Τ.Ε.Ι. ΠΑΤΡΑΣ

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ

ΤΜΗΜΑ:ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ

“ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΑΙ ΣΥΣΧΕΤΙΣΗ”

ΒΑΜΒΑΤΣΙΚΟΥ ΓΕΩΡΓΙΑ
ΔΗΜΗΤΡΙΟΥ ΚΩΣΤΑΝΤΙΝΟΣ

ΕΠΙΒΛΕΠΩΝ ΚΑΘΗΓΗΤΗΣ: ΑΝΤΩΝΟΠΟΥΛΟΥ ΗΡΑ

ΠΑΤΡΑ 2014

Περίληψη

Στη παρούσα εργασία γίνεται μια σύντομη περιγραφή των περιγραφικών μέτρων στατιστικής ανάλυσης και εν συνεχεία, δίνεται ιδιαίτερη έμφαση στο μοντέλο της απλής γραμμικής παλινδρόμησης. Αφού παρουσιάζεται όλο το θεωρητικό υπόβαθρο που σχετίζεται με τις ανωτέρω έννοιες, γίνεται εφαρμογή του μοντέλου σε διάφορες μεταβλητές προκειμένου να διαπιστωθεί η ύπαρξη ή μη της γραμμικής τους εξάρτησης.

Πίνακας Περιεχομένων

1. Εισαγωγή

Στην επιστημονική διάλεκτο, ο όρος «στατιστική», έχει ευρύτερη σημασία. Σημαίνει, την επιστήμη που έχει σαν αντικείμενο όχι μόνο τη συγκέντρωση, επεξεργασία και παρουσίαση, αλλά και μελέτη και ανάλυση των παρατηρήσεων ή μετρήσεων που αναφέρονται σε χαρακτηριστικές ιδιότητες ενός συγκεκριμένου αντικειμένου ή γεγονότος, οποιαδήποτε και αν είναι η φύση του. Έτσι η στατιστική περιλαμβάνει τόσο τις μεθόδους συλλογής, οργάνωσης, επεξεργασίας και παρουσίασης των συγκεντρωθέντων στοιχείων, όσο και τις μεθόδους ανάλυση και ερμηνείας αριθμητικών δεδομένων για την εξαγωγή λογικών και τεκμηριωμένων συμπερασμάτων που θα βοηθούσαν στη λήψη ορθών αποφάσεων.

Σύμφωνα με τα παραπάνω, θα μπορούσε να ειπωθεί ότι η στατιστική είναι η επιστήμη που ασχολείται με τις επιστημονικές μεθόδους συλλογής, επεξεργασίας, παρουσίασης ανάλυσης και ερμηνείας αριθμητικών δεδομένων, ώστε να καταλήξει στη διατύπωση συμπερασμάτων, τα οποία είναι χρήσιμα στη λήψη ορθών αποφάσεων [6].

Ο όρος στατιστική έχει τις ρίζες του, στη λατινική έκφραση ***statisticum collegium*** (διάλεξη για υποθέσεις της πολιτείας), από την οποία προήρθε η Ιταλική λέξη ***statista***, που σημαίνει πολιτικός, και η Γερμανική λέξη ***Statistik***, η οποία αρχικά αναφερόταν στην ανάλυση των δεδομένων για την πολιτεία. Πήρε την έννοια της συλλογής και ταξινόμησης δεδομένων γενικά στις αρχές του δεκάτου ένατου αιώνα[1].

Η πρώτη στατιστική γραφή εντοπίζεται σε βιβλίο του 9ου αιώνα με τον τίτλο "*Κρυπτογραφημένα μηνύματα*", συγγραφή του Al-Kindi (801–873 π.Χ). Στο βιβλίο του, ο Al-Kindi έδωσε αναλυτική περιγραφή του πώς χρησιμοποιείται η στατιστική και η ανάλυση συχνότητας, ώστε κάποιος να μπορέσει να δημιουργήσει κρυπτογραφημένα μηνύματα. Αυτό υπήρξε το ορόσημο για τη γέννηση της στατιστικής και κρυπτανάλυσης [7,8].

[1] http://en.wikipedia.org/wiki/Statistics

Ωστόσο, είναι γεγονός ότι πολλούς αιώνες πριν οι άνθρωποι ασχολήθηκαν με την ανάλυση στατιστικών δεδομένων, ακόμα και αν δεν υπήρχε τότε η θεσμοθετημένη έννοια της στατιστικής επιστήμης. Χαρακτηριστικό παράδειγμα αποτελεί το γεγονός ότι η πρώτη εξακριβωμένη απογραφή πληθυσμού, πραγματοποιήθηκε στη Κίνα από τον αυτοκράτορα Υ-άο το έτος 2238 π.Χ. Στην Αγγλία αντίστοιχα πραγματοποιήθηκε το 1805, από τον Γουλιέλμο τον κατακτητή, η πρώτη καθολική απογραφή του πληθυσμού και του πλούτου.

Το 1583 γράφεται από τον Sansovino το πρώτο βιβλίο στατιστικού περιεχομένου και λίγο αργότερα, κατά το διάστημα 1606 – 1681, εισάγεται από τον Könring η στατιστική στην ανώτερη εκπαίδευση.

Μερικοί επιστήμονες θέτουν αφετηρία της στατιστικής το 1663, με τη έκδοση του βιβλίου *Φυσικές και Πολιτικές παρατηρήσεις της Θνησιμότητας* από τον John Graunt [9]. Τριάντα χρόνια μετά (1693), ο Άγγλος αστρονόμος Halley, χρησιμοποίησε τα ληξιαρχικά βιβλία γεννήσεων και θανάτων της πόλης Breslaou, για να παρουσιάσει το πρώτο πίνακα θνησιμότητας.

Πρόσφατες αιτήσεις της στατιστικής σκέψης, περιτριγυρίζονται από τις ανάγκες της πολιτείας να χτίσει πολιτική στα δημοκρατικά και οικονομικά δεδομένα. Σήμερα η στατιστική είναι ευρέως διαδομένη στη πολιτική, στις επιχειρήσεις, και στις φυσικές και κοινωνικές επιστήμες. Μια απλή απαρίθμηση των εφαρμογών της στατιστικής, φανερώνει το γεγονός ότι χρησιμοποιείται σε όλους σχεδόν τους τομείς της ανθρώπινης δραστηριότητας.

Η στατιστική είναι απαραίτητη στις διοικούντες αρχές, για τη λήψη ορθών αποφάσεων, οι οποίες με τη σειρά τους συμβάλουν στην ανάπτυξη/πρόοδο ενός οργανισμού, κράτους, επιχείρησης, κτλ. Αυτός είναι και ο λόγος για τον οποίο, οι σημερινές σύγχρονες επιχειρήσεις, χρησιμοποιούν τις στατιστικές μεθόδους και γενικότερα τα εργαλεία που τους παρέχει η στατιστική επιστήμη, για τη λήψη επιχειρηματικών αποφάσεων.

Στις ενότητες που ακολουθούν, θα εστιάσουμε σε ένα συγκεκριμένο τομέα της στατιστικής επιστήμης και πιο συγκεκριμένα στη συσχέτιση μεταξύ δύο μεταβλητών. Ακολούθως θα αναλυθούν οι έννοιες της απλής γραμμικής παλινδρόμησης και του συντελεστή συσχέτισης, καθώς και θα παρουσιαστούν υπολογιστικά προβλήματα στα οποία θα εφαρμόσουμε τις μεθόδους αυτές.

2. Βασικές Έννοιες

2.1. Περιγραφικά Μέτρα

Στο σημείο αυτό θα δοθούν οι ορισμοί και θα αναλυθούν κάποιες από τις βασικές έννοιες της στατιστικής. Τα περιγραφικά μέτρα αποτελούν μεθόδους περιγραφής των δεδομένων, που μπορούν να εκφράζονται είτε ως συναρτήσεις των δεδομένων του δείγματος (στατιστικά), είτε ως συναρτήσεις των δεδομένων του πληθυσμού (παράμετροι).

Τα περιγραφικά μέτρα θα μπορούσαν να χωριστούν σε δύο κύριες κατηγορίες, στα μέτρα θέσης και στα μέτρα διασποράς.

2.1.1. Μέτρα θέσης

Οι μέθοδοι περιγραφής των δεδομένων που αποτελούν τα μέτρα θέσης περιγράφονται στις ακόλουθες υπό-ενότητες.

2.1.1.1. Αριθμητικός Μέσος (mean)

Ο αριθμητικός μέσος ενός συνόλου παρατηρήσεων X_1, X_2, …, X_N, οι οποίες αποτελούν το σύνολο του πληθυσμού που τελεί υπό έρευνα, δίνεται από την ακόλουθη εξίσωση:

$$\mu = \sum_{i=1}^{N} X_i \Big/ N$$

όπου:

- N: Το πλήθος του πληθυσμού
- X_i: Οι παρατηρήσεις

Με αντίστοιχο τρόπο προκύπτει και ο ορισμός του δειγματικού μέσου που εκφράζεται από την ακόλουθη εξίσωση

$$\overline{X} = \sum_{i=1}^{n} \chi_i / n$$

όπου:

- n: Το μέγεθος του δείγματος
- χ_i: Οι παρατηρήσεις

Στο σημείο αυτό, παρουσιάζεται μια σημαντική ιδιότητα του μέσου σύμφωνα με την οποία ισχύει

$$\sum_{i=1}^{n} (X_i - \overline{X}) = 0$$

2.1.1.2. ***Διάμεσος*** *(median)*

Σε ένα σύνολο παρατηρήσεων X_1, X_2, …, X_N, ο διάμεσος Μ είναι η τιμή της μέσης παρατήρησης όταν οι παρατηρήσεις είναι διατεταγμένες σε αύξουσα ή φθίνουσα τάξη. Κατά την εξαγωγή της διαμέσου από ένα πλήθος n παρατηρήσεων διακρίνουμε δύο περιπτώσεις:

i. n = άρτιος αριθμός: Στην περίπτωση αυτή ο διάμεσος προκύπτει από το ημιάθροισμα των δυο μέσων παρατηρήσεων.
ii. n = περιττός αριθμός: Στην περίπτωση αυτή ο διάμεσος είναι η μέση παρατήρηση.

2.1.1.3. ***Επικρατούσα Τιμή*** *(mode)*

Από ένα σύνολο αταξινόμητων δεδομένων, η επικρατούσα τιμή είναι το δεδομένο εκείνο το οποίο έχει την μεγαλύτερη συχνότητα της κατανομής των συχνοτήτων. Με άλλα λόγια είναι εκείνη η τιμή που συναντάται με μεγαλύτερη συχνότητα.

2.1.2. Μέτρα Διασποράς

Οι μέθοδοι περιγραφής των δεδομένων που αποτελούν τα μέτρα θέσης περιγράφονται ακολούθως.

2.1.2.1. Εύρος Δεδομένων (range)

Το εύρος δεδομένων είναι εκείνο το μέτρο διακύμανσης των δεδομένων που επιστρέφει ως αποτέλεσμα τη διαφορά της ελάχιστης παρατήρησης από τη μέγιστη. Παρόλο ότι ο υπολογισμός του συγκεκριμένου μέτρου είναι πολύ εύκολος, διαθέτει ένα μεγάλο μειονέκτημα που είναι το γεγονός ότι δεν αποτελεί συνάρτηση όλων των δεδομένων με συνέπεια να χάνονται πληροφορίες που βασίζονται στις τιμές των δεδομένων.

Η έκφραση που μας επιστρέφει τη τιμή του εύρους δεδομένων, δίνεται από την ακόλουθη εξίσωση:

$$E.\Delta. = X_{max} - X_{min}$$

όπου:

- X_{max}: η μέγιστη παρατήρηση,
- X_{min}: η ελάχιστη παρατήρηση

2.1.2.2. Μέση απόκλιση (average deviation)

Σε ένα σύνολο παρατηρήσεων X_1, X_2, …, X_N, η μέση απόκλιση δίνεται από την εξίσωση:

$$M.A. = \sum_{i=1}^{N} |X_i - \overline{X}| \Big/ N$$

2.1.2.3. *Δειγματική Τυπική Απόκλιση (sample standard deviation)*

Η δειγματική τυπική απόκλιση (S) μαζί με τη δειγματική διασπορά (S^2) αποτελούν τα βασικότερα μέτρα διασποράς. Και τα δύο εφαρμόζονται κατά κόρον στη στατιστική ανάλυση δεδομένων. Η τυπική απόκλιση ενός δείγματος $X_1, X_2, \ldots, X_n$, δίνεται από την εξίσωση:

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

Αν θεωρηθούν οι παρατηρήσεις $X_1, X_2, \ldots, X_N$, οι οποίες είναι όλες οι τιμές μιας μεταβλητής x οι οποίες περιέχονται στον πληθυσμό Π, τότε η τυπική απόκλιση (σ) της μεταβλητής x για αταξινόμητα δεδομένα δίνεται από την εξίσωση:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

2.1.2.4. *Δειγματική Διασπορά (sample variance)*

Πριν οριστεί η δειγματική διασπορά, θα πρέπει να δοθεί η έννοια της διασποράς. Ως διασπορά (ή διακύμανση) ορίζεται η μεταβλητότητα των παρατηρήσεων γύρω από τον αριθμητικό μέσο. Με άλλα λόγια, διασπορά είναι ο μέσος όρος των τετραγώνων των αποστάσεων των παρατηρήσεων από τον μέσο του δείγματος. Επειδή όμως η απόκλιση μιας παρατήρησης X_i από τη μέση τιμή ορίζεται ως $X_i - \overline{X}$, είναι προφανές ότι το άθροισμα αυτών των αποκλίσεων θα ισούται με το μηδέν, όπως άλλωστε προκύπτει από τις ιδιότητες του δειγματικού μέσου, σύμφωνα με τις οποίες ισχύει η έκφραση:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$$

Η δειγματική διασπορά (S^2) ενός δείγματος $X_1, X_2, \ldots, X_n$, προκύπτει από τον τύπο της τυπικής απόκλισης υψωμένο στο τετράγωνο. Άρα λοιπόν, η σχέση που προκύπτει περιγράφεται από την εξίσωση:

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

Αν θεωρηθούν οι παρατηρήσεις $X_1, X_2, \ldots, X_N$, οι οποίες είναι όλες οι τιμές μιας μεταβλητής x οι οποίες περιέχονται στον πληθυσμό Π, τότε η τυπική απόκλιση (σ^2) της μεταβλητής x για αταξινόμητα δεδομένα δίνεται από την εξίσωση:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

2.2. Ροπές και Μέτρα Ασυμμετρίας – Κύρτωσης

Τα περιγραφικά μέτρα που εξετάστηκαν νωρίτερα, χρησιμοποιούνται σχεδόν πάντα σε διάφορες πρακτικές εφαρμογές και συνήθως είναι επαρκή για την αντιμετώπιση πολλών στατιστικών προβλημάτων. Υπάρχουν όμως αρκετές φορές, που πέραν της θέσης και της διασποράς, απαιτείται ο υπολογισμός και η χρήση ορισμένων άλλων παραμέτρων που αφορούν τη μορφή και το σχήμα της κατανομής συχνοτήτων.

Το σχήμα των κατανομών εξαρτάται από τους συντελεστές ασυμμετρίας και κύρτωσης, οι οποίοι είναι συναρτήσεις των ροπών

2.2.1. Ροπές

Η τιμή των ροπών εξαρτάται από εντοπισμό της «αρχής». Αν ως «αρχή» λογίζεται η τιμή της μεταβλητής $X=0$ *ή* $X=\mu$, τότε έχουμε *ροπές περί την αρχή ή περί το μέσο* μ.

Έστω ένα τυχαίο δείγμα από τις παρατηρήσεις $X_1, X_2, \ldots, X_N$, από το πληθυσμό μεγέθους N της μεταβλητής X. Η δειγματική ροπή περί την αρχή (X=0), τάξης k ορίζεται:

i. Για αταξινόμητα δεδομένα:

$$G' = \frac{1}{n}\sum_{i=1}^{n} X_i^k \, , k = 1,2,3, \ldots$$

ii. Για ταξινομημένα δεδομένα με συχνότητες f_i:

$$G' = \frac{1}{n}\sum_{i=1}^{n} f_i X_i^k$$

Έστω τώρα ένα δείγμα από τις παρατηρήσεις $X_1, X_2, \ldots, X_N$, της ποσοτικής μεταβλητής X. Η δειγματική κεντρική ροπή περί το μέσο μ (X=μ), τάξης k ορίζεται:

i. Για αταξινόμητα δεδομένα:

$$G = \frac{1}{n}\sum_{i=1}^{n} (X_i - \mu)^k$$

ii. Για ταξινομημένα δεδομένα με συχνότητες f_i:

$$G = \frac{1}{n}\sum_{i=1}^{n} f_i (X_i - \mu)^k$$

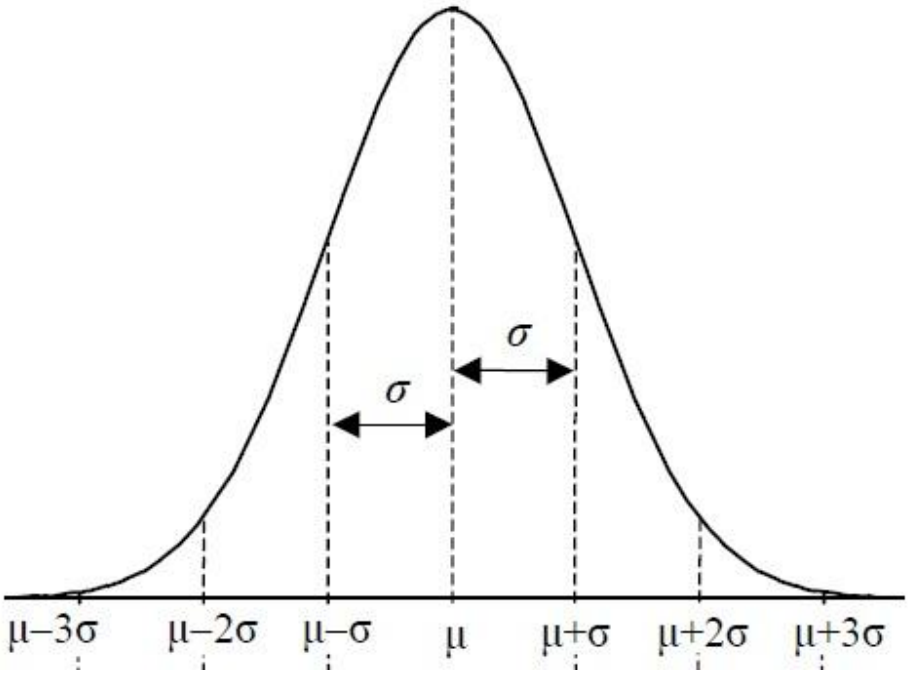
2.2.2. Μέτρα Ασυμμετρίας – Κύρτωσης

Η μορφολογία της κατανομής αφορά την ασυμμετρία αυτής, δηλαδή το βαθμό απόκλισης της καμπύλης συχνοτήτων μιας κατανομής από μια πρότυπη συμμετρική καμπύλη (κανονική κατανομή).

Όσον αφορά τη κύρτωση, μετράει το πόσο «λεπτή» ή «πλατιά» είναι η καμπύλη συχνότητας μιας κατανομής σε σχέση με τη καμπύλη συχνότητας μιας κανονικής κατανομής.

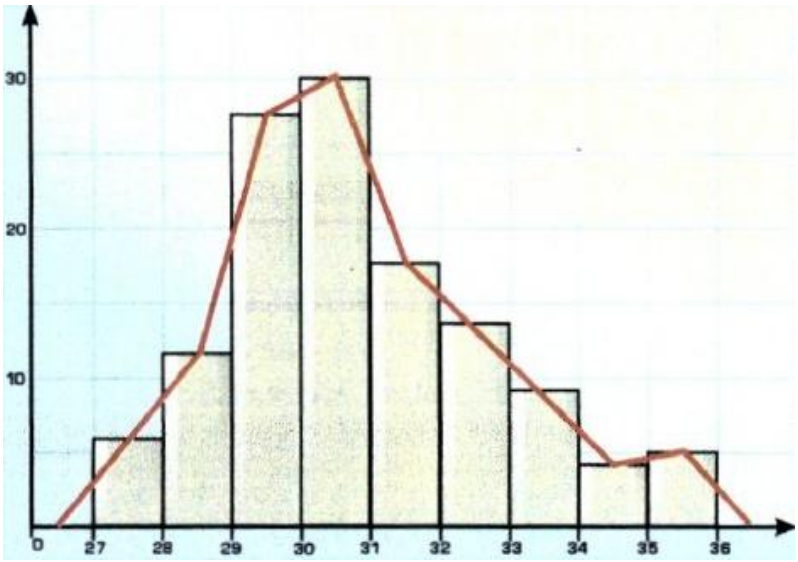
Για να γίνει καλύτερα κατανοητή η έννοια της ασυμμετρίας, πρέπει να γίνει καλύτερα αντιληπτή η έννοια της συμμετρικής κατανομής.

Μια κατανομή είναι συμμετρική, όταν όλες οι τιμές της τοποθετούνται συμμετρικά γύρω από τη μέση αριθμητική τιμή (βλ. Εικόνα 1).



Εικόνα 1: Συμμετρική Κατανομή

Από την άλλη γραφικά, στην ασύμμετρη κατανομή δεν τοποθετούνται οι τιμές της συμμετρικά γύρω από τη μέση αριθμητική τιμή (βλ. Εικόνα 2).



Εικόνα 2: Ασύμμετρη Κατανομή

Για τον προσδιορισμό της ύπαρξης ή όχι συμμετρίας (και του βαθμού αυτής), χρησιμοποιείται ο συντελεστής ασυμμετρίας S_k [1,2]. Για τον υπολογισμό του συντελεστή ασυμμετρίας, χρησιμοποιούνται διάφοροι τύποι.

Σύμφωνα με τον Pearson [2] ο τύπος είναι:

$$S_k = \frac{\overline{X} - Mode}{S} = \frac{\mu - ET}{\sigma}$$

όπου:

- μ: ο αριθμητικός μέσος
- ET: η επικρατούσα τιμή
- σ: η τυπική απόκλιση
- Το πεδίο τιμών του S_k είναι:

$$-1 \leq S_k \leq 1$$

Όσον αφορά τις τιμές που παίρνει το S_k διακρίνουμε τρεις περιπτώσεις:

[2] http://en.wikipedia.org/wiki/Karl_Pearson

- $S_k = 0$: Συμμετρική κατανομή
- $S_k > 0$: Θετική ασυμμετρία
- $S_k < 0$: Αρνητική ασυμμετρία

Όσο οι τιμές που παίρνει το S_k, τείνουν προς τα άκρα του, τόσο εντείνεται η ασυμμετρία.

Οι Kendall και Stuart, έδωσαν ένα άλλο δημοφιλές μέτρο ασυμμετρίας, το οποίο εφαρμόζεται σε όλες τις κατανομές των οποίων οι ροπές υπάρχουν για k=1,2,3,4 [3].

$$S_k = \frac{\sqrt{\beta_1}(\beta_2 + 3)}{2(5\beta_2 - 6\beta_1 - 9)}$$

όπου:

- β_1 ο συντελεστής ασυμμετρίας, ο οποίος δίνεται από τη σχέση

$$\beta_1 = \frac{G_3}{G_2^{3/2}}$$

- β_2 ο συντελεστής ασυμμετρίας, ο οποίος δίνεται από τη σχέση

$$\beta_1 = \frac{G_4}{G_2^2}$$

- Όταν $\beta_1 = 0$, τότε η κατανομή είναι συμμετρική.

2.3. Έλεγχος υποθέσεων

Ο στατιστικός έλεγχος υποθέσεων (hypothesis testing) είναι μια συμπερασματική διαδικασία/μέθοδος που προσφέρει η Στατιστική Συμπερασματολογία και βρίσκει εφαρμογή σε στοχαστικά προβλήματα απόφασης μεταξύ δύο εναλλακτικών υποθέσεων. Η μία υπόθεση έχει επικρατήσει να συμβολίζεται με H_0 και ονομάζεται μηδενική υπόθεση (null hypothesis), και η άλλη με H_1 και ονομάζεται εναλλακτική

υπόθεση (alternative hypothesis). Αναγκαία προϋπόθεση για τη σωστή εφαρμογή των στατιστικών ελέγχων και κυρίως για τη σωστή ερμηνεία των αποτελεσμάτων τους, είναι η κατανόηση της λογικής και του νοήματός τους[3].

Ο έλεγχος υποθέσεων έχει ως αποτέλεσμα μια απόφαση για τη τιμή ενός συντελεστή/παραμέτρου ενός στατιστικού πληθυσμού. Συνήθως, αυτός ο έλεγχος αφορά τον μέσο μ ή τη διακύμανση σ^2. Κάποιοι άλλοι έλεγχοι, ασχολούνται είτε με τη σύγκριση δύο πληθυσμών, είτε με το βαθμό τυχαιότητας των δεδομένων, είτε με το είδος συσχέτισης δύο μεταβλητών (simple linear regression).

Ο απώτερος σκοπός του ελέγχου υποθέσεων, είναι η λήψη απόφασης για τις παραμέτρους ενός πληθυσμού, έπειτα από εξέταση που έχει προηγηθεί στις παρατηρήσεις του πληθυσμού αυτού.

Η υπόθεση είναι μια δήλωση για έναν ή περισσότερους πληθυσμούς ή τις παραμέτρους τους. Στον έλεγχο υποθέσεων έχουμε στην ουσία την αντιπαραβολή δύο υποθέσεων. Πιο συγκεκριμένα, την αντιπαραβολή της μηδενικής και της εναλλακτικής υπόθεσης. Η μηδενική υπόθεση, είναι η υπόθεση που ελέγχεται για την ορθότητά της. Η εναλλακτική υπόθεση, είναι η υπόθεση η οποία διατίθεται όταν απορρίπτεται η μηδενική υπόθεση.

Ο έλεγχος υποθέσεων διακρίνεται από κάποια στάδια, τα οποία αναλύονται ακολούθως [5]:

2.3.1. Δήλωση υποθέσεων

Έστω ότι θέλουμε να πραγματοποιηθεί ένας έλεγχος για τον μέσο μ ενός πληθυσμού. Το συμπέρασμα που θέλουμε να εξαχθεί, τίθεται πάντα ως εναλλακτική υπόθεση. Έστω λοιπόν ότι το ζητούμενο είναι να αποδειχτεί ότι $\mu \neq \mu_0$.

[3] http://www.aua.gr/gpapadopoulos/files/hypoth-tests-4.pdf

Σύμφωνα με τα όσα αναφέρθηκαν στη προηγούμενη παράγραφο θα έχουμε ότι, η μηδενική υπόθεση είναι H_0: $\mu=\mu_0$ και η εναλλακτική υπόθεση είναι H_1: $\mu\neq\mu_0$. Ο έλεγχος αυτής της μορφής, καλείται έλεγχος δύο διευθύνσεων.

2.3.2. Προσδιορισμός στατιστικού του ελέγχου

Στο στάδιο αυτό εξετάζεται το στατιστικό του ελέγχου, το οποίο είναι μια συνάρτηση των δεδομένων, που το αποτέλεσμα που επιστρέφει χρησιμεύει στην απόρριψη ή όχι της μηδενικής υπόθεσης. Οι διάφορες μορφές των στατιστικών όταν γίνεται ένας έλεγχος για τη τιμή του μέσου μ του πληθυσμού είναι:

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$

2.3.3. Ορισμός επιπέδου σημαντικότητας

Κατά τον έλεγχο υποθέσεων, η μηδενική υπόθεση μπορεί να είναι είτε αληθής είτε ψευδής. Κατ' επέκταση, η απόφαση που θα ληφθεί μπορεί να είναι απόρριψη της υπόθεσης H_0 ή μη απόρριψη της. Ακολούθως περιγράφονται δύο καταστάσεις οι οποίες οδηγούν σε ισάριθμους ορισμούς.

Η απόρριψη της μηδενικής απόφασης H_0 δεδομένου ότι η H_0 είναι αληθής, είναι μια μη σωστή απόφαση. Η πιθανότητα α = P[να απορριφθεί η H_0 / η H_0 είναι αληθής], λέγεται επίπεδο σημαντικότητας.

Η μη απόρριψη της μηδενικής απόφασης H_0 δεδομένου ότι η H_0 είναι ψευδής, είναι μια μη σωστή απόφαση. Αν β = P[να μην απορριφθεί η H_0 / η H_0 είναι ψευδής] τότε ισχύει η πιθανότητα 1-β = P[να απορριφθεί η H_0 / η H_0 είναι ψευδής], η οποία καλείται δύναμη του ελέγχου.

Σκοπός του ελέγχου υποθέσεων είναι, αρχικά να οριστεί το μέγεθος του σφάλματος α, και εν συνεχεία να ελαχιστοποιηθεί το σφάλμα β δεδομένης της τιμής του α. Συνήθως χρησιμοποιούνται οι τιμές α = 0.05, α = 0.01 και α = 0.1.

2.3.4. Κανόνας απόφασης

Ο χώρος απόρριψης της H_0 είναι το σύνολο όλων των τιμών του στατιστικού για τις οποίες η υπόθεση H_0 θα απορριφθεί. Ο χώρος αποδοχής της H_0 είναι το σύνολο των υπόλοιπων τιμών του στατιστικού του ελέγχου, αν αφαιρεθεί από τη κατανομή του στατιστικού το χώρο απόρριψης της H_0. Όταν απορρίπτεται η H_1, δεν είναι ότι γίνεται δεκτή η H_0 αλλά στην ουσία απορρίπτεται η υπόθεση H_1.

Οι τιμές του στατιστικού του ελέγχου που χωρίζουν το χώρο απόρριψης της H_0 από το χώρο αποδοχής της, καλούνται κριτικές τιμές. Η τιμή του επιπέδου σημαντικότητας α, καθορίζει τη θέση των χώρων απόρριψης και αποδοχής της H_0.

Ο κανόνας απόφασης δηλώνει ότι, αν η τιμή του στατιστικού του ελέγχου βρίσκεται στο χώρο απόρριψης της H_0, τότε η υπόθεση H_0 απορρίπτεται. Από την άλλη πλευρά, αν η τιμή του στατιστικού βρίσκεται στο χώρο αποδοχής της H_0, τότε η υπόθεση H_0 δεν απορρίπτεται. Αν η τιμή του στατιστικού είναι ίση με τη κριτική τιμή, τότε η υπόθεση H_0 απορρίπτεται.

2.3.5. Συλλογή δεδομένων

Στο σημείο αυτό, συλλέγονται όλα τα δεδομένα και η τιμή του στατιστικού του ελέγχου. Το δείγμα από το οποίο προέρχονται τα δεδομένα, πρέπει να είναι τυχαίο. Έπειτα εκτελούνται όλοι οι κατάλληλοι υπολογισμοί.

2.3.6. Λήψη αποφάσεων

Αμέσως μετά από το στάδιο της συλλογής δεδομένων και της εκτέλεσης των κατάλληλων υπολογισμών, έρχεται η λήψη της απόφασης. Εφαρμόζοντας το κανόνα απόφασης, παρατηρείται αν η τιμή του στατιστικού του ελέγχου είναι μεγαλύτερη ή μικρότερη της κριτικής τιμής και αναλόγως απορρίπτεται ή δεν απορρίπτεται η υπόθεση H_0.

2.3.7. Εξαγωγή συμπερασμάτων

Στο τελευταίο στάδιο υπάρχει η εξαγωγή συμπερασμάτων. Αν η μηδενική υπόθεση H_0 απορριφθεί, συμπεραίνεται ότι η εναλλακτική υπόθεση H_1 είναι αληθής. Αν η μηδενική υπόθεση H_0 δεν απορριφθεί, συμπεραίνεται ότι η μηδενική υπόθεση H_0 μπορεί και να είναι αληθής.

3. Συσχέτιση Δύο Μεταβλητών

Το κλασσικό μοντέλο της απλής γραμμικής παλινδρόμησης (που θα εξεταστεί στη συνέχεια), απαιτεί μόνο ότι η εξαρτημένη μεταβλητή Y δέχεται τυχαίες τιμές. Στο μοντέλο αυτό η ανεξάρτητη μεταβλητή X παίρνει ορισμένες τιμές.

Ωστόσο, είναι εφικτό να βρεθεί η ευθεία γραμμικής παλινδρόμησης ακόμα και όταν η μεταβλητή X δέχεται τυχαίες τιμές, μέσω του *συντελεστή συσχέτισης*.

3.1. Συντελεστής συσχέτισης ρ

Το μοντέλο συσχέτισης, περιγράφει το πώς συσχετίζονται δυο μεταβλητές X και Y, οι οποίες θεωρούνται τυχαίες. Οι μεταβλητές αυτές έχουν μια κατανομή από κοινού, η οποία καλείται δισδιάστατη κανονική κατανομή, με παραμέτρους τις σ_1, σ_2, μ_1, μ_2, και ρ.

Ο συντελεστής ρ, καλείται *συντελεστής συσχέτισης* ***(correlation coefficient)*** και εκφράζει την ένταση της σχέσης μεταξύ των τυχαίων μεταβλητών X και Y. Ο συντελεστής ρ, δίνεται από την εξίσωση:

$$\rho = \frac{E[(x-\mu_1)(y-\mu_2)]}{\sqrt{E[(x-\mu_1)^2]E[(y-\mu_2)^2]}}$$

όπου:

- $\mu_1 = E(x)$ και
- $\mu_2 = E(y)$.

Επιπροσθέτως, ο *δειγματικός συντελεστής συσχέτισης* ***(sample correlation coefficient)*** r, είναι ο εκτιμητής του συντελεστή συσχέτισης ρ και προκύπτει από τις ακόλουθες εκφράσεις:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{[n\sum {x_i}^2 - (\sum x_i)^2][n\sum {y_i}^2 - (\sum y_i)^2]}}$$

ή εναλλακτικά

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

όπου:

$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{1}{n}(\sum_{i=1}^{n} x_i)^2$$

$$S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2$$

Ένα άλλο μέρος που εκφράζει την ένταση μεταξύ των μεταβλητών X και Y, είναι η *συνδιακύμανση (covariance)*. Θεωρώντας ότι ισχύουν οι σχέσεις: μ_1 = E(x) και μ_2 = E(y), τότε η συσχέτιση cov(X,Y) των μεταβλητών X και Y, εκφράζεται από την εξίσωση:

$$cov(X,Y) = E[(X-\mu_1)(Y-\mu_2)]$$

Το δεξί μέρος της ανωτέρω εξίσωσης, παρατηρούμε ότι ισοδυναμεί με τον αριθμητή της εξίσωσης που προσδιορίζει τον συντελεστή συσχέτισης ρ.

Επιπλέον, ο εκτιμητής της συνδιακύμανσης cov(X,Y) που καλείται δειγματική συνδιακύμανση εκφράζεται από τη σχέση:

$$c_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})$$

όπου:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Το μέγεθος του δείγματος n προσδιορίζεται από το ζεύγος (x_i,y_i). Στη περίπτωση που εξετάζεται ένας πληθυσμός (και όχι ένα δείγμα), στην εξίσωση της δειγματικής συνδιακύμανσης ο παρανομαστής (n-1) αντικαθιστάται από το μέγεθος του πληθυσμού.

3.1.1. Ιδιότητες του συντελεστή συσχέτισης r

Ο δειγματικός συντελεστής συσχέτισης r διακρίνεται από τις ακόλουθες ιδιότητες που τον χαρακτηρίζουν:

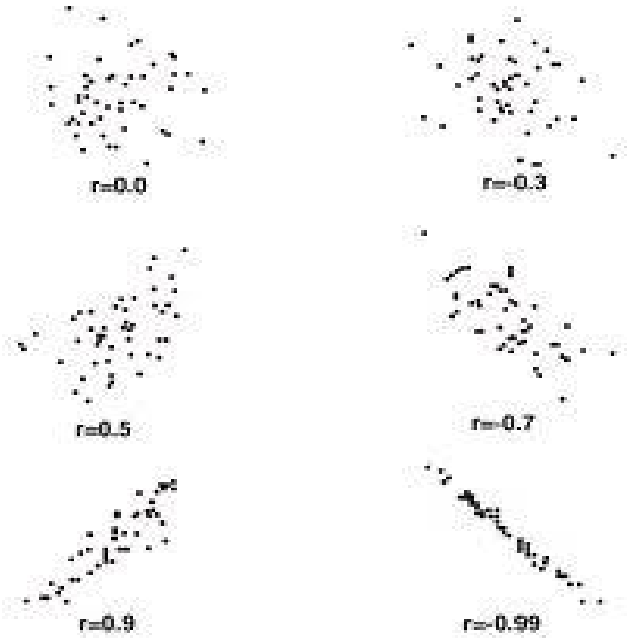
i. Η έκφραση $-1 \leq r \leq 1$, ισχύει πάντα.
ii. Ο συντελεστής συσχέτισης r έχει πάντα το ίδιο πρόσημο με τον εκτιμητή $\hat{\beta}$ (κλίση παλινδρόμησης).
iii. Ισχύουν πάντα οι ακόλουθες εξισώσεις:

$$\hat{\beta} = r\frac{\sqrt{S_{yy}}}{\sqrt{S_{xx}}}$$

$$s_e = \sqrt{\frac{S_{yy}(1-r^2)}{n-2}}$$

Στο σημείο αυτό παρατίθεται το στικτό διάγραμμα κάποιων δεδομένων για συγκεκριμένες καταστάσεις του δειγματικού συντελεστή συσχέτισης r. Ιδιαίτερο ενδιαφέρον παρουσιάζουν τα διαγράμματα για εκείνες τις τιμές του r που τείνουν προς τα άκρα του διαστήματος [-1,1] και τα οποία φανερώνουν ισχυρή (αρνητική και θετική αντίστοιχα) σχέση μεταξύ των δύο μεταβλητών, καθώς επίσης και για εκείνες τις τιμές

που το r τείνει και προς το 0 και για τις οποίες είναι φανερό ότι δεν υπάρχει γραμμική εξάρτηση μεταξύ των δύο μεταβλητών (Εικόνα 3).



Εικόνα 3: Στικτά Διαγράμματα

3.1.2. Έλεγχος υποθέσεων για το συντελεστή ρ

Προκειμένου να καθοριστεί το μέγεθος του δειγματικού συντελεστή συσχέτισης r που υποδηλώνει σημαντική σχέση μεταξύ των δυο μεταβλητών X και Y, πρέπει να γίνει έλεγχος υποθέσεων για το συντελεστή ρ, ο οποίος είναι ο συντελεστής συσχέτισης του πληθυσμού των ζευγών (X,Y). Οι μεταβλητές X και Y είναι δύο τυχαίες μεταβλητές.

Το στατιστικό είναι ο δειγματικός συντελεστής συσχέτισης r. Στη περίπτωση που η τιμή που επιστρέφει ο r είναι μεγαλύτερη από τη κατάλληλη τιμή, τότε η υπόθεση H_0 : ρ = 0 απορρίπτεται.

3.1.3. Συντελεστής προσδιορισμού r^2

Ο συντελεστής προσδιορισμού r^2 (coefficient of determination), εκφράζει το ποσοστό της παρατηρούμενης διακύμανσης της μεταβλητής Y, το οποίο μπορεί να αναλυθεί με το απλό μοντέλο γραμμικής παλινδρόμησης που καθορίζει μια προσεγγιστική σχέση μεταξύ των μεταβλητών X και Y.

Ο συντελεστής προσδιορισμού r^2, προσδιορίζεται από την ακόλουθη έκφραση:

$$r^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2}$$

Επιπλέον όπως είναι λογικό, ισχύει ότι ο συντελεστής προσδιορισμού ισούται με το τετράγωνο του δειγματικού συντελεστή συσχέτισης:

$$r^2 = \frac{(S_{xy})^2}{S_{xx}S_{yy}}$$

3.1.4. Παράδειγμα συντελεστή συσχέτισης και συντελεστή προσδιορισμού

Το παράδειγμα που ακολουθεί αναφέρεται στην επίδραση της αύξησης της τιμής του πετρελαίου, πάνω στη τιμή λιανικής πώλησης της βενζίνης [5]. Το ζητούμενο είναι ο υπολογισμός του συντελεστή συσχέτισης (r) και του συντελεστή προσδιορισμού (r^2).

Στον πίνακα που ακολουθεί, εμφανίζονται τα δεδομένα μας.

Y **(τιμή μονάδας βενζίνης ανα λίτρο)**	X **(τιμή μονάδας πετρελαίου ανα βαρέλι)**
16,80	4,15
25,30	10,38
30,20	10,89
35,30	11,96
42,40	12,46
65,20	17,72
75,00	28,07
80,20	36,11

Με τη χρήση κάποιου υπολογιστικού εργαλείου (όπως το excel), θα υπολογίσουμε τις απαραίτητες συναρτήσεις που αναλύθηκαν νωρίτερα, προκειμένου να υπολογίσουμε τον συντελεστή συσχέτισης και τον συντελεστή προσδιορισμού.

Ακολούθως παρουσιάζονται τα αποτελέσματα των επιμέρους αθροισμάτων για το ανωτέρω παράδειγμα:

$$\sum_{i=1}^{8} X_i = 370{,}40$$

$$\sum_{i=1}^{8} Y_i = 131{,}74$$

$$\sum_{i=1}^{8} X_i Y_i = 7.768{,}32$$

$$\sum_{i=1}^{8} X_i^2 = 21.186{,}30$$

$$\sum_{i=1}^{8} Y_i^2 = 2.947{,}71$$

Εν συνεχεία έχουμε ότι:

$$S_{xx} = \sum_{i=1}^{8} x_i^2 - \frac{1}{8}\left(\sum_{i=1}^{8} x_i\right)^2 = 4.036{,}78$$

$$S_{yy} = \sum_{i=1}^{8} y_i^2 - \frac{1}{8}(\sum_{i=1}^{8} y_i)^2 = 778{,}28$$

$$S_{xy} = \sum_{i=1}^{8} x_i y_i - \frac{1}{8}\sum_{i=1}^{8} x_i \sum_{i=1}^{8} y_i = 1.668{,}76$$

Με βάση τις τιμές που προέκυψαν από τους υπολογισμούς, έχουμε ότι ο συντελεστής συσχέτισης είναι:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = 0{,}94$$

Ακολούθως υπολογίζουμε και τον συντελεστή προσδιορισμού, ο οποίος είναι:

$$r^2 = \frac{(S_{xy})^2}{S_{xx}S_{yy}} = 0{,}89$$

3.2. Συντελεστής συσχέτισης του Spearman

Στη περίπτωση που οι μεταβλητές X και Y είναι ποσοτικές (δηλαδή παίρνουν πραγματικές τιμές), τότε το καλύτερο στατιστικό μέτρο για τη μέτρηση του βαθμού εξάρτησης είναι ο συντελεστής συσχέτισης r ή ο συντελεστής προσδιορισμού r^2, οι οποίοι εξετάστηκαν στις προηγούμενες παραγράφους. Στη περίπτωση ωστόσο που οι μεταβλητές X και Y είναι ποιοτικές (δηλαδή δεν είναι ποσοτικά μετρήσιμα μεγέθη, αλλά κάθε μεταβλητή επιτρέπει τη διάταξη των n μονάδων του πληθυσμού κατά τάξη μεγέθους), για να μελετηθεί η συσχέτιση μεταξύ τους δεν χρησιμοποιούνται οι αρχικές τιμές των μεταβλητών X και Y, αλλά η σειρά κατάταξης των μονάδων του προς έρευνα πληθυσμού ως προς τη μεταβλητή X και Y.

Ένα μέτρο συσχέτισης μεταξύ τάξεων, είναι ο συντελεστής συσχέτισης του Spearman (r_s), ο οποίος περιγράφεται από την εξίσωση:

$$r_s = \frac{CC_{uv}}{\sqrt{CC_{uu}CC_{vv}}}$$

όπου:

$$CC_{uv} = \sum_{i=1}^{n}(u_i - \overline{u})(v_i - \overline{v}) = \sum_{i=1}^{n} u_i v_i - \frac{1}{n}\sum_{i=1}^{n} u_i \sum_{i=1}^{n} v_i$$

$$CC_{uu} = \sum_{i=1}^{n}(u_i - \overline{u})^2 = \sum_{i=1}^{n} u_i^2 - \frac{1}{n}(\sum_{i=1}^{n} u_i)^2$$

$$CC_{vv} = \sum_{i=1}^{n}(v_i - \overline{v})^2 = \sum_{i=1}^{n} v_i^2 - \frac{1}{n}(\sum_{i=1}^{n} v_i)^2$$

- u_i: η τάξη της i παρατήρησης στο 1^o δείγμα,
- v_i: η τάξη της i παρατήρησης στο 2^o δείγμα και
- n: ο αριθμός από τα ζεύγη παρατηρήσεων.

Η μορφή με την οποία συναντάται συχνότερα ο συντελεστής συσχέτισης του Spearman είναι:

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

όπου $d_i = u_i - v_i$: οι διαφορές στη σειρά κατάταξης των μονάδων του πληθυσμού των μεταβλητών X και Y.

3.2.1. **Παράδειγμα συντελεστή συσχέτισης του** Spearman

Για να γίνει καλύτερα κατανοητός ο συντελεστής του Spearman, δίνεται το ακόλουθο παράδειγμα:

Έστω ότι δέκα φοιτητές, εξετάστηκαν σε δύο διαγωνίσματα (X και Y), και βαθμολογήθηκαν ως εξής:

Φοιτητές	**Διαγώνισμα Χ**	**Διαγώνισμα Υ**
Α	5	6
Β	3	2
Γ	4	5
Δ	6	4
Ε	7	9
Ζ	10	7
Η	8	10
Θ	1	3
Ι	2	1
Κ	9	8

Εν συνεχεία ταξινομούμε τα δεδομένα μας για τη μεταβλητή Χ όπως εμφανίζονται στον ακόλουθο πίνακα.

Φοιτητές	**Διαγώνισμα Χ**	**Σειρά κατάταξης της μεταβλητής Χ**
Ζ	10	1
Κ	9	2
Η	8	3
Ε	7	4
Δ	6	5
Α	5	6

Γ	4	7
Β	3	8
Ι	2	9
Θ	1	10

Ομοίως δουλεύουμε και για τη ταξινόμηση των δεδομένων για τη μεταβλητή Υ.

Φοιτητές	**Διαγώνισμα Υ**	**Σειρά κατάταξης της μεταβλητής Υ**
Η	10	1
Ε	9	2
Κ	8	3
Ζ	7	4
Α	6	5
Γ	5	6
Δ	4	7
Θ	3	8
Β	2	9
Ι	1	10

Με βάση τους δύο ανωτέρω πίνακες που εξήχθησαν, κατασκευάζουμε τον ακόλουθο πίνακα:

Φοιτητές	**Σειρά κατάταξης της μεταβλητής X**	**Σειρά κατάταξης της μεταβλητής Y**	**di=|Xi-Yi|**	**di*di**
Z	1	4	3	9
K	2	3	1	1
H	3	1	2	4
E	4	2	2	4
Δ	5	7	2	4
A	6	5	1	1
Γ	7	6	1	1
B	8	9	1	1
I	9	10	1	1
Θ	10	8	2	4
				30

Άρα λοιπόν ο συντελεστής συσχέτισης του Spearman είναι:

$$r_s = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2 = 1 - \frac{6}{10(100-1)}30 = 0{,}82$$

3.3. **Συντελεστής συσχέτισης του** Kendall

Ο συντελεστής συσχέτισης του Kendall εκφράζεται από την εξίσωση

$$t_k = \frac{2A}{C_2^n}$$

όπου

- Α: ο αριθμός των αντιστρόφων και
- C_2^n: ο αριθμός των δυνατών συγκρίσεων, δηλαδή των μονάδων του πληθυσμού ανά δύο.

Η αντιστροφή ισχύει όταν:

- $x_i < x_j$ και $y_i > y_j$ ή
- $x_i > x_j$ και $y_i < y_j$

Η διαδικασία υπολογισμού των αντιστροφών απλοποιείται, αν οι μονάδες της μιας μεταβλητής τεθούν κατά φυσική διάταξη και υπολογιστούν οι αντίστροφες στις μονάδες της άλλης μεταβλητής, δηλαδή οι περιπτώσεις στις οποίες ένας βαθμός προηγείται ενός άλλου μικρότερου.

3.3.1. Παράδειγμα συντελεστή συσχέτισης του Kendall

Έστω ότι έχουμε τα ακόλουθα δεδομένα σχετικά με τη βαθμολογία 10 μαθητών σε 2 μαθήματα:

Φοιτητές	**Διαγώνισμα Χ**	**Διαγώνισμα Υ**
Α	15	7
Β	5	15
Γ	12	20
Δ	19	18
Ε	17	9
Ζ	14	10
Η	15	12

Θ	9	18
I	16	16
K	15	14

Όπως και στον συντελεστή του Spearman έτσι και στον συντελεστή του Kendall, ταξινομούμε τα δεδομένα μας διαδοχικά. Έτσι προκύπτει:

Φοιτητές	**Διαγώνισμα X**	**Σειρά κατάταξης για X**
Δ	19	1
E	17	2
I	16	3
A	15	4
H	15	4
K	15	4
Z	14	5
Γ	12	6
Θ	9	7
B	5	8

και

Φοιτητές	Διαγώνισμα Υ	Σειρά κατάταξης για Υ
Γ	20	1
Δ	18	2
Θ	18	2
Ι	16	3
Β	15	4
Κ	14	5
Η	12	6
Ζ	10	7
Ε	9	8
Α	7	9

Με βάση τους 2 πίνακες που προηγούνται κατασκευάζουμε το τελευταίο πίνακα απ' όπου θα εξάγουμε τις αντιστροφές.

Φοιτητές	Σειρά κατάταξης για Χ	Σειρά κατάταξης για Υ
Δ	1	2
Ε	2	8
Ι	3	3
Α	4	9
Η	4	6
Κ	4	5

Z	5	7
Γ	6	1
Θ	7	2
Β	8	4

Με βάση τον ανωτέρω πίνακα προκύπτουν 23 αντιστροφές, από τις ακόλουθες συγκρίσεις:

1. (Ε, Ι)
2. (Ε, Η)
3. (Ε, Κ)
4. (Α, Ζ)
5. (Ε, Ζ)
6. (Ζ, Γ)
7. (Κ, Γ)
8. (Η, Γ)
9. (Α, Γ)
10. (Ι, Γ)
11. (Ε, Γ)
12. (Δ, Γ)
13. (Ζ, Θ)
14. (Κ, Θ)
15. (Η, Θ)
16. (Α, Θ)
17. (Ι, Θ)
18. (Ε, Θ)

19. (Z, B)
20. (K, B)
21. (H, B)
22. (A, B)
23. (E, B)

Άρα λοιπόν ο συντελεστής του Kendall ισούται με:

$$t_k = \frac{2A}{C_2^n} = \frac{2 * 23}{C_2^{10}} = \frac{46}{115} = 0{,}4$$

4. Απλή Γραμμική Παλινδρόμηση

Σε πολλά πραγματικά προβλήματα είναι απαραίτητη η μελέτη συγχρόνως δύο ή και περισσότερων μεταβλητών, για την εξαγωγή χρήσιμων συμπερασμάτων και την ορθή λήψη αποφάσεων (πχ εκ μέρους των διοικούντων μιας επιχείρησης).

Το βασικό πρόβλημα, το οποίο πραγματεύεται το παρών κεφάλαιο είναι, η πρόβλεψη των τιμών μιας μεταβλητής με βάση τις τιμές μιας άλλης μεταβλητής (πρόβλημα γραμμικής παλινδρόμησης).

Σύμφωνα με το προσδιοριστικό μοντέλο (deterministic model), μια ακριβής γραμμική σχέση δύο μεταβλητών x και y εκφράζεται από την εξίσωση

$$y = \alpha x + \beta$$

Από την άλλη, σύμφωνα με το πιθανοθεωρητικό μοντέλο (probabilistic model), η σχέση μεταξύ δύο μεταβλητών x και y εκφράζεται από την εξίσωση

$$y = \alpha + \beta x + \varepsilon$$

όπου:

- Η μεταβλητή **x** καλείται, *ανεξάρτητη* μεταβλητή (independent variable),
- Η μεταβλητή **y** καλείται, *εξαρτημένη* μεταβλητή (dependent variable),
- Το **ε** είναι μια *συνεχής τυχαία* μεταβλητή, που υποθέτουμε ότι έχει μέση τιμή Ε(ε)=0 και διακύμανση σ^2.

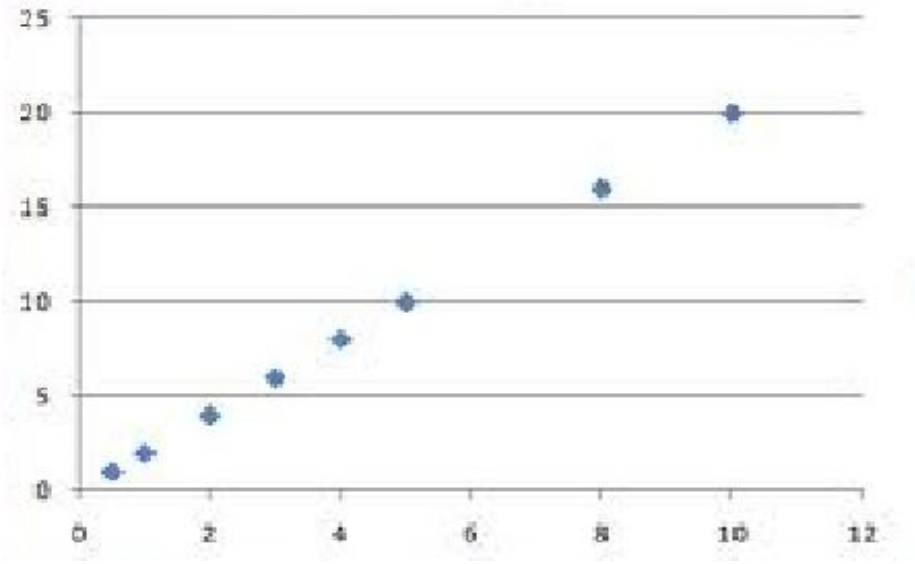
Το μοντέλο της απλής γραμμικής παλινδρόμησης (simple linear regression model), που θα εξεταστεί και στο παρών κεφάλαιο, είναι στην ουσία ένα πιθανοθεωρητικό μοντέλο της μορφής:

$$y = \alpha + \beta x + \varepsilon$$

όπου:

- Η ανεξάρτητη μεταβλητή **x** παίρνει καθορισμένες τιμές,
- Η εξαρτημένη μεταβλητή **y** λαμβάνει τιμές με βάση τις τιμές που παίρνει η **x**,
- Η τυχαία μεταβλητή **ε**, έχει κανονική κατανομή με μέση τιμή Ε(ε)=0 και διακύμανση σ^2,
- Οι παράμετροι α, β είναι άγνωστοι και γίνεται χρήση εκτιμητών με βάση τα διαθέσιμα δεδομένα.

Η γραμμική εξάρτηση μιας μεταβλητής y από μια μεταβλητή x, θα μπορούσε εύκολα να διαφανεί σε κάποιες περιπτώσεις με τη χρήση του στικτού διαγράμματος (Εικόνα 4).



Εικόνα 4: **Στικτό διάγραμμα γραμμικής εξάρτησης δυο μεταβλητών**

Όπως παρατηρείται και στο διάγραμμα, υπάρχει μια γραμμική εξάρτηση μεταξύ των μεταβλητών x και y. Πιο συγκεκριμένα καθώς αυξάνονται οι τιμές που δέχεται η μεταβλητή x, αυξάνονται και οι τιμές που επιστέφει η μεταβλητή y.

4.1. Συνθήκες της Απλής Γραμμικής Παλινδρόμησης

Όπως έχει ήδη αναφερθεί, για τον προσδιορισμό ενός μοντέλου και τη μετέπειτα ερμηνεία της συμπεριφοράς της μεταβλητής Y με βάση τη συμπεριφορά των μεταβλητών $X_1, X_2, \ldots, X_n$, η μεταβλητή Y θα καλείται εξαρτημένη μεταβλητή και οι

μεταβλητές X_1, X_2, …, X_n, θα καλούνται ανεξάρτητες ή ερμηνευτικές μεταβλητές. Η εκτίμηση ενός στατιστικού μοντέλου πραγματοποιείται με βάση την ανάλυση στατιστικών δεδομένων, δηλαδή με βάση τις παρατηρήσεις της εξαρτημένης μεταβλητής Y σε επιλεγμένα επίπεδά της ή των ερμηνευτικών μεταβλητών. Για το σκοπό αυτό έχουν αναπτυχθεί στατιστικές τεχνικές και οι οποίες αναφέρονται ως ανάλυση παλινδρόμησης [10].

Για τη τυχαία μεταβλητή Y με ερμηνευτικές μεταβλητές X_1, X_2, …, X_n, η γενική μορφή ενός μοντέλου παλινδρόμησης είναι:

$$Y = f(X_1, X_2, \dots, X_n) + \varepsilon$$

όπου

$$E(Y) = f(X_1, X_2, \dots, X_n) \quad και \quad E(\varepsilon) = 0$$

Η τυχαία μεταβλητή εκφράζεται ως το άθροισμα μιας σχέσης που εκφράζει τη μέση τιμή της εξαρτημένης μεταβλητής (ως συνάρτηση των ερμηνευτικών μεταβλητών X_1, X_2, …, X_n) και ενός τυχαίου όρου. Η συνάρτηση f (X_1, X_2, …, X_n,) καλείται συνάρτηση παλινδρόμησης Y επί των X_1, X_2, …, X_n. Στη περίπτωση τώρα που το μοντέλο που εξετάζεται είναι μιας μορφής όπου η μεταβλητή Y είναι γραμμική συνάρτηση των παραμέτρων του μοντέλου, τότε βρισκόμαστε στη περίπτωση του γραμμικού μοντέλου. Η πιο απλή μορφή μιας τέτοιας σχέσης είναι:

$$Y = \alpha + \beta X + \varepsilon$$

όπου

$E(\varepsilon) = 0$

και οι α, β είναι σταθερές (με εκτιμητές $\hat{a}$ και $\hat{\beta}$).

Η μέση τιμή της Υ για ορισμένη τιμή της Χ, βρίσκεται πάνω σε μια ευθεία με σταθερό όρο α και κλίση β. Όταν το μοντέλο περιέχει περισσότερες από μια ερμηνευτικές μεταβλητές, καλείται πολυμεταβλητό γραμμικό μοντέλο και έχει τη γενική μορφή:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

όπου

$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon$$

Το *α* είναι ο σταθερός όρος της συνάρτησης παλινδρόμησης, δηλαδή η τιμή της Υ όταν $X_1 = X_2 = \ldots = X_n = 0$. Ο συντελεστής β είναι η μεταβολή της Υ όταν η ερμηνευτική μεταβλητή Χ αυξηθεί κατά μια μονάδα και οι υπόλοιπες μεταβλητές παραμένουν σταθερές[4].

Όσον αφορά τη διαδικασία εκτίμησης του μοντέλου, μελετώνται τα ακόλουθα στάδια:

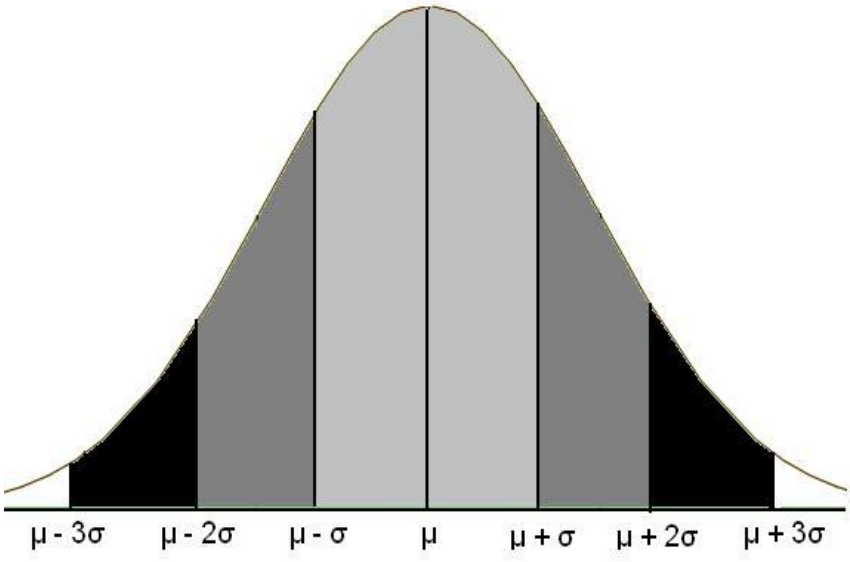
- Επιλογή των ερμηνευτικών μεταβλητών με βάση την αντίστοιχη θεωρία ή τη γνώση που κατέχουμε για τη διαδικασία παραγωγής των τιμών της Υ. Κάποιες από αυτές μπορούν να κριθούν ως περιττές ή να προστεθούν επιπλέον ερμηνευτικές μεταβλητές.
- Επιλογή του κατάλληλου μαθηματικού τύπου για το προσδιορισμό μέρους της σχέσης της Υ με την ερμηνευτική μεταβλητή.
- Επιλογή τιμών για τις παραμέτρους του μοντέλου, έτσι ώστε αυτό να εκτιμά τις παρατηρήσεις Y_1, Y_2, …, Y_n, ορθότερα σύμφωνα με κάποιο κριτήριο. Στο γραμμικό μοντέλο, όταν η συμπεριφορά του τυχαίου μέρους ικανοποιεί ορισμένες συνθήκες, επιλέγεται ως κριτήριο καλής προσαρμογής, η ελαχιστοποίηση του αθροίσματος για όλες τις τιμές της μεταβλητής Υ των τετραγωνικών σφαλμάτων εκτίμησης. Ισοδύναμα επιλέγεται ως μέθοδος εκτίμησης η μέθοδος των ελαχίστων τετραγώνων.

[4] http://www.stat-athens.aueb.gr/~jpan/diatrives/Mpouras/chapter6.pdf

- Τα αποτελέσματα που εξάγονται από το προηγούμενο στάδιο, γενικεύονται σ' όλες τις παρατηρήσεις Y_i της μεταβλητής Y, και οι οποίες θα μπορούσαν να είχαν συλλεχθεί με τον ίδιο τρόπο που συλλέχθηκαν οι παρατηρήσεις των δεδομένων μας.
- Έλεγχος της καταλληλότητας του εκτιμημένου μοντέλου. Στην ουσία εξετάζεται το κατά πόσο οι υποθέσεις που έγιναν για τη συνάρτηση παλινδρόμησης και για τον όρο σφάλματος, στηρίζονται από τα δεδομένα ή αν θα πρέπει να υπάρξουν τροποποιήσεις στην εξειδίκευση του μοντέλου. Οι τροποποιήσεις αυτές, μπορεί να υπαγορεύουν κάποια άλλη μέθοδο εκτίμησης (πέραν των ελαχίστων τετραγώνων).

Για να είναι ορθό το μοντέλο της απλής γραμμικής παλινδρόμησης, θα πρέπει να ισχύουν οι ακόλουθες συνθήκες:

- Τα σφάλματα ε_i που αντιστοιχούν στη ποσότητα $\varepsilon_i = y_i - \alpha - \beta x_i$, i=1(1)n, πρέπει να έχουν κανονική κατανομή με μέσο $E(\varepsilon_i)=0$ και διακύμανση σ^2. Πρέπει να έχουν μια γραφική απεικόνιση, όπως αυτή αποδίδεται στην ακόλουθη εικόνα.



Εικόνα 3: Κανονική Κατανομή

Το 1809 ο Γερμανός μαθηματικός Carl F. Gauss, διαπίστωσε ότι τα σφάλματα που γίνονται σε αστρονομικές παρατηρήσεις μπορούν να περιγραφούν ικανοποιητικά από τη κανονική κατανομή. Εν συνεχεία, διαπιστώθηκε ότι τα τυχαία σφάλματα (όχι τα συστηματικά) που εμφανίζονται σε διάφορες μετρήσεις ακολουθούν με ικανοποιητική προσέγγιση κανονική κατανομή. Για αυτό το λόγο, η κανονική κατανομή, καλείται επίσης και κατανομή σφαλμάτων (law of errors ή και Gaussian distribution)[5].

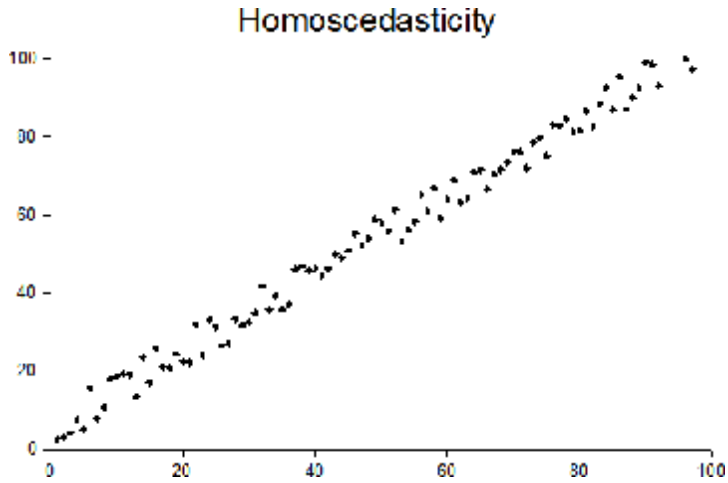
Το ιδιαίτερο χαρακτηριστικό γνώρισμα της κανονικής κατανομής, είναι η γραφική απεικόνισή της με την κανονική καμπύλη, η οποία διαθέτει κωδωνοειδή μορφή, είναι συμμετρική και τόσο το αριστερό όσο και το δεξί άκρο της τείνουν να ακουμπήσουν ασυμπτωτικά τον οριζόντιο άξονα (xx΄).

Το σύνολο των εφαρμογών που ακολουθούν τη κανονική κατανομή, βασίζεται στο *Κεντρικό Οριακό Θεώρημα* της θεωρίας πιθανοτήτων, σύμφωνα με το οποίο κάθε ποσότητα της οποίας η τιμή μπορεί να θεωρηθεί ότι διαμορφώνεται από ένα μεγάλο αριθμό ανεξάρτητων παραγόντων ή μεταβλητών, ακολουθεί προσεγγιστικά τη κανονική κατανομή. Οι ανεξάρτητες μεταβλητές είναι αυτοί που παίζουν σημαντικό ρόλο σε μια παρατήρηση και όπως είναι λογικό την επηρεάζουν σε διαφορετικό βαθμό, χωρίς ωστόσο να επηρεάζει ο ένας τον άλλο.

- Η ανεξάρτητη μεταβλητή x πρέπει να λαμβάνει ορισμένες ή τυχαίες τιμές. Στη περίπτωση που οι τιμές της μεταβλητής x είναι τυχαίες, τότε προκύπτει το μοντέλο απλής γραμμικής παλινδρόμησης που χρησιμοποιείται για την ανάλυση συσχέτισης (όπως θα παρουσιαστεί σε επόμενη ενότητα)
- Για κάθε τιμή της ανεξάρτητης μεταβλητής x, υπάρχει ένας συγκεκριμένος πληθυσμός της μεταβλητής y. Πρακτικά αυτό σημαίνει ότι, για κάθε τιμή x_i, της μεταβλητής x, η εξαρτημένη μεταβλητή y ακολουθεί τη κατανομή $f(y/x_i)$ (όπου οι κατανομές $f(y/x_i)$ είναι κανονικές κατανομές) με μέσο $\mu_{y/xi}$ και διακύμανση σ^2, η οποία είναι και σταθερή για κάθε τιμή της x_i. Η παραπάνω συνθήκη

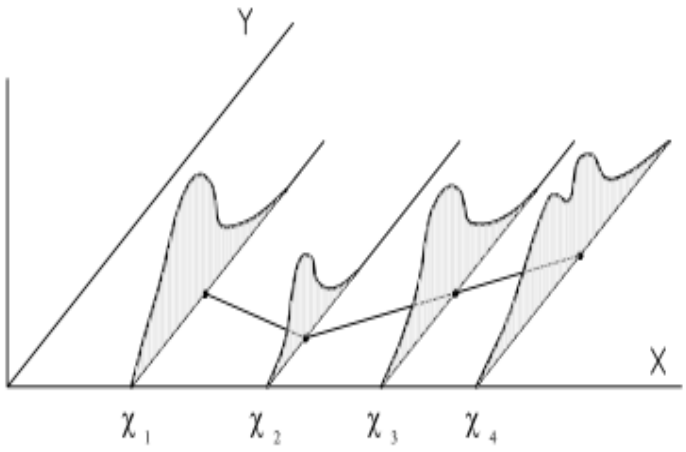
[5] http://www.aua.gr/gpapadopoulos/files/normal010-2.pdf

συναντάται συχνά και ως υπόθεση ομοσκεδαστικότητας – σταθερότητα διασποράς (homoscedasticity – variance stability).



Εικόνα 4: Ομοσκεδαστικότητα

Στην εικόνα που ακολουθεί βλέπουμε ένα παράδειγμα παραβίασης της συνθήκης ομοσκεδαστικότητας, αφού όπως παρατηρούμε η διασπορά της Y στο επίπεδο x_2, είναι μικρότερη από τη διασπορά της Y στο επίπεδο x_1.



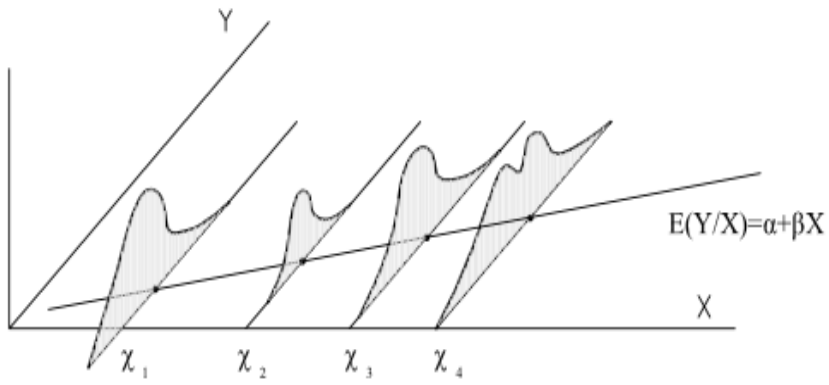
Εικόνα 5: Παραβίαση συνθήκης ομοσκεδαστικότητας

- Για κάθε τιμή της μεταβλητής x_i, i=1(1)n, οι μέσες τιμές $f(y/x_i)$ των πληθυσμών της μεταβλητής y, βρίσκονται σε μια ευθεία γραμμή. Αυτό το φαινόμενο

ονομάζεται και υπόθεση γραμμικότητας (linearity) και εκφράζεται από την εξίσωση:

$$E\left(\frac{Y}{x_i}\right) = \alpha + \beta x_i$$

Όπου οι παράμετροι α και β εκτιμώνται από το δείγμα (x_i, y_i). Με άλλα λόγια, υποθέτουμε ότι οι μέσες τιμές της Y (για τα διάφορα επίπεδα της X), είναι γραμμικές συναρτήσεις της X (βρίσκονται σε ευθεία γραμμή – βλ. εικόνα).



Εικόνα 6: Υπόθεση γραμμικότητας

Να σημειωθεί ότι στο μοντέλο $Y = \alpha + \beta X + \varepsilon$, οι μόνες τυχαίες μεταβλητές είναι η Y και η ε.

- Τέλος, θα πρέπει να ισχύει η συνθήκη της ανεξαρτησίας (independence). Όπως έχει ήδη αναφερθεί η τιμές της μεταβλητής y είναι ανεξάρτητες, δηλαδή οι τιμές που προκύπτουν για τη μεταβλητή y από τη τιμή x_i, δεν εξαρτώνται από τις τιμές που προκύπτουν για την μεταβλητή y από τις τιμές x_j.

4.2. Σημειακή Εκτίμηση της Απλής Γραμμικής Παλινδρόμησης

Στην ενότητα αυτή θα εξεταστούν η εκτίμηση των παραμέτρων α, β, με τη μέθοδο τον ελαχίστων τετραγώνων και εν συνεχεία η εκτίμηση της διακύμανσης (σ^2).

4.2.1. Εκτίμηση των Παραμέτρων α,β, με τη Μέθοδο των Ελαχίστων Τετραγώνων

Στο σημείο αυτό παρουσιάζεται η μεθοδολογία που ακολουθείται, για την εκτίμηση των παραμέτρων α και β στο απλό μοντέλο γραμμικής παλινδρόμησης, που εκφράζεται από την εξίσωση:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad i = 1(1)n$$

Με βάση τα ζεύγη τιμών (x_i, y_i), όπου αντιπροσωπεύουν τις n τιμές των μεταβλητών X και Y, και σύμφωνα με τη μέθοδο των ελαχίστων τετραγώνων, θα γίνει η εκτίμηση των συντελεστών α και β.

Αρχικά, υποθέτουμε ότι οι τυχαίες ποσότητες ε_i, ακολουθούν τη κανονική κατανομή με μέση τιμή $E(\varepsilon_i)=0$ και διακύμανση σ^2, καθώς και ότι είναι στατιστικά ανεξάρτητες. Επιπλέον στη περίπτωση που ισχύει το μοντέλο της απλής γραμμικής παλινδρόμησης, οι μέσες τιμές E(y) για κάποια τιμή x_i, βρίσκονται σε κάποια ευθεία γραμμή της μορφής:

$$E\left(\frac{Y}{x_i}\right) = \alpha + \beta x_i$$

Εν συνεχεία, θέτοντας τη ποσότητα:

$$\hat{y}_i = E\left(\frac{Y}{x_i}\right)$$

Έπειτα με τη χρήση της μεθόδου ελαχίστων τετραγώνων, αναζητούμε τους εκτιμητές των συντελεστών α και β που τους συμβολίζουμε $\hat{\alpha}$ και $\hat{\beta}$ αντιστοίχως. Η ευθεία στην οποία θα γίνει η εκτίμηση των συντελεστών είναι η:

$$\hat{y}_i = \alpha + \beta x_i$$

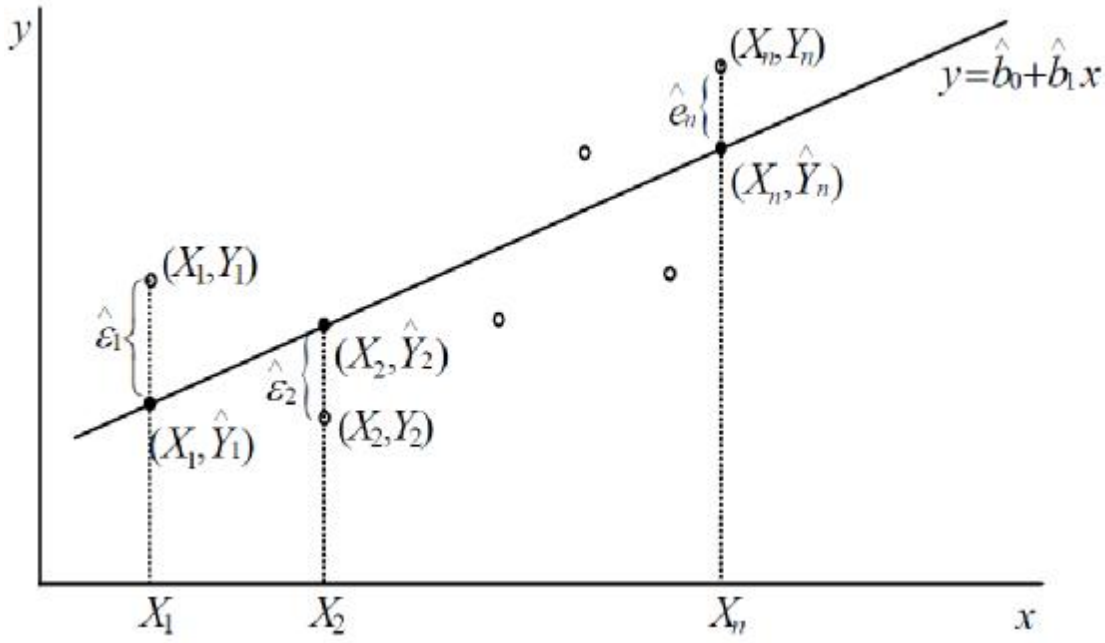
Οι προβλέψεις των Y_i (Y predicted) ή προσαρμοσμένες τιμές των Y_i πάνω στην εκτιμημένη ευθεία γραμμικής παλινδρόμησης[6], καλούνται οι εκτιμήσεις των:

$$E(Y_i) = \alpha + \beta X_i$$

Η εξίσωση που περιγράφει αυτές τις προβλέψεις είναι:

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i = \bar{Y} + \hat{\beta}(X_i - \bar{X})$$

Οι ακριβείς τιμές και οι προβλέψεις αυτών παρουσιάζονται στην ακόλουθη εικόνα:



Εικόνα 7: **Πραγματικές τιμές και προβλέψεις – εκτιμήσεις αυτών**

[6] http://www.unipi.gr/faculty/mbouts/statprog/SPSS_lesson9-10.pdf

4.2.1.1. *Μέθοδος των ελαχίστων τετραγώνων*

Στο σημείο αυτό θα παρουσιαστεί η μέθοδος των ελαχίστων τετραγώνων σύμφωνα με την οποία θα γίνει ο προσδιορισμός μιας εκτίμησης $\hat{y}_i = \hat{a} + \hat{\beta}x_i$ της ευθείας $E\left(\frac{Y}{x_i}\right) = \alpha + \beta x_i$, όπου $\hat{a}, \hat{\beta}$ οι εκτιμήτριες των α και β αντίστοιχα.

Το άθροισμα των τετραγώνων των κατακόρυφων διαφορών από τα σημεία (x_i, y_i), i=1(1)n, προς την ευθεία $\hat{y}_i = \alpha + \beta x_i$, δίνεται από την έκφραση:

$$G(\alpha, \beta) = \sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

Οι εκτιμητές $\hat{a}$ και $\hat{\beta}$ των συντελεστών α και β αντίστοιχα, είναι εκείνες οι τιμές των α και β για τις οποίες ελαχιστοποιείται η ποσότητα *G(α,β)*. Οι τιμές $\hat{a}$ και $\hat{\beta}$ καλούνται εκτιμητές ελαχίστων τετραγώνων και η ευθεία ελαχίστων τετραγώνων δίνεται εκφράζεται από την εξίσωση:

$$\hat{y}_i = \hat{a} + \hat{\beta}x_i$$

Επιπροσθέτως, τα σφάλματα e_i, δίνονται από τη σχέση:

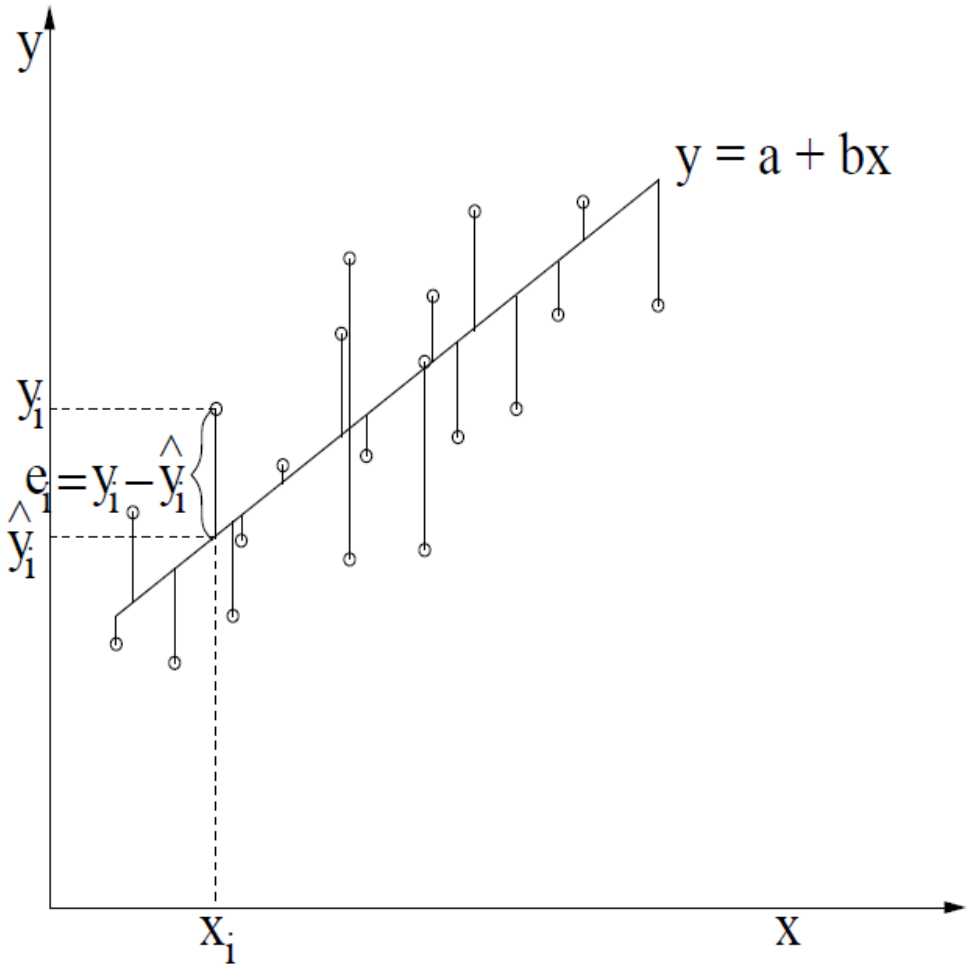
$$e_i = y_i - \hat{y}_i$$

Σε σχέση με τα σφάλματα πρέπει να ειπωθεί ότι, για κάθε τιμή x_i, εκτιμάται η προσεγγιστική τιμή $\hat{y}_i$, η οποία (έστω και με πολύ μικρή απόκλιση) είναι διαφορετική από τη πραγματική τιμή y_i.

Η διαφορά της προσεγγιστικής από τη πραγματική τιμή, όπως αναφέρθηκε καλείται σφάλμα. Στη γενική μορφή τους τα σφάλματα εκφράζονται από τη την ισότητα:

$$e_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{\beta}x_i$$

Γραφικά αυτό μεταφράζεται, ως η κατακόρυφη απόσταση της πραγματική τιμής από την ευθεία ελαχίστων τετραγώνων και η οποία ονομάζεται σφάλμα ελαχίστων τετραγώνων ή υπόλοιπο (residual).



Εικόνα 8: Ευθεία ελαχίστων τετραγώνων και υπόλοιπα

Στις ενότητες που ακολουθούν, θα διαφανεί πως με τη χρήση των υπολοίπων, προκύπτει η εκτίμηση της διασποράς του σφάλματος.

4.2.1.2. Εύρεση των εκτιμητών

Το ζητούμενο εδώ είναι να βρεθούν οι τιμές των εκτιμητών $\hat{\alpha}$ και $\hat{\beta}$ που ελαχιστοποιούν τη ποσότητα $G(\alpha,\beta)$.

Βρίσκοντας τις μερικές παραγώγους της *G(α,β)*, ως προς α και β και εξισώνοντας με το μηδέν, οδηγούμαστε στις ακόλουθες δύο εξισώσεις, που καλούνται *κανονικές εξισώσεις*.

$$\sum_{i=1}^{n} y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\alpha}\sum_{i=1}^{n} x_i + \hat{\beta}\sum_{i=1}^{n} x_i^2$$

Επιλύοντας το σύστημα των κανονικών εξισώσεων προκύπτουν οι λύσεις των εκτιμητών $\hat{\alpha}$ και $\hat{\beta}$, οι οποίες είναι:

$$\hat{\beta} = \frac{\sum xy - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

όπου:

$$\bar{y} = \frac{\sum y_i}{n}$$

$$\bar{x} = \frac{\sum x_i}{n}$$

Εναλλακτικά οι παραπάνω εξισώσεις μπορούν να γραφτούν στη μορφή:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\alpha} = \frac{\sum y_i - \hat{\beta}\sum x_i}{n}$$

όπου:

$$S_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

Επομένως, η εκτίμηση ελαχίστων τετραγώνων $\hat{y}_i = \hat{a} + \hat{\beta} x_i$ της ευθείας παλινδρόμησης από το δείγμα των n ζευγών παρατηρήσεων, είναι η:

$$\hat{y}_i = \hat{a} + \hat{\beta} x_i = \bar{y} - \hat{\beta}\bar{x} + \hat{\beta} x_i = \bar{y} + \widehat{\beta}(x_i - \bar{x})$$

ή εναλλακτικά

$$\hat{y}_i = \bar{y} + \frac{S_{xy}}{S_{xx}}(x_i - \bar{x})$$

4.2.2. Εκτίμηση της διασποράς των σφαλμάτων παλινδρόμησης

Ένα μέσο για την αξιολόγηση της καλής προσαρμογής της εξίσωσης $\hat{y}_i = \hat{a} + \hat{\beta} x_i$ στο διάγραμμα διασποράς είναι το μέσο τετραγωνικό σφάλμα.

Υποθέτοντας ότι οι παρατηρήσεις – ζεύγη (x_i, y_i), i=1(1)n, ακολουθούν το μοντέλο της απλής γραμμικής παλινδρόμησης $Y_i = \alpha + \beta X_i + \varepsilon_i$, όπου οι τυχαίες μεταβλητές είναι μεταξύ τους ανεξάρτητες και ακολουθούν τη κανονική κατανομή, με $E(\varepsilon_i)$ και $V(\varepsilon_i)=\sigma^2$. Η ποσότητα σ^2, εκφράζεται από την εξίσωση:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{n}(y_i - \mu_{Y/x})^2$$

όπου:

$\mu_{Y/x} = E(Y/x) = \alpha + \beta x$.

Η εκτιμήτρια της $\mu_{Y/x}$ είναι η ευθεία $\hat{y} = \hat{a} + \hat{\beta}x$, όπου όπως έχει ήδη αναφερθεί οι συντελεστές $\hat{a}$ *και* $\hat{\beta}$ είναι οι εκτιμητές των συντελεστών α και β αντίστοιχα. Ως εκ τούτου, ο εκτιμητής της διακύμανσης σ^2,δίνεται από την εξίσωση:

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

όπου:

$\hat{y} = \hat{a} + \hat{\beta}x$

$e_i = y_i - \hat{y}_i$, το υπόλοιπο (residual).

Πρέπει να τονιστεί, ότι ο λόγος που ο παρανομαστής στον εκτιμητή της διακύμανσης είναι *n-2* είναι, ότι εκτιμώνται δύο παράμετροι για τον υπολογισμό του s_e^2, οι α και β.

Σύμφωνα με τα παραπάνω, ακολουθεί η διατύπωση εξισώσεων που σχετίζονται με το τετραγωνικό σφάλμα.

Το άθροισμα των τετραγώνων των σφαλμάτων (Error sum of squares – SSE) εκφράζεται από την εξίσωση:

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

ή εναλλακτικά με από την εξίσωση:

$$SSE = \sum_{i=1}^{n} {y_i}^2 - \hat{a}\sum_{i=1}^{n} y_i - \hat{\beta}\sum_{i=1}^{n} x_i y_i$$

Ακολούθως , το μέσο τετραγωνικό σφάλμα s_e^2, εκφράζεται από την εξίσωση:

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

Και η τυπική απόκλιση του εκτιμητή αντίστοιχα από την εξίσωση:

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

Μεγάλες τιμές του μέσου τετραγωνικού σφάλματος φανερώνουν μεγάλες κατακόρυφες αποκλίσεις (δηλαδή υπόλοιπα), των πραγματικών τιμών y_i από τις εκτιμούμενες τιμές $\hat{y}_i$. Η τυπική απόκλιση του εκτιμητή, χρησιμεύει για τη κατασκευή διαστημάτων εμπιστοσύνης, καθώς και στον έλεγχο υποθέσεων για τις παραμέτρους α και β, τα οποία θα αναλυθούν στις ενότητες που ακολουθούν.

4.3. Διάστημα Εμπιστοσύνης των Παραμέτρων α και β

Όπως αναφέρθηκε και στο Κεφάλαιο 2, σκοπός του ελέγχου υποθέσεων είναι η βοήθεια που προσφέρει, μέσω της εξέτασης των δεδομένων ενός δείγματος, στη λήψη αποφάσεων για κάποιο στατιστικό πληθυσμό.

Με τη χρήση του διαστήματος εμπιστοσύνης για τη παράμετρο α ή β, μπορεί να ελεγχθεί η υπόθεση $H_0 : \beta = 0$ ή $H_0 : \alpha = 0$, αντιστοίχως. Η έκφραση για το διάστημα εμπιστοσύνης είναι:

(εκτιμητής παραμέτρου θ) ± *(παράγοντας αξιοπιστίας)* x *(τυπική απόκλιση εκτιμητή),*

με συντελεστή εμπιστοσύνης: 100 (1 - α)%.

Τα διάστημα εμπιστοσύνης με συντελεστή για τις παραμέτρους α και β είναι:

$$\hat{\alpha} \pm (t_{\frac{a}{2},n-2}) \cdot s_e \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{S_{xx}}}$$

$$\hat{\beta} \pm (t_{\frac{a}{2},n-2}) \cdot \frac{s_e}{\sqrt{S_{xx}}}$$

Στη περίπτωση που το διάστημα εμπιστοσύνης που κατασκευάζεται για τη παράμετρο β, περιέχει τη τιμή 0, τότε συμπεραίνεται ότι η υπόθεση H_0 : β = 0 δεν μπορεί να απορριφθεί. Αναλόγως και για τη παράμετρο α, αν το διάστημα εμπιστοσύνης περιέχει τη τιμή 0, τότε συμπεραίνεται ότι η υπόθεση H_0 : α = 0 δεν μπορεί να απορριφθεί.

4.4. Εκτίμηση παραμέτρων και πρόβλεψη με τη χρήση του μοντέλου γραμμικής παλινδρόμησης

Στη περίπτωση που κριθεί ότι το μοντέλο απλής γραμμικής παλινδρόμησης είναι ικανοποιητικό, τότε μπορεί να χρησιμοποιηθεί για την εκτίμηση παραμέτρων και τη πρόβλεψη της τιμής $\hat{y}$ για μια τιμή x = x_0.

Για μια τιμή x = x_0, συνηθίζεται να χρησιμοποιείται το μοντέλο για εκτίμηση του μέσου E(y) ή για πρόβλεψη της τιμής $\hat{y} = \hat{\alpha} + \hat{\beta} x_0$.

Άρα λοιπόν, για μια τιμή x = x_0, ο μέσος του Y δίνεται από τη εξίσωση:

$$E(Y|x_0) = \alpha + \beta x_0$$

Αντίστοιχα, η προβλεπόμενη τιμή του Y δίνεται από την εξίσωση:

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_0$$

Επιπροσθέτως, η τυπική απόκλιση της κατανομής δειγματοληψίας του εκτιμητή $\hat{y}$ του μέσου $E(Y|x_0)$ δίνεται από τη σχέση [5]:

$$\sigma_y = s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

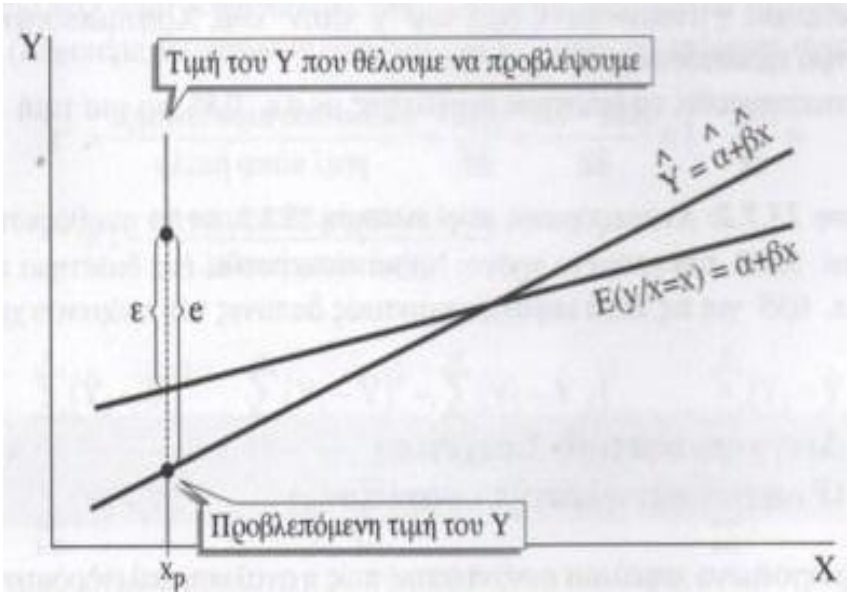
όπου:

s_e είναι ο εκτιμητής της πραγματικής τυπικής απόκλισης σ του τυχαίου σφάλματος ε, στο μοντέλο y = α + β x + ε.

Επιπλέον, για μια ορισμένη τιμή x_0, η τυπική απόκλιση του σφάλματος της πρόβλεψης $y - \hat{y}$ δίνεται από την έκφραση:

$$\sigma_{(y-\hat{y})} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Τα όσα αναφέρουμε ανωτέρω, απεικονίζονται στην εικόνα που ακολουθεί:



Εικόνα 9: Εκτίμηση και πρόβλεψη

4.4.1. Διαστήματα εμπιστοσύνης και πρόβλεψης για x = x_0

Στην υπο-ενότητα αυτή εξετάζεται για μια δεδομένη τιμή x = x_0, το διάστημα εμπιστοσύνης για τον μέσο της y, καθώς και το διάστημα πρόβλεψης για την $\hat{y}$.

Το διάστημα εμπιστοσύνης για τη τιμή $E(Y|x_0)$ με συντελεστή 100(1-α)% δίνεται από την έκφραση:

$$\hat{y} \pm t_{a/2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

όπου:

$t_{a/2} = t_{\frac{a}{2}, n-2}$

Αντίστοιχα, για $x = x_0$, το διάστημα πρόβλεψης για τη νέα παρατήρηση $\hat{y}$, δίνεται από την έκφραση:

$$\hat{y} \pm t_{a/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

4.5. Η χρήση του συντελεστή προσδιορισμού r^2

Όπως αναφέρθηκε και στο προηγούμενο κεφάλαιο ο συντελεστής προσδιορισμού r^2 (coefficient of determination), εκφράζει το ποσοστό της παρατηρούμενης διακύμανσης της μεταβλητής Y, το οποίο μπορεί να αναλυθεί με το απλό μοντέλο γραμμικής παλινδρόμησης που καθορίζει μια προσεγγιστική σχέση μεταξύ των μεταβλητών X και Y. Με άλλα λόγια ο συντελεστής προσδιορισμού, εκφράζει το ποσοστό της συνολικής μεταβλητότητας των y_i που «απορροφάται» από τη παλινδρόμηση.

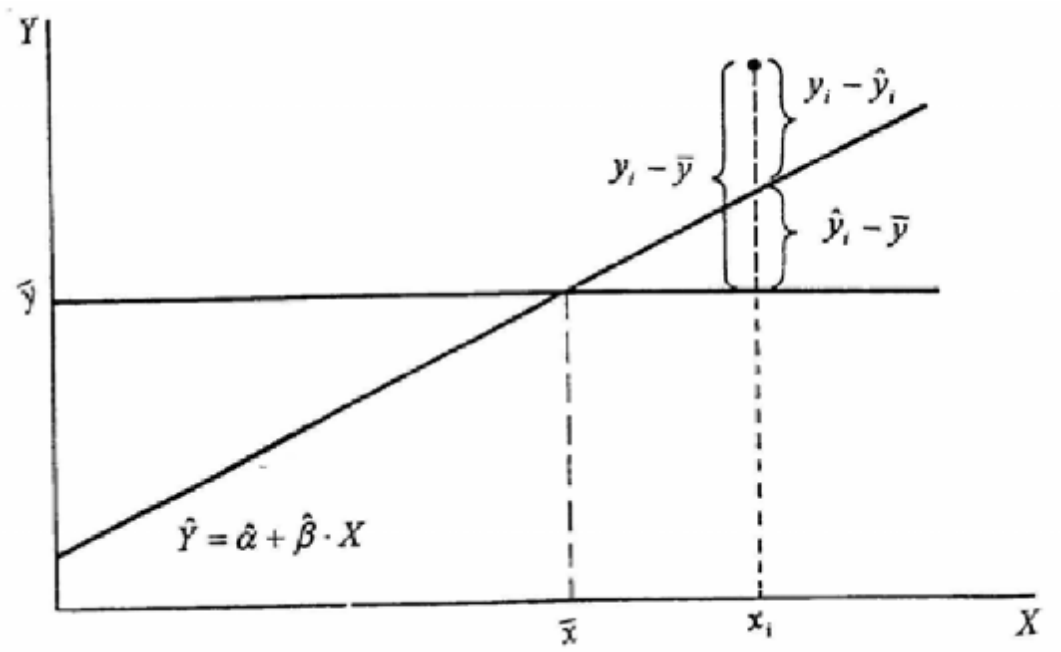
Ο συντελεστής προσδιορισμού r^2, προσδιορίζεται από την ακόλουθη έκφραση:

$$r^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \overline{y})^2}$$

Εναλλακτικά, ο συντελεστής προσδιορισμού μπορεί να εκφραστεί ως το τετράγωνο του δειγματικού συντελεστή συσχέτισης:

$$r^2 = \frac{(S_{xy})^2}{S_{xx} S_{yy}}$$

Οι τιμές που δέχεται ο συντελεστής προσδιορισμού ανήκουν στο κλειστό διάστημα **[0,1].** Έστω τώρα ότι χρησιμοποιούμε το ακόλουθο γράφημα, για να απεικονίσουμε την ευθεία ελαχίστων τετραγώνων ως εκτίμηση της ευθείας παλινδρόμησης.



Εικόνα 10: **Ευθεία ελαχίστων τετραγώνων**

Λαμβάνοντας ως δεδομένο ότι όλα τα σημεία $T_1(x_1,y_1)$, $T_2(x_2,y_2)$,…, $T_n(x_n,y_n)$, βρίσκονται επί της ευθείας ελαχίστων τετραγώνων, προκύπτει ότι $y_i = \hat{y}_i$ και συνεπώς ισχύει ότι:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 0$$

Από την ανωτέρω σχέση προκύπτει ότι ο συντελεστής συσχέτισης ισούται με τη μονάδα.

Αντίστοιχα, όταν η κλίση της ευθείας ελαχίστων τετραγώνων ισούται με το μηδέν ($\hat{\beta} = 0$), ο συντελεστής συσχέτισης θα ισούται με το μηδέν.

Όπως αναφέρθηκε και νωρίτερα ο συντελεστής προσδιορισμού έχει εύρος τιμών στο διάστημα [0,1]. Όσο πιο κοντά στο 1 βρίσκεται η τιμή που έχει ο συντελεστής προσδιορισμού, τόσο καλύτερη είναι η εκτίμηση της ευθείας παλινδρόμησης από την ευθεία ελαχίστων τετραγώνων.

4.6. Παράδειγμα μοντέλου απλής γραμμικής παλινδρόμησης

Τα όσα αναφέρθηκαν νωρίτερα θα τα δούμε στο παράδειγμα που ακολουθεί. Έστω ότι δίνεται ακόλουθος πίνακας, που περιέχει τις τιμές 2 προϊόντων σε διάστημα 10 ετών.

Έτος	X	Y
2003	12	13
2004	15	17
2005	17	19
2006	9	13
2007	20	22
2008	23	25
2009	18	20
2010	9	12
2011	10	9
2012	13	14

Με βάση τα παραπάνω δεδομένα, να γίνουν τα ακόλουθα:

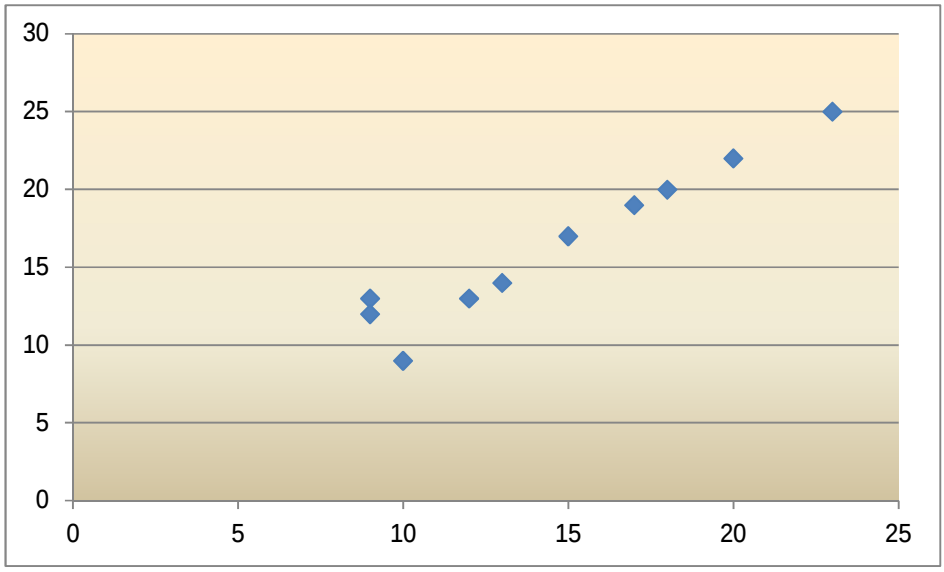
i. Το στικτό διάγραμμα.
ii. Να βρεθούν οι συντελεστές $\hat{\alpha}$ και $\hat{\beta}$, της ευθείας $\hat{y} = \hat{\alpha} + \hat{\beta}x$
iii. Να υπολογιστούν οι προβλέψεις $\hat{y}_i$.
iv. Να υπολογιστούν τα σφάλματα ε_i.
v. Να γίνει ο έλεγχος υποθέσεων για
    a. H_0: $\alpha = 0$ έναντι της υπόθεσης H_1: $\alpha \neq 0$, στο επίπεδο σημαντικότητας α=0,05.
    b. H_0: $\beta = 0$ έναντι της υπόθεσης H_1: $\beta \neq 0$, στο επίπεδο σημαντικότητας α=0,05.
vi. Με τη χρήση το μοντέλου γραμμική παλινδρόμησης και για $x_0 = 16$, να βρεθούν:
    a. Η προβλεπόμενη τιμή του $\hat{y}$.

b. Το διάστημα εμπιστοσύνης για την $E(Y| x_0 = 16)$, με συντελεστή εμπιστοσύνης 95%.
c. Το διάστημα πρόβλεψης για τη τιμή $\hat{y}$, με συντελεστή εμπιστοσύνης 95%.

vii. Να υπολογιστεί ο συντελεστής προσδιορισμού r^2.

Λύση:

i. Το στικτό διάγραμμα των δεδομένων μας είναι:



Από το στικτό διάγραμμα, διαφαίνεται ότι τα σημεία (X_i, Y_i) μπορούν αν προσαρμοστούν στην ευθεία γραμμικής παλινδρόμησης.

ii. Για τον προσδιορισμό των συντελεστών της ευθείας γραμμικής παλινδρόμησης αρχικά υπολογίζουμε τα επιμέρους αθροίσματα, όπως παρουσιάζονται στον ακόλουθο πίνακα:

Έτος	**X**	**Y**	**X2**	**Y2**	**XY**
2003	12	13	144	169	156
2004	15	17	225	289	255
2005	17	19	289	361	323

2006	9	13	81	169	117
2007	20	22	400	484	440
2008	23	25	529	625	575
2009	18	20	324	400	360
2010	9	12	81	144	108
2011	10	9	100	81	90
2012	13	14	169	196	182
Αθροίσματα	**146**	**164**	**2342**	**2918**	**2606**

Εν συνεχεία, από τον παραπάνω πίνακα υπολογίζουμε τα Sxx, Syy, Sxy καθώς και τη μέση τιμή των X, Y.

$$S_{xx} = \sum_{i=1}^{10} x_i^2 - \frac{1}{10}\left(\sum_{i=1}^{10} x_i\right)^2 = 210{,}40$$

$$S_{yy} = \sum_{i=1}^{10} y_i^2 - \frac{1}{10}(\sum_{i=1}^{10} y_i)^2 = 228{,}40$$

$$S_{xy} = \sum_{i=1}^{10} x_i y_i - \frac{1}{10}\sum_{i=1}^{10} x_i \sum_{i=1}^{10} y_i = 211{,}60$$

$$\bar{X} = \frac{1}{10}\sum_{i=1}^{10} x_i = 14{,}60$$

$$\bar{Y} = \frac{1}{10}\sum_{i=1}^{10} y_i = 16{,}40$$

Άρα λοιπόν οι συντελεστές $\hat{\alpha}$ και $\hat{\beta}$ είναι:

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{211{,}60}{210{,}40} = 1{,}0057$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 16{,}40 - 1{,}0057 * 14{,}60 = 1{,}7167$$

Άρα λοιπόν η ευθεία γραμμικής παλινδρόμησης είναι:

$$\hat{y} = 1{,}7167 + 1{,}0057x$$

iii. Με βάση την ευθεία γραμμικής παλινδρόμησης που προέκυψε από το ερώτημα ii, υπολογίζουμε τις προβλέψεις του y για κάθε τιμή του x. Έτσι προκύπτει ο ακόλουθος πίνακας:

X	**Y**	**Ypred**
12	13	13,78517
15	17	16,80228
17	19	18,81369
9	13	10,76806
20	22	21,8308
23	25	24,84791
18	20	19,81939
9	12	10,76806
10	9	11,77376
13	14	14,79087

iv. Τα σφάλματα *ε*, προκύπτουν από τον υπολογισμό της διαφοράς της προσεγγιστικής τιμής από την ακριβή, που δίνεται από την έκφραση:

$$e_i = y_i - \hat{y}_i$$

Οι τιμές των σφαλμάτων που προκύπτουν για κάθε τιμή του X σε απόλυτη τιμή, δίδονται στον ακόλουθο πίνακα:

X	**ε**
12	0,785171
15	0,197719
17	0,186312

9	2,231939
20	0,169202
23	0,152091
18	0,180608
9	1,231939
10	2,773764
13	0,790875

v. Στο ερώτημα αυτό θα γίνει ο έλεγχος των υποθέσεων για τις δυο παραμέτρους.

a. Αρχικά θα γίνει ο έλεγχος υποθέσεων για την παράμετρο α όπου:

H_0: α = 0

H_1: α ≠ 0

Υπολογίζοντας τη τυπική απόκλιση (s_e) του εκτιμητή έχουμε:

$$s_e = \sqrt{\frac{S_{yy} - \hat{\beta} S_{xy}}{n-2}} = \sqrt{\frac{228{,}40 - 1{,}0057 * 210{,}40}{8}} = 1{,}3961$$

Έπειτα υπολογίζουμε το στατιστικό μέσω του οποίου θα γίνει ο έλεγχος:

$$t = \frac{\hat{\alpha} - 0}{s_e \sqrt{\frac{1}{n} + \frac{\bar{\bar{x}}^2}{S_{xx}}}} = \frac{1{,}7167 - 0}{1{,}3961 \sqrt{\frac{1}{10} + \frac{14{,}60^2}{210{,}40}}} = 1{,}1655$$

Στο τελευταίο στάδιο, ελέγχουμε αν η στατιστική τιμή που βρέθηκε είναι μεγαλύτερη ή όχι από τη κριτική τιμή. Η κριτική τιμή ισούται με:

$$t_{\frac{a}{2}, n-2} = t_{0.025,8} = 2{,}306$$

Με βάση τη κριτική τιμή, απορρίπτουμε την H_0 αν t > 2,306 ή t < -2,306.

Άρα για 1,1655 < 2,306 δεν μπορούμε να απορρίψουμε την H_0: α = 0.

b. Έπειτα συνεχίζουμε με τον έλεγχο υποθέσεων για τη παράμετρο β, όπου:

H_0: β = 0

H_1: β ≠ 0

Υπολογίζοντας τη τυπική απόκλιση (s_e) του εκτιμητή έχουμε:

$$s_e = \sqrt{\frac{S_{yy} - \hat{\beta} S_{xy}}{n-2}} = \sqrt{\frac{228{,}40 - 1{,}0057 * 211{,}60}{8}} = 1{,}3961$$

Έπειτα υπολογίζουμε το στατιστικό μέσω του οποίου θα γίνει ο έλεγχος:

$$t = \frac{\hat{\beta} - 0}{s_e / \sqrt{S_{xx}}} = \frac{1{,}0057 - 0}{1{,}3961 / \sqrt{210{,}40}} = 10{,}4489$$

Στο τελευταίο στάδιο, ελέγχουμε αν η στατιστική τιμή που βρέθηκε είναι μεγαλύτερη ή όχι από τη κριτική τιμή. Η κριτική τιμή ισούται με:

$$t_{\frac{a}{2}, n-2} = t_{0.025,8} = 2{,}306$$

Με βάση τη κριτική τιμή, απορρίπτουμε την H_0 αν t > 2,306 ή t < -2,306.

Άρα για 10,4489 > 2,306 απορρίπτουμε την H_0: β = 0 και συνεπώς υπάρχει μια γραμμική σχέση μεταξύ των μεταβλητών X και Y διότι β ≠ 0.

vi. Σύμφωνα με τα δεδομένα της άσκησης αλλά και τα αριθμητικά αποτελέσματα των προηγούμενων ερωτημάτων έχουμε ότι:

a. Η προβλεπόμενη τιμή του $\hat{y}$ για x_0 = 16, είναι:

$$\hat{y} = \hat{\alpha} + \hat{\beta} x_0 = 1{,}7167 + 1{,}0057 * 16 = 17{,}8080$$

b. Για τον υπολογισμό του διαστήματος εμπιστοσύνης για την E(Y| x_0 = 16), αρχικά υπολογίζουμε τη τυπική απόκλιση της κατανομής δειγματοληψίας του εκτιμητή $\hat{y}$ του μέσου E(Y| x_0 = 16), η οποία δίνεται από την εξίσωση:

$$\sigma_y = s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Άρα λοιπόν έχουμε ότι:

$$\sigma_y = s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 1{,}3961 \sqrt{\frac{1}{10} + \frac{(16 - 14{,}6)^2}{210{,}40}} = 0{,}4616$$

Εν συνεχεία, γνωρίζουμε ότι ο συντελεστής εμπιστοσύνης είναι 95%, άρα προκύπτει ότι:

$$t_{\frac{a}{2},n-2} = t_{0{,}025,8} = 2{,}306 \ (\beta\lambda.\, \pi\acute{\iota}\nu\alpha\kappa\alpha \ II \ \pi\alpha\rho\alpha\rho\tau\acute{\eta}\mu\alpha\tau\sigma\varsigma)$$

Το διάστημα εμπιστοσύνης για την E(Y| x_0 = 16), με συντελεστή εμπιστοσύνης 95%, προκύπτει από τον υπολογισμό της έκφρασης:

$$\hat{y} \pm t_{a/2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Συνεπώς έχουμε ότι:

$$\hat{y} \pm t_{a/2} s_e \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 17{,}8080 \pm 2{,}306 * 0{,}4616$$

Εν τέλει, το διάστημα εμπιστοσύνης είναι (16,7435 , 18,8724).

c. Για τον υπολογισμό του διαστήματος πρόβλεψης για την $\hat{y}$ αρχικά υπολογίζουμε την τυπική απόκλιση του σφάλματος της πρόβλεψης $y - \hat{y}$ για $x_0 = 16$, η οποία δίνεται από την εξίσωση:

$$\sigma_{(y-\hat{y})} = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Άρα λοιπόν έχουμε ότι:

$$\sigma_y = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 1{,}3961 \sqrt{1 + \frac{1}{10} + \frac{(16 - 14{,}6)^2}{210{,}40}} = 1{,}4704$$

Εν συνεχεία, γνωρίζουμε ότι ο συντελεστής εμπιστοσύνης είναι 95%, άρα προκύπτει ότι:

$$t_{\frac{a}{2},n-2} = t_{0{,}025,8} = 2{,}306 \ (\beta\lambda.\, \pi\acute{\iota}\nu\alpha\kappa\alpha \ II \ \pi\alpha\rho\alpha\rho\tau\acute{\eta}\mu\alpha\tau\sigma\varsigma)$$

Το διάστημα πρόβλεψης για την $\hat{y}$, με συντελεστή εμπιστοσύνης 95%, προκύπτει από τον υπολογισμό της έκφρασης:

$$\hat{y} \pm t_{a/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}$$

Συνεπώς έχουμε ότι:

$$\hat{y} \pm t_{a/2} s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} = 17{,}8080 \pm 2{,}306 * 1{,}4704$$

Εν τέλει, το διάστημα πρόβλεψης είναι (14,4171 , 21,1988).

vii. Για τον υπολογισμό του συντελεστή προσδιορισμού θα κάνουμε χρήση του τύπου:

$$r^2 = \frac{(S_{xy})^2}{S_{xx}S_{yy}}$$

Σύμφωνα λοιπόν και με τα αριθμητικά δεδομένα που έχουμε ήδη εξάγει από τα προηγούμενα ερωτήματα, προκύπτει ότι:

$$r^2 = \frac{(S_{xy})^2}{S_{xx}S_{yy}} = \frac{211{,}60^2}{210{,}40 * 228{,}40} = 0{,}9317 = 93{,}17\%$$

Από τη τιμή που προκύπτει για τον συντελεστή προσδιορισμού, συμπεραίνουμε ότι η ευθεία ελαχίστων τετραγώνων εκτιμά άκρως ικανοποιητικά την ευθεία παλινδρόμησης και η οποία δίνεται από τη σχέση:

$$\hat{y} = 1{,}7167 + 1{,}0057x$$

5. Εφαρμογή με Χρήση του Προγράμματος SPSS

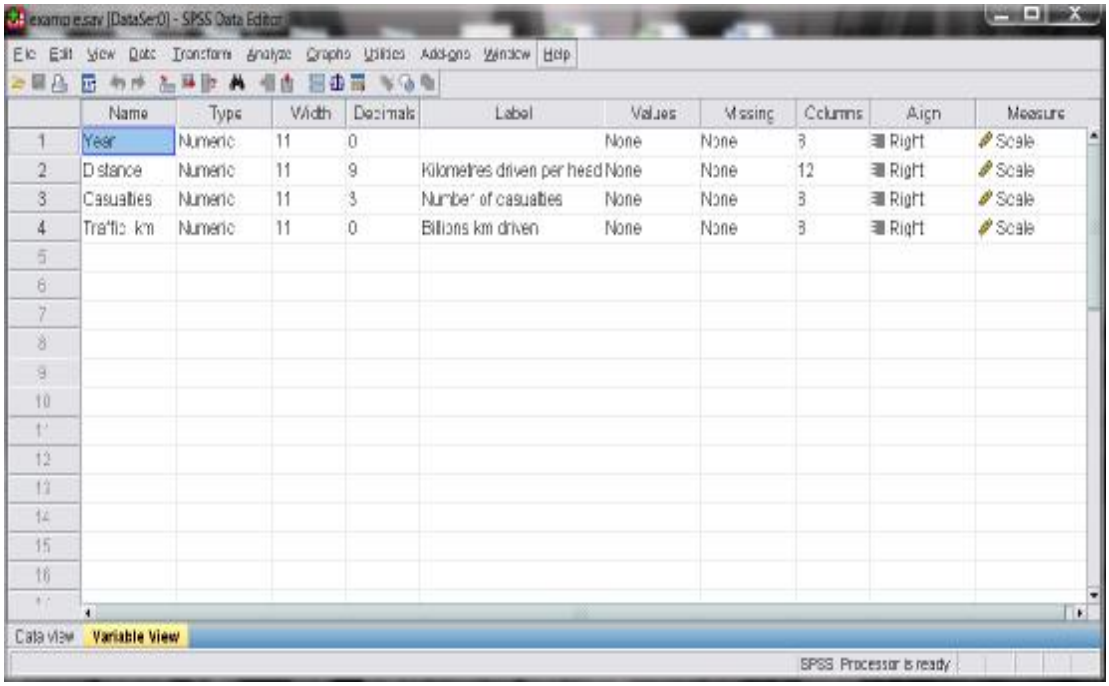
Στο κεφάλαιο αυτό θα εφαρμοστούν τα όσα εξετάσαμε νωρίτερα σε θεωρητικό επίπεδο, με τη χρήση του SPSS, σε ένα παράδειγμα που αποτελείται από τέσσερεις μεταβλητές. Τα δεδομένα που χρησιμοποιήθηκαν, εξήχθησαν από τον δικτυακό τόπο: http://staff.bath.ac.uk/pssiw/stats2/page16/page16.html.

5.1. Τα Δεδομένα

Οι μεταβλητές οι οποίες θα εξεταστούν είναι οι ακόλουθες:

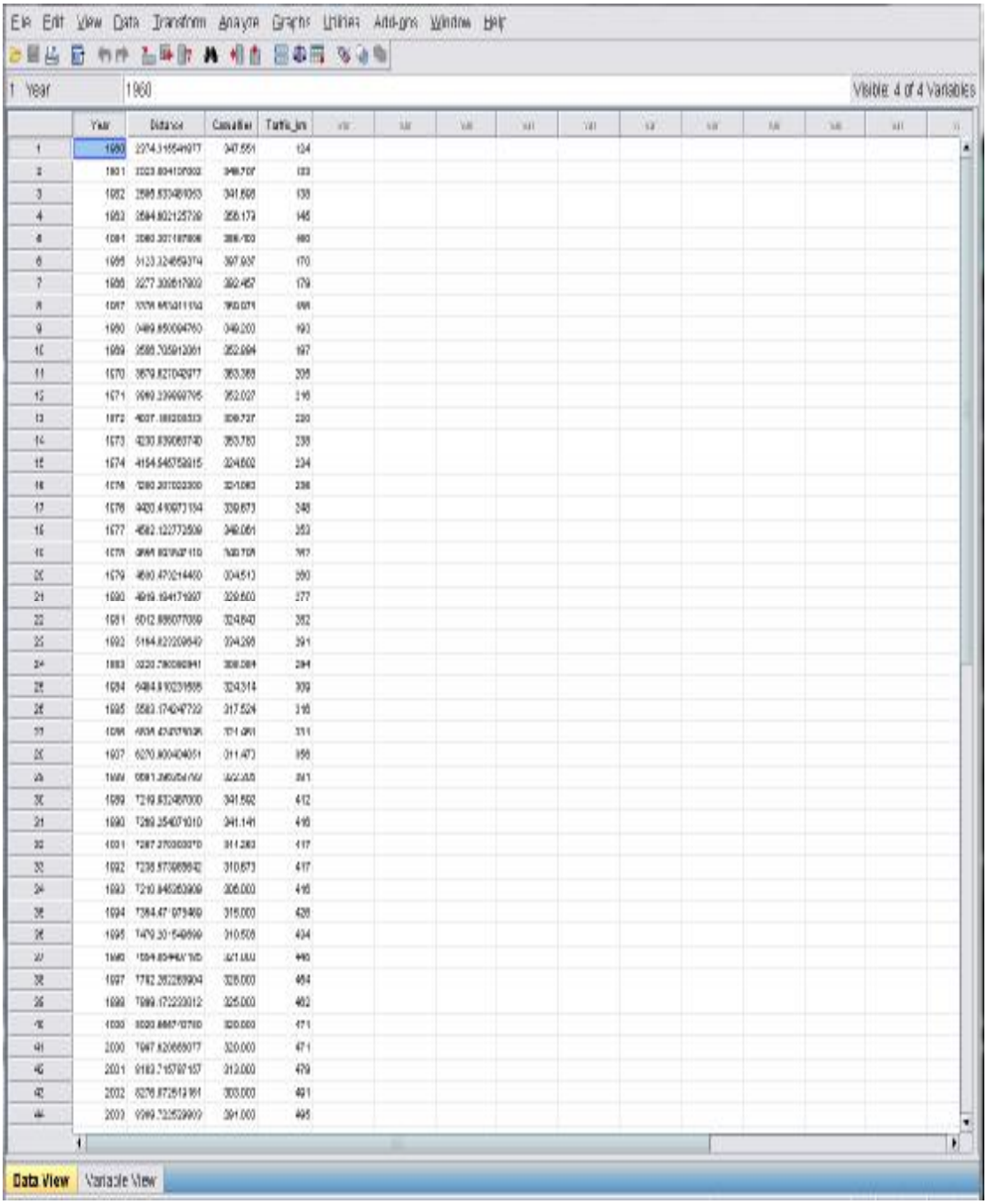
- Year: Το κάθε έτος για το οποίο εμφανίζονται το δεδομένα
- Distance: Ο μέσος όρος των χιλιομέτρων που καλύφθηκαν ανά άτομο κατά έτος.
- Casualties: Ο αριθμός των ατυχημάτων (σε χιλιάδες) κατά έτος.
- Traffic_km: Τα χιλιόμετρα (σε δισεκατομμύρια) που καλύφθηκαν από τους οδηγούς κατά έτος.

Οι ποσοτικές μεταβλητές αυτές εισάγονται στη καρτέλα «Variable View» διαμορφώνοντας ανάλογα τα αντίστοιχα πεδία (βλ. Εικόνα 11).



Εικόνα 11: Εισαγωγή μεταβλητών στη καρτέλα «Variable View»

Μετά από τον χαρακτηρισμό των μεταβλητών, ακολουθεί η εισαγωγή των δεδομένων ανά μεταβλητή. Το σύνολο των δεδομένων έχει μήκος 44 στοιχείων, τα οποία εισάγονται στη καρτέλα «Data View» (βλ. Εικόνα 12).



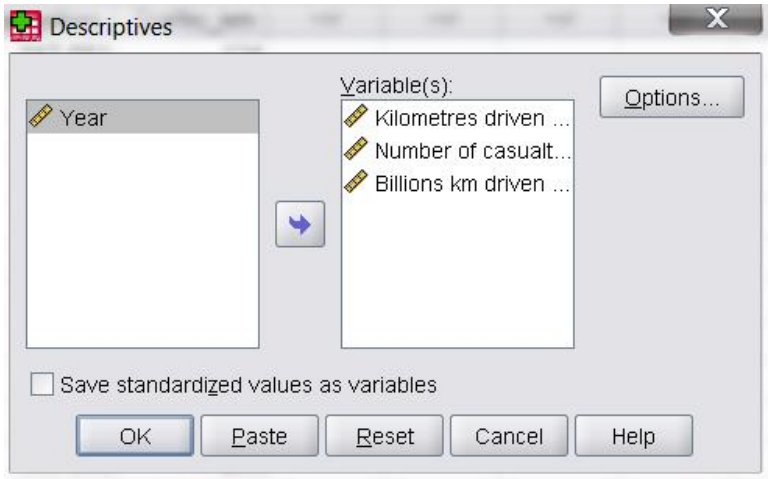
Εικόνα 12: Τα δεδομένα των μεταβλητών στο «Data View»

5.2. Τα Περιγραφικά Μέτρα των Μεταβλητών

Στο σημείο αυτό θα εξεταστούν τα περιγραφικά μέτρα για τις μεταβλητές μας (πλην των ετών).

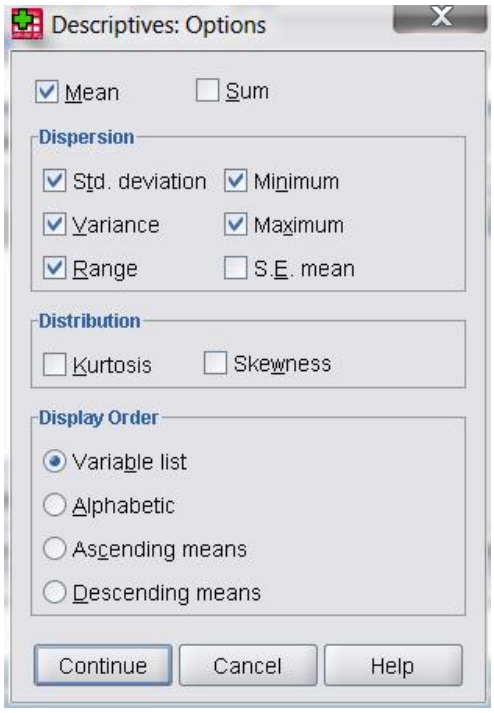
Αυτό επιτυγχάνεται αν από το μενού επιλέξουμε «Analyze» και εν συνεχεία την εντολή «Descriptives…» από τη λίστα «Descriptive Statistics».

Στο νέο παράθυρο διαλόγου που μας ανοίγει, επιλέγουμε εκείνες τις μεταβλητές για τις οποίες επιθυμούμε να γίνει ανάλυση των περιγραφικών μέτρων (βλ Εικόνα 13).



Εικόνα 13: Το παράθυρο διαλόγου «Descriptives»

Εν συνεχεία, επιλέγουμε το «Options…» για να οδηγηθούμε στο νέο παράθυρο διαλόγου «Descriptive: Options», απ' όπου μπορούμε να επιλέγουμε εκείνα τα μέτρα για τα οποία επιθυμούμε να γίνει ανάλυση (βλ. Εικόνα 13).



Εικόνα 14: **Παράθυρο διαλόγου** «Descriptives: Options»

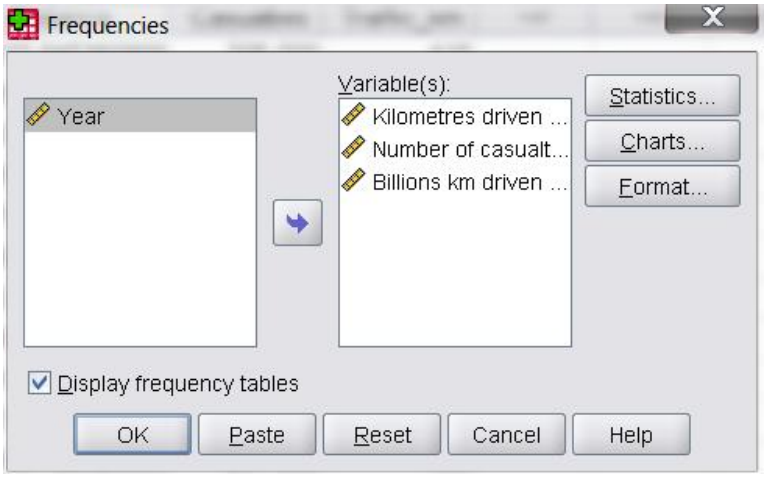
Αφού λοιπόν επιλεχθούν οι μεταβλητές και τα μέτρα για τα οποία ζητείται ανάλυση, το SPSS βγάζει το επιθυμητό αποτέλεσμα το οποίο εμφανίζεται στην ¨έξοδο (Output)¨. Με βάση τις επιλογές μας, το αποτέλεσμα που εξάγεται παρουσιάζεται στον ακόλουθο πίνακα.

Descriptive Statistics

	N	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Kilometres driven per head	44	5.9354E3	2.3743E3	8.30972E3	5.400646E3	1.90942748E3	3645913.339
Number of casualties (1000s)	44	106.937	291.000	397.937	335.09614	23.909957	571.686
Billions km driven	44	371	124	495	307.88	116.460	13562.844
Valid N (listwise)	44						

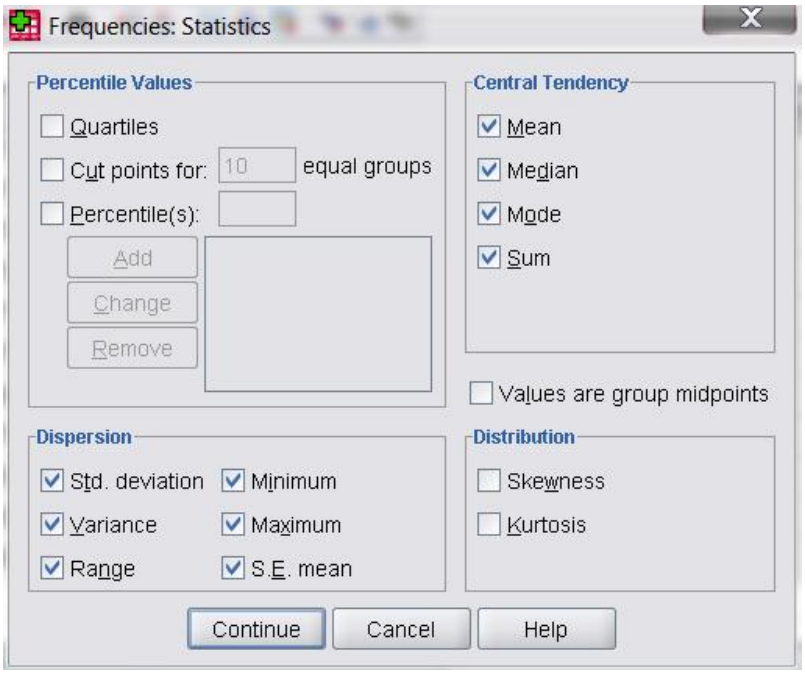
Στον ανωτέρω πίνακα παρουσιάζεται για κάθε μεταβλητή, το πλήθος των δεδομένων (N), το εύρος των δεδομένων (Range), η ελάχιστη και μέγιστη τιμή (Minimum and Maximum Value), ο αριθμητικός μέσος (Mean), η τυπική απόκλιση (Standard Deviation) και η διασπορά (Variance).

Εναλλακτικά, μπορεί να ακολουθηθεί διαφορετική διαδρομή προκειμένου να εμφανιστούν τα περιγραφικά μέτρα. Σύμφωνα με την εναλλακτική προσέγγιση, η διαδρομή που μπορεί να επιλεγεί είναι: Analyze **à** Descriptive Statistics **à** Frequencies…, ανοίγοντας το παράθυρο διαλόγου «Frequencies» και προσδιορίζοντας τις μεταβλητές που θα αναλυθούν (βλ. Εικόνα 15).



Εικόνα 15: Παράθυρο διαλόγου «Frequencies»

Εν συνεχεία επιλέγουμε «Statistics…», για να γίνει ο προσδιορισμός των μέτρων που θέλουμε να υπολογιστούν (βλ. Εικόνα 16).



Εικόνα 16: **Παράθυρο διαλόγου** «Frequencies: Statistics»

Το αποτέλεσμα που προκύπτει από την αναπαραγωγή της ανωτέρω μεθοδολογίας παρουσιάζεται στον ακόλουθο πίνακα:

Statistics

		Kilometres driven per head	Number of casualties (1000s)	Billions km driven
N	Valid	44	44	44
	Missing	1	1	1
	Mean	5.40064631E3	335.09614	307.88
	Std. Error of Mean	2.87857024E2	3.604562	17.557
	Median	5.08845464E3	328.30000	286.60
	Mode	2.37431554E3	320.000	124[a]
	Std. Deviation	1.90942748E3	23.909957	116.460
	Variance	3645913.339	571.686	13562.844
	Range	5.93540698E3	106.937	371
	Minimum	2.37431554E3	291.000	124
	Maximum	8.30972253E3	397.937	495
	Sum	2.37628437E5	14744.230	13547

a. Multiple modes exist. The smallest value is shown

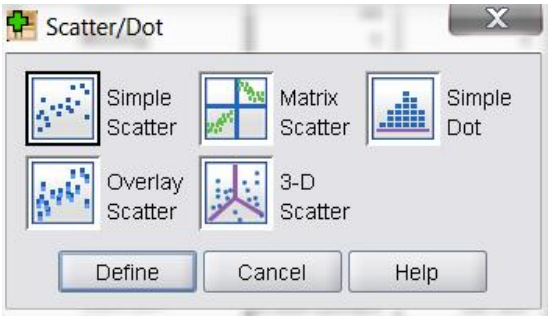
Από τον ανωτέρω πίνακα, απορρέει ο ακόλουθος σχολιασμός[7]:

a) Ο μέσος όρος θνησιμότητας στη διάρκεια των 44 ετών, είναι 335,09614.
b) Η διάμεσος είναι 328,30, γεγονός που σημαίνει ότι το 50% των παρατηρήσεων είναι πάνω από αυτό τον αριθμό, ενώ το υπόλοιπο 50% κάτω.
c) Η τιμή που συναντάται περισσότερες φορές είναι το 320.
d) Η τυπική απόκλιση είναι 23,909957.
e) Η διακύμανση είναι 571,686.
f) Το εύρος μεταβολής είναι 106,937.
g) Η ελάχιστη τιμή είναι 291.
h) Η μέγιστη τιμή είναι 397,937.

5.3. **Συσχέτιση μεταβλητών** Year – Distance

Στο σημείο αυτό θα εξεταστεί το κατά πόσο υπάρχει συσχέτιση μεταξύ του της απόστασης που καλύπτουν οι άνθρωποι και του χρόνου.

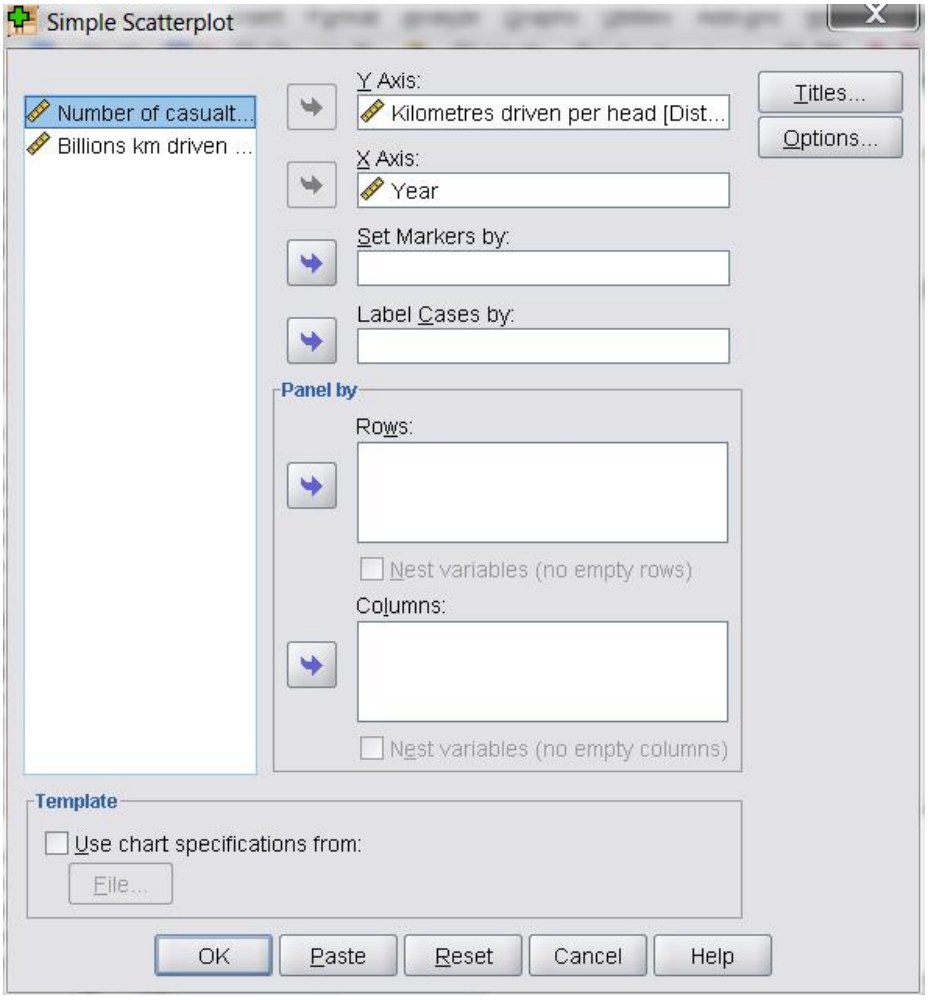
Αρχικά θα κατασκευαστεί το διάγραμμα διασποράς (scatterplot) μεταξύ των δύο μεταβλητών. Για τη κατασκευή του διαγράμματος διασποράς στο SPSS, επιλέγουμε: Graphs **à** Legacy Dialogs **à** Scatter/Dot… οπού οδηγεί στο αντίστοιχο παράθυρο διαλόγου (βλ. Εικόνα 17).



Εικόνα 17: Παράθυρο διαλόγου «Scatter/Dot»

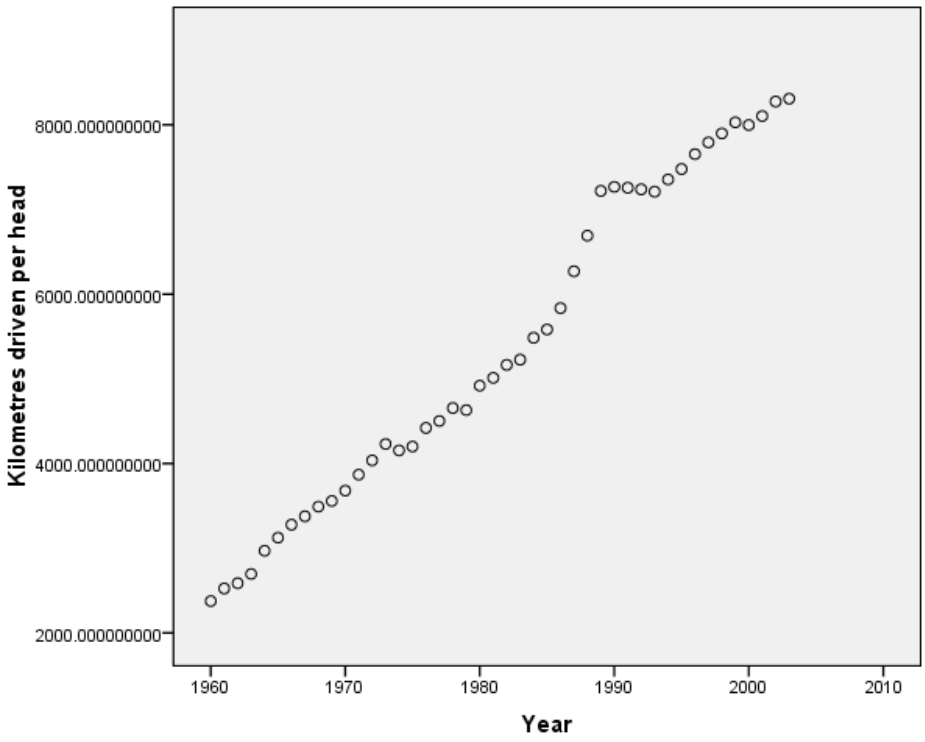
[7] Εξετάζεται η περίπτωση του αριθμού θνησιμότητας κατά έτος. Ανάλογος είναι και ο σχολιασμός για τις άλλες δύο μεταβλητές.

Από τις επιλογές που προσφέρονται, επιλέγεται η «Simple Scatter» στην οποία ορίζεται η μεταβλητή «Year» στον άξονα των X, και η μεταβλητή «Distance» στον άξονα των Y (βλ. Εικόνα 18).



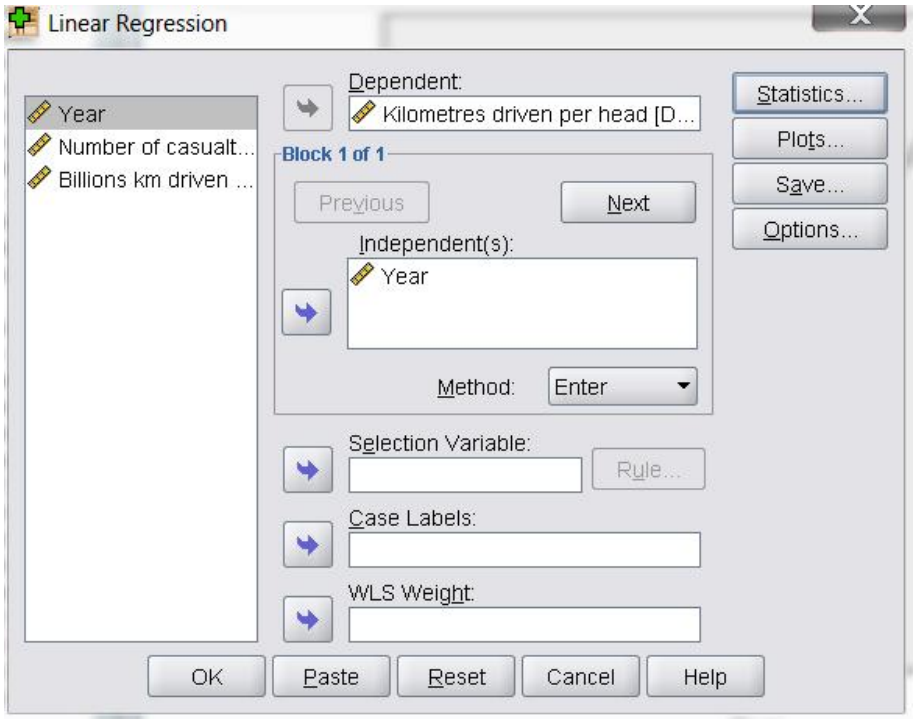
Εικόνα 18: **Παράθυρο διαλόγου** «Simple Scatterplot»

Από την ανωτέρω διαδικασία προκύπτει το ακόλουθο διάγραμμα διασποράς για τις μεταβλητές «Year» και «Distance».



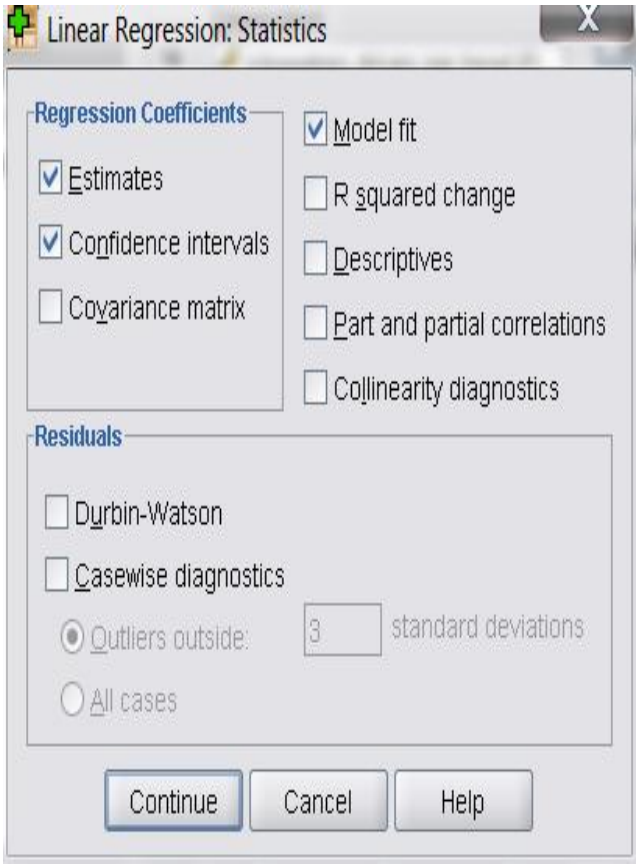
Από το ανωτέρω διάγραμμα είναι σαφές ότι υπάρχει γραμμική σχέση μεταξύ των μεταβλητών και επομένως μπορεί να εφαρμοστεί το μοντέλο της απλής γραμμικής παλινδρόμησης.

Για την εφαρμογή του μοντέλου της απλής γραμμικής παλινδρόμησης στο SPSS, επιλέγεται η διαδρομή Analyze **à** Regression **à** Linear… απ' όπου εμφανίζεται το ακόλουθο παράθυρο διαλόγου.

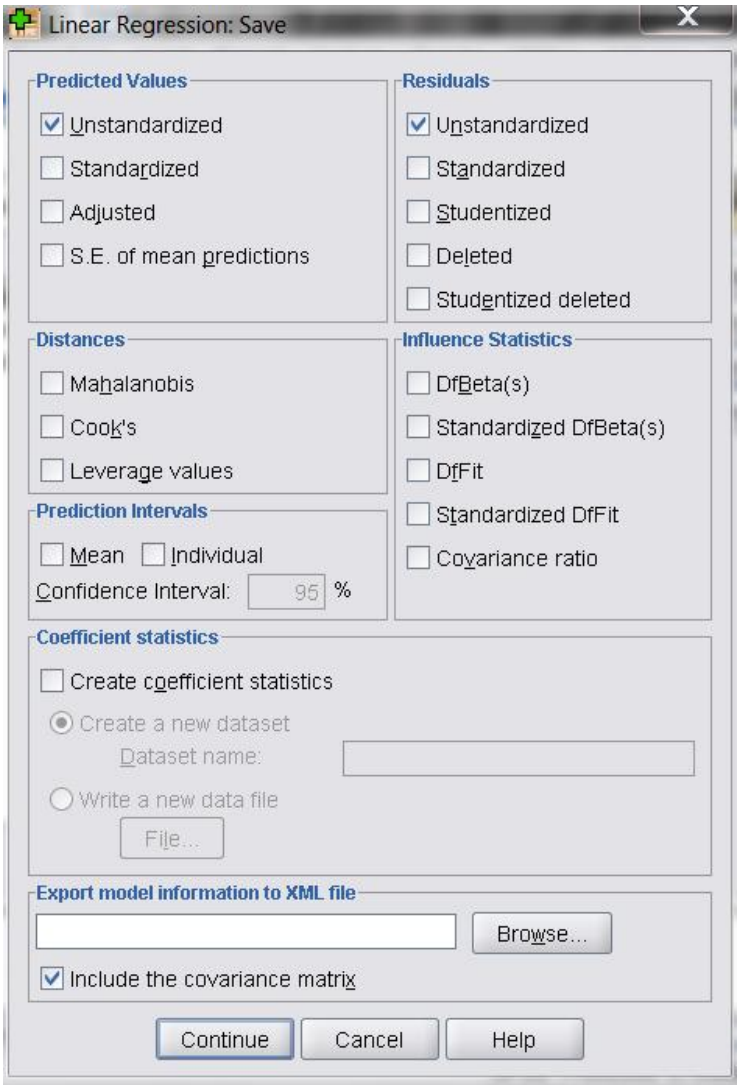


Εικόνα 19: **Παράθυρο διαλόγου** «Linear Regression»

Στο παράθυρο διαλόγου της εικόνας 19, αρχικά επιλέγουμε «Statistics» όπου επιλέγουμε το «Confidence intervals», ώστε να εμφανιστούν στην έξοδο και τα διαστήματα εμπιστοσύνης (βλ. Εικόνα 20). Έπειτα, επιλέγουμε «Save» και επιλέγουμε «Unstandardized», από τις κατηγορίες «Predicted values» και «Residuals» (βλ Εικόνα 21).



Εικόνα 20: **Παράθυρο διαλόγου** «Linear Regression: Statistics»



Εικόνα 21:Παράθυρο διαλόγου «Linear Regression: Save»

Σύμφωνα με τις προτιμήσεις που ορίσαμε παραπάνω, οι πίνακες οι οποίοι θα πάρουμε ως έξοδο από το SPSS είναι:

Variables Entered/Removed[b]

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

a. All requested variables entered.

b. Dependent Variable: Kilometres driven per head

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.991[a]	.983	.983	2.52573780E2

a. Predictors: (Constant), Year

b. Dependent Variable: Kilometres driven per head

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.541E8	1	1.541E8	2415.527	.000[a]
	Residual	2679327.614	42	63793.515		
	Total	1.568E8	43			

a. Predictors: (Constant), Year

b. Dependent Variable: Kilometres driven per head

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	-286619.002	5941.759		-48.238	.000	-298609.958	-274628.047
	Year	147.373	2.999	.991	49.148	.000	141.322	153.424

a. Dependent Variable: Kilometres driven per head

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2.2321E3	8.56916E3	5.4006E3	1.89304077E3	44
Residual	-4.0174E2	7.13988E2	-2.65E-11	2.49619599E2	44
Std. Predicted Value	-1.674	1.674	.000	1.000	44
Std. Residual	-1.591	2.827	.000	.988	44

a. Dependent Variable: Kilometres driven per head

Από τους παραπάνω πίνακες σημειώνουμε τα εξής σημαντικά:

i. Στον πίνακα «Coefficients», ο συντελεστής Beta (Standardized Coefficients: Beta) είναι η εκτίμηση του b_1, αφού εφαρμοστεί το μοντέλο παλινδρόμησης:

$$Y_i = b_0 + b_1 X_i' + \varepsilon_i$$

όπου τα X_i' είναι οι τυποποιημένες τιμές των X_i.

ii. Οι σημειακές εκτιμήσεις των συντελεστών b_0 και b_1 είναι -286619,002 και 147,373 αντίστοιχα.

iii. Τα διαστήματα εμπιστοσύνης είναι (-298609,958 , -274628,047) και (141,322 , 153,424) για τα b_0 και b_1 αντίστοιχα.

iv. Το p-value είναι 0 άρα η υπόθεση $b_0 = 0$ και $b_1 = 0$ απορρίπτεται. Επομένως, η μεταβλητή ***Distance*** εξαρτάται από τη μεταβλητή ***Year***.

v. Όσον αφορά τον πίνακα ANOVA, το p-value για τον έλεγχο: $H_0: b_1 = 0$,
$H_1: b_1 \neq 0$
ισούται με το μηδέν. Ο έλεγχος της συγκεκριμένης υπόθεσης γίνεται μέσω της τιμής F και στην ουσία είναι ισοδύναμος με τον έλεγχο που γίνεται μέσω της t1 (όπως είδαμε στο iv).

vi. Σε σχέση με την εκτίμηση της διασποράς των σφαλμάτων, από τον πίνακα ANOVA, παρατηρούμε ότι ισούται με 63793,515.

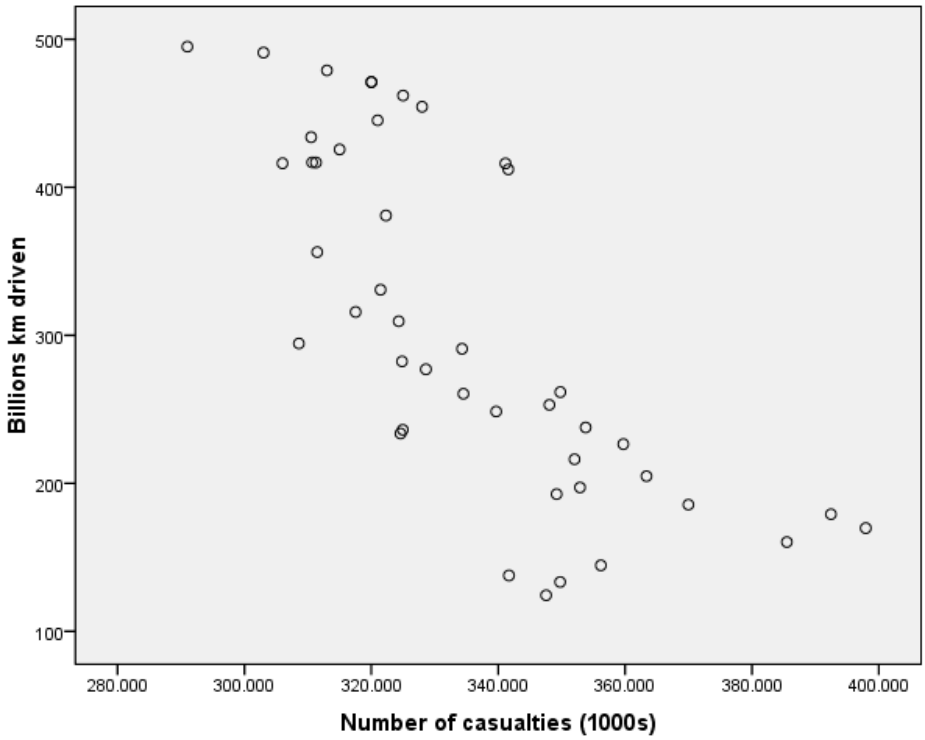
vii. Τέλος, στον πίνακα ***Residuals Statistics***, δίνονται οι προσαρμοσμένες τιμές των Y_i (δηλαδή, οι προβλέψεις των Y_i), καθώς και τα κατάλοιπα[8].

5.4. **Συσχέτιση μεταβλητών** Casualties – Traffic_km

Στο σημείο αυτό θα εξεταστεί το κατά πόσο υπάρχει συσχέτιση μεταξύ των ατυχημάτων που πραγματοποιούνται κάθε χρόνο και των χιλιομέτρων που καλύπτονται συνολικά από τους οδηγούς για το αντίστοιχο έτος.

Ακλουθώντας την ίδια διαδικασία με νωρίτερα (Ενότητα 5.3), αρχικά κατασκευάζεται το διάγραμμα διασποράς (scatterplot) μεταξύ των δύο μεταβλητών.

[8] Εμφανίζονται οι τιμές με βάση κάποια μέτρα (πχ ελάχιστη, μέγιστη τιμή), ο πλήρης πίνακας δίνεται εμφανίζεται στο Data View στο SPSS.



Στο ανωτέρω διάγραμμα δεν καθίσταται απόλυτα σαφές αν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών και επομένως προχωράμε στην εφαρμογή του μοντέλου της απλής γραμμικής παλινδρόμησης για να εξετάσουμε τη συσχέτιση.

Ακολουθώντας την ίδια διαδικασία με νωρίτερα, οι πίνακες οι οποίοι θα πάρουμε ως έξοδο από το SPSS είναι:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Number of casualties (1000s)[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Billions km driven

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.754[a]	.569	.559	77.364

a. Predictors: (Constant), Number of casualties (1000s)

b. Dependent Variable: Billions km driven

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	331822.105	1	331822.105	55.440	.000[a]
	Residual	251380.175	42	5985.242		
	Total	583202.280	43			

a. Predictors: (Constant), Number of casualties (1000s)

b. Dependent Variable: Billions km driven

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	1539.027	165.758		9.285	.000	1204.514	1873.541
	Number of casualties (1000s)	-3.674	.493	-.754	-7.446	.000	-4.670	-2.678

a. Dependent Variable: Billions km driven

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	77.01	469.89	307.88	87.845	44
Residual	-145.945	130.416	.000	76.459	44
Std. Predicted Value	-2.628	1.844	.000	1.000	44
Std. Residual	-1.886	1.686	.000	.988	44

a. Dependent Variable: Billions km driven

Από τους παραπάνω πίνακες σημειώνουμε τα εξής σημαντικά:

i. Στον πίνακα «Coefficients», ο συντελεστής Beta (Standardized Coefficients: Beta) είναι η εκτίμηση του b_1, αφού εφαρμοστεί το μοντέλο παλινδρόμησης:

$$Y_i = b_0 + b_1 X_i' + \varepsilon_i$$

όπου τα X_i' είναι οι τυποποιημένες τιμές των X_i.

ii. Οι σημειακές εκτιμήσεις των συντελεστών b_0 και b_1 είναι 1539,027 και -3,674 αντίστοιχα.

iii. Τα διαστήματα εμπιστοσύνης είναι (1204,514 , 1873,541) και (-4,670 , -2,678) για τα b_0 και b_1 αντίστοιχα.

iv. Το p-value είναι 0 άρα η υπόθεση $b_0 = 0$ και $b_1 = 0$ απορρίπτεται. Επομένως, η μεταβλητή ***Traffic_km*** εξαρτάται από τη μεταβλητή ***Casualties***.

v. Όσον αφορά τον πίνακα ANOVA, το p-value για τον έλεγχο: $H_0: b_1 = 0$,
$H_1: b_1 \neq 0$
ισούται με το μηδέν. Ο έλεγχος της συγκεκριμένης υπόθεσης γίνεται μέσω της τιμής F και στην ουσία είναι ισοδύναμος με τον έλεγχο που γίνεται μέσω της t1 (όπως είδαμε στο iv).

vi. Σε σχέση με την εκτίμηση της διασποράς των σφαλμάτων, από τον πίνακα ANOVA, παρατηρούμε ότι ισούται με 5985,242.

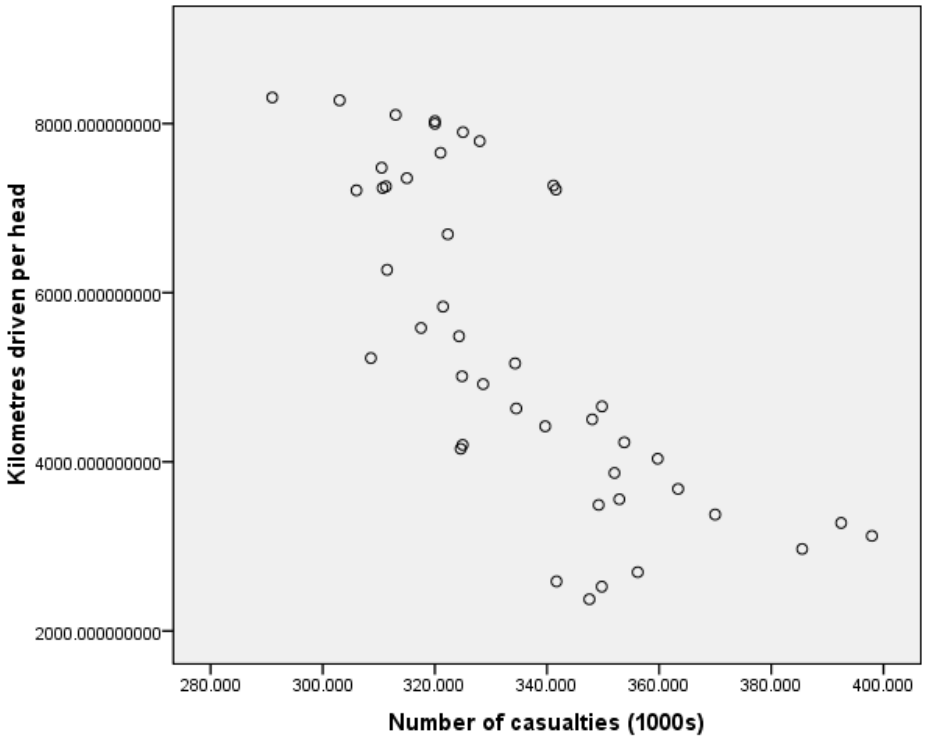
vii. Τέλος, στον πίνακα ***Residuals Statistics***, δίνονται οι προσαρμοσμένες τιμές των Y_i (δηλαδή, οι προβλέψεις των Y_i), καθώς και τα κατάλοιπα[9].

5.5. **Συσχέτιση μεταβλητών** Casualties – Distance

Στο σημείο αυτό θα εξεταστεί το κατά πόσο υπάρχει συσχέτιση μεταξύ των ατυχημάτων που πραγματοποιούνται κάθε χρόνο και της απόστασης που διανύει κατά μέσο όρο ο κάθε οδηγός για το αντίστοιχο έτος.

[9] Εμφανίζονται οι τιμές με βάση κάποια μέτρα (πχ ελάχιστη, μέγιστη τιμή), ο πλήρης πίνακας δίνεται εμφανίζεται στο Data View στο SPSS.

Ακολουθώντας την ίδια διαδικασία με νωρίτερα (Ενότητα 5.3), αρχικά κατασκευάζεται το διάγραμμα διασποράς (scatterplot) μεταξύ των δύο μεταβλητών.



Στο ανωτέρω διάγραμμα δεν καθίσταται απόλυτα σαφές αν υπάρχει γραμμική σχέση μεταξύ των μεταβλητών και επομένως προχωράμε στην εφαρμογή του μοντέλου της απλής γραμμικής παλινδρόμησης για να εξετάσουμε τη συσχέτιση.

Ακολουθώντας την ίδια διαδικασία με νωρίτερα, οι πίνακες οι οποίοι θα πάρουμε ως έξοδο από το SPSS είναι:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Number of casualties (1000s)[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Kilometres driven per head

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.755[a]	.570	.560	1.26684536E3

a. Predictors: (Constant), Number of casualties (1000s)

b. Dependent Variable: Kilometres driven per head

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8.937E7	1	8.937E7	55.685	.000[a]
	Residual	6.741E7	42	1604897.183		
	Total	1.568E8	43			

a. Predictors: (Constant), Number of casualties (1000s)

b. Dependent Variable: Kilometres driven per head

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	25605.181	2714.299		9.433	.000	20127.503	31082.859
	Number of casualties (1000s)	-60.295	8.080	-.755	-7.462	.000	-76.601	-43.989

a. Dependent Variable: Kilometres driven per head

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1.6116E3	8.05941E3	5.4006E3	1.44164466E3	44
Residual	-2.4171E3	2.23208E3	...	1.25202795E3	44
Std. Predicted Value	-2.628	1.844	.000	1.000	44
Std. Residual	-1.908	1.762	.000	.988	44

a. Dependent Variable: Kilometres driven per head

Από τους παραπάνω πίνακες σημειώνουμε τα εξής σημαντικά:

i. Στον πίνακα «Coefficients», ο συντελεστής Beta (Standardized Coefficients: Beta) είναι η εκτίμηση του b_1, αφού εφαρμοστεί το μοντέλο παλινδρόμησης:

$$Y_i = b_0 + b_1 X_i' + \varepsilon_i$$

όπου τα X_i' είναι οι τυποποιημένες τιμές των X_i.

ii. Οι σημειακές εκτιμήσεις των συντελεστών b_0 και b_1 είναι 25605,181 και -60,295 αντίστοιχα.

iii. Τα διαστήματα εμπιστοσύνης είναι (20127,503 , 31082,859) και (-76,601 , -43,989) για τα b_0 και b_1 αντίστοιχα.

iv. Το p-value είναι 0 άρα η υπόθεση $b_0 = 0$ και $b_1 = 0$ απορρίπτεται. Επομένως, η μεταβλητή *Distance* εξαρτάται από τη μεταβλητή *Casualties*.

v. Όσον αφορά τον πίνακα ANOVA, το p-value για τον έλεγχο: $H_0: b_1 = 0$,
$H_1: b_1 \neq 0$
ισούται με το μηδέν. Ο έλεγχος της συγκεκριμένης υπόθεσης γίνεται μέσω της τιμής F και στην ουσία είναι ισοδύναμος με τον έλεγχο που γίνεται μέσω της t1 (όπως είδαμε στο iv).

vi. Σε σχέση με την εκτίμηση της διασποράς των σφαλμάτων, από τον πίνακα ANOVA, παρατηρούμε ότι ισούται με 1604897,183.

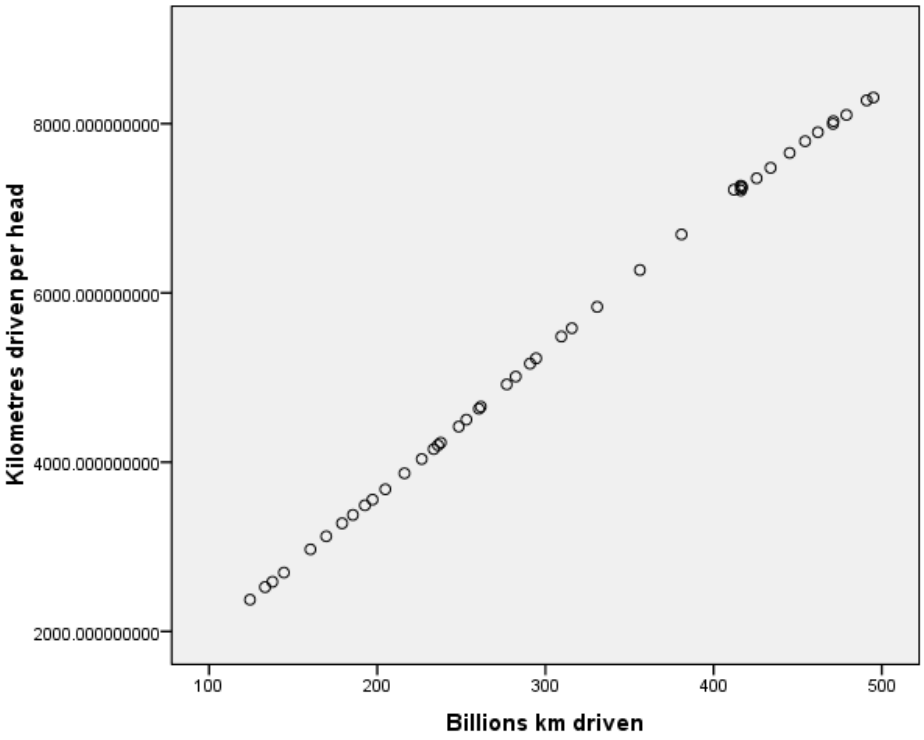
vii. Τέλος, στον πίνακα ***Residuals Statistics***, δίνονται οι προσαρμοσμένες τιμές των Y_i (δηλαδή, οι προβλέψεις των Y_i), καθώς και τα κατάλοιπα[10].

5.6. **Συσχέτιση μεταβλητών** Traffic_km – Distance

Στο σημείο αυτό θα εξεταστεί το κατά πόσο υπάρχει συσχέτιση μεταξύ των συνολικών χιλιομέτρων που πραγματοποιούνται κάθε χρόνο από τους οδηγούς και της απόστασης που διανύει κατά μέσο όρο ο κάθε οδηγός για το αντίστοιχο έτος.

[10] Εμφανίζονται οι τιμές με βάση κάποια μέτρα (πχ ελάχιστη, μέγιστη τιμή), ο πλήρης πίνακας δίνεται εμφανίζεται στο Data View στο SPSS.

Ακλουθώντας την ίδια διαδικασία με νωρίτερα (Ενότητα 5.3), αρχικά κατασκευάζεται το διάγραμμα διασποράς (scatterplot) μεταξύ των δύο μεταβλητών.



Από το ανωτέρω διάγραμμα καθίσταται απόλυτα σαφές ότι υπάρχει γραμμική σχέση μεταξύ των μεταβλητών και επομένως προχωράμε στην εφαρμογή του μοντέλου της απλής γραμμικής παλινδρόμησης για να εξετάσουμε τη συσχέτιση.

Ακολουθώντας την ίδια διαδικασία με νωρίτερα, οι πίνακες οι οποίοι θα πάρουμε ως έξοδο από το SPSS είναι:

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Billions km driven[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Kilometres driven per head

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	1.000[a]	.999	.999	5.57852655E1

a. Predictors: (Constant), Billions km driven

b. Dependent Variable: Kilometres driven per head

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.566E8	1	1.566E8	50335.404	.000[a]
	Residual	130703.826	42	3111.996		
	Total	1.568E8	43			

a. Predictors: (Constant), Billions km driven

b. Dependent Variable: Kilometres driven per head

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	354.823	24.011		14.777	.000	306.366	403.279
	Billions km driven	16.389	.073	1.000	224.356	.000	16.241	16.536

a. Dependent Variable: Kilometres driven per head

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	2.3929E3	8.46726E3	5.4006E3	1.90863137E3	44
Residual	-1.5754E2	1.11456E2	...	5.51327838E1	44
Std. Predicted Value	-1.576	1.607	.000	1.000	44
Std. Residual	-2.824	1.998	.000	.988	44

a. Dependent Variable: Kilometres driven per head

Από τους παραπάνω πίνακες σημειώνουμε τα εξής σημαντικά:

i. Στον πίνακα «Coefficients», ο συντελεστής Beta (Standardized Coefficients: Beta) είναι η εκτίμηση του b_1, αφού εφαρμοστεί το μοντέλο παλινδρόμησης:

$$Y_i = b_0 + b_1 X_i' + \varepsilon_i$$

όπου τα X_i' είναι οι τυποποιημένες τιμές των X_i.

ii. Οι σημειακές εκτιμήσεις των συντελεστών b_0 και b_1 είναι 354,823 και 16,389 αντίστοιχα.

iii. Τα διαστήματα εμπιστοσύνης είναι (306,366 , 403,279) και (16,241 , 16,536) για τα b_0 και b_1 αντίστοιχα.

iv. Το p-value είναι 0 άρα η υπόθεση $b_0 = 0$ και $b_1 = 0$ απορρίπτεται. Επομένως, η μεταβλητή ***Distance*** εξαρτάται από τη μεταβλητή ***Traffic_km***.

v. Όσον αφορά τον πίνακα ANOVA, το p-value για τον έλεγχο: $H_0: b_1 = 0$,

$H_1: b_1 \neq 0$

ισούται με το μηδέν. Ο έλεγχος της συγκεκριμένης υπόθεσης γίνεται μέσω της τιμής F και στην ουσία είναι ισοδύναμος με τον έλεγχο που γίνεται μέσω της t1 (όπως είδαμε στο iv).

vi. Σε σχέση με την εκτίμηση της διασποράς των σφαλμάτων, από τον πίνακα ANOVA, παρατηρούμε ότι ισούται με 3111,996.

vii. Τέλος, στον πίνακα ***Residuals Statistics***, δίνονται οι προσαρμοσμένες τιμές των Y_i (δηλαδή, οι προβλέψεις των Y_i), καθώς και τα κατάλοιπα[11].

[11] Εμφανίζονται οι τιμές με βάση κάποια μέτρα (πχ ελάχιστη, μέγιστη τιμή), ο πλήρης πίνακας δίνεται εμφανίζεται στο Data View στο SPSS.

6. Συμπεράσματα

Στις ενότητες που προηγήθηκαν, έγινε μια λεπτομερή περιγραφή των περιγραφικών μέτρων στατιστικής ανάλυσης, καθώς και των μοντέλων και συντελεστών που χρησιμοποιούνται για τον έλεγχο ύπαρξης συσχέτισης μεταξύ δύο μεταβλητών. Ακολούθησε πλήθος παραδειγμάτων απ' όπου διαφάνηκε η εφαρμογή της θεωρίας σε ρεαλιστικά προβλήματα.

Στην τελευταία ενότητα, εφαρμόστηκε το μοντέλο της απλής γραμμικής παλινδρόμησης σε τέσσερεις μεταβλητές (με διαφόρους συνδυασμούς), όπου και διαπιστώθηκε η μεταξύ τους εξάρτηση. Στις περιπτώσεις των ενοτήτων 5.3 και 5.6 η γραμμική εξάρτηση των μεταβλητών ήταν εμφανής από το πρώτο κιόλας στάδιο, όπου αφορούσε τη κατασκευή του διαγράμματος διασποράς (scatterplot). Αντιθέτως, στην ανάλυση των μεταβλητών που εξετάστηκαν στις ενότητες 5.4 και 5.5, ήταν απαραίτητη η εφαρμογή του μοντέλου της απλής γραμμικής παλινδρόμησης για την διαπίστωση ύπαρξης ή μη γραμμικής συσχέτισης των μεταβλητών.

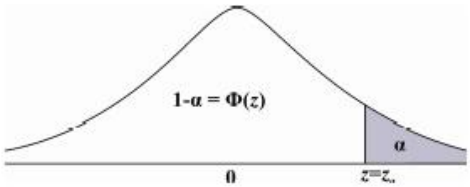
Όπως έγινε αντιληπτό, η χρήση του στατιστικού εργαλείου – πακέτου SPSS, διευκολύνει πολύ την ανάλυση τέτοιων προβλημάτων, μειώνοντας κατά πολύ τον υπολογιστικό χρόνο που χρειάζεται ο χρήστης και ταυτόχρονα παρέχοντας σε αυτόν συγκεντρωτικούς πίνακες και διαγράμματα για την εξαγωγή χρήσιμων συμπερασμάτων.

Βιβλιογραφία

1. J. L. Rodgers and W. A. Nicewander. Thirteen ways to look at the correlation coefficient. The American Statistician, 42(1):59–66, February 1988.
2. Stigler, Stephen M., "Francis Galton's Account of the Invention of Correlation". *Statistical Science* 4 (2): 73–79, 1989.
3. Kendall, M.G., Stuart, A., *The Advanced Theory of Statistics, Volume 2: Inference and Relationship*, Griffin, 1973.
4. Andrew Hansen, Eric Peterson, Todd Walter - Per Enge, Correlation Structure of Ionospheric Estimation and Correction for WAAS, Stanford University.
5. Χρήστος Κ. Φράγκος, Στατιστική Επιχειρήσεων για τις οικονομικές και κοινωνικές επιστήμες, Εκδ. Σταμούλης, Αθήνα, 1998.
6. Π. Α. Κιόχος και Α. Π. Κιόχος, Στατιστική για τις Επιχειρήσεις και την Οικονομία, Εκδ. Ελένη Κιόχου, Αθήνα, 2010.
7. Ibrahim A. Al-Kadi "The origins of cryptology: The Arab contributions", *Cryptologia*, 16(2), pp. 97–126, 1992.
8. Singh, Simon, *The code book : the science of secrecy from ancient Egypt to quantum cryptography* (1st Anchor Books έκδοση). New York: Anchor Books, 2000.
9. Willcox, Walter, *The Founder of Statistics*. Review of the International Statistical Institute 5(4):321–328, 1938.
10. Ι. Πανάρετος, Εκτιμητική και Έλεγχοι Υποθέσεων, Αθήνα, 1997.
11. Ι. Γ. Χαλκιάς, Στατιστική – Μέθοδοι ανάλυσης για επιχειρηματικές αποφάσεις, Εκδ. Rosili, 2010.
12. Μ. Κούτρας και Χ. Ευαγγελάρας, Ανάλυση παλινδρόμησης, Εκδ. Σταμούλη, 2011.

Παράρτημα – Στατιστικοί Πίνακες

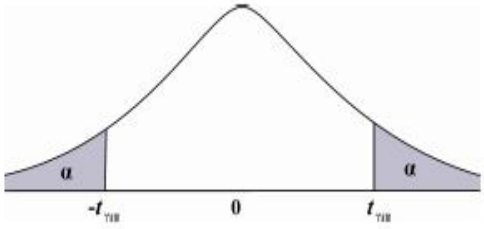
Πίνακας Ι: Η τυπική κανονική κατανομή



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586
0.1	0.53983	0.54380	**0.54776**	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490
0.7	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84850	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92786	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574
2.2	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899
2.3	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520
2.6	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736
2.8	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99897	0.99900

α	0.0005	0.001	0.005	0.01	0.025	0.05	0.10
z_α	3.29	3.09	2.576	2.326	1.960	1.645	1.282

Πίνακας ΙΙ: Η Κατανομή του t κατά Student



ν	α = 0.10	α = 0.05	α = 0.025	α = 0.01	α = 0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576