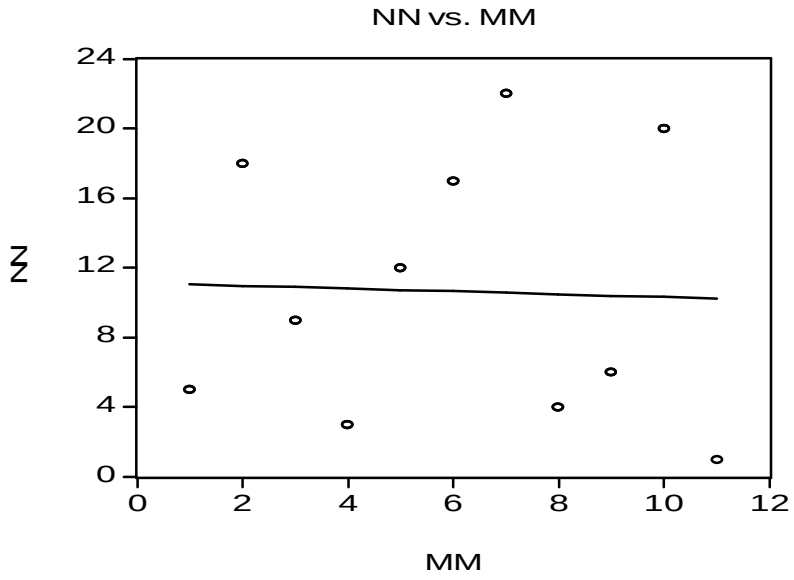
Τέλεια Θετική Συσχέτιση



Τέλεια Αρνητική Συσχέτιση

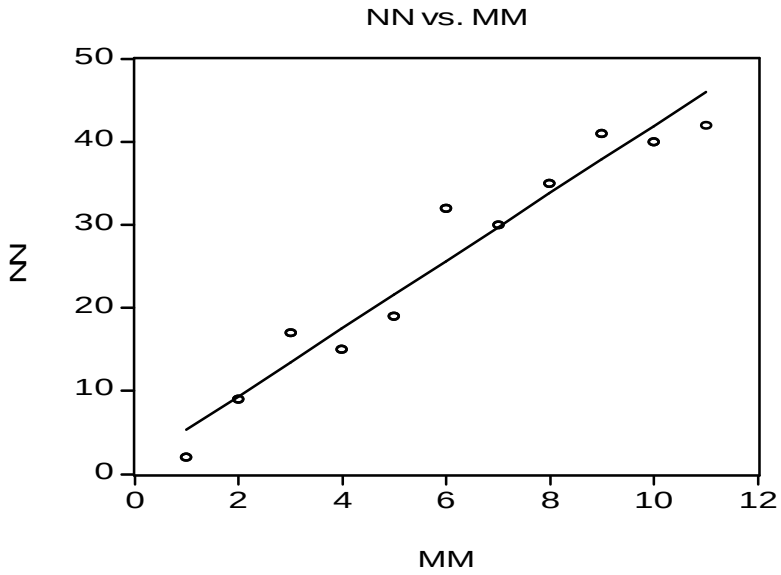


- **Μηδενική Συσχέτιση**



-

Ισχυρή Θετική Συσχέτιση



Ιδιότητες

• Ο συντελεστής γραμμικής συσχέτισης είναι καθαρός αριθμός και δεν έχει μονάδες μέτρησης.

• $-1 \leq r \leq 1$ Όταν παίρνει την τιμή -1, σημαίνει ότι υπάρχει πλήρης(τέλεια) συσχέτιση και μάλιστα οι τιμές της μιας Μεταβλητής αυξάνουν, ενώ οι τιμές της άλλης μεταβλητής μειώνονται. Ομοίως η τιμή+1 σημαίνει πλήρης(τέλεια) συσχέτιση των δύο μεταβλητών και μάλιστα οι τιμές και των δύο βαίνουν αύξουσες ή φθίνουσες. Και στις δύο αυτές ακραίες τιμές του συντελεστή γραμμικής συσχέτισης ισχύει ανάμεσα στις δύο μεταβλητές X και Y η ποσοτική(συναρτησιακή, μαθηματική σχέση $Y = \alpha + \beta \cdot X$

• Αντίστροφα, όταν οι μεταβλητές Χ και Υ συνδέονται με τη σχέση Υ=α+ β·Χ, τότε r = -1 ανβ>0 καιr=1 ανβ<0.

• Αν r=0 τότε οι μεταβλητές Χ και Υ λέγονται ασυσχέτιστες.

Εδώ θα πρέπει να θυμηθούμε άλλο πράγμα εννοούμε με τον όρο ανεξάρτητες μεταβλητές και άλλο προτίμα με τον όρο ασυσχέτιστες

- Αν r = ±1 υπάρχει τέλεια γραμμική συσχέτιση.

  Αν −0,3≤ r < 0,3 δεν υπάρχει γραμμική συσχέτιση. Αυτό, όμως, δεν σημαίνει

  ότι δεν υπάρχει άλλου είδους συσχέτιση μεταξύ των δύο μεταβλητών.

  Αν −0,5 < r ≤ −0,3 ή 0,3≤ r < 0,5 υπάρχει ασθενής γραμμική συσχέτιση.

  Αν −0,7 < r ≤ −0,5 ή 0,5 ≤ r < 0,7 υπάρχει μέση γραμμική συσχέτιση.

  Αν −0,8 < r ≤ −0,7 ή 0,7 ≤ r < 0,8 υπάρχει ισχυρή γραμμική συσχέτιση.

  Αν −1< r ≤ −0,8 ή 0,8 ≤ 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 αύξηση της μιας μεταβλητής αντιστοιχεί σε αύξηση ή σε μείωση της άλλης 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

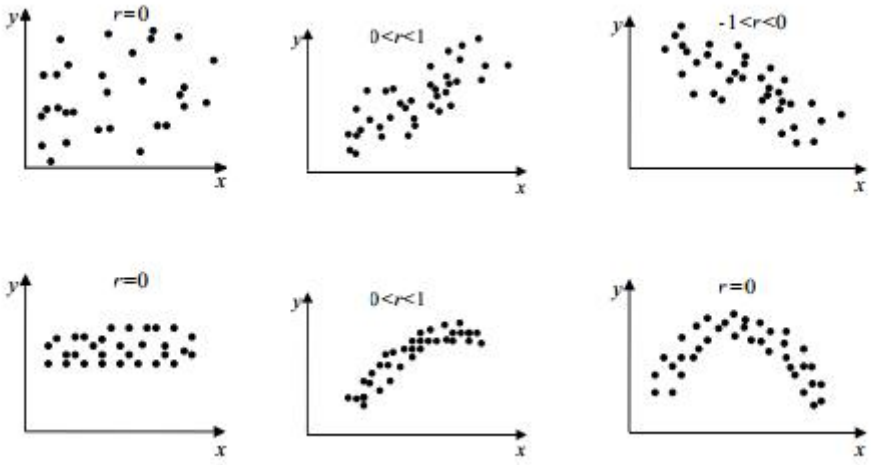
μόνο που το διάγραμμα διασποράς (στικτό διάγραμμα) έχει σχήμα επιμήκους

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.



Όταν χρησιμοποιείται ο συντελεστής συσχέτισης Pearson r πρέπει να λαμβάνεται υπόψη ότι :

1. Επηρεάζεται από ακραίες παρατηρήσεις. Αν έχετε άτομα που έχουν πολύ μεγάλες ή πολύ μικρές τιμές είναι πολύ πιθανό αυτές να (δια)στρεβλώνουν σημαντικά το μέγεθος του συντελεστή συσχέτισης.

2. Επηρεάζεται, όπως και οι περισσότεροι δείκτες, από το μέγεθος του δείγματος. Πολύ μικρά ή πολύ μεγάλα δείγματα δημιουργούν προβλήματα στην ερμηνεία της πιθανότητας και της στατιστικής σημαντικότητας.

3. Η αξιοπιστία του δείκτη "πλήττεται" από τη μη-τήρηση της προϋπόθεσης αναφορικά με τη γραμμικότητα της σχέσης. Εξετάζεται η γραμμικότητα (τουλάχιστον) από το διάγραμμα σχεδιασμού για την ύπαρξη μη γραμμικών σχέσεων.

4. Επηρεάζεται, όπως και οι περισσότεροι δείκτες, από την αξιοπιστία των μετρήσεων. Αν οι μετρήσεις έχουν μεγάλες τιμές σε στατιστικό σφάλμα, το πιθανότερο είναι ότι το σφάλμα αυτό θα "φορτίσει" και τον συντελεστή συσχέτισης με απρόβλεπτες συνέπειες (συνήθως συμπιέζει τις τιμές του δείκτη προς τα κάτω16).
5. Είναι ακατάλληλος όταν οι μεταβλητές δεν είναι συνεχείς αλλά διακριτές. Στην δεύτερη περίπτωση συντελεστές όπως ο Φ ή ο point-biserial είναι καταλληλότεροι, ανάλογα με το αν η μία ή και οι δύο μεταβλητές είναι κατηγορικές.
6. Επηρεάζεται από την ύπαρξη στρεβλών κατανομών, ειδικά όταν αυτές αποκλίνουν κατά πολύ από την κανονικότητα. Αν καταπατείται κάποια από τις προϋποθέσεις εξετάστε 17 την χρήση μη παραμετρικών κριτηρίων, τα οποία συζητώνται περιεκτικά παρακάτω.
7. Δεν μπορεί να χρησιμοποιηθεί για τη διερεύνηση αιτιωδών σχέσεων. Αυτό δεν σημαίνει ότι δεν υπάρχουν αιτιακές σχέσεις μεταξύ των μεταβλητών που μελετήθηκαν, αλλά ότι η χρήση του συντελεστή συσχέτισης δεν επιτρέπει αυτό το συμπέρασμα. Το μόνο συμπέρασμα που μπορεί να προκύψει από τη χρήση του συντελεστή συσχέτισης είναι ότι οι δύο μεταβλητές συνδιακυμαίνονται. Μόνο η μελλοντική χρήση πειραματικών σχεδίων μπορεί να επιβεβαιώσει αν οι αρχικές αυτές συνάφειες έχουν αιτιακό χαρακτήρα.

Ένα παράδειγμα για το πως λειτουργεί ο συντελεστής pearson είναι το παρακάτω

5.5 ΕΦΑΡΜΟΓΗ

Να υπολογιστεί και να ερμηνευτεί ο συντελεστής συσχέτισης *r* μεταξύ των μεταβλητών *X* και *Y* με βάση τις παρακάτω τιμές:

x	10	13	17	21	25	28	30
y	21	24	29	25	36	33	40

ΛΥΣΗ

Για τον υπολογισμό του συντελεστή συσχέτισης μεταξύ των X και Y διευκολύνει ο παρακάτω πίνακας:

x	y	x^2
10 13 17 21 25 28 30	21 24 29 25 36 33 40	100 169 289 441 625 784 900
$\Sigma x = 144$	$\Sigma y = 208$	$\Sigma x^2 = 3308$

y^2	xy
441 576 841 625 1296 1089 1600	210 312 493 525 900 924 1200
$\Sigma y^2 = 6468$	$\Sigma xy = 4564$

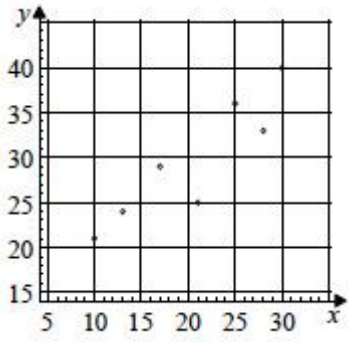
ν = 7

Ο συντελεστής συσχέτισης υπολογίζεται από τη σχέση:

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

$$= \frac{7(4564) - (144)(208)}{\sqrt{7(3308) - (144)^2}\sqrt{7(6468) - (208)^2}} \approx 0{,}9 .$$

Η υψηλή τιμή του r μας δείχνει ότι υπάρχει πολύ έντονη θετική γραμμική συσχέτιση Μεταξύ των μεταβλητών Χ και Υ ,όπως εξάλλου μπορούμε να το διαπιστώσουμε και Από το αντίστοιχο δίαγραμμα διασποράς.



Πριν προχωρήσουμε παρακάτω θα ήταν οφέλιμο να αναφέρουμε τον πίνακα συσχετίσεων r καθώς και οτιδήποτε αφορά τον συγκεκριμένο πίνακα.

ΚΕΦΑΛΑΙΟ 6

6.1 ΠΙΝΑΚΑΣ ΣΥΣΧΕΤΙΣΕΩΝ R

Ο πίνακας συσχετίσεων είναι ο πίνακας που περιέχει ως στοιχεία του τους συντελεστές συσχέτισης του Pearson για κάθε ζευγάρι.Από όσα είδαμε και αναφέρουμε και παραπάνω ο συντελεστής Pearson μετράει μόνο τη γραμμική συσχέτιση ανάμεσα στις μεταβλητές οπότε είναι κατάλληλος μόνο για ζεύγη ποσοτικών μεταβλητών

$$R = \begin{bmatrix} 1 & r_{12} & r_{1p} \\ r_{21} & 1 & r_{2p} \\ \cdots & \cdots & \cdots \\ r_{p1} & r_{p2} & 1 \end{bmatrix} \quad \text{όπου} \qquad r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}},\ j,k = 1,2,\ldots,p$$

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

Τιμές -1 και 1 σημαίνουν απόλυτα γραμμική σχέση των δυο μεταβλητών.Το πρόσημο υποδηλώνει την ύπαρξη θετικής ή αρνητικής σχέσης.Η θετική σχέση ερμηνεύεται εως εξής: Όσο αυξάνεται η τιμή της μιας μεταβλητής τόσο αυξάνεται και η τιμή της άλλης ένω στην αρνητική σχέση όσο αυξάνεται η τιμή της μιας μεταβλητής μειώνεται η τιμή της άλλης.

6.2 ΕΛΕΓΧΟΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΣΥΝΤΕΛΕΣΤΗ ΣΥΣΧΕΤΙΣΗΣ R

Εφόσον ο συντελεστής συσχέτισης r αποτελεί εκτίμηση του πληθυσμιακού ρ θα
πρέπει να αξιολογηθεί για να διαπιστωθεί αν η εκτίμηση είναι καλή. Θα πρέπει να

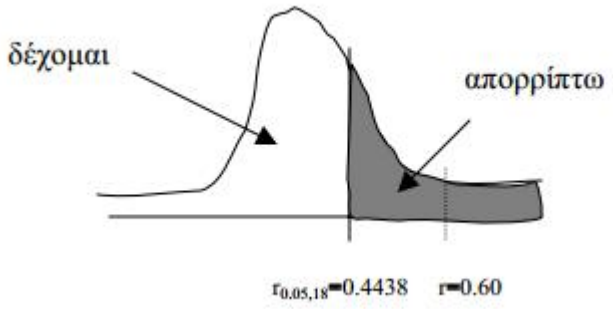
ελεγχθεί με κάποια πιθανότητα αν ο εκτιμηθής συντελεστής είναι στατιστικά σημαντικός
και άρα αξιόπιστος.
Για να προχωρήσουμε σε ελέγχους υποθέσεων θα πρέπει να γνωρίζουμε την
κατανομή δειγματοληψίας του r.Το γεγονός ότι έχουμε θεωρήσει κανονική κατανομή για
την από κοινού κατανομή των (x,y) δίνει την δυνατότητα να γνωρίζουμε την κατανομή
δειγματοληψίας του συντελεστή r. Η μορφή κατανομής του r διαφοροποιείται με το αν ο
πληθυσμιακός ρ είναι ίσος με το 0 ή διάφορος του 0.
Για ρ=0 αποδεικνύεται ότι η κατανομή είναι συμμετρική γύρο από το 0 με διακύμανση
εξαρτώμενη από το μέγεθος του δείγματος n και τύπο

$$\text{var}(r) = \frac{1-p^2}{n-2}$$

Οι θεωρητικές τιμές του r έχουν υπολογιστεί και παρουσιάζονται σε ειδικούς πίνακες με επίπεδο σημαντικότητας α και βαθμούς ελευθερίας n-2.
Ο στατιστικός έλεγχος αναφέρεται στο αν η τιμή του r που εκτιμήθηκε από τα στοιχεία του δείγματος διαφέρει σημαντικά από το ρ=0 δηλαδή
H0: ρ=0 vs H1: ρ≠0.
Τότε αν |r| > rα,n-2 τότε ο εκτιμηθείς συντελεστής είναι στατιστικά σημαντικός αν |r| < rα,n-2 τότε ο εκτιμηθείς συντελεστής δεν είναι στατιστικά σημαντικός

Παράδειγμα

Έστω δείγμα από 20 παρατηρήσεις με δειγματικό συντελεστή συσχέτισης r=0.60.
Για επίπεδο σημαντικότητας α=0.05 και βαθμούς ελευθερίας n=20-2=18 έχουμε από τους πίνακες $r_{0,05,18} = 0,4438$. Αρα



Άρα απορρίπτουμε την υπόθεση Η0 με αποτέλεσμα ο συντελεστής συσχέτισης r=0.60 να
είναι στατιστικά σημαντικός.
Επίσης μπορεί να χρησιμοποιηθεί το στατιστικό τεστ με βάση τον τύπο

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \to t_{n-2,a/2}$$

όπου ακολουθεί t-κατανομή με n-2 βαθμούς ελευθερίας. Κατά συνέπεια μπορούν να
χρησιμοποιηθούν οι πίνακες της t-κατανομής, η οποία ακολουθεί ασυμπτωτικά την
κανονική κατανομή.
Ο στατιστικός έλεγχος αναφέρεται στο αν η τιμή του r που εκτιμήθηκε από τα
στοιχεία του δείγματος διαφέρει σημαντικά από το ρ=0 δηλαδή Η0: ρ=0 vs Η1: ρ≠0.
Τότε αν |t| > tα/2,n-2 τότε ο εκτιμηθείς συντελεστής είναι στατιστικά σημαντικός

αν |t| < rα/2,n-2 τότε ο εκτιμηθείς συντελεστής δεν είναι στατιστικά σημαντικός.

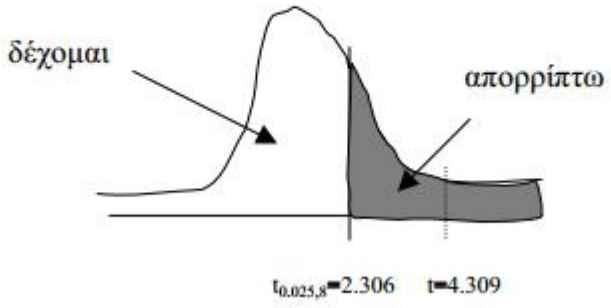
Παράδειγμα

Έστω δείγμα από 10 παρατηρήσεις με δειγματικό συντελεστή συσχέτισης r=0.836. Το στατιστικό τεστ δίνεται από τον παρακάτω τύπο

$$t = \frac{t}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,836}{\sqrt{\frac{1-0,836^2}{10-2}}} = 4,309$$

Για επίπεδο σημαντικότητας α=0.025 και βαθμούς ελευθερίας n=10-2=8 έχουμε από

τους πίνακες t0.025,8=2.306. Αρα



Αρα απορρίπτουμε την υπόθεση Η0 με αποτέλεσμα ο συντελεστής συσχέτισης r=0.836
να είναι στατιστικά σημαντικός.

Για ρ≠0

Η κατανομή δεν είναι συμμετρική και άρα δεν μπορεί να χρησιμοποιηθεί η t-
κατανομή. Αποδεικνύεται ότι

$$z_r = \frac{1}{2}\ln\frac{1+r}{1-r} \to N(\frac{1}{2}\ln\frac{1+p}{1-p}, \frac{1}{n-3})$$

Άρα μπορούν να χρησιμοποιηθούν οι πίνακες της κανονικής κατανομής για
τυποποιημένες τιμές z.

Παράδειγμα

Έστω δείγμα από 19 παρατηρήσεις με δειγματικό συντελεστή συσχέτισης
r=0.762. Επιθυμούμε να ελέγξουμε αν η τιμή αυτή είναι στατιστικά σημαντική από την
υποθετική του πληθυσμιακού ρ=0.5. Άρα θα ελέγξουμε Η0: ρ=0.5 vs Η1: ρ≠0.5.

Από τα δεδομένα έχουμε

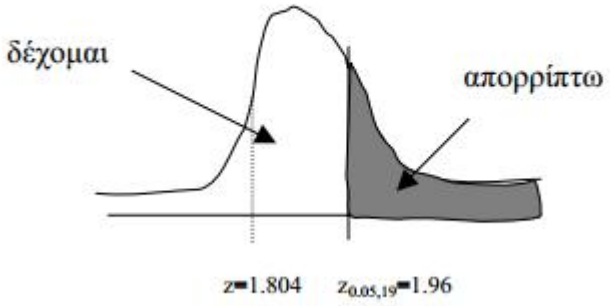
$$z_r = \frac{1}{2}\ln\frac{1+r}{1-r} = ... = 1$$

$$E[z_r] = \frac{1}{2}\ln\frac{1+p}{1-p} = \ldots = 0,549$$

$$V[z_r] = \sqrt{\frac{1}{n-3}} = \ldots = 0,25$$

Άρα έχουμε $z = \frac{z_r - E[z_r]}{V[z_r]} = \frac{1-0,549}{0,25} = 1,804$.

Για επίπεδο σημαντικότητας α=0.05 και βαθμούς ελευθερίας n=19 έχουμε από τους πίνακες z0.05,19=1.96.
Αρα



Άρα δεχόμαστε την υπόθεση Η0 με αποτέλεσμα ο πληθυσμιακός συντελεστής
συσχέτισης ρ=0.5 να μην διαφέρει σημαντικά από το δειγματικό συντελεστή συσχέτισης $r = 0,762$.

6.3 ΜΗ-ΠΑΡΑΜΕΤΡΙΚΟΙ ΣΥΝΤΕΛΕΣΤΕΣ ΣΥΣΧΕΤΙΣΗΣ (Spearman)

Όταν οι παραμετρικές προϋποθέσεις δεν ικανοποιούνται (δηλ. η κανονικότητα και η γραμμικότητα, το εύρος των παρατηρήσεων και η ύπαρξη ισοδιαστημικής κλίμακας), τότε πρέπει να χρησιμοποιηθούν

εναλλακτικοί στατιστικοί δείκτες για την ανίχνευση σχέσεων μεταξύ μεταβλητών. Ένας από αυτούς είναι και ο δείκτης συσχέτισης του Spearman, που υπολογίζεται μετατρέποντας τα δεδομένα σε “σειρές” με βάση το μέγεθος τους (δηλ. οι αρχικές τιμές μπαίνουν σε σειρά με βάση το μέγεθός τους: πρώτος, δεύτερος, τρίτος, κλπ.). Αυτό έχει σαν αποτέλεσμα οι αποστάσεις μεταξύ των παρατηρήσεων να χάνουν τη σημασία τους και να αξιολογείται η σειρά των συμμετεχόντων στην πρώτη μεταβλητή σε σχέση με τη σειρά που αυτοί έχουν στην δεύτερη μεταβλητή, κ.ό.κ. Το μέγεθος της συμφωνίας ή όχι της σειράς στις δύο μεταβλητές εκφράζει και το πρόσημο αλλά και το μέγεθος της σχέσης. Για παράδειγμα, αν κάποιος που είναι πρώτος στις επιδόσεις είναι και πρώτος στη δημοτικότητα και ακολουθείται από άτομα μικρότερων επιδόσεων και μικρότερης δημοτικότητας, θα αναδείξει μια θετική συνάφεια μεταξύ επίδοσης και δημοτικότητας.

Η εξίσωση για τον υπολογισμό του συντελεστή συσχέτισης Spearman έχει ως εξής:

$$1-\frac{6\sum D^2}{N\left(N^2-1\right)}$$

όπου N είναι ο αριθμός των ζευγαριών και D η διαφορά στη σειρά μεταξύ πρώτης και δεύτερης μέτρησης (ζεύγη μετρήσεων). Στο SPSS, η χρήση του δείκτη Spearman είναι πολύ απλή αφού το μόνο που χρειάζεται είναι η επιλογή “Spearman” στο μενού CorrelateÆBivariate. Άλλος σχετικός μη παραμετρικός δείκτης είναι ο “τ” του Kendall ο οποίος όμως δε θα συζητηθεί επί του παρόντος.

Στο παραπάνω παράδειγμα η εφαρμογή του συντελεστή Spearman έδωσε τα εξής αποτελέσματα:

Correlations

			Study	Grades
Spearman's rho	Study	Correlation Coefficient	1.000	.985**
		Sig. (2-tailed)	.	.000
		N	10	10
	Grades	Correlation Coefficient	.985**	1.000
		Sig. (2-tailed)	.000	.
		N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

Όπως φαίνεται στην Εικόνα 19, η συνάφεια μεταξύ των δύο μεταβλητών ήταν επίσης πολύ δυνατή r = .985 και στατιστικά σημαντική (p < .001). Δεν περιμένουμε να δούμε πολύ μεγάλες διαφορές μεταξύ των 2 δεικτών, ειδικά αν η καταπάτηση των προϋποθέσεων δεν έχει γίνει σε μεγάλο βαθμό. Στην παρούσα περίπτωση ο δείκτης Spearman είναι λίγο μεγαλύτερος από τον δείκτη Pearson, αν και συνήθως περιμένουμε πιο "συντηρητικά" ευρήματα από τη χρήση μη παραμετρικών δοκιμασιών.

Τέλος θα αναλύσουμε τους συντελεστές που απαρτίζουν την γραμμική παλινδρόμηση καθώς και τους συντελεστές συσχέτισης.

6.4 ΕΚΤΙΜΗΣΗ ΤΩΝ ΣΥΝΤΕΛΕΣΤΩΝ Α ΚΑΙ Β ΤΟΥ ΥΠΟΔΕΙΓΜΑΤΟΣ ΤΗΣ ΑΠΛΗΣ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.

Όταν αποδεχόμαστε ότι για το φαινόμενο που μελετάμε ισχύει το υπόδειγμα $Y_t = ax_t + b + e_t$ της απλής γραμμικής παλινδρόμησης οφείλουμε να εκτιμήσουμε τους άγνωστους συντελεστές α και β καθώς και τη διακύμανση s^2. Υποθέτουμε ότι έχουμε διαθέσιμες Τ παρατηρήσεις τις (x_t, y_t) και έστω $\hat{a}$ και $\hat{b}$ μια εκτίμηση των συντελεστών a και b. Συμβολίζουμε με $\hat{y}_t$ την αντίστοιχη εκτίμηση του υποδείγματος $\hat{y}_t = \hat{a}x_t + b$ δηλαδή την εκτίμηση της υπό συνθήκη μαθηματικής ελπίδας $E(y_t / x)$.

Γενικότερα όμως για κάθε παρατήρηση t, η τιμή Y_t που υπολογίζουμε από το υπόδειγμα που εκτιμήσαμε δεν συμπίπτει με την τιμή y_t που παρατηρήσαμε. Έτσι έχουμε μια πρώτη εκτίμηση e_t που είναι $\hat{e}_t = y_t - \hat{y}_t$. Αυτή η τιμή $\hat{e}_t$ διαφέρει από την πραγματική τιμή e_t. Ενώ το πραγματικό υπόδειγμα που υποθέτουμε ότι ισχύει είναι: $y_t = ax_t + b + e_t$. Το υπόδειγμα που εκτιμούμε με τη βοήθεια του υποδείγματος είναι : $y_t = \hat{a}x_t + b + \hat{e}_t = \hat{y}_t + \hat{e}_t$. Αρκεί λοιπόν να προσδιορίσουμε τους συντελεστές $\hat{a}$ και b τους οποίους συνήθως υπολογίζουμε με την βοήθεια του αθροίσματος των τετραγώνων των διαφορών.

Έστω ότι το νέφος των Τ σημείων του διαγράμματος διασποράς δυο μεταβλητών Χ και Υ σε ορθογώνιο σύστημα αξόνων είναι: $A_1(x_1, y_1), A_2(x_2, y_2), ..., A_T(x_T, y_T)$. Τα σημεία δηλαδή

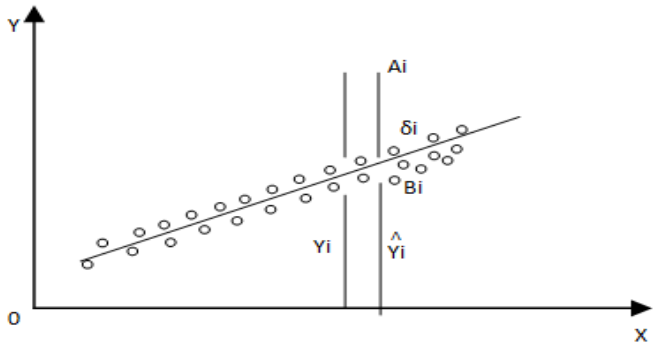
$A_t(x_t, y_t) : t = 1, 2, ...T$ αντιστοιχούν στις πραγματικές τιμές των μεταβλητών X και Y .Έστω επίσης ότι η θέση των σημείων αυτών είναι τέτοια που περνά ΄΄όσο γίνεται πιο κοντά΄΄ από αυτά. Δηλαδή στην περίπτωση αυτή προσπαθούμε να προσδιορίσουμε τους συντελεστές a και b της συναρτησιακής σχέσης $y = ax + b$ οι οποίοι θα προσδιορίσουν ακριβώς τη ζητούμενη ευθεία (Ε).Αν $A_t(x_t, y_t)$ είναι ένα τυχαίο σημείο του διαγράμματος διασποράς και $B_t(x_t, \hat{y}_t)$ το σημείο της ευθείας αυτής που έχει τετμημένη $x = x_t$ τότε η απόκλιση του B_t ανήκει στην (Ε) έχουμε $\hat{y}_t = \hat{a}x_t + b$ και $e_t = y_t - \hat{a}x_t - b$.Το άθροισμα των τετραγώνων των αποστάσεων των σημείων A_t από την ευθεία (Ε) είναι ίσο με το άθροισμα των τετραγώνων των αποκλίσεων των B_t από τα A_t.

Δηλαδή έχουμε: $\Delta = d_1^{\,2} + d_2^{\,2} + ... + d_t^{\,2}$ ή

$\Delta = (y_1 - \hat{a}x_1 - b)^2 + (y_2 - \hat{a}x_2 - b)^2 + ... + (y_t - \hat{a}x_t - b)^2$

$\underset{(1)}{\Delta} = \sum_{t=1}^{T} (y_t - \hat{a}x_t - \hat{b})^2$. Η παράσταση $\Delta(1)$ θα παίρνει την τιμή της για εκείνες τις τιμές των $\hat{a}$ και b που οι μερικές της παράγωγοι ως προς $\hat{a}$ και b είναι ίσες με το μηδέν.

Δηλαδή όταν $\frac{\partial_{\Delta}}{\partial_a} = 0$ ή $\frac{\partial \sum_{t=1}^{T} (y_t - \hat{a}x_t - \hat{b})^2}{\partial_a} = 0$ (1.1)



$\frac{\partial_{\Delta}}{\partial_b} = 0$ ή $\frac{\partial \sum_{t=1}^{T} (y_t - \hat{a}x_t - \hat{b})^2}{\partial_b} = 0$ (1.2)

Στη συνέχεια θα δούμε τους τρόπους με τους οποίους μπορούμε από τις δύο σχέσεις (1.1), (1.2) να προσδιορίσουμε τους συντελεστές a και

b στην περίπτωση που έχουμε απλά δεδομένα καθώς και όταν τα δεδομένα είναι ταξινομημένα σε πίνακα διπλής εισόδου.

Το πρόβλημα της στοχαστικής εξάρτησης εντοπίζεται στην εύρεση μιας καμπύλης η οποία διέρχεται πολύ κοντά από ορισμένα σημεία.

$y = j\,(x)$:η εξίσωση μιας καμπύλης η οποία να διέρχεται πολύ κοντά από ρ ορισμένα σημεία $(K_1 / X_{1,}Y_1),(K_2 / X_2, Y_1),.....,(K_P / X_P, Y_{1/P})$.Για να έχει νόημα η έκφραση <<πολύ κοντά>> θα πρέπει να βρούμε κάποιο μέτρο το οποίο εκφράζει την απόσταση των ρ σημείων από οποιαδήποτε καμπύλη του επιπέδου.

$\Lambda_1, \Lambda_2, ..., \Lambda_P$ τα σημεία οποιασδήποτε καμπύλης (γ) που έχουν τις ίδιες τετμημένες $x_1, x_2, ..., x_r$ παίρνουμε

$A = (K_1\Lambda_1)^2 + (K_2\Lambda_2)^2 + ... + (K_P\Lambda_P)^2$.Μια καμπύλη θα θεωρείται τόσο πιο κοντά στα σημεία $K_1, K_2, ..., K_r$ όσο πιο μικρό είναι το άθροισμα A δηλαδή $(K_1\Lambda_1)^2 + (K_2\Lambda_2)^2 + ... + (K_P\Lambda_P)^2 =$ ελάχιστο.

Παρατηρούμε ότι υπάρχουν γενικά πολλές καμπύλες με διαφορετικά σχήματα που διέρχονται κοντά από τα ρ αυτά σημεία. Γι' αυτό ακριβώς προσδιορίζουμε από την αρχή ανάλογα με τη θέση που έχουν τα ρ σημεία το είδος της καμπύλης που θα τοποθετήσουμε ανάμεσα τους. Δηλαδή παίρνουμε αυθαίρετα τη μορφή εξίσωσης $y = j\,(x)$ και μετά προσδιορίζουμε τα διάφορα σημεία της με τη βοήθεια των συντεταγμένων των δεδομένων σημείων. Με τις προϋποθέσεις αυτές το παραπάνω γενικό πρόβλημα λύνεται με μια μέθοδο που λέγεται **μέθοδος ελάχιστων τετραγώνων** και η καμπύλη που βρίσκουμε **καμπύλη ελάχιστων τετραγώνων.**

6.5 ΕΠΑΓΩΓΙΚΗ ΑΞΙΟΛΟΓΗΣΗ ΤΩΝ ΣΥΝΤΕΛΕΣΤΩΝΑ ΚΑΙ Β ΓΡΑΜΜΙΚΗΣ ΤΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Ο προσδιορισμός των συντελεστών της ευθείας των ελάχιστων τετραγώνων για τα δειγματικά δεδομένα

$\hat{y} = \hat{a} + \hat{b}c$

αποτελεί το πρώτο βήμα για την εκτίμηση των συντελεστών της πληθυσμιακής ευθείας της παλινδρόμησης η οποία ορίζεται από την εξίσωση $m_{y/x} = a + bc$.

Ο δειγματικός συντελεστής $\hat{a}$ της ευθείας των ελάχιστων τετραγώνων αποδεικνύεται ότι αποτελεί εκτίμηση του πληθυσμιακού συντελεστή a , ενώ ο συντελεστής $\hat{b}$ αποτελεί εκτίμηση του πληθυσμιακού συντελεστή b . Αν υποθέσουμε ότι λαμβάνονται όλα τα δυνατά δείγματα μεγέθους n από τον αντίστοιχο πληθυσμό και για κάθε ένα από αυτά υπολογίζεται η ευθεία των ελάχιστων τετραγώνων τότε οι εκτιμήσεις των συντελεστών a και b που προκύπτουν από κάθε δείγμα ακολουθούν την κανονική κατανομή με μέσες τιμές τις πληθυσμιακές παραμέτρους a και b αντίστοιχα. Οι τυπικές αποκλίσεις αυτών των κατανομών – δηλαδή τα τυπικά σφάλματα των εκτιμήσεων $\hat{a}$ και $\hat{b}$ - είναι ίσες αντίστοιχα με

$$se(\hat{b}) = \frac{s_{y/x}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}} \text{ και } se(\hat{a}) = s_{y/x}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Τα τυπικά σφάλματα των συντελεστών $\hat{a}$ και $\hat{b}$ εξαρτώνται και τα δύο από την ποσότητα $s_{y/x}$, δηλαδή την κοινή τυπική απόκλιση των υπό-πληθυσμών της Y που ορίζονται για τις διάφορες τιμές της X . Στην πράξη αυτή η ποσότητα είναι κατά κανόνα άγνωστη και εκτιμάται από την αντίστοιχη δειγματική τυπική απόκλιση $s_{y/x}$, όπου

$$s_{y/x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} .$$

Στον υπολογισμό της δειγματικής τυπικής απόκλισης $s_{y/x}$ υπεισέρχεται το άθροισμα των τετραγώνων των αποκλίσεων των δειγματικών τιμών y_i από τις εκτιμώμενες από ην ευθεία των ελάχιστων τετραγώνων τιμών $\hat{y}_i$, δηλαδή το άθροισμα των τετραγώνων των σφαλμάτων. Το άθροισμα αυτό όπως ήδη αναφέρθηκε είναι η ποσότητα που ελαχιστοποιείται κατά τον υπολογισμό της ευθείας των ελάχιστων τετραγώνων. Η ποσότητα $s_{y/x}$, η οποία αποτελεί εκτίμηση της $s_{y/x}$ ονομάζεται τυπική απόκλιση της παλινδρόμησης (*standard deviation from regression)*. Κατά τον υπολογισμό της ευθείας των ελάχιστων τετραγώνων που προσδιορίζει τη σχέση της ενεργειακής πρόληψης με την ηλικία , η τυπική απόκλιση της παλινδρόμησης είναι ίση με

$$s_{y/x} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}} = 544,32 .$$

Η παραπάνω τιμή μπορεί να χρησιμοποιηθεί ως σημειακή εκτίμηση της $s_{y/x}$ κατά τον υπολογισμό του τυπικού σφάλματος του συντελεστή $\hat{b}$

$$s\hat{e}(\hat{b}) = \frac{s_{y/x}}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = 6,59$$ και του τυπικού σφάλματος του συντελεστή $\hat{a}$

$$s\hat{e}(\hat{a}) = s_{y/x}\sqrt{\frac{1}{n} + \frac{\overline{c}^2}{\sum_{i=1}^{n}(c_i - \overline{c})^2}} = 359,12.$$

Η κλίση της ευθείας παλινδρόμησης είναι ο πιο σημαντικός συντελεστής υποδείγματος διότι προσδιορίζει τη μέση μεταβολή της εξαρτημένης μεταβλητής Y για κάθε μονάδα αύξησης της X . Βασιζόμενοι στην κλίση $\hat{b}$ της ευθείας των ελάχιστων τετραγώνων , μπορούμε να ελέγξουμε την τιμή κλίσης b της πληθυσμιακής ευθείας της παλινδρόμησης ως προς μια καθορισμένη αριθμητική τιμή b_0 . Δηλαδή να ελέγξουμε τη μηδενική υπόθεση

$$H_0 : b = b_0$$

Έναντι της εναλλακτικής

$$H_A : b \neq b_0 .$$

Ο παραπάνω έλεγχος καταλήγει στον προσδιορισμό της πιθανότητας να πάρουμε μια τιμή για δειγματικό συντελεστή $\hat{b}$ τόσο ακραία όσο και η τιμή b_0 , υπό την προϋπόθεση ότι ισχύει η μηδενική υπόθεση (δηλαδή ότι η δειγματοληπτική κατανομή του συντελεστή $\hat{b}$ έχει μέση τιμή τη b_0). Ο έλεγχος γίνεται με τη βοήθεια της ποσότητας

$$t = \frac{\hat{b} - b_0}{\hat{s}e(\hat{b})} .$$

Η οποία εφόσον ακολουθεί την κατανομή t με $n-2$ βαθμούς ελευθερίας. Μπορούμε να υπολογίσουμε την πιθανότητα p του ελέγχου και να τη συγκρίνουμε με το προκαθορισμένο επίπεδο σημαντικότητας

προκειμένου να απορρίψουμε ή να μην απορρίψουμε τη μηδενική υπόθεση H_0. Συνήθως ο έλεγχος του πληθυσμιακού συντελεστή b γίνεται ως προς την τιμή 0 τότε $m_{y/x} = a + 0_c = a$

Και επομένως η μέση τιμή κάθε υπό-πληθυσμού της Y ανεξαρτήτως της τιμής της X στην οποία αντιστοιχεί , ισούται με το συντελεστή a της παλινδρόμησης. Δηλαδή δεν υπάρχει γραμμική σχέση μεταξύ της X και Y . Για παράδειγμα της εξάρτησης των τιμών της ενεργειακής πρόσληψης από την ηλικία αυτό θα σήμαινε ότι η μέση ενεργειακή πρόσληψη των ατόμων όλων των ηλικιών είναι η ιδία.

Ο έλεγχος του συντελεστή της παλινδρόμησης b ως προς την τιμή 0 είναι ισοδύναμος με τον έλεγχο του πληθυσμιακού συντελεστή συσχέτισης r με την τιμή 0. Και αυτό διότι αποδεικνύεται εύκολα ότι ο δειγματικός συντελεστής $\hat{b}$ συνδέεται με το δειγματικό συντελεστή συσχέτισης r δια μέσου της σχέσης $\hat{b} = r(\frac{s_y}{s_x})$

Όπου s_x και s_y είναι οι τυπικές αποκλίσεις των δειγματικών τιμών x_i και y_i αντίστοιχα.

Προκειμένου να πραγματοποιήσουμε έναν αμφίπλευρο έλεγχο ως προς τη τιμή 0, για το συντελεστή b της παλινδρόμησης των τιμών της ενεργειακής πρόσληψης επί της ηλικίας , υπολογίζουμε την ποσότητα $t = \frac{\hat{b} - b_0}{\hat{se}(\hat{b})} = \frac{-27 - 0}{6,59} = -4,10$.

Για μια κατανομή t με 40-2=38 βαθμούς ελευθερίας η πιθανότητα να προκύψει μια τιμή τόσο ακραία όσο η -4,10 ή η 4,10 είναι $p\langle 2(0,0005) = 0,0001$ και επομένως η μηδενική υπόθεση ότι ο συντελεστής b είναι ίσος με μηδέν απορρίπτεται. Δηλαδή στον πληθυσμό των ενηλίκων από τον οποίο προέρχεται το τυχαίο δείγμα των 40 ατόμων, υπάρχει σημαντική σχέση μεταξύ της ημερήσιας ενεργειακής πρόσληψης και της ηλικίας .Συμφώνα με τη σχέση αυτή η ημερήσια ενεργειακή πρόληψη ελαττώνεται όσο η ηλικία.

ΚΕΦΑΛΑΙΟ 7

Για να παρουσιάσουμε την θεωρητική ανάλυση σε πρακτικό επίπεδο θα πρέπει να δώσουμε κάποια παραδείγματα .Όπως είπαμε παραπάνω στην απλή γραμμική παλινδρόμηση χρησιμοποιούμε μονο μια μεταβλήτη χ και μια δεύτερη μεταβλητή Υ η οποία μπορεί να προσεγγίσει απο μια συνάρτηση του χ π.χ η Υ να εκφράζεται μέσω της χ Υ= 3χ+5.

Χ:ανεξάρτητη μεταβλητή (independent variable)

Υ:εξαρτημένη μεταβλητή (dependent variable)

Στην ανάλυση απλής παλινδρόμησης υπάρχει μόνο μία ανεξάρτητη μεταβλητή για αυτό και καλείται απλή παλινδρόμηση ενώ αν υπάρχουν περισσότερες απο μια ανεξάρτητες μεταβλητές λέγεται πολλαπλή παλινδρόμηση.Ένα απλό παράδειγμα απλής γραμμικής παλινδρόμησης είναι η εύρεση σχέση μεταξύ της συνολικής παραγωγής ενός αγρού και της ποσότητας λιπάσματος που χρησιμοποιήθηκε.Σε αντίθεση με την πολλαπλή παλινδρόμηση στην ανάλυση της οποίας η εύρεση σχέσης μεταξύ της συνολικής παραγωγής ενός αγρού και της ποσότητας λιπάσματος που χρησιμοποιήθηκε ,της υγρασίας και θερμοκρασιάς της περιοχής.

Η ανάλυση γραμμικής παλινδρόμησης μπορεί να χρησιμοποιηθεί σε πολλούς τομείς όπως οι κλινικές μελέτες, στην ιατρική , στην εκπαίδευση, στην τεχνολογία , στη γεωργία , στις παραγωγικές εργοστασιακές μονάδες.Για παράδειγμα στις κλινικές μελέτες μπορεί να χρησιμοποιηθεί η ανάλυση της απλής γραμμικής παλινδρόμησης απο τους ερευνητές όσον αφορά στον καθορισμό απο πρίν στις δόσεις ενός φαρμάκου (ανεξάρτητη μεταβλητή) που δίνει στους ασθενείς και στην συνέχεια μέτρα τις αντιδράσεις τους στο φάρμακο (εξαρτημένη μεταβλητή).Με την παλινδρόμηση προσδιορίζεται η σχέση δόσης-αντίδρασης για το συγκεκριμένο φάρμακο δηλαδη για δεδομένη δόση να προβλέπει την αντιδράση.Πολλές φορές ο διαχωρισμός της ανεξάρτητης και εξαρτημένης είναι αρκέτα δύσκολη ορισμένες φορές.

Για παράδειγμα σε ένα δείγμα 20 μαθητών μετράμε το βάρος και το ύψος τους.Αν αυτό που μας ενδιαφέρει είναι το ΄΄τι συμβαίνει με το βάρος των παιδίων όταν αλλάζει το ύψος τους΄΄, τότε θεωρούμε ως ανεξάρτητη μεταβλητή χ το ύψος και ως εξαρτημένη μεταβλητή Υ το βάρος.Άρα

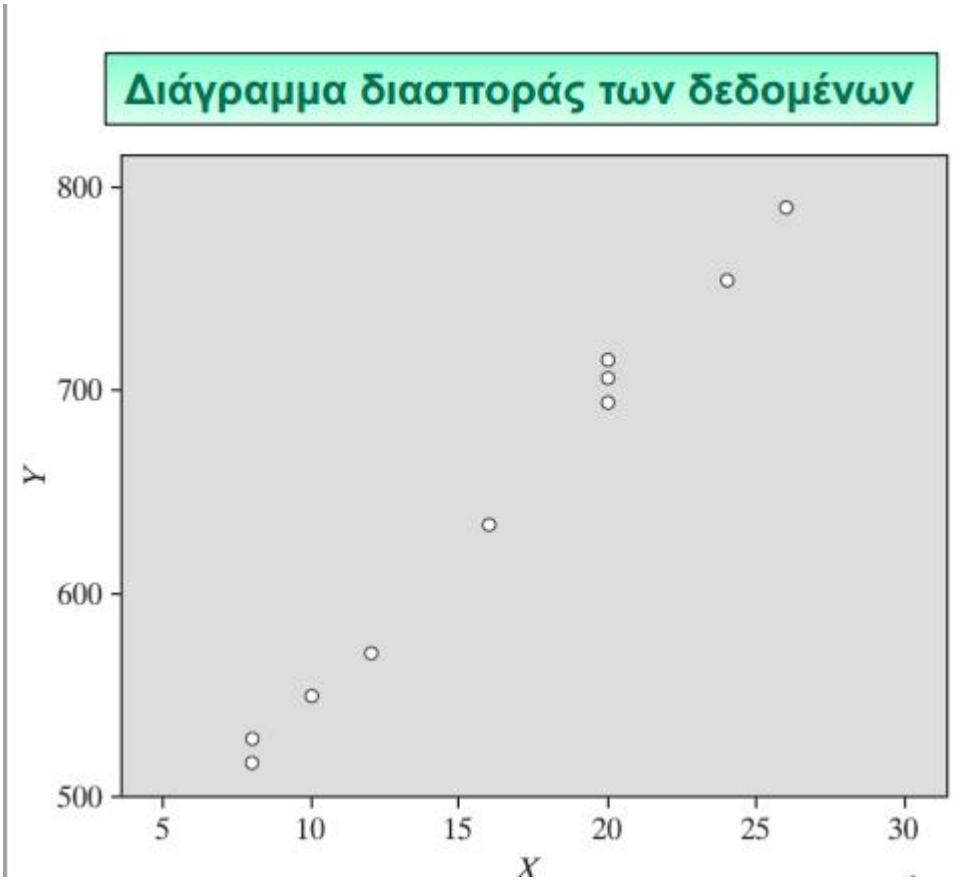
ενδιαφερόμαστε για την παλινδρόμηση τους βάρους (Υ0 πάνω στο ύψος (Χ).

Αντίθετα αν μας ενδιαφέρει το ΄΄τι συμβαίνει με το ύψος των παιδιών όταν αλλάζει το βάρος τους΄΄ τότε θεωρούμε ανεξάρτητη Χ το βάρος και ως εξαρτημένη μεταβλητή Υ το ύψος τότε έχουμε παλινδρόμηση του ύψους (Υ) πάνω στο βάρος (Χ).Όπως θα δούμε και στο παρακάτω παράδειγμα για την εύρεση του κατάλληλου μοντέλου για την περιγραφή της σχέσης μεταξύ δύο μεταβλητών οι οποίες μας ενδιαφέρουν ξεκινάμε με την κατασκευή του διαγράμματος διασποράς (scatter plot) στο επίπεδο των παρατηρήσεων που διαθέτουμε. Όπως έχουμε δει και στην παραπάνω θεωρητική ανάλυση που έχουμε κάνει σε αυτύ του είδους το διάγραμμα οι τιμές της μεταβλητής Χ τοποθετούνται στον οριζοντιο άξονα και της μεταβλητής Υ στον κατακόρυφο άξονα.

7.1 ΄ΕΝΑ ΑΠΛΟ ΠΑΡΑΔΕΙΓΜΑ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.

Ένας αγρότης ενδιαφέρεται να προσδιορίσει τον τρόπο με τον οποίο η ποσότητα *X* του λιπάσματος που χρησιμοποιείται σε ένα αγροτεμάχιο επηρεάζει την παραγωγή *Y* του αγροκτήματος. Για το σκοπό αυτό πειραματίζεται με *ν*=10 όμοια αγροτεμάχια (ίδιου εμβαδού, σε περιοχές που επικρατούν παρόμοιες κλιματολογικές συνθήκες κλπ) έτσι ώστε οι όποιες διαφοροποιήσεις παρατηρούνται στην παραγωγή των αγρών να οφείλονται κατά κύριο λόγο στις διαφορετικές ποσότητες λιπάσματος που χρησιμοποιήθηκαν. Στο διπλανό πίνακα δίνεται η παραγωγή *Y* (σε χιλιάδες κιλά) για *ν*=10 όμοια αγροτεμάχια καθώς και η ποσότητα *X* του λιπάσματος που χρησιμοποιήθηκε στο καθένα (σε εκατοντάδες κιλά).

i	x_i	y_i
1	20	706
2	10	550
3	26	790
4	8	517
5	20	694
6	16	634
7	20	715
8	12	571
9	8	529
10	24	754



Πριν μπούμε στο πρακτικό κομματι της εργασίας θα πρέπει να δώσουμε τον ορισμό του κριτηρίου ελέγχου καθώς και των βαθμών ελευθερίας.Ας ξεκινήσουμε με το κριτήριο ελέγχου.

7.2 ΕΠΙΛΟΓΗ ΚΡΙΤΗΤΡΙΟΥ ΕΛΕΓΧΟΥ

Το κριτήριο ελέγχου παίζει σημαντικό ρόλο στην εργασία μας για αυτό και θα χρειαστεί να αναλυθεί.

- Χωρίζεται σε δύο μέρη τα οποία είναι 1. Η σύγκριση της ίδιας μεταβλητής σε δυο δείγματα και 2. Η σύγκριση της ίδιας μεταβλητής σε περισσότερα απο δυο δείγματα.Κάθε κατηγορία χωρίζεται δε ανεξάρτητα και εξαρτημένα δεδομένα.Για την $1^{η}$ κατηγορία όταν η μεταβλητή εινάι κατηγορική και τα δείγματα είναι ανεξάρτητα για να γίνει το Test x^2 ομοιογένειας θα πρέπει: α)ο έλεγχος να γίνεται σε δείγματα μεγέθους >30 ,β)οι θεωρητικές τιμές που υπολογίζονται στον πίνακα συνάφειας να είναι >1 και γ)το 80%να είναι > του 5.Όταν η μεταβλητή είναι κατηγορική και τα δείγματα είναι εξαρτημένα το test ελέγχει να είναι στατιστικά σημαντικές οι αλλαγές που επήλθαν απο την επίδραση

κάποιου ξένου παράγοντα.Οι προυποθέσεις που θα πρέπει να υπάρχουν είναι:

- α)οι μεταβλητές να είναι δίτιμες (dichotomous) και β)οι μεταβλητές να έχουν τις ίδιες τιμές (value labels).

  Όταν η μεταβλητή είναι ποσοτική και τα δείγματα είναι ανεξάρτητα έχουμε T-test.Οι προυποθέσεις που πρέπει να ισχύουν για να γίνει αυτός ο έλεγχος είναι α)τα δύο δείγματα θα πρέπει να έχουν κανονικές κατανομές .Αυτό δείχνει ότι θα πρέπει να έχει προηγηθεί έλεγχος κανονικότητας.Αν δεν ακολουθούν κανονική κατανομή προτείνεται η χρήση άλλου ελέγχου.Ανάλογα με το εανν οι διασπορές των δυο δειγμάτων είναι ίσες ή οχι επιλέγεται και ειδική εκδοχή του t-test .Ό έλεγχος ισότητας των διασπορών γίνεται με τη χρήση της F-κατανομής (F-test).

  Όταν η μεταβλητή είναι ποσοτική καιτα δείγματα είναι εξαρτημένα έχουμε T-test ζευγαρωτών παρατηρήσεων.Οι προυποθέσεις που πρέπει να ισχύουν είναι οι εξής:α)τα δύο δείγματα θα πρέπει να έχουν κανονικές κατανομές.Αυτό σημαίνει ότι θα πρέπει να προηγηθεί έλεγχος κανονικότητας αλλίως ακολουθείται κάποιος άλλος έλεγχος.

  Στην δεύτερη κατηγορία όταν η μεταβλητή είναι κατηγορική και τα δείγματα ανεξάρτητα έχουμε τον Πίνακα Συνάφειας.Ο πίνακας έχει η γραμμές (όπου n= ο αριθμός των δειγμάτων) και κ στήλες (όπου κ = οι τιμές της μεταβλητής).Μπορεί να ισχύει και το αντιστροφο. x^2 test ομοιογένειας.Προσοχή το x^2 εμφανίζεται στις συχνότητες και όχι στα ποσοστά.Οι προυποθέσεις που πρέπει να ισχύουν είναι α) σε όλα τα κελία του πίνακα συνάφειας να έχουμε θεωρητικές συχνότητες >1 και για το 80%>5.

  Όταν η μεταβλητή είναι κατηγορική και τα δείγματα εξαρτημένα ο Πίνακας Συναφειας έχει n γραμμές (όπου n= ο αριθμός περιπτώσεων) και κ στήλες (όπου κ= ο αριθμός των δειγμάτων).Προυπόθεση είναι οι μεταβλητές να είναι δίτιμες(dichotomous).

- Το κριτήριο χρησιμοποιείται και στην Ανάλυση Διασποράς ANOVA και πρέπει να ισχύουν οι προυποθέσεις α)η μεταβλητή των παρατηρήσεων θα πρέπει να ακολουθεί κανονική κατανομή ,β)η μεταβλητή παρουσιάζει την ίδια διασπορά σε όλους τους πληθυσμούς.Προσοχή πρέπει να δοθεί στον υπολογισμό και την εκτίμηση των υπολοίπων.Όταν η μεταβλητή είναι συνεχής και τα δείγματα εξαρτημένα σε γενικές γραμμές υιοθετείται η Ανάλυση Διασποράς.

- Σύγκριση δυο μεταβλητών στο ίδιο δείγμα.

Όταν η εξαρτημένη μεταβλητή είναι συνεχής και η ανέξαρτητη επίσης συνεχής έχουμε γραμμική παλινδρόμηση και υπολογισμό του συντελεστή γραμμικής συσχέτισης pearson.Προυπόθεση είναι ότι η κατανομή των παρατηρήσεων θα πρέπει να ακολουθεί την κανονική κατανομή.

Όταν η εξαρτημένη μεταβλητή είναι συνεχής και η ανεξάρτητη κατηγορική έχουμε Ανάλυση Διασποράς με έναν παράγοντα και υπολογισμό του συντελεστή του.

Παρατήση:Ανάλογα με την κατανομή της εξαρτημένης μεταβλητής χρησιμοποιείται παραμετρικό ή μη παραμετρικό test.

Όταν η εξαρτημένη μεταβλητή είναι διατακτική και η ανεξάρτητη

είναι συνεχής.

Μη παραμετρική παλινδρόμηση και υπολογισμός του συντελεστή

Spearman.

Όταν η εξαρτημένη μεταβλητή είναι διατακτική και η ανεξάρτητη

είναι επίσης διατακτική.

Μη παραμετρική παλινδρόμηση και υπολογισμός του συντελεστή

Spearman.

Όταν η εξαρτημένη μεταβλητή είναι διατακτική και η ανεξάρτητη

είναι κατηγορική.

x^2 test.

Όταν η εξαρτημένη μεταβλητή είναι κατηγορική και η ανεξάρτητη

είναι συνεχής.

x^2 test.

Όταν η εξαρτημένη μεταβλητή είναι κατηγορική και η ανεξάρτητη

είναι διατακτική.

x^2 test.

Όταν η εξαρτημένη μεταβλητή είναι κατηγορική και η ανεξάρτητη

είναι επίσης κατηγορική.

x^2 test.

Σε αυτό το σημείο θα πρέπει να αναλύσουμε και το δείκτη Durbin-Watson τον οποίο θα χρησιμοποιήσουμε και παρακάτω στο πρακτικό κομματί της εργασίας αυτής.

7.3 ΔΕΙΚΤΗΣ DURBIN-WATSON

Είναι ένα στατιστικό τέστ που χρησιμοποιείται για να δούμε αν υπάρχει αυτοσυσχέτιση (μια σχέση μεταξύ των τιμών χωρίζονται μεταξύ τους απο μια δεδομένη χρονική υστέρηση) στα κατάλοιπα απο μια ανάλυση παλινδρόμησης.Αυτό το στατιστικό στοιχείο εφαρμόζεται στα κατάλοιπα των ελάχιστων τετραγώνων παλινδρόμησης.

Αν e_t είναι η υπολειμματική συνδέονται με την παρατήρηση σε χρόνο t , τότε το στατιστικό αποτέλεσμα της δοκιμής είναι

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2},$$

όπου T είναι ο αριθμός των παρατηρήσεων. Επειδή d είναι περίπου ίση με 2 (1 - R), όπου *το* R είναι η αυτοσυσχέτιση δείγμα των καταλοίπων, [1] d = 2, δείχνει ότι δεν υπάρχει αυτοσυσχέτιση.Η τιμή του D βρίσκεται πάντα μεταξύ 0 και 4. Αν ο Durbin-Watson στατιστική είναι ουσιωδώς μικρότερο από 2, υπάρχουν ενδείξεις θετικής σειριακής συσχέτισης. Ως πρόχειρη κανόνα του αντίχειρα, εάν Durbin-Watson είναι μικρότερη από 1,0, μπορεί να υπάρχει λόγος ανησυχίας. Μικρές τιμές του d δείχνουν διαδοχικές όρους σφάλματος είναι, κατά μέσο όρο, σε αξία κοντά το ένα στο άλλο, ή συσχετίζεται θετικά. Εάν D είναι> 2, οι διαδοχικές όρων σφάλματος, κατά μέσο όρο, πολύ διαφορετική αξία σε ένα από το άλλο, δηλ., αρνητικώς συσχετίζονται. Στις παλινδρομήσεις, αυτό μπορεί να συνεπάγεται μια υποτίμηση του επιπέδου στατιστικής σημασίας .

Για να ελέγξετε για **θετική αυτοσυσχέτιση** σε σημασία α , η στατιστική δοκιμή d συγκρίνεται με άνω και κάτω κρίσιμες τιμές ($\delta_{L,\alpha}$ και $\delta_{U,\alpha}$):

- Αν $d < d_{L,\alpha}$, υπάρχουν στατιστικά στοιχεία που δείχνουν ότι οι όροι σφάλματος θετική συσχέτιση,.
- Αν $d > d_{U,\alpha}$, δεν υπάρχει **καμία** στατιστικά στοιχεία που αποδεικνύουν ότι οι όροι σφάλματος θετική συσχέτιση,.
- Αν $d_{L,\alpha} < d < d_{U,\alpha}$, η δοκιμή είναι ασαφές.

Θετική σειριακή συσχέτιση είναι σειριακή συσχέτιση κατά την οποία ένα θετικό λάθος για μία παρατήρηση αυξάνει τις πιθανότητες ενός θετικού σφάλματος για μια άλλη παρατήρηση.

Για να ελέγξετε για την **αρνητική αυτοσυσχέτιση** σε σημασία α , το στατιστικό αποτέλεσμα της δοκιμής (4 - δ) η σύγκριση με το χαμηλότερο και το ανώτερο κρίσιμες τιμές ($\delta_{L,\alpha}$ και $\delta_{U,\alpha}$):

- Αν (4 - δ) < $d_{L,\alpha}$, υπάρχουν στατιστικά στοιχεία που δείχνουν ότι οι όροι σφάλματος αρνητική συσχέτιση,.
- Αν (4 - d)> $d_{U,\alpha}$, δεν υπάρχει **καμία** στατιστικά στοιχεία ότι οι όροι σφάλματος αρνητική συσχέτιση,.
- Αν $d_{L,\alpha}$ <(4 - δ) < $d_{U,\alpha}$, η δοκιμή είναι ασαφές.

Αρνητική σειριακή συσχέτιση υποδηλώνει ότι ένα θετικό λάθος για μία παρατήρηση αυξάνει την πιθανότητα ενός αρνητικού σφάλματος για μια άλλη παρατήρηση και ένα αρνητικό σφάλμα για μία παρατήρηση αυξάνει τις πιθανότητες ενός θετικού σφάλματος για ένα άλλο.

7.4 ΒΑΘΜΟΙ ΕΛΕΥΘΕΡΙΑΣ (Degrees of Freedom)

Η τιμή n-1 που απαιτείται για τον υπολογισμό της δειγματικής διακύμανσης αναφέρεται ως βαθμοί ελευθερίας. Η ονομασία οφείλεται στο ότι αν πρόκειται να διαλέξουμε n τιμές που θα πρέπει να έχουν ένα δεδομένο μέσο, ο αριθμός των τιμών που μπορούμε να διαλέξουμε ελεύθερα και αυθαίρετα είναι n-1. Αν ο μέσος των n τιμών είναι καθορισμένος τότε και το άθροισμά τους είναι καθορισμένο οπότε η τελευταία τιμή θα προκύπτει ως η διαφορά του αθροίσματος των n τιμών μείον το άθροισμα των n-1 τιμών που έχουν επιλεγεί αυθαίρετα. Δοθέντος ότι η δειγματική διακύμανση υπολογίζεται μέσω των αποκλίσεων των n τιμών των δεδομένων από τον μέσο τους, λέμε ότι έχει n-1 βαθμούς ελευθερίας.

7.5 ΧΡΗΣΗ ΤΟΥ SPSS

7.5.1 ΠΛΗΡΟΦΟΡΙΕΣ ΓΙΑ ΤΟ SPSS

Μερικές πληροφορίες για το spss πριν μπούμε στη διαδικασία της άσκησης.Η πρώτη έκδοση του spss(Statistical Package for the social Siences) εμφανίστηκε για πρώτη φορά στην αγορά του 1968 απο τους Norman H Nie και C.Hadlai Hull.Είναι το πιο διαδεδομένο πρόγραμμα στατιστικής ανάλυσης στην κοινωνία των επιστημών.Μεταξύ του 2009 και 2010 το πρώτο λογισμικό του spss ονομαζόταν PASW(Predictive Analytics SoftWare)Statistics. Το 2009 του Ιουλίου η εταιρεία ανακοίνωσε ότι αποκτήθηκε απο την IBM για 1,2 δισεκατομμύρια δολλάρια και απο τον Ιανουάριο του 2010 μετονομάστηκε σε SPSS: An IBM Company.

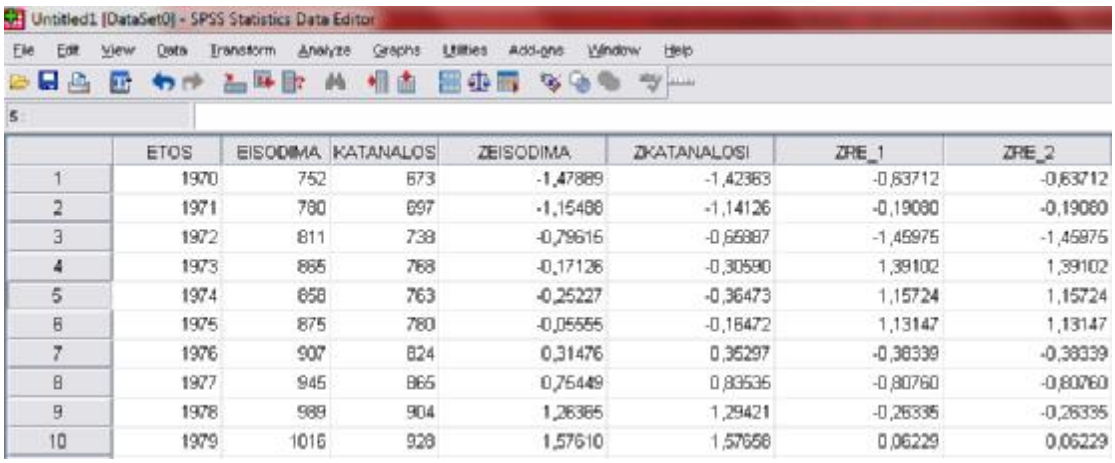
Με τη χρήση του spss δημιουργούμε γραφήματα και πίνακες που μας δείχνουν αν υπάρχει συσχέτιση μεταξύ της εξαρτημένης μεταβλητής και κάθε ανεξάρτητης μεταβλητής ξεχωριστά καθώς επίσης με τα δεδομένα που εξάγουμε απο το spss κάνουμε έλεγχο υποθέσεων για να δούμε αν υφίσταται παλινδρόμηση και αν όντως υφίσταται ορίζουμε την εξίσωση παλινδρόμησης και βρίσκουμε αν έχουμε ασθενή,μέτρια ή ισχυρή γραμμική συσχέτιση στο μοντέλο μας.

Θα παρουσιάσουμε ένα απλό παράδειγμα για να δούμε στη πράξη τις παραπάνω πληροφορίες που παραθέσαμε.Μεταβλητές είναι το Διαθέσιμο Ατομικό Εισόδημα και τα Ατομικά Έξοδα Κατανάλωσης.

7.5.2 ΑΣΚΗΣΗ ΜΕ SPSS

Τα δεδομένα που θα χρησιμοποιηθούν στην παρακάτω ανάλυση στο spss έχουν παρθεί απο τα λογιστικά αρχεία τοπικής εργοστασιακής μονάδας και μας δειχνούν την χρονολογία καθώς και το εισόδημα και επιπλεόν την κατανάλωση που έγινε στο εισόδημα συγκεκριμένο μήνα του χρόνου απο το ίδιο άτομο.Το εισόδημα αυξάνεται κατα χρονιά αφού αυξάνεται και η διατήρηση της θέσης.Αυτό που μας ενδιαφέρει είναι η συσχέτιση και η σχέση που υπάρχει ανάμεσα στο εισόδημα και την κατανάλωση.Ο μήνας που αναφέρεται το δείγμα ειναι ο Οκτώμβρης.

7.5.3 ΕΙΣΑΓΩΓΗ ΔΕΔΟΜΕΝΩΝ


Untitled1 [DataSet0] - SPSS Statistics Data Editor

File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help

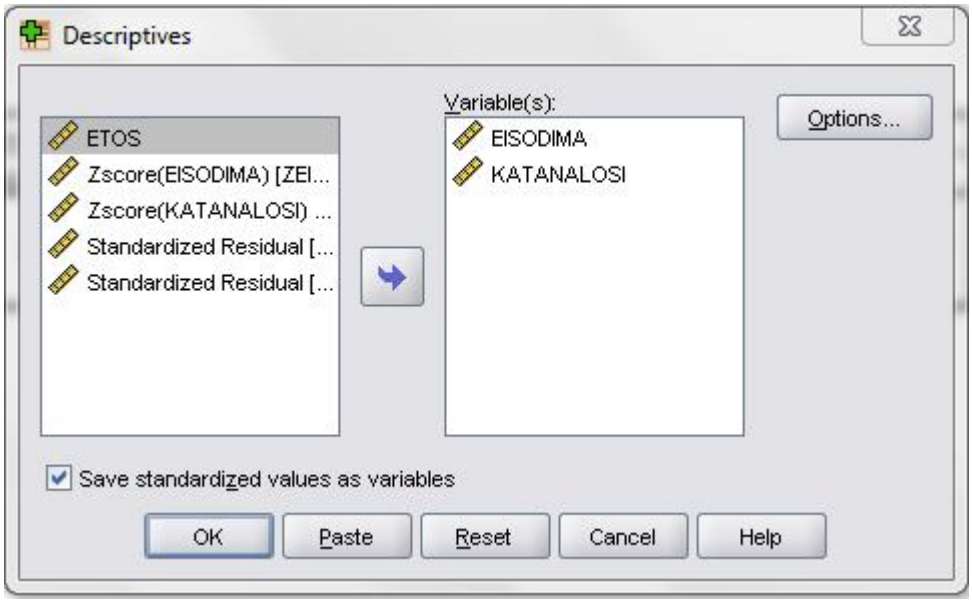
5

	ETOS	EISODIMA	KATANALOS	ZEISODIMA	ZKATANALOSI	ZRE_1	ZRE_2
1	1970	752	673	-1,47889	-1,42363	-0,63712	-0,63712
2	1971	780	697	-1,15488	-1,14126	-0,19080	-0,19080
3	1972	811	738	-0,79616	-0,66887	-1,45975	-1,45975
4	1973	865	768	-0,17126	-0,30590	1,39102	1,39102
5	1974	858	763	-0,25227	-0,36473	1,15724	1,15724
6	1975	875	780	-0,05555	-0,16472	1,13147	1,13147
7	1976	907	824	0,31476	0,35297	-0,38339	-0,38339
8	1977	945	865	0,75449	0,83535	-0,80760	-0,80760
9	1978	989	904	1,26365	1,29421	-0,26335	-0,26335
10	1979	1016	928	1,57610	1,57658	0,06229	0,06229

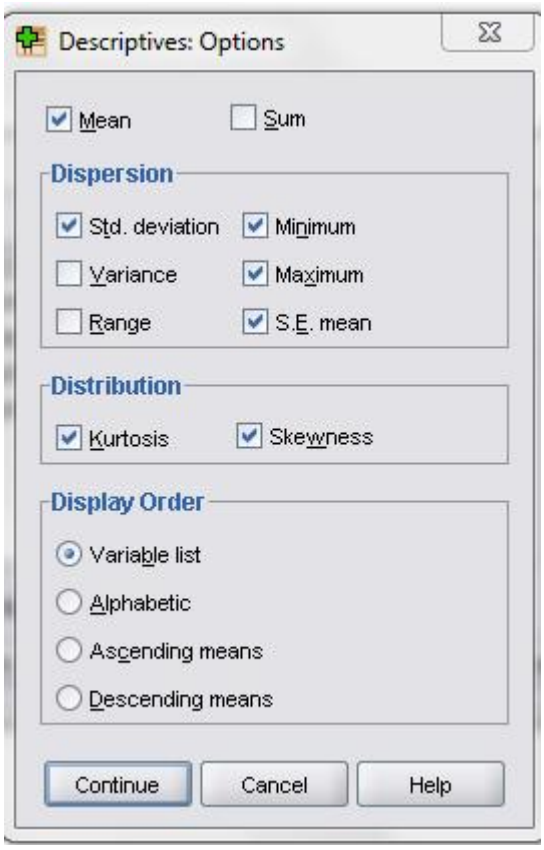
Για να ξεκινήσουμε την ανάλυση το πρώτο βήμα είναι το εξής:

Descriptives Statistics → Descriptives

Το παρακάτω πινακάκι μας δείχνει τη συνέχεια της διαδικασίας



Και συνεχίζουμε με



Continue και O.K

Μας δίνει το αποτέλεσμα

Descriptives

[DataSet0]

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic	Std. Error	Statistic	Std. Error
EISODIMA	10	752	1016	879,80	27,327	86,416	,152	,687	-,842	1,334
KATANALOSI	10	673	928	794,00	26,877	84,994	,274	,687	-,955	1,334
Valid N (listwise)	10									

Την ανάλυση μας θα την ξεκινήσουμε με τους *Ελέγχους Υποθέσεων* .

Θα δούμε τους στοιχειώδεις έλεγχους υποθέσεων που απαρτίζουν μια ανάλυση δεδομένων (συνεχών και κατηγορικών δεδομένων).Χρησιποποιείται για να ελέγξουμε το μέσο του δείγματος ως προς μια ισότητα.

$H_0 : m = c$
$H_1 : m \neq c$

όπου c είναι ένας σταθερός αριθμός.

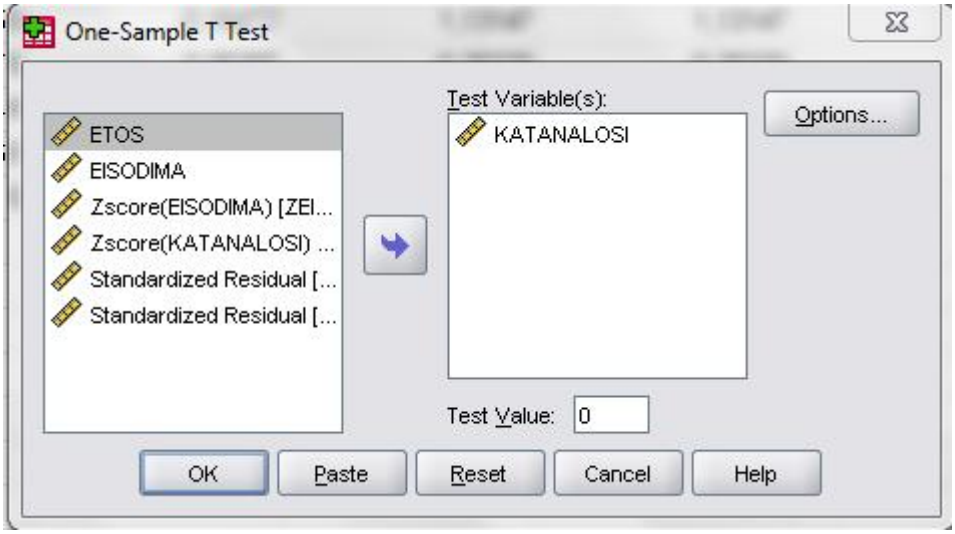
H_0 : ο μέσος ισούται με την σταθερά c

H_1 :ο μέσος είναι διαφορετικός απο τη σταθερά c .

Σε περίπτωση που το p-value<0,05 (sig<0,05) τότε απορρίπτουμε την H_0 .Προυπόθεση είναι ότι τα δεδομένα πρέπει να ακολουθούν την κανονική κατανομή.

Η διαδικασιά στο spss είναι η εξής:

Analyze $\rightarrow$ ***Compare Means*** $\rightarrow$ ***One sample T-test***



Βάζουμε στο T-test variable(s) για την οποία θέλουμε να ελέγξουμε αν ο μέσος ισούται με κάποια συγκεκριμένη τιμή.

One-Sample Test

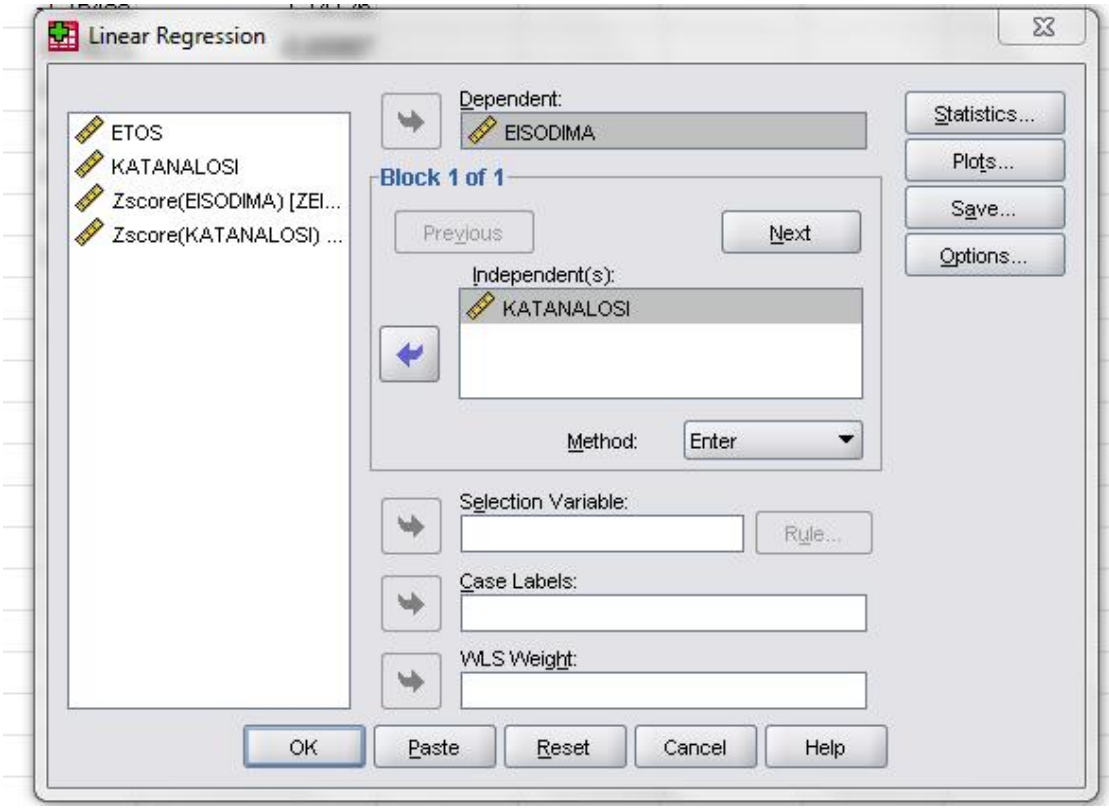
	Test Value = 0					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
KATANALOSI	29,541	9	,000	794,000	733,20	854,80

Από τον παραπάνω πίνακα βλέπουμε ότι εφόσον p-value=0,000<0,05 η μηδενική υπόθεση απορρίπτεται.Επομένως ο μέσος της υπό μελέτης μεταβλητής δεν ισούται με 0.

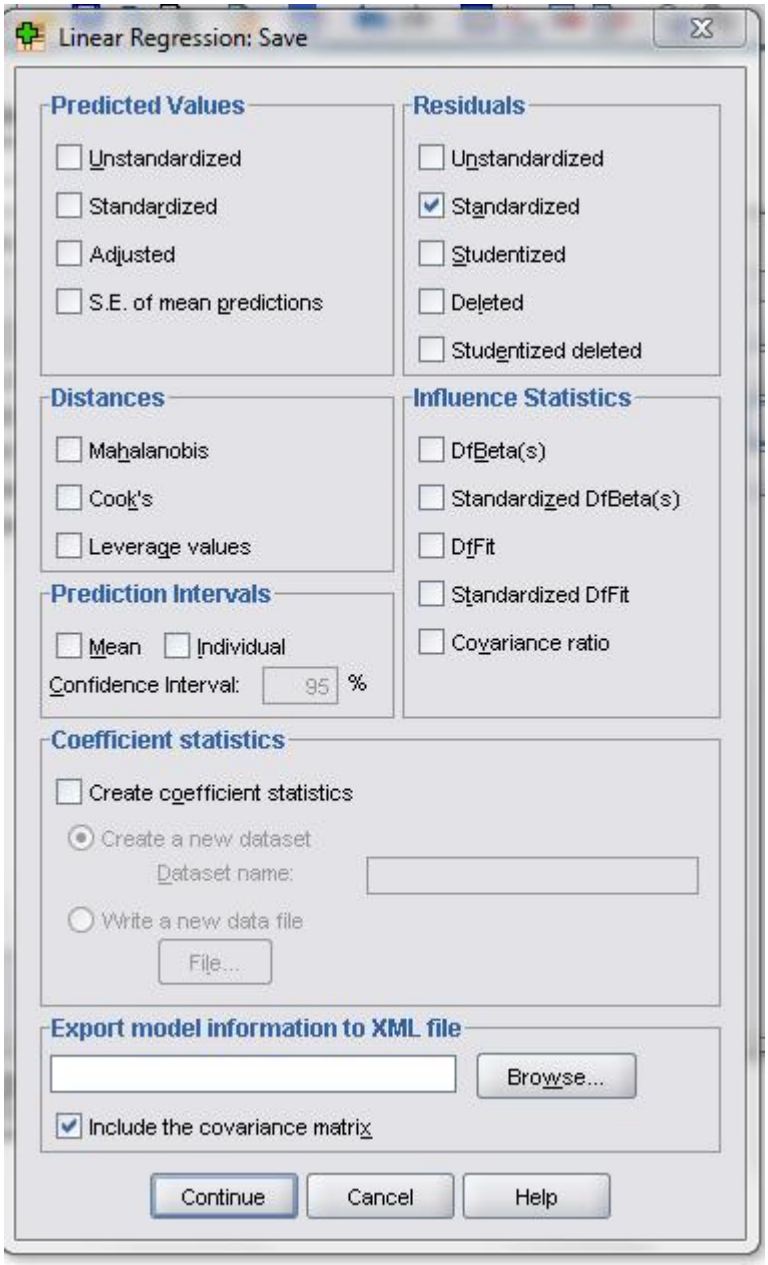
Αφού κάναμε τον έλεγχο t-test συνεχίζουμε με τον *Έλεγχο Κανονικότητας*.Μια απο τις προυποθέσεις που θα πρέπει να ισχύουν για να μπορέσουμε να εκτιμήσουμε σωστά ένα γραμμικό μοντέλο είναι η υπόθεση ότι τα κατάλοιπα ακολουθούν κανονική κατανομή με μέσο 0 και διακύμανση γνωστή.

Η διαδικασία στο spss είναι η ακόλουθη:

Analyze → Regression → Linear

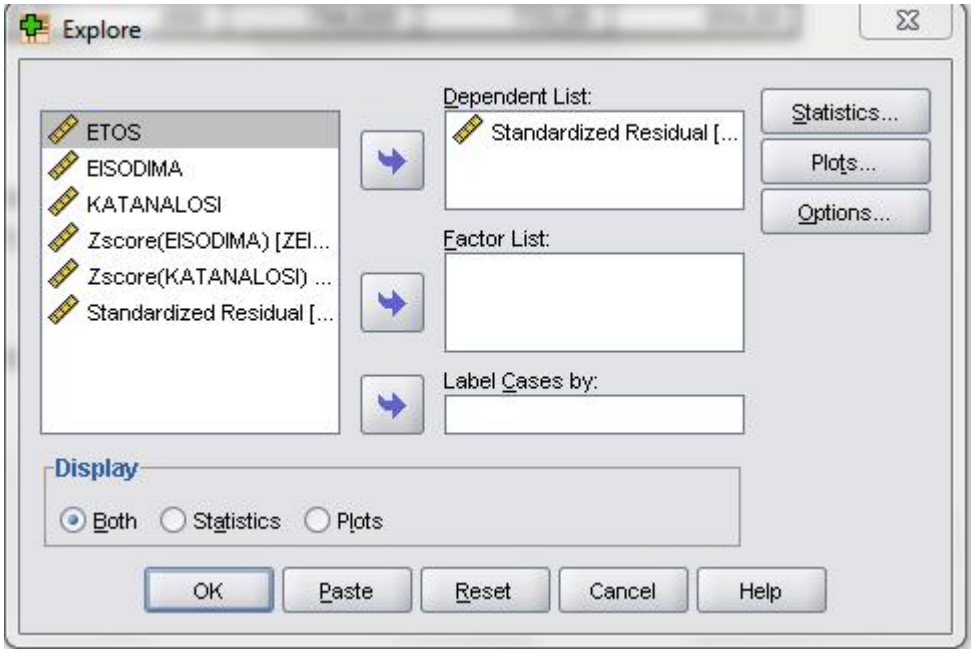


Συνεχίσουμε με την εντολή save και έχουμε τα εξής:

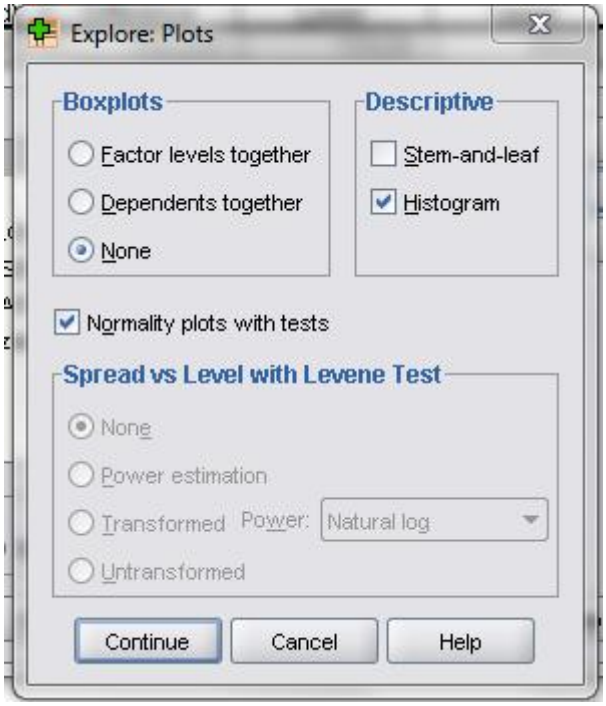


Πατάμε continue και μετά o.k

Συνεχίζουμε με ***Analyze*** → ***Descriptive Statistics*** → ***Explore***



Συνεχίζουμε με



Συνεχίζουμε με continue και o.k.

Αμέσως μετά βλέπουμε τον πίνακα Test of Normality

Tests of Normality

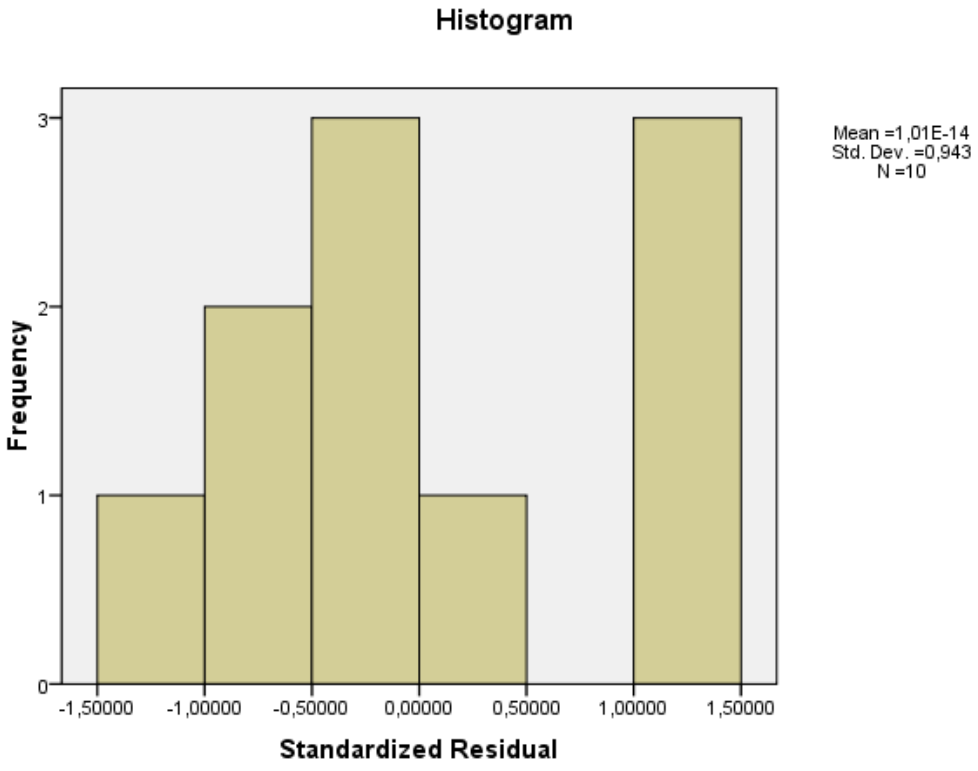
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Standardized Residual	,185	10	,200*	,924	10	,390

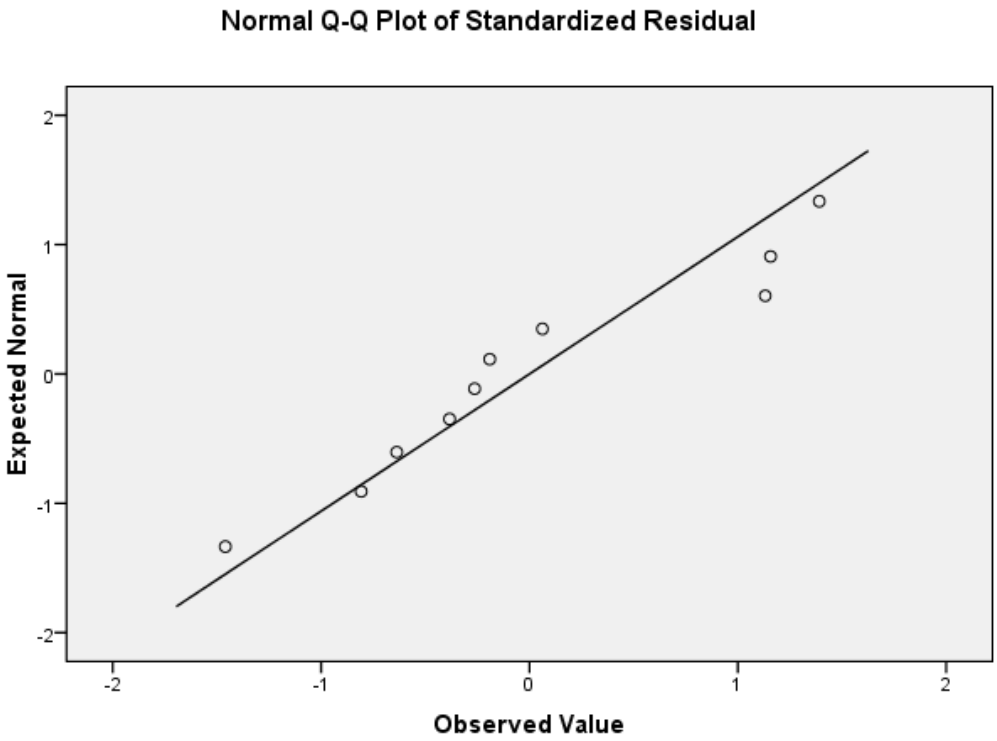
a. Lilliefors Significance Correction

*. This is a lower bound of the true significance.

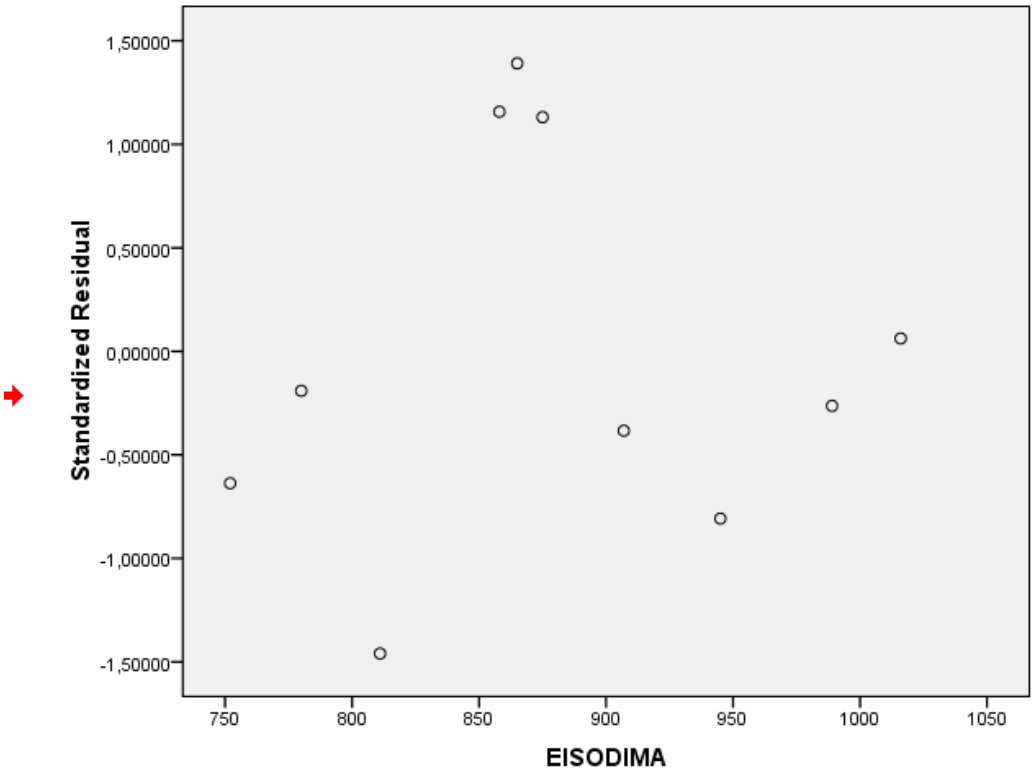
Η μηδενική απόφαση γίνεται αποδεκτή απο τους ελέγχους αφού p-value=0,20>0,05.Άρα ισχύει η υπόθεση κανονικότητας.

Σε αυτό το σημείο θα πρέπει να αναφέρουμε τον όρο *Ομοσκεδαστικότητα* δηλαδη έχουμε ίσες διακυμάνσεις .





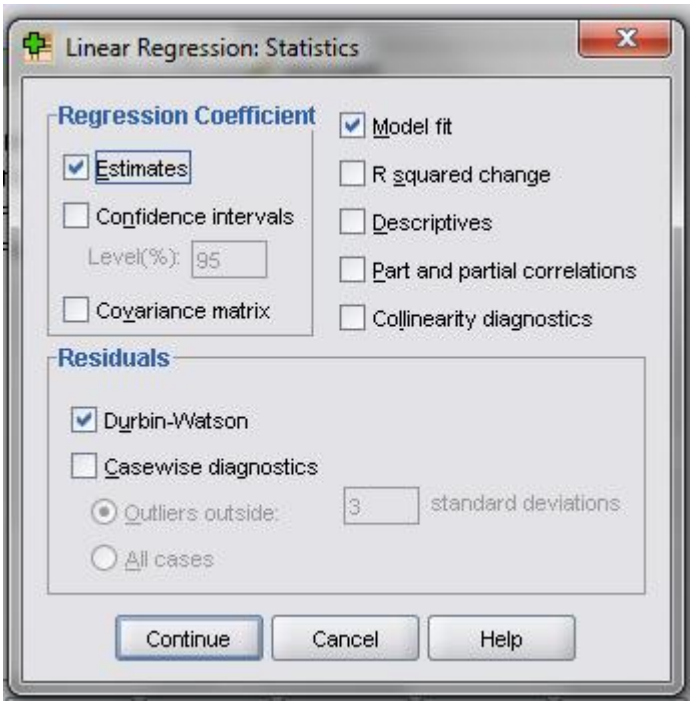
Απο τα διαγράμματα βλέπουμε ότι τα κατάλοιπα ακολουθούν μια πολυ φυσιολογική ροή χωρις ανατροπές και πολύ μεγάλες αποστάσεις μεταξύ τους.



Από το παραπάνω διάγραμμα παρατηρούμε ότι παραβιάζεται η υπόθεση της γραμμικότητας, αφού δεν βλέπουμε να υπάρχει κάποια γραμμική τάση στα σημεία.
Για να υπολογίσουμε τον δείκτη Durbin-Watson:

Analyze → Regression → Linear

Και συνεχίζουμε απο το μενού ***Statistics***



Συνεχίζουμε με continue και o.k.

Απο όλη αυτή τη διαδικάσια θα πάρουμε τον πίνακα *Model Summary ο οποίος μας δίνει την τιμή του δείκτη* Durbin-Watson.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	,996[a]	,992	,991	8,287	1,609

a. Predictors: (Constant), KATANALOSI

b. Dependent Variable: EISODIMA

Στον παραπάνω πίνακα αυτοό μας ενδιαφέρει είναι η τιμή του Durbin-Watson η οποία δεν είναι κοντά στο 2 άρα παραβιάζεται η υπόθεση της ανεξαρτησίας καταλοίπων.

Απο τον πίνακα *ANOVA* που ειναι αποτέλεσμα της ίδιας διαδικασίας με τον πίνακα *Model Summary*.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	66660,267	1	66660,267	970,780	,000[a]
	Residual	549,333	8	68,667		
	Total	67209,600	9			

a. Predictors: (Constant), KATANALOSI

b. Dependent Variable: EISODIMA

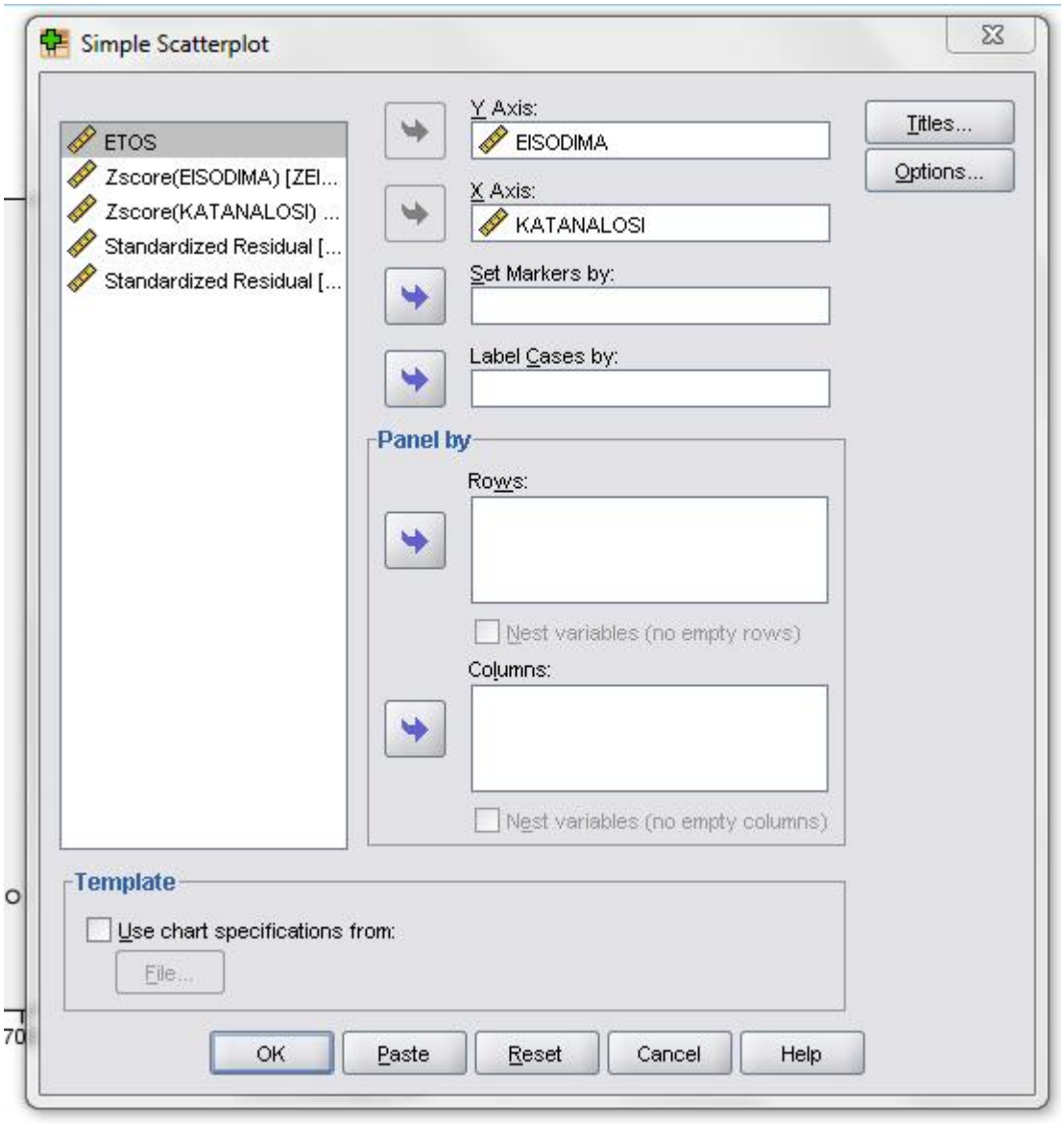
Αυτό που μας ενδιαφέρει στο παραπάνω πίνακα είναι ότι το p-value του ελέγχου είναι (0,00<0,05).Άρα η μηδενική υπόθεση απορρίπτεται.Επομένως το μοντέλο μας προσαρμόζεται καλά στα δεδομένα μας.

Σε αυτό το σημείο θα δουμε το *Διάγραμμα Διασποράς*.

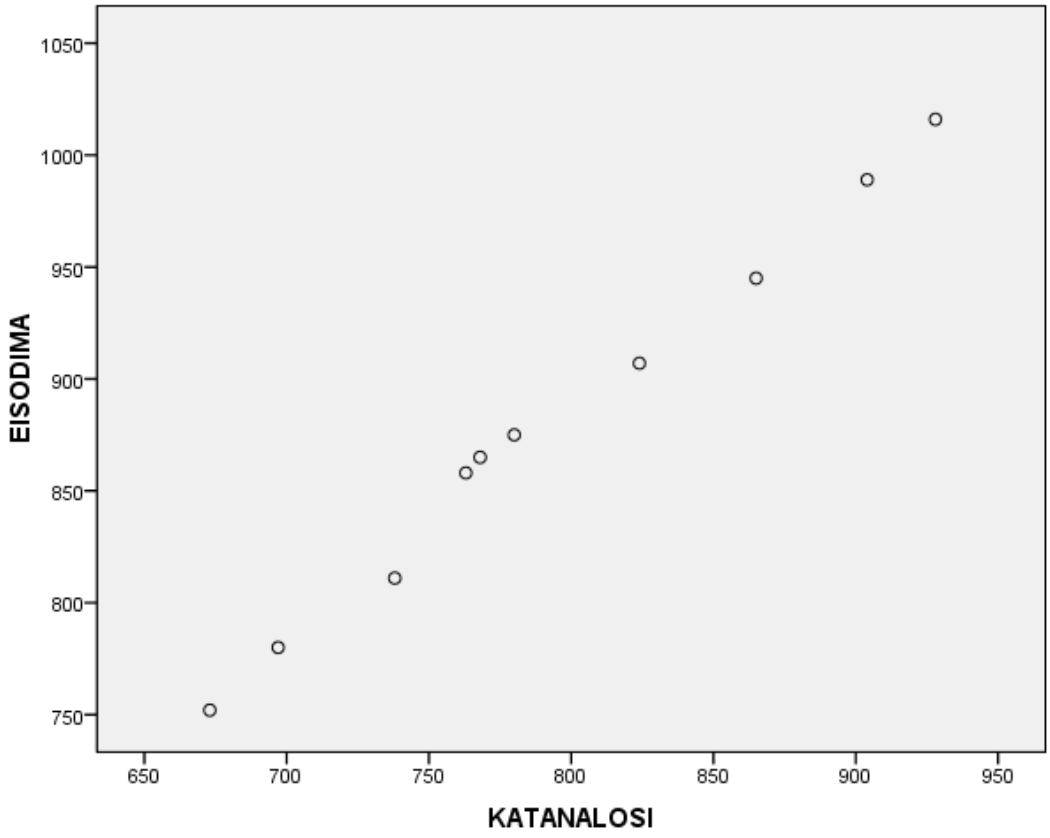
Η διαδικασία είναι η εξής:

Grahps → Scatter → Simple scatter

Και συνεχίζουμε για να φτάσουμε στο διάγραμμα με το εξής



Στη συνέχεια πατάμε ο.k και έχουμε το παρακάτω αποτέλεσμα

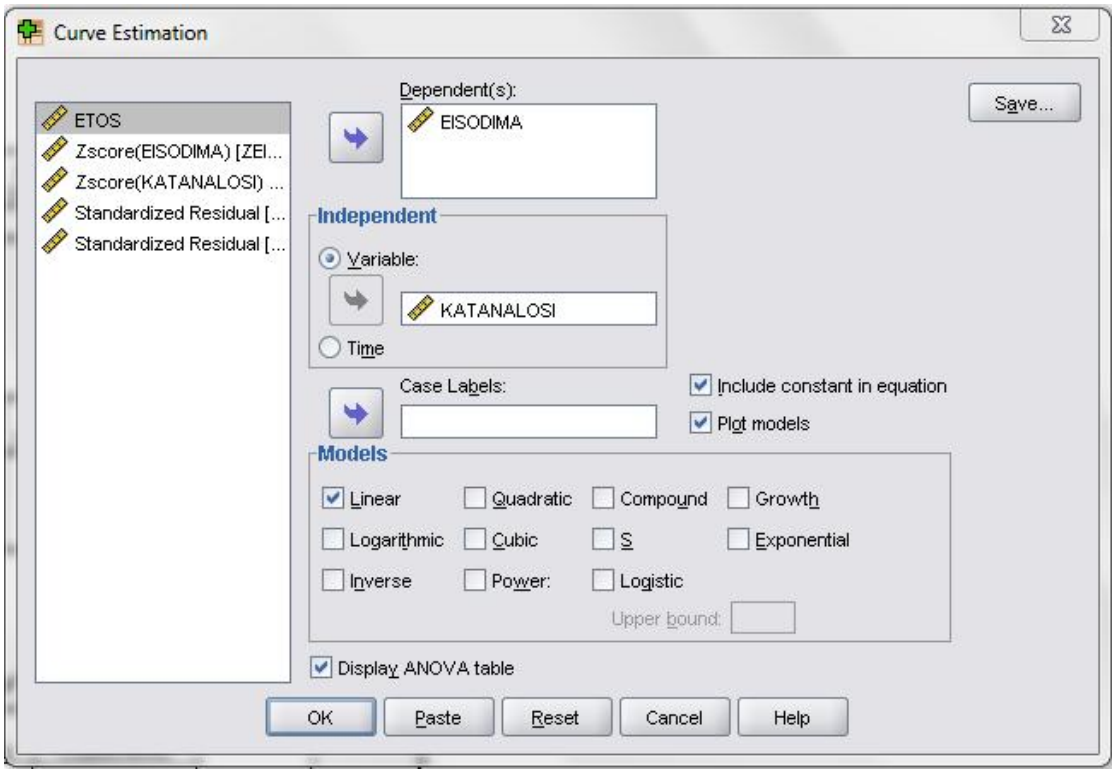


Το διάγραμμα διασποράς περιέχει απο ότι βλέπουμε τις μεταβλητές εισόδημα και κατανάλωση.αυτό που θέλουμε να δούμε με αυτό το διάγραμμα είναι αν αυτές οι δύο μεταβλητές συσχετίζονται μεταξύ τους.Αν συσχετίζονται τα σημεία θα πέσουν κατά μήκος μιας γραμμής και όσο πιο σφιχτά τα σημεία αγκαλίαζουν αυτή τη γραμμή τόσο καλή είναι η σχέση μεταξύ τους.Το παραπάνω σχήμα με βάση την παραπάνω ανάλυση μας δείχνει μια πολύ ισχυρή συσχέτιση μεταξύ των δυο μεταβλητών.

Αφού έχουμε αναλύσει και το δίαγραμμα διασποράς θα περάσουμε στην απλή γραμμική παλινδρόμηση.

Η διαδικασία στο spss είναι η εξής

Analyze →Regression →Curve Estimate



Συνεχίζουμε με o.k. και τα αποτελέσματα που παίρνουμε αρχικά αναλύουμε τον πίνακα *ANOVA.*

ANOVA

	Sum of Squares	df	Mean Square	F	Sig.
Regression	66660,267	1	66660,267	970,780	,000
Residual	549,333	8	68,667		
Total	67209,600	9			

The independent variable is KATANALOSI.

Αυτο που μας ενδιαφέρει στο συγκεκριμένο πίνακα είναι η τιμή του p-value(sig) και απο ότι βλέπουμε το sig=0,000<0,05.Επομένως απορρίπτεται η μηδενική υπόθεση άρα το γραμμικό μοντέλο προσαρμόζεται απόλυτα στα δεδομένα μας.

Αμέσως μετά αναλύουμε τον πίνακα *Model Summary*

Model Summary

R	R Square	Adjusted R Square	Std. Error of the Estimate
,996	,992	,991	8,287

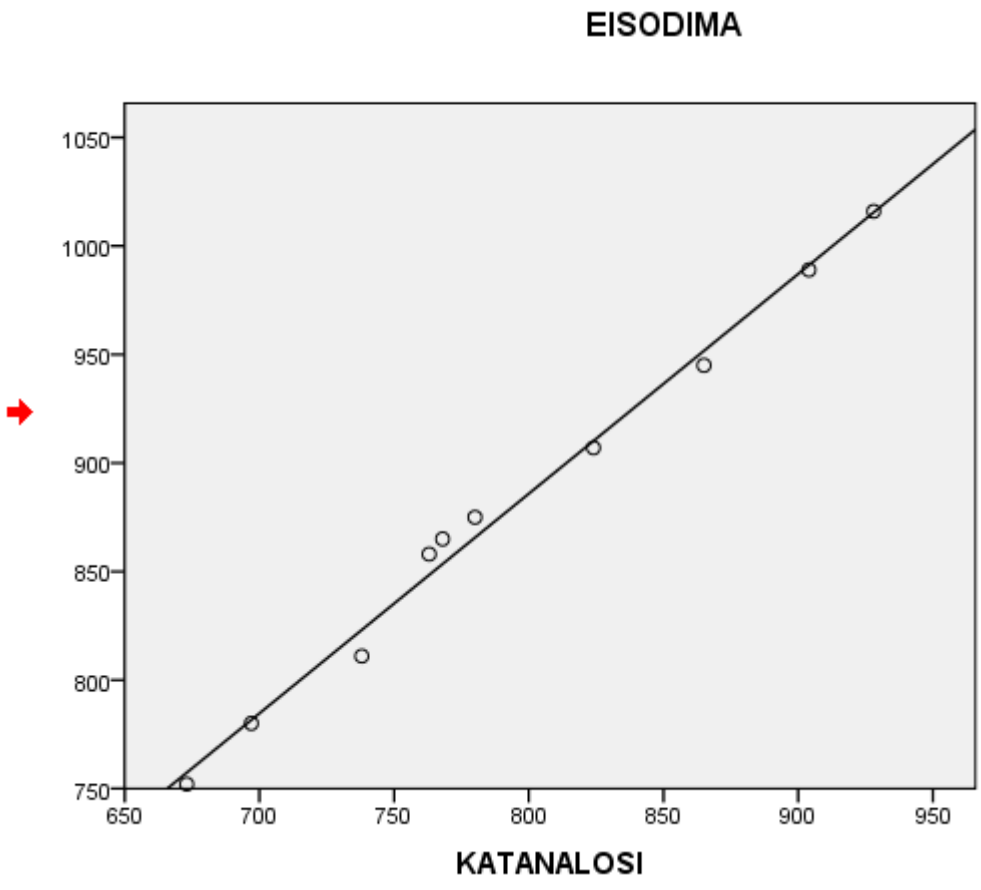
The independent variable is KATANALOSI.

Βλέπουμε απο το R Square ότι το μοντέλο εξηγείται το 99,2% της συνολικής διακύμανσης.Η ερμηνεία του μοντέλου είναι ότι όσο αυξάνεται η κατανάλωση κατα μία μονάδα, το εισόδημα θα μεταβάλλεται κατά β.

ΕΞΙΣΩΣΗ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

$$y = a + b\,x$$
$$y = 1,013 + 75,822x$$
$$y = 1,013 + 75,822 * 0,001$$
$$y = 1,013 + 0,075$$
$$y = 1,088 \approx 1,5$$

Από την πρόβλεψη βλέπουμε ότι για την αύξηση μιας μονάδας κατανάλωσης το εισόδημα μεταβάλλεται κατά 1,5 μονάδα.



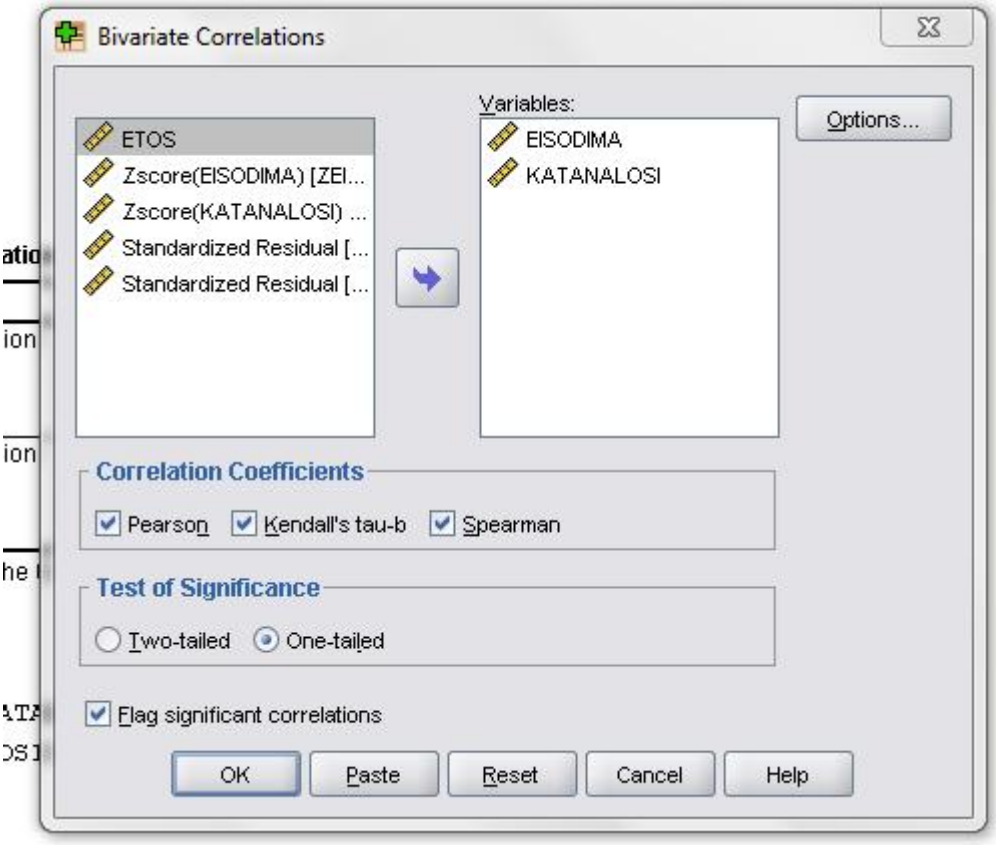
Από το παραπάνω διάγραμμα βλέπουμε την πλήρη γραμμική παλινδρόμηση αφού το νέφος των σημείων ακολουθεί τη γραμμή η οποία δείχνει μια ισχυρή σχέση μεταξύ τους αφού τα σημεία βρίσκονται κοντά το ένα στο άλλο και δεν έχουν μεγάλες αποστάσεις μεταξύ τους.

Αφού έχουμε ολοκληρώσει την ανάλυση της απλής γραμμικής παλινδρόμησης θα πρέπει να συνεχίσουμε στην ανάλυση της συσχέτισης καθώς και των συντελεστών συσχετισης (Pearson,Kendall,Speatman's).

Θα ξεκινήσουμε με τον ***Pearson.***Η διαδικασία στο spss είναι η παρακάτω:

Analyze →Correlate →Bivariate

Και καθώς συνεχίζει η διαδικασία στο spss έχουμε



Πατάμε o.k και παίρνουμε τα αποτελέσματα στο output.

Το πινακάκι που αφορά τον δείκτη Pearson είναι το παρακάτω

Correlations

[DataSet0]

Correlations

		EISODIMA	KATANALOSI
EISODIMA	Pearson Correlation	1	,996**
	Sig. (2-tailed)		,000
	N	10	10
KATANALOSI	Pearson Correlation	,996**	1
	Sig. (2-tailed)	,000	
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

Από τον παραπάνω πίκανα Correlation παίρνουμε

Μέγεθος Σχέσης=0,996

Πιθανότητα=0,00

Μέγεθος Δείγματος=10

Από τον παραπάνω πίνακα φαίνεται πως η σχέση του είσοδήματος και της κατανάλωσης για το δείγμα των 10 τιμών που αξιολογήθηκαν ήταν 0,996 η οποία όπως φαίνεται και απο την παρατηρούμενη πιθανότητα p=0,000<0,001 δεχόμαστε την εναλλακτική υπόθεση σύμφωνα με την οποία υπάρχει στατιστικά σημαντική συνάφεια μεταξύ ΄΄εισοδήματος΄΄ και ΄΄κατανάλωσης΄΄.Η παρατηρούμενη σχέση είναι τόσο ισχυρή που εντοπίστηκε παρά το μικρό μέγεθος του δείγματος.

Συνεχίζουμε με την ανάλυση του δείκτη Spearmann και βλέπουμε το πινακάκι που τον αφορά

Καθώς και το πινακάκι που αφορά το δείκτη Kendall

Nonparametric Correlations

[DataSet0]

Correlations

			EISODIMA	KATANALOSI
Kendall's tau_b	EISODIMA	Correlation Coefficient	1,000	1,000**
		Sig. (2-tailed)	.	.
		N	10	10
	KATANALOSI	Correlation Coefficient	1,000**	1,000
		Sig. (2-tailed)	.	.
		N	10	10
Spearman's rho	EISODIMA	Correlation Coefficient	1,000	1,000**
		Sig. (2-tailed)	.	.
		N	10	10
	KATANALOSI	Correlation Coefficient	1,000**	1,000
		Sig. (2-tailed)	.	.
		N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

Για τον δείκτη Spearman η ανάλυση είναι η εξής

Μέγεθος σχέσης=1.000

Πιθανότητα=0,000

Μέγεθος δείγματος=10

Όπως φαίνεται στην εικόνα η συνάφεια των δύο μεταβλητών είναι αρκετά δυνατή r=0,100 και στατιστικά σημαντική (p<0,001).δεν περιμέναμε να δούμε μεγάλες διαφορές μεταξύ των δύο δεικτών.Στην παρούσα περίπτωση ο δείκτης Spearman είναι ελάχιστα μεγαλύτερες απο το δείκτη Pearson.

Ο δείκτης Kendall δεν θα αναλυθεί παραπάνω σε αυτήν την εργασία.Χρησιμοποιείται για κατηγορικές μεταβλητες οι οποίες όμως είναι υποχρωτικά σε κλίμακα διάταξης.Το κομμάτι αυτό δεν θα αναλυθεί σε αυτή την εργασία.

7.6 ΣΥΜΠΕΡΑΣΜΑΤΑ

Για να καταλήξουμε σε ένα συμπέρασμα αφού έχουμε δεί ολή τη διαδικασία ξεκινώντας απο τη γραμμική παλινδρόμηση μπορούμε να πουμε ότι το δείγμα πάνω στο οποίο δουλέψαμε προσαρμόστηκε απόλυτα στα πρότυπα του απλού γραμμικού υποδείγματος κάθως μας έδωσε αρκέτα σωστά νούμερα καθώς και αψογη συμπεριφορά όσον αφορα τα διαγράμματα που χρησιμοποιήθηκαν για την ανάλυση της παλινδρόμησης.

Για να συνεχίσουμε και να ολοκληρώσουμε την ερευνά μας όσον αφορά τη γραμμική παλινδρόμηση σε αυτήν την εργασια θα πρέπει να αναφερθούμε και στην κακή χρηση της γραμμικής παλινδρόμησης καθώς και σε τι αποτελέσματα μας οδηγει.

ΟΙ ΠΑΡΑΒΙΑΣΕΙΣ ΤΗΣ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Όπως ήδη αναφέρθηκε, το εκάστοτε υπόδειγμα που στηρίζεται στην απλή γραμμική παλινδρόμηση. Οι υποθέσεις της απλής γραμμικής παλινδρόμησης είναι γνωστές στο χώρο της επιστήμης της στατιστικής και της οικονομετρίας. Η παραβίαση αυτών των - 152 - υποθέσεων δημιουργεί σοβαρά προβλήματα στις εκτιμήσεις και στην αξιοπιστία των συμπερασμάτων. Τα προβλήματα που θα παρουσιάσουμε και που δημιουργούνται από την παραβίαση των υποθέσεων του γραμμικού υποδείγματος είναι το πρόβλημα της αυτοσυσχέτισης, το πρόβλημα της κανονικότητας, το πρόβλημα της εξειδίκευσης, και το πρόβλημα της ετεροσκεδαστικότητας:

α) Το πρόβλημα της αυτοσυσχέτισης. Αυτοσυσχέτιση υπάρχει όταν η διακύμανση του διαταρακτικού όρου δεν είναι σταθερή και η συνδιακύμανση όλων των

διαταρακτικών όρων δεν ισούται με το μηδέν. Σε αυτή την περίπτωση έχουμε το φαινόμενο της αυτοσυσχέτισης ή αυτοπαλινδρόμησης. β) Σύμφωνα με τον ορισμό της κανονικότητας η εξαρτημένη μεταβλητή της γραμμικής παλινδρόμησης κατανέμεται

κανονικά, όπως συμβαίνει και με τους εκτιμητές των συντελεστών της παλινδρόμησης.

Επίσης, οι έλεγχοι των υποθέσεων, όπως και οι μέθοδοι εκτίμησης βασίζονται στην

κανονική κατανομή ή στις παράγωγές της. Όταν δεν ισχύει κάτι από τα προηγούμενα

δεν έχουμε κανονικότητα.

γ) Ετεροσκεδαστικότητα υπάρχει όταν οι διαταρακτικοί όροι δεν έχουν την ίδια διακύμανση και, τέλος,

δ) η εξειδίκευση του υποδείγματος αναφέρεται τόσο στην περιγραφή των ερμηνευτικών μεταβλητών, όσο και στην διατύπωση του διαταρακτικού όρου. Επειδή δεν υπάρχουν κριτήρια για την επιλογή του πιο κατάλληλου υποδείγματος, με το πρόβλημα της εξειδίκευσης αναφερόμαστε στην παράλειψη π.χ. μιας ερμηνευτικής μεταβλητής ή σε μη σωστή μορφή του μοντέλου προς ανάλυση.

Στην επόμενη ενότητα παρουσιάζεται το ερευνητικό μέρος της εργασίας, το

οποίο περιλαμβάνει τον τρόπο συλλογής των δεδομένων και την ανάλυσή τους.

Για την μεγαλύτερη χρονική περίοδο παρατηρούνται οι παραβιάσεις της απλής

γραμμικής παλινδρόμησης, με την μέθοδο των ελαχίστων τετραγώνων, για την

εκτίμηση των παραμέτρων του υποδείγματος της παλινδρόμησης, στοιχεία που

δείχνουν ότι όσο μεγαλώνει το ποσοστό των παραβιάσεων, τόσο δεν ισχύει η υπόθεση της αποτελεσματικής αγοράς για την περίοδο ανάλυσης. εκαετία πριν.

Τέλος, πρέπει να επισημάνουμε ότι τις περισσότερες φορές η τυχαία συμπεριφορά των τιμών των μετοχών οφείλεται ως ένα βαθμό στην ανταγωνιστικότητα της αγοράς και στην προσφορά και ζήτηση του χρήματος. Επίσης, η μικρή χρονική περίοδος έρευνας, καθώς και η έλλειψη χρόνου, μας οδήγησαν στο να αρκεστούμε σε έναν περιγραφικό έλεγχο των παραβίασεων των βασικών υποθέσεων της απλής γραμμικής παλινδρόμησης.

Θα πρέπει σε αυτό το σημείο να αναφερθούμε και στους συντελεστες συσχέτισης και στα συμπεράσματα που μπορούμε να βγάλουμε σε ότι αφορά τους συντελεστές συσχέτισης σε αυτήν την εργασία.

Οσόν λοιπόν αφορά τους συντελεστές συσχέτισης καθώς και τα διαγράμματα διασποράς μας δείχνουν το μέγεθος και τη φύση της συσχέτισης.απο όσα είδαμε η μοντελοποιήση της συσχέτισης δεν είναι απλή .Απαιτούνται έλεγχοι του μοντέλου, δυνατότητα ερμηνείας

του,ειδαγωγή νέων μεταβλητων.Οι μεταβλητές που αναλύσαμε είναι δυο οπότε δημιουργήσαμε μια μήτρα συσχετισης 2χ2.Η συσχέτιση των μεταβλητών μας δίνει και το ακριβές επίπεδο σημαντικότητας και μάλιστα επίπεδο σημαντικότητας διπλής ουράς.

Μια άλλη πλευρα ανάλυσης αφορά την κλίση της γραμμής της διασπορας η οποία χαρακτηρίζει και την συσχέτιση.στη δική μας περίπτωση η κλίση της διασποράς δείχνει μια μάλλον ευθεία γραμμή.Ένδειξη οτι ύπαρχει γραμμική (linear) παρά καμπυλόγραμμη(curvilinear) συσχέτιση.αν η συσχέτιση είναι καμπυλόγραμμη οι συντελεστές Pearson ή Spearman μπορεί να είναι παραπλανητική.Αν η ευθεία πηγαίνει απο κάτω δεξία προς τα επάνω αριστερά έχουμε θετική συσχέτιση.

Ένας συντελεστής συσχέτισης πότε δεν πρέπει να παρουσιάζεται χωρις να παρουσιάζεται και να εξετάζεται το γράφημα διασποράς για τυχόν προβλήματα όπως μη γραμικές σχέσεις και έντονα αποκλίνουσες τιμές.Σε μια μελέτη πρέπει πάντοτε να συμπεριλαμβάνουμε γραφήματα τέτοιου είδους.Δυστυχώς τα άρθα των περιοδικών και τα βιβλία έχουν συνήθως περιορισμούς και κόστος που δεν περιλαμβάνουν τα απαραίτητα στοιχεία.

Τελικό συμπέρασμα ο κατάλληλος μεθοδολογικός σχεδιασμός και ο περιορισμός στο μεγαλύτερο δυνατό βαθμό του τυχαίου και του συστηματικού σφάλματος και ιδιαίτερα των συγχυτών αποτέλουν το πρώτο και πλέον καθοριστικό βήμα στη διεξαγωγή μιας μελέτης και την εξαγωγή ασφαλών συμπερασμάτων.Εξίσου σημαντική βεβαίως είναι και η στατιστική επεξεργασία των δεδομενών με την επιλογή της κατάλληλης μεθόδου ανάλογα με το είδος τοσό της ερευνητικής υποθέσεων όσο και δεδομένων.Σημειώνεται ότι η ύπαρξη των πολυμεταβλητών μαθηματικών μοντέλων επιτρέπει την εξουδετέρωση εως ένα βαθμό της σύγχυσης που προκαλείται απο διάφορα χαρακτηριστικά ,αρκεί ο αριθμός των παρατηρήσεων να είναι επαρκής.Η γνώση του σχεδιασμού μιας μελέτης και δευτερευόντως της ανάλυσης δεδομένων είναι ζήτημα πρωταρχικής σημασίας για τους επιστήμονες υγείας και αν σωστά θεωρηθεί ότι η ανάλυση των δεδομένων μπορεί να αποτελέσει αντικείμενο και των στατιστικών,τότε σίγουρα ο ερευνητικός σχεδιασμός αποτελεί αντικείμενο των επιστημόνων υγείας.

Σε αρκετές περιπτώσεις , η παρουσίαση των αποτελεσμάτων περιορίζεται στην αναφορά των παρατηρούμενων επιπέδων στατιστικής σημαντικότητας (τιμές p) αγνοώντας τον καθοριστικό ρόλο που παίζει η παράθεση των διαστημάτων εμπιστοσύνης στην ερμηνεία του τυχαίου

σφάλματος.Επιπλεέον οι τιμές p δεν προσδιορίζουν το μέγεθος της σχέσης ανάμεσα στο μελετώμενο προσδιοριστή και την έκβαση , αποτελώντας απλά ένα μέτρο της συμβατότητας μεταξύ της μηδενικής υπόθεσης και των δεδομένων μιας μελετης .Τα μέτρα σχέσης είναι εκείνα που καθορίζουν το μέγεθος της σχέσης μεταξύ προσδιορισμού και έκβασης και γι'αύτο πρέπει να καθορίζεται με σαφήνεια ο τρόπος υπολογισμού τους.

ΒΙΒΛΙΟΓΡΑΦΙΑ

<u>ΠΗΓΕΣ:</u>

- Γιάννης Παπαδημητρίου,Στατιστική τεύχος 1.Επαγωγική Στατιστική.
- Γιάννης Παπαδημητρίου,Στατιστική τεύχος 2.Επαγωγική Στατιστική.
- Χρήστος Κίτσας,Εισαγωγή στην Εφαρμοσμένη Στατιστική .
- Γεώργιος Δημ.Πέκος. Ασκήσεις Στατιστικής.
- Δρο Παναγιώτου Ν. Γεωβργανη,Στατιστική Εφαρμοσμένη στις επιστήμες της συμπεριφοράς,τόμος Α΄ →Περιγραφική Στατιστική.
- Πέτρος Α Κίοχος,Στατιτική.

ΔΙΑΔΙΚΤΥΟ

- http://www.teipat.gr
- http://wikipedia.org/wiki/spss
- http://www.actuar.aegean.gr
- http://users.auth.gr/dkugiu/Teach/DataAnalysis/Chp5.pdf
- http://www.lib.teiher.gr/webnotes/seyp/SPSS/Kef08.pdf
- http://users.auth.gr/dkugiu/Teach/CivilEngineer/regression.pdf
- http://www.aua.gr/gpapadopoulos/files/sisxetisi091.pdf
- http://www.math.ntua.gr/~fouskakis/EPIPSI/05.pdf
- http://users.uoi.gr/hyepilab/assets/pdfs/biomathematics/Linear_Regression.pdf
- http://7.nsa-virtualeducation.com/images/l.notes3.pdf